PHONETIC ANALYSIS OF DYSARTHRIC SPEECH TEMPO AND APPLICATIONS TO ROBUST PERSONALISED DYSARTHRIC SPEECH RECOGNITION

Feifei Xiong[†], *Jon Barker*[†], *Heidi Christensen*[†‡]

[†]Speech and Hearing Group (SPandH), Dept. of Computer Science, University of Sheffield, UK
[‡]Centre for Assistive Technology and Connected Healthcare (CATCH), University of Sheffield, UK
{f.xiong,j.p.barker,heidi.christensen}@sheffield.ac.uk

ABSTRACT

Improving the accuracy of personalised speech recognition for speakers with dysarthria is a challenging research field. In this paper, we explore an approach that non-linearly modifies speech tempo to reduce mismatch between typical and atypical speech. Speech tempo analysis at the phonetic level is accomplished using a forced-alignment process from traditional GMM-HMM in automatic speech recognition (ASR). Estimated tempo adjustments are applied directly to the acoustic features rather than to the time-domain signals. Two approaches are considered: i) adjusting dysarthric speech towards typical speech for input into ASR systems trained with typical speech, and ii) adjusting typical speech towards dysarthric speech for data augmentation in personalised dysarthric ASR training. Experimental results show that the latter strategy with data augmentation is more effective, resulting in a nearly 7% absolute improvement in comparison to baseline speaker-dependent trained system evaluated using UASpeech corpus. Consistent recognition performance improvements are observed across speakers, with greatest benefit in cases of moderate and severe dysarthria.

Index Terms— Dysarthria, Speech tempo, Phonetics, Data augmentation, Personalised speech recognition

1. INTRODUCTION

Dysarthria is a speech impairment caused by damage to the parts of the central or peripheral nervous system that control the muscles involved in speech production, i.e., respiration, phonation, and articulation. Consequences for speech include increased respiration frequency, inadequate pauses, breathy or hoarse voice, reduced speech rate, deviations in pitch and volume, hyper- or hyponasality, and misarticulated sounds [1], all of which can disrupt speech communication. At the same time, since dysarthria is often associated with severe physical disabilities like cerebral palsy, for this group of people, speech-enabled and hands-free interfaces often provide a more attractive and efficient means of access in comparison to hardwired switches, keyboards and remote controls [2, 3, 4, 5, 6].

Recent advances in the robustness of automatic speech recognition (ASR) technology mean that speech can now be practically used as a machine interface in everyday environments [7]. However, the state-of-the-art is tailored towards people with typical speech pronunciation. For people with *dysarthric* (i.e., disordered) speech, satisfactory recognition performance is seldom achievable due to the high inter- and intra-speaker variability inherent in dysarthric speech [8, 9]. In addition, and more crucially, the difficulty in collecting dysarthric speech data [10], means that the resources needed to train models that match an individual's voice characteristics are not available.

Research has been conducted recently for improving ASR performance by better modelling of the dysarthric speech variability. For instance, articulatory information has been applied to support acoustic features for improving acoustic modelling of dysarthric speech, provided that the use of speech production knowledge can be beneficial for the capture of inter-speaker variability [11, 12]. It has been demonstrated that adaptation in traditional GMM-HMM is effective to model intra-speaker variability of dysarthric speech at individual severity levels: A combination system consisting of state-transition interpolation and maximum a-posteriori (MAP) adaptation was introduced in [13], and a comparative study of the fundamental training and adaptation techniques was carried out in [9]; speaker adaptive training (SAT) with maximum likelihood linear regression concatenation with a MAP adaptation was reported to give the best performance in [14]. Additionally, an automatic method for adapting the pronunciation lexicon was introduced in [15, 16]. The use of the current advanced deep neural networks (DNNs) in robust acoustic modelling was exploited in [17, 18]. Further, unsupervised learning has been applied to ASR which has the advantage that neither the transcription of the training data nor a linguistic pronunciation lexicon is required. To this end, HMM-based self-organizing units and acoustic unit descriptors, that are phone-like units, were proposed in [19] and [20] to achieve the unsupervised training of (severe) dysarthric speech recognisers.

Other work has focused on dysarthric speech data collection. However, in contrast to typical speech, dysarthric speech is far more difficult to collect: dysarthric speakers are sparse in the general population and they are often not able to speak for a long time. The current American English datasets – including Whitaker [21], Nemours [22], TORGO [23] and UASpeech [24] – are all small compared to the datasets for typical speech used to train modern state-of-the-art ASR systems.

