Handling acoustic variation in dysarthric speech recognition systems through model combination

Enno Hermann[1,2]*, Mathew Magimai Doss*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École polytechnique fédérale de Lausanne, Switzerland
{enno.hermann, mathew}@idiap.ch

Abstract

Developing automatic speech recognition (ASR) systems that recognise dysarthric speech as well as control speech from unimpaired speakers remains challenging. Including more highly variable dysarthric speech during training can also negatively affect the performance on control speakers, which is not desirable when developing speech recognisers for a wider audience. In this work, we analyse how the acoustic variability of dysarthric speech affects ASR systems and propose the combination of multiple acoustic models trained on different subsets of speakers to mitigate this effect. This approach shows improvements for both dysarthric and control speakers on the Torgo and UA-Speech corpora.

Index Terms: speech recognition, pathological speech, dysarthria

1. Introduction

Automatic speech recognition (ASR) systems that can recognise the dysarthric speech of Parkinson's or amyotrophic lateral sclerosis (ALS) patients, could improve their quality of life by facilitating interaction with electronic devices and controlling home automation systems. This allows them to avoid input methods like buttons or touch screens that are not designed for their needs.

There are three main challenges for the automatic recognition of dysarthric speech. First, dysarthric speech differs from typical speech in a number of aspects, depending on the patient and their individual pathology. These can include a lower speech rate, different and less distinct articulation, and others. ASR systems, which are usually trained mostly on typical speech, need to be able to model these changes.

Second, dysarthric speech itself has a lot of variability. Speech characteristics can vary significantly between different speakers and even for the same speaker over time because of medication or therapy.

Third, the large variety of different pathologies and their individual manifestations in patients results in a lack of training data with sufficient coverage. Recording speech for an extended period of time can also be strenuous for patients and existing dysarthric speech databases are therefore relatively small.

Previous works have handled the data scarcity by adapting models trained on typical speech [1, 2], transforming pathological speech to resemble unimpaired speech [3, 4], and generating artificial dysarthric speech data through data augmentation [5, 4].

In this work we mainly target the first two issues and aim to overcome the acoustic- and pronunciation-level mismatch between dysarthric and typical speech through model combination. We argue that simply adding more typical speech data through data augmentation methods is not sufficient because data scarcity is not the only reason for poor performance of ASR systems on dysarthric speech. The differences of dysarthric speech, which are hard to quantify, also need to be accounted for.

2. Proposed approach

The variation in dysarthric speech w.r.t control speech lies at acoustic level as well as at pronunciation level. Dealing with those two variations separately is not a trivial task for two main reasons. First, the lexicon is based on control speech, i.e. highly intelligible, typically native speech. Second, the set of acoustic units is determined by phonotactic constraints enforced by the lexicon, even though dysarthric speech may be unintelligible and the phoneme sequences not clearly identifiable. The analyses presented in Section 4 demonstrate these issues.

We propose to train separate acoustic models on different subsets of the data and combine them by dynamic acoustic model selection during decoding. Thus, each model is specialised for the acoustic characteristics of its training speakers, but with model combination we can still recognise a variety of speech conditions without requiring prior information about the speaker. This approach could further be extended to also handle pronunciation variation by combining models trained with different acoustic subword units like phonemes and graphemes.

3. Experiments

3.1. Databases

We used the UA-Speech [6] and Torgo [7] dysarthric speech corpora for our experiments to ensure that the results are not specific to a single dataset.

The UA-Speech corpus consists of recordings of isolated words made with 7 microphones from 15 dysarthric speakers (about 40 hours in total) and 13 control speakers without any speech impairment (about 30 hours in total). We use the resegmented version of the corpus from [4] where excessive portions of silence have been removed via forced alignment. The set of words is divided into three distinct blocks, two of which (Block 1 and 3) are commonly used as the training set, while models are evaluated on Block 2. We group the results by speaker severity based on perceptual speech intelligibility ratings provided with the corpus in the same way as previous works [4].

The Torgo corpus contains about 15 hours of recordings from 15 speakers made with an array and a headset microphone. There are 8 mostly severely dysarthric speakers (6 hours of speech in total) and 7 control speakers (9 hours of speech in total). About 75% of the utterances are isolated words, most of them minimal pairs, the remaining 25% are sentences. We report our results separately for the isolated words and the sentences so that they are more informative because the data is so different.

Due to the small amount of speakers and data, the corpus is commonly evaluated in a leave-one-out cross validation setup, where for each test speaker a separate system is trained on the remaining speakers. For brevity, we then average the results across dysarthric and control speakers. Nevertheless, there can be considerable variations between individual speakers depending on their severity, as we show in the more detailed breakdowns for UA-Speech.

3.2. Baseline systems

We trained all our ASR models with the open-source Kaldi speech recognition toolkit [8]. We followed the typical development pipeline to train a hidden Markov model (HMM)/Gaussian mixture model (GMM) baseline with speaker adaptive training (SAT) on 39-dimensional MFCC+Δ+$\Delta\Delta$ features and use these to train neural network models with a lattice-free maximum mutual information (LF-MMI) objective function [9]. The UA-Speech models use position-dependent phonemes as lexical units, the Torgo models do not. For UA-Speech, we used the recipe of [4][1] as a basis and train a neural network acoustic model with 6 convolutional neural network (CNN) layers followed by 9 factorised time-delay neural network (TDNN) layers. The Torgo recipe is based on [10][2] and the acoustic model has 12 factorised TDNN layers. The LF-MMI systems on both corpora are trained with two additional, speed-perturbed copies of the training data as a form of data augmentation [11]. We did not make any modifications to these existing recipes to avoid overfitting on these small datasets.

For UA-Speech and the Torgo isolated word task, we use a decoding grammar that only contains the possible output words and restricts the output to a single word. For the Torgo sentence task, we use a bigram language model trained on all sentences in the corpus. The substantial textual overlap between speakers makes it infeasible to exclude all test sentences from the language model in a cross-validation setup. While this makes the sentence task easier, it is still challenging with dysarthric speech and the main focus of our work is on acoustic modelling..

3.3. Model combination

We combine the trained acoustic models by computing the union of the decoding lattices with subsequent minimum Bayes risk (MBR) decoding as it is implemented in Kaldi [12] through the `lattice-combine` binary. During lattice combination, the total probability of each path is normalised, which allows to combine any two models even when they are trained with different acoustic unit sets as in our case. ROVER [13] is another common way to combine decoding hypotheses from multiple ASR systems and has been used in a similar way to combine different pronunciation models [14].

4. Results and discussion

4.1. Baselines

Table 1 shows the word error rates (WERs) of the baseline models on the isolated word and sentence recognition tasks. As previous work [10] has shown, ASR systems trained with a LF-MMI loss function bring significant improvements over traditional subspace GMM (SGMM) based systems for most dysarthric speakers. For best results, control speech should be added to the training data. However, LF-MMI models then perform worse on control speakers than GMMs or a system trained on control speakers alone for both tasks. This might not be desirable, for example when developing general purpose ASR systems where the target audience is more likely not known a priori.

[1] https://github.com/ffxiong/uaspeech/tree/master/s5_segment
[2] https://github.com/idiap/torgo_asr

Table 1: *WER results on Torgo, averaged over dysarthric and control speakers, respectively. The acoustic models were trained either on both dysarthric and control speakers, or only one of those sets.*

	Training data	Isolated		Sentences	
		Dys	Con	Dys	Con
SGMM [10]	Both	56.1	19.4	41.5	4.4
LF-MMI	Both	**49.2**	24.0	**25.9**	7.9
	Dysarthric	55.0	41.9	42.1	18.4
	Control	52.9	**17.9**	48.7	**2.9**