In this work we explore an approach that allows typical speech to be used in the development of dysarthric ASR, by applying speech tempo transformations to reduce typical vs. dysarthric speech mismatch. As suggested in [25], speech tempo is an important characterization of dysarthric speech which affects ASR performance. Dysarthric speech tempo is analysed at the phonetic level using a forced-alignment process from traditional GMM-HMM models trained incorporating a phonetic lexicon. This phoneme-based speech tempo ratio between typical and individual dysarthric speech is then used in tempo adjustment for two potential applications for robust personalised dysarthric speech recognition: i) the speech

This research has been supported by the DeepArt Project sponsored by Google, as well as been partly supported under the European Union's H2020 Marie Skłodowska-Curie programme TAPAS (Training Network for PAthological Speech processing; Grant Agreement No. 766287).

tempo of dysarthric speech can be adjusted to better match that observed in typical speech, thereby making the speech better matched to that modelled by acoustic models trained only on typical speech; ii) the speech tempo of typical speech can be adjusted towards that observed for dysarthric speech and the altered typical speech data can be applied during training in a data augmentation setup to train personalised/speaker-dependent acoustic models (emphasized by [26] to deal with severe dysarthria). This is in contrast to [27] where data augmentation was based on speed and tempo perturbation in the signal domain using an empirical selection of the perturbation ratio at a speaker-independent level.

In the remainder of this paper, we first describe the phonetic analysis based on ASR forced-alignment with UASpeech training data in Section 2 with a defined phoneme-based speech tempo ratio between typical and dysarthric speech. This ratio will be applied to speech tempo adjustment of dysarthric test speech or of typical speech to augment existing dysarthric training data for robust dysarthric speech recognition in Section 3. Experimental results will be presented in Section 4 before Section 5 concludes the paper.

2. PHONETIC ANALYSIS

Phonetic knowledge is required in traditional ASR systems to link the acoustic representation and the word sequence output. To this end, the speech region corresponding to a specific phoneme is aligned by a forced-alignment process using the GMM-HMM model. The individual speech tempo property, at the phonetic level, can then be analysed in an automatic manner using a speaker-dependent (SD) trained ASR model that provides the detailed phoneme-alignment information.

2.1. Data

The UASpeech corpus [24] is employed for speech tempo analysis of both typical and dysarthric speakers. It consists of data from 15 dysarthric speakers with cerebral palsy and 13 control (typical) speakers. There are 3 blocks of words for each speaker, and each block consists of 10 digits, 26 international radio alphabets, 19 computer commands, 100 common words and 100 distinct uncommon words, which were not repeated across blocks. The speech data is at sampling rate of 16 kHz, and all 7 microphones' recordings are included. Following previously published work using UASpeech for ASR (e.g., [13, 9]), CTL (typical/control) and DYS (disordered/dysarthric) datasets are divided into training and test data with a $2:1$ split, using blocks 1 and 3 for training and block 2 for test.

Note that the original recordings of UASpeech always contain very long initial and trailing segments of silence in each utterance, as well as some un-recognisable words that do not match the transcripts [14, 12]. In order to clean up the redundant data portion for more meaningful ASR experiments, we re-segmented all UASpeech data using the trained SD GMM-HMM model of each CTL and DYS speaker (with all 3 blocks) to decode the speech data itself again with a biased language model (cf. cleanup scheme in [7]). Re-segmented UASpeech data is summarized in Table 1, and baseline ASR performance w.r.t. multi-speaker training (MST) with SAT GMM-HMM for the DYS test set [12][1] benefits from this cleanup scheme, particularly for the MST-DYS set.

Sets(#Spk)	Re-segment	Block 1 & 3	Block 2	WER
CTL	✗	46410 (22.7 h)	23205 (11.1 h)	57.42
#13	✓	46403 (19.8 h)	23205 (9.7 h)	56.86
DYS	✗	49204 (44.3 h)	24731 (21.7 h)	48.60
#15	✓	49204 (27.3 h)	24727 (13.4 h)	44.91

Table 1. The number of utterances (and duration in hours) in UASpeech training and test set with and without re-segment, as well as the baseline ASR performance in terms of averaged word error rate (WER) with DYS test set (15 speakers) using block 2.