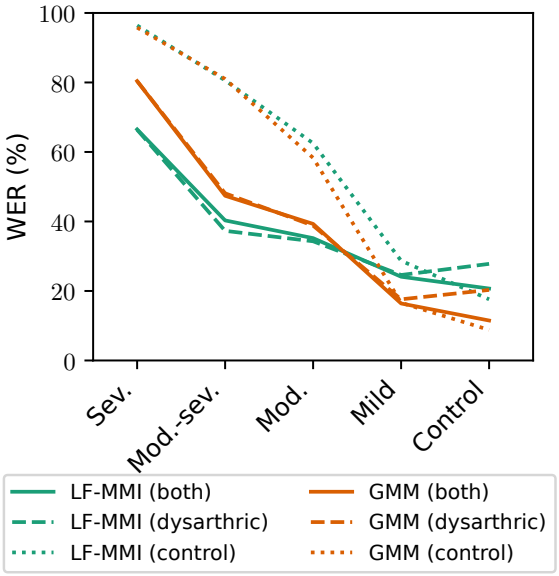
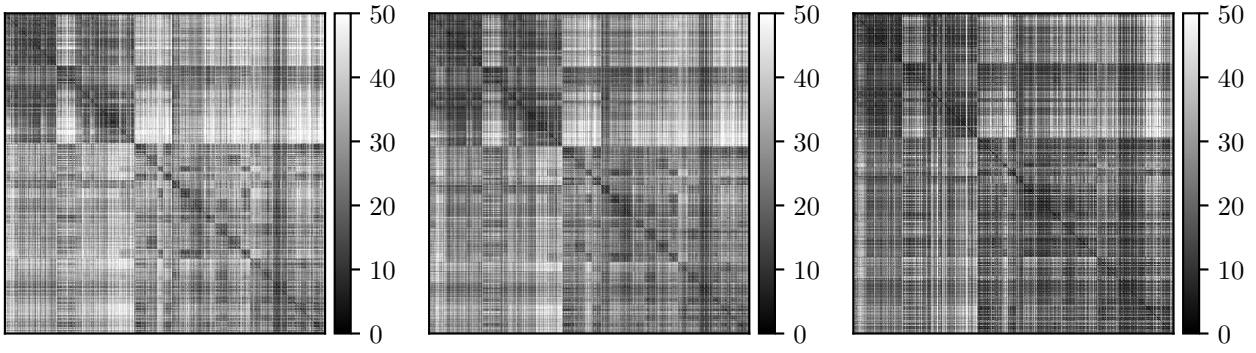


Figure 1: *WER results on UA-Speech: Training only on dysarthric speakers, only on control speakers, or both sets for GMM (orange) and LF-MMI (green) systems.*

We observe similar patterns when we repeat these experiments on UA-Speech in Figure 1. The LF-MMI system overall performs better than the GMM, except for the mildly dysarthric and control speakers. In this case there is not a big performance difference on dysarthric speech between training on only dysarthric or all of the speakers, probably because UA-Speech contains much more data than Torgo. However, on control speech there is still a 3% absolute drop in WER when training on both sets of speakers compared to a control-only system.

4.2. Analysis of acoustic models for dysarthric speech

In order to better understand the behaviour described in Section 4.1, to illustrate the challenges that acoustic models face in recognising dysarthric speech and to support our approach, we analysed the acoustic subword units used by these models for the Torgo corpus. We compare the LF-MMI acoustic models trained only on dysarthric speakers, only on unimpaired control speak-



(a) *Control* (b) *Control + dysarthric* (c) *Dysarthric*

Figure 2: *Confusion matrices of acoustic units for 3 different models, trained on control, dysarthric, or both sets of speakers on the Torgo corpus. Units are grouped by manner of articulation and then by phoneme, from consonants in the upper left, to semi-vowels and vowels in the bottom right corner. Darker regions indicate higher similarity and lower Kullback-Leibler (KL) divergence. Values above 50 are clipped. More dysarthric speech data results in a less discriminative acoustic unit space where units are more similar to each other.*

ers, and on the combined speaker set. We follow the approach of Razavi and Doss [15, 16] to compare two sets of acoustic units, represented as GMMs, where we compute the KL divergence between each unit to obtain a confusion matrix.