2.2. Forced-alignment

For SD GMM-HMM training (with blocks 1 and 3) as the basis of forced-alignment process, 13-dimensional MFCCs incorporating a spliced context window of length 9 frames are used, and these are subsequently transformed to a 40-dimensional vector via linear discriminant analysis and maximum likelihood linear transform (cf. [28]). SAT is employed based on feature-space maximum likelihood linear regression (fMLLR) [29]. A uniform language model is generated based on the transcriptions of speech files, as well as a word grammar network containing a silence model at the start and one following single word, denoted as $<$ sil $>$ word. The UASpeech phone set is listed in Table 2, and we further group the phonemes into 4 types of vowels and 9 types of consonants based on similar duration (cf. [30]) for speech tempo analysis at the phonetic level (denoted as *phoneme-based*).

Vowels #16	(V1) short vowels	AH AO AX EH IH UH
	(V2) medium vowels	AE
	(V3) long vowels	AA ER IY UW
	(V4) diphthongs	AW AY EY OW OY
Consonants #24	(C1) glides	L R W Y
	(C2) unvoiced stops	K P T
	(C3) voiced stops	B D G
	(C4) nasals	M N NG
	(C5) unvoiced fricatives	F S SH TH
	(C6) voiced fricatives	DH V Z ZH
	(C7) unvoiced affricates	CH
	(C8) voiced affricates	JH
	(C9) aspirates	HH

Table 2. The grouping of phonemes according to UASpeech phone set for speech tempo analysis at the phonetic level.

2.3. Speech Tempo Analysis

As shown in Fig. 1, speech tempo in the CTL set is relatively consistent across the typical speakers in general. On average (across about 1.8 hours' data for each speaker), the length of vowels is approximately 80 ms longer than that of consonants. By contrast, speech tempo varies dramatically across dysarthric speakers in the DYS set, which were grouped in 4 severity levels based on a subjective estimate of perceptual speech intelligibility ratings [24], namely *Severe*, *Moderate-Severe*, *Moderate* and *Mild*. Roughly speaking, averaged phoneme duration seems to be proportional to the dysarthria severity, i.e., the higher the severity, the longer the phoneme. However, this does not always hold for speakers in group *Severe*, likely because it is difficult for people with severe dysarthria to pronounce words with explicit standard phonemes.

Due to the observed high inter- and intra-speaker variability in terms of speech tempo, it is necessary to individually analyse the dysarthric speech tempo and its relationship to typical speech at a phoneme level. Based on the averaged length of vowels and consonants across all utterances of each speaker, *phoneme-based* speech tempo ratio between specific dysarthric and typical speaker can be

[1] We have released our baseline Kaldi scripts for UASpeech with and without re-segments in `https://github.com/ffxiong/uaspeech`

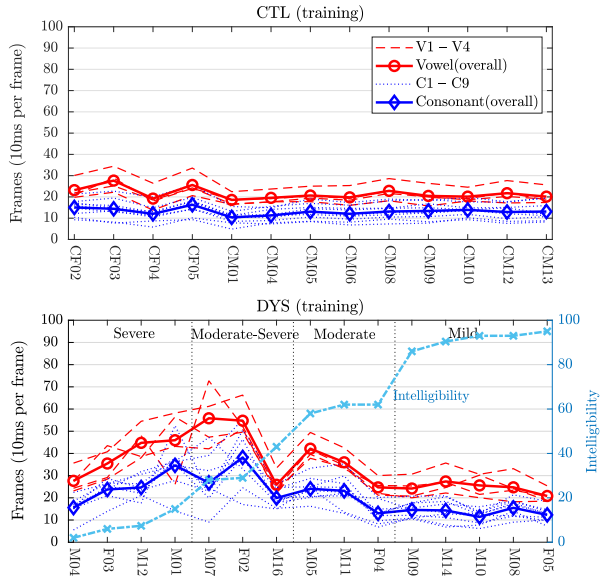


Fig. 1. Averaged duration of the phonetic groups (Table 2) across all utterances in CTL and DYS (with an additional speech intelligibility rating in right y-axis) training data for each speaker.

defined as

$$\mathcal{R}_{d\leftarrow c}(p) = \frac{T_d(p)}{T_c(p)}\,, \tag{1}$$

where $T_d(p)$ and $T_c(p)$ denote the duration of the specific phoneme-group $p \in (\mathrm{V1}-\mathrm{V4}, \mathrm{C1}-\mathrm{C9})$ from dysarthric speaker d and typical speaker c, respectively. When averaged over all p, $\overline{\mathcal{R}_{d\leftarrow c}}$ represents the *speaker-based* ratio in terms of general speaking rate.

3. SPEECH TEMPO ADJUSTMENT FOR ASR

Speech tempo adjustment using ratio $\mathcal{R}_{d\leftarrow c}(p)$ can be applied in the test and the training stage of dysarthric speech recognition.

3.1. Test Stage

To better match the ASR model trained using typical speech alone, dysarthric test speech can be adjusted towards typical speech before decoding. This can be done in either the signal or the feature domain, as depicted in Fig. 2. Typically, tempo adjustment is performed in the signal domain, e.g., via WSOLA algorithm [31] (implemented in SoX[2] with *tempo* function) with preserved pitch and spectral envelope. Essentially, tempo changes in signals will result in the time warping in MFCCs as well, indicating that tempo adjustment can be directly performed in the feature domain via a simple interpolation incorporating a sequence downsample operation.

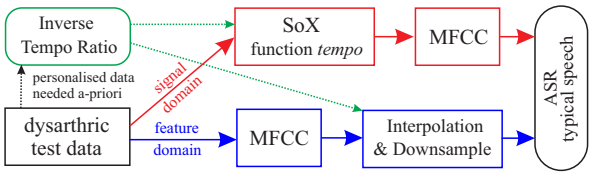


Fig. 2. Speech tempo adjustment of the dysarthric test data for ASR trained using typical speech in the signal and the feature domain.

[2] http://sox.sourceforge.net

Due to the lack of alignment knowledge in dysarthric test data, it is not possible to apply phoneme-based tempo ratios. Instead, a speaker-based tempo ratio $\overline{\mathcal{R}_{d\leftarrow c}}$ can be determined using a small amount of speech from the target dysarthric speaker, assuming that the personalised speaking rate is fairly constant. The inverse tempo ratio $1/\overline{\mathcal{R}_{d\leftarrow c}}$ is then applied to normalise the dysarthric speech.

3.2. Training Stage

DNN-HMM based ASR generally benefits from training data augmentation as DNN generalization will be enhanced. However, any training-test mismatch introduced by data augmentation must be minimized. It is therefore desirable to augment the training data by simulating dysarthric speech that matches the characteristics of the speaker to be recognised, and this can be achieved by tempo adjustment of the typical training speech. Using the available training data alignments, the phoneme-based tempo ratio in (1) can be applied.

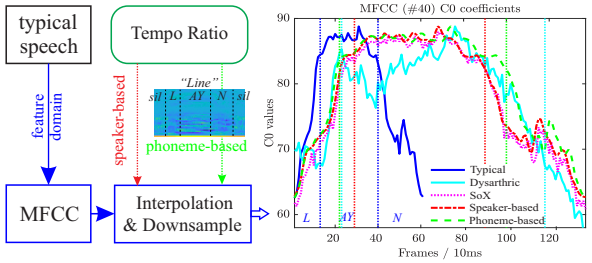


Fig. 3. Speech tempo adjustment of typical speech towards dysarthric done in the feature domain for data augmentation in ASR training stage. Right panel shows one example with different tempo adjustment schemes w.r.t. C0 coefficient in MFCCs.

As illustrated in the right panel of Fig. 3, tempo adjustment in the feature domain provides more smoothed MFCCs than that in the signal domain (omitted in left panel, cf. Fig. 2), when comparing speaker-based tempo adjustment with SoX using the same ratio $\overline{\mathcal{R}_{d\leftarrow c}}$. Also, it can be clearly observed that phoneme-based tempo adjustment provides a more accurate match of phoneme duration than other schemes when compared to MFCCs calculated from real dysarthric speech, particularly with vowels. Note that phoneme-based tempo adjustment in the signal domain using SoX is possible, but excluded, as distortion was introduced at the boundary between phonemes for concatenation that degraded ASR performance in our pilot experiments.