Before training neural network acoustic models in Kaldi, a new decision tree for the specific HMM topology of LF-MMI is usually built, which determines the set of acoustic units. After this, we estimate simple Gaussian distributions without mixtures for these units, which we can use to compare two acoustic unit sets. The KL divergence $D_{\mathrm{KL}}(f||g)$ between two multivariate Gaussian distributions $\mathcal{N}_f(\boldsymbol{\mu}_f, \Sigma_f)$ and $\mathcal{N}_g(\boldsymbol{\mu}_g, \Sigma_g)$ with mean vectors $\boldsymbol{\mu}$, covariance matrices Σ, and dimensionality d is [17]

$$D_{\mathrm{KL}}(f||g) = \frac{1}{2}\log\frac{|\Sigma_g|}{|\Sigma_f|} + \frac{1}{2}\mathrm{Tr}(\Sigma_g^{-1}\Sigma_f) + \frac{1}{2}(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g)^T\Sigma_g^{-1}(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g) - \frac{d}{2} \quad . \quad (1)$$

We computed the KL divergences between all units for each of the three models to obtain the confusion matrices in Figure 2. We observe that the KL divergences, and thus the state and phoneme discriminability, are the highest for the control speech model. The phonemes are most confusable in the dysarthric model and the combined one falls between the two. Dysarthric speakers are less intelligible and have articulation difficulties, so it makes sense that acoustic subword units derived from dysarthric speech are more confusable than those from control speech.

To better quantify these differences and demonstrate their effect on word discrimination, we compared HMM state sequences of word pairs. We picked a subset of 854 unique words from the Torgo corpus and the state sequences corresponding to their pronunciation for a given model and decision tree. For each possible pair of two different words, we then computed the dynamic time warping (DTW) distance between the state sequences, with the KL divergence between the corresponding Gaussians as the local distance. The pair generation and DTW computation is based on code of [18][3].

Figure 3 shows kernel density estimates of the distributions of DTW distances between word pairs for the same 3 models as above. The dysarthric speech model is worse at discriminating words than the control speech one, and the combined model is again in the middle, which confirms our previous analysis.

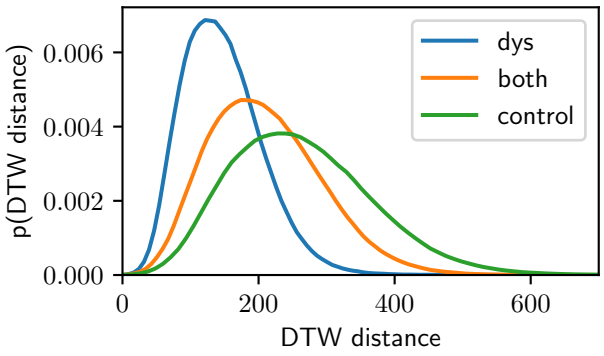


Figure 3: *Distributions of DTW distances between word pairs with acoustic units from models trained only on dysarthric, only on control, or on both sets of speakers.*

These analyses explain that the lower performance on control speech with models trained also on dysarthric speech is due to the dysarthric speech adding too much variability for the neural network training. This affects the acoustic unit space and reduces the model's discriminative power. Indeed, Table 1 shows that while including control speech into LF-MMI training is crucial to perform well on dysarthric speakers, we observe the lowest WER on control speech when training without any dysarthric speech data, beating also the SGMM. Similarly, it was previously shown that data augmentation by speed perturbation — which we use in all our LF-MMI experiments — helps for dysarthric speakers, but not introducing this additional source of variability is better for control speakers when training on both sets [10].