4. EXPERIMENTAL SETUP AND RESULTS

The setting for SAT GMM-HMM training is the same as described in Section 2.2. Hybrid DNN-HMM training is then applied using chain model with time-delayed neural network (TDNN) [32], which integrates the advantages from long temporal context extraction of speech frames and connectionist temporal classification, but at the cost of the requirement of a larger amount of training data. To this end, speed perturbation is employed in the baseline to generate additional 2 copies of the original training data by adjusting the pitch and tempo (together) via SoX resampling algorithm with ratios of 0.9 and 1.1 [33].

Firstly, speech tempo adjustment is applied to the test stage for ASR model based on multi-speaker training (MST) using typical speech data in CTL training set (cf. Table 1), and speaker-based tempo ratio $\overline{\mathcal{R}_{d\leftarrow c}}$ is determined by calculating averaged tempo from CTL training data (all 13 speakers) and DYS training data of each dysarthric speaker. In general, according to the estimated

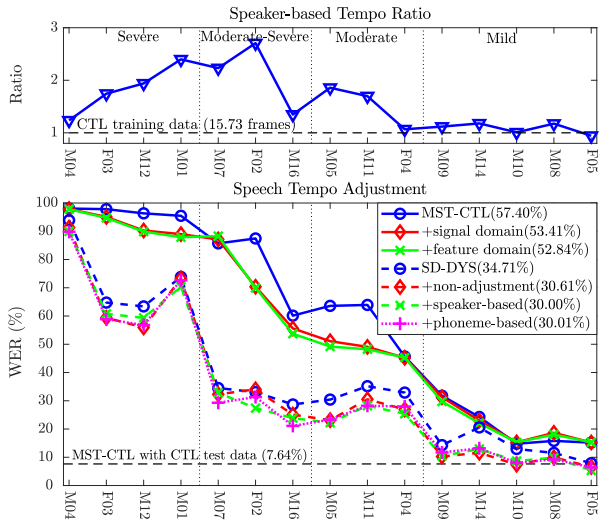
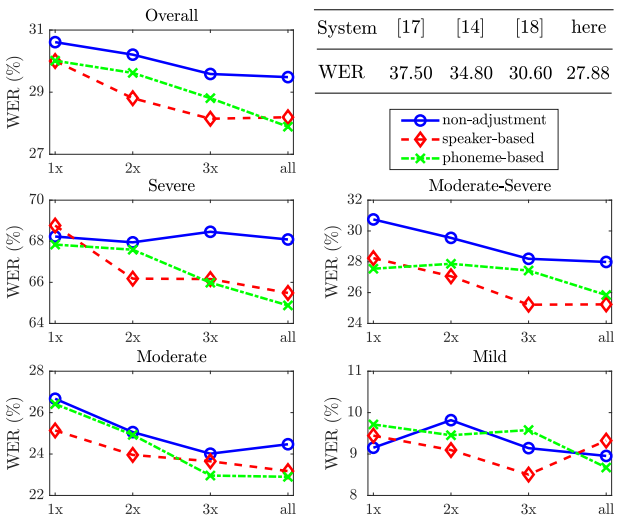


Fig. 4. Speaker-based tempo ratio between DYS and CTL training data and applications to tempo adjustment in DYS test set with MST-CTL TDNN model. For comparison, performance with SD-DYS TDNN model, as well as data augmentation with equal amount (1x) of augmented data to original SD training data is included.



System	[17]	[14]	[18]	here
WER	37.50	34.80	30.60	27.88

Fig. 5. Effectiveness of different tempo adjustment schemes for a rising-scaled data augmentation ('all' denotes the case with all the available CTL training data which is around $4-5$x), and performance comparison to other state-of-the-art dysarthric ASR systems.

speaker-based tempo ratio, the DYS test data is required to speed up within 3 times to match better the ASR model trained with typical speech (CTL). As shown in Fig. 4, tempo adjustment in the feature domain slightly outperforms the process in the signal domain, indicating that it is more effective to directly perform adjustment on features than on time-domain signals. Note that in the following experiments, results using tempo adjustment in the signal domain are omitted. On average, 4.6% absolute WER reduction can be achieved by tempo adjustment in the test stage and this benefit becomes more noticeable for speakers with moderate and severe dysarthria.

On the other hand, performance with tempo adjustment on the test stage is still far from comparable to results using speaker-dependent model (SD-DYS), suggesting a personalised speech recogniser especially for people with moderate and severer dysarthria. By this, it further shows that data augmentation using typical speech from CTL training set (randomly sorted) with an equal amount to original SD-DYS training data can bring 4% absolute accuracy improvement on average, indicating that DNNs can make good use of external data within the same vocabulary and leverage well between dysarthric and typical speech even though they show very different characteristics. Compared to the case without tempo adjustment, the proposed speaker-based and phoneme-based speech tempo adjustment in the feature domain achieve further reduced WERs, and a more consistent improvement across all 15 DYS speakers can be observed, particularly for groups *Severe* and *Moderate-Severe*.

Training	Severe	Mod.-Severe	Moderate	Mild	Overall
SD-DYS	72.65	32.25	32.70	13.44	34.71
+non-adjustment	68.22	30.74	26.66	**9.15**	30.61
+speaker-based	68.76	28.23	**25.13**	9.44	**30.00**
+V1−V4	69.33	29.13	25.47	9.62	30.50
+C1−C9	70.69	29.54	27.41	9.40	31.16
+phoneme-based	**67.83**	**27.55**	26.41	9.71	30.01

Table 3. Averaged WERs for 4 DYS Groups in terms of data augmentation (1x) with different speech tempo adjustment schemes.

Table 3 further summarizes the detailed WERs in terms of each DYS group when data augmentation is applied, to pinpoint the individual advantage of different tempo adjustment schemes. Phoneme-based tempo adjustment outperforms other schemes for groups *Moderate-Severe* and *Severe*, indicating that this dynamic tempo adjustment is necessary for dysarthric speech simulation that better matches real dysarthric speech with a high inter-phoneme variability of speech tempo (cf. Fig. 1). Further, it is not sufficient to solely adjust vowels or consonants in this phoneme-based tempo adjustment and it seems that vowels are more important to concern than consonants.

The effectiveness of the proposed tempo adjustment in data augmentation is further tested using more augmented data (from CTL training set) on a rising scale w.r.t. the original SD training data from each dysarthric speaker. Fig. 5 illustrates that with more augmented data for SD-DYS training, a consistent performance improvement can be achieved when augmented data is generated by the proposed tempo adjustment, while performance becomes easily saturated or even degrades when typical speech data that exhibits large mismatch to original dysarthric training data is directly adopted for data augmentation. This is particularly noticeable for groups *Moderate* to *Severe*, and when all the available CTL training data is used, phoneme-based scheme yields the best overall WER of 27.88%, which is also better than the reported best results from other state-of-the-art dysarthric ASR systems [17, 14, 18].

5. CONCLUSIONS

This paper presented two approaches for improving dysarthric speech recognition performance based on modification of typical speech via speech tempo adjustment operating at the phonetic level and in the feature domain. Results showed that data augmentation using temporally modified typical speech is an effective strategy and can improve personalised dysarthric ASR performance for moderate to severe dysarthric speakers. The benefits of data augmentation were seen to increase as more simulated data was used until eventual saturation. To improve performances further, more work is needed to better model the mapping of typical to dysarthric speech dynamics at every single phoneme level.

6. REFERENCES

[1] F. L. Darley, A. E. Aronson, and J. R. Brown, "Clusters of deviant speech dimensions in the dysarthrias," *Journal of Speech and Hearing Research*, vol. 12, no. 3, pp. 462–496, 1969.

[2] M. Fried-Oken, "Voice recognition device as a computer interface for motor and speech impaired people," *Archives of Physical Medicine and Rehabilitation*, vol. 66, no. 10, pp. 678–681, 1985.

[3] J. Noyes and C. Frankish, "Speech Recognition Technology for Individuals with Disabilities," *Augmentative and Alternative Communication*, vol. 8, no. 4, pp. 297–303, 1992.

[4] K. Rosen and S. Yampolsky, "Automatic Speech Recognition and a Review of Its Functioning with Dysarthric Speech," *Augmentative and Alternative Communication*, vol. 16, no. 1, pp. 48–60, 2000.

[5] M. S. Hawley, S. P. Cunningham, P. D. Green, P. Enderby, R. Palmer, S. Sehgal, and P. O'Neill, "A Voice-Input Voice-Output Communication Aid for People with Severe Speech Impairment," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 21, no. 1, pp. 23–31, 2013.

[6] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain, "homeService: Voice-enabled assistive technology in the home using cloud-based automatic speech recognition," in *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*, Grenoble, France, Aug. 2013, pp. 29–34.