4.3. Model combination

Table 2 shows the results of model combination of the different LF-MMI systems for Torgo from Table 1. We find that the

[3] github.com/kamperh/recipe_bucktsong_awe_py3/tree/master/samediff

Table 2: *WER results on Torgo, averaged over dysarthric and control speakers, respectively. Lattice combination of the LF-MMI models from Table 1.*

Combinations of systems from Table 1	**Isolated**		**Sentences**	
	Dys	Con	Dys	Con
Dysarthric + Both	44.0	26.9	27.1	10.3
Control + Both	45.1	**16.6**	27.3	4.3
Dysarthric + Control	48.0	21.3	38.1	5.9
All 3 models	**42.2**	19.1	29.0	7.4

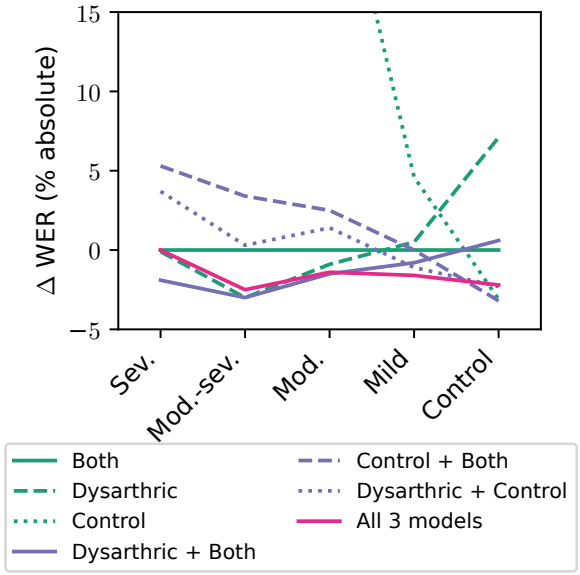


Figure 4: *WER results on UA-Speech: Lattice combination of the LF-MMI systems from Figure 1, which were trained on only dysarthric speakers, only control speakers, or both sets. Shown is the absolute WER difference to the model trained on both dysarthric and control speech (solid green line).*

combined systems improve substantially on dysarthric speakers for isolated word recognition and are better than any of the individual results. Combining all three systems is best with a WER of 42.2% compared to the previous best of 49.2% in the system trained on both dysarthric and control speech. There are no improvements for dysarthric speech on the sentence task, but as long as the system trained on both sets of speakers is included, the results are close to the previous one. As hypothesised, the model combination approach allows to also improve on the control speakers, in one instance even outperforming the previous best result. As long as the model trained only on control speakers is included in the ones to combine, the results are better than from that system alone.

Similarly, Figure 4 shows absolute WER improvements over the baseline LF-MMI model trained on both dysarthric and control speakers of UA-Speech from Figure 1. We observe that when we include the system trained on only control speakers in the ones to combine, the good performance on those speakers is again maintained. The performance on dysarthric speakers can even improve slightly and we see the best results (38.0% WER across the dysarthric speakers, 18.5% on control) when combining all three LF-MMI systems from Figure 1.

4.4. Severity-conditioned models

It appears tempting to take this model combination strategy further and not only combine separate acoustic models on dysarthric and control speakers. We also considered training 5 separate severity-specific models on only the data of the respective severity (or the control speech), bridging the gap to fully speaker-dependent systems. We evaluate each of the 5 models on the corresponding test data, which assumes knowing the test speaker's severity in advance. As Table 3 shows, except on the less variable control and mildly dysarthric speech, these individual models perform worse than the models trained on all dysarthric or all data. We hypothesise that this is because the *dysarthric* and *both* models benefit both from having more data overall and from similarities between the different severity levels.