[7] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proceedings of Interspeech*, Hyderabad, India, Sept. 2018, pp. 1561–1565.

[8] M. S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. O'Neill, and R. Palmer, "A speech-controlled environmental control system for people with severe dysarthria," *Medical Engineering and Physics*, vol. 29, no. 5, pp. 586–593, 2007.

[9] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Proceedings of Interspeech*, Portland, Oregon, USA, Sept. 2012, pp. 1776–1779.

[10] 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*, South Lake Tahoe, USA, Dec. 2014, pp. 254–259.

[11] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 947–960, 2011.

[12] F. Xiong, J. Barker, and H. Christensen, "Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition," in *ITG Conference on Speech Communication*, Oldenburg, Germany, Oct. 2018.

[13] H. V. Sharma and M. Hasegawa-Johnson, "State transition interpolation and MAP adaptation for HMM-based dysarthric speech recognition," in *Proceedings of workshop on Speech and Language Processing for Assistive Technology (SLPAT)*, Los Angeles, California, USA, June 2010, pp. 72–79.

[14] S. Sehgal and S. Cunningham, "Model Adaptation and Adaptive Training for the Recognition of Dysarthric Speech," in *Proceedings of workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, Dresden, Germany, Sept. 2015.

[15] K. F. Mengistu and F. Rudzicz, "Adapting Acoustic and Lexical Models to Dysarthric Speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 4924–4927.

[16] H. Christensen, P. Green, and T. Hain, "Learning speaker-specific pronunciations of disordered speech," in *Proceedings of Interspeech*, Lyon, France, Aug. 2013, pp. 1159–1163.

[17] H. Christensen, M. B. Aniol, P. Bell, P. Green, T. Hain, S. King, and P. Swietojanski, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech," in *Proceedings of Interspeech*, Lyon, France, Aug. 2013, pp. 3642–3645.

[18] J. Yu, X. Xie, S. Liu, S. Hu, M. W. Y. LAM, X. Wu, K. H. Wong, X. Liu, and H. Meng, "Development of the CUHK dysarthric speech recognition system for the UASpeech corpus," in *Proceedings of Interspeech*, Hyderabad, India, Sept. 2018, pp. 2938–2942.

[19] M. Siu, H. Gish, A. Chan, W. Belfield, and S. Lowe, "Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery," *Computer Speech & Language*, vol. 28, no. 1, pp. 210–223, 2013.

[20] O. Walter, V. Despotovic, R. Haeb-Umbach, J. F. Gemmeke, B. Ons, and H. Van hamme, "An evaluation of unsupervised acoustic model training for a dysarthric speech interface," in *Proceedings of Interspeech*, Singapore, Sept. 2014, pp. 1013–1017.

[21] J. R. Deller, M. S. Liu, L. J. Ferrier, and P. Robichaud, "The Whitaker database of dysarthric (cerebral palsy) speech," *The Journal of the Acoustical Society of America*, vol. 93, no. 6, pp. 3516–3518, 1993.

[22] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The Nemours database of dysarthric speech," in *International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, USA, Oct. 1996, pp. 1962–1965.

[23] 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.

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

[25] F. Rudzicz, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech & Language*, vol. 27, no. 6, pp. 1163–1177, 2013.

[26] E. Sanders, M. Ruiter, L. Beijer, and H. Strik, "Automatic recognition of Dutch dysarthric speech A pilot study," in *Proceedings of Interspeech*, Denver, Colorado, USA, Sept. 2002, pp. 661–664.

[27] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," in *Proceedings of Interspeech*, Hyderabad, India, Sept. 2018, pp. 471–475.

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Big Island, HI, USA, July 2011.

[29] D. Povey and K. Yao, "A Basis Method for Robust Estimation of Constrained MLLR," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 4460–4463.

[30] D. R. Hill, "Low-level articulatory synthesis: A working text-to-speech solution and a linguistic tool," *Canadian Journal of Linguistics*, vol. 62, no. 3, pp. 1–40, 2017.

[31] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Minneapolis, USA, Apr. 1993, pp. 554–557.

[32] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proceedings of Interspeech*, San Francisco, USA, Sept. 2016, pp. 2751–2755.

[33] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proceedings of Interspeech*, Dresden, Germany, Sept. 2015, pp. 3586–3589.