Table 3: *WER results on UA-Speech: We train 5 separate acoustic models for each severity and the control speakers and evaluate them on the corresponding test data. This is better than combining these 5 models. Neither of these beat the LF-MMI systems from Figure 1 (green).*

Model	Sev.	Mod.-sev.	Mod.	Mild	Control
Both	66.5	40.3	35.2	**24.1**	20.7
Dysarthric	**66.4**	**37.3**	**34.3**	24.6	27.8
Control	96.4	80.6	62.5	28.7	**17.6**
5 separate	73.6	41.8	39.3	24.8	**17.6**
5 combined	79.1	49.5	42.1	25.8	22.4

However, in this case there is also no improvement from combining these 5 models. This is easily explained because the individual models do not generalise well to other data and yield high error rates on speakers of different severity, which does not leave much room for improvement in the combined model.

5. Conclusion

Combining automatic speech recognition (ASR) systems trained on different groups of speakers can improve recognition results on dysarthric speech and partially offset the drop in performance on control speech observed when training models on only dysarthric or both dysarthric and control speakers compared to a model trained on control speakers only. We found this to be the case for isolated word recognition on the Torgo and UA-Speech corpora, but on the Torgo sentence task we did not see further improvements on dysarthric speech. Model combination thus provides a good method to handle the acoustic variations between dysarthric and control speakers and could pave the way for ASR systems that can deal well with a wider range of speech conditions. This approach could also be extended to handle pronunciation variation and a combination of the two is a good direction for future research.

6. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No 766287 (TAPAS).

7. References

[1] K. T. Mengistu and F. Rudzicz, "Adapting Acoustic and Lexical Models to Dysarthric Speech," in *Proc. ICASSP*, 2011, pp. 4924–4927.

[2] 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 *Proc. Interspeech*, 2013, pp. 3642–3645.

[3] C. Bhat, B. Das, B. Vachhani, and S. K. Kopparapu, "Dysarthric Speech Recognition Using Time-delay Neural Network Based Denoising Autoencoder," in *Proc. Interspeech*, 2018, pp. 451–455.

[4] F. Xiong, J. Barker, and H. Christensen, "Phonetic Analysis of Dysarthric Speech Tempo and Applications to Robust Personalised Dysarthric Speech Recognition," in *Proc. ICASSP*, 2019, pp. 5836–5840.

[5] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating Dysarthric Speech for Training Data Augmentation in Clinical Speech Applications," in *Proc. ICASSP*, 2018, pp. 6009–6013.

[6] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric Speech Database for Universal Access Research," in *Proc. Interspeech*, 2008, pp. 1741–1744.

[7] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources & Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[8] 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 *Proc. ASRU*, 2011.

[9] 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 *Proc. Interspeech*, 2016, pp. 2751–2755.

[10] E. Hermann and M. Magimai.-Doss, "Dysarthric Speech Recognition with Lattice-Free MMI," in *Proc. ICASSP*, 2020, pp. 6109–6113.

[11] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio Augmentation for Speech Recognition," in *Proc. Interspeech*, 2015, pp. 3586–3589.

[12] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes Risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.

[13] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proc. ASRU*, 1997, pp. 347–354.

[14] E. Fosler-Lussier, "Contextual Word and Syllable Pronunciation Models," in *Proc. ASRU*, 1999.

[15] M. Razavi and M. Magimai.-Doss, "An HMM-based formalism for automatic subword unit derivation and pronunciation generation," in *Proc. ICASSP*, 2015, pp. 4639–4643.

[16] M. Razavi, R. Rasipuram, and M. Magimai.-Doss, "Towards weakly supervised acoustic subword unit discovery and lexicon development using hidden Markov models," *Speech Communication*, vol. 96, pp. 168–183, 2018.

[17] J.-L. Durrieu, J.-P. Thiran, and F. Kelly, "Lower and upper bounds for approximation of the Kullback-Leibler divergence between Gaussian Mixture Models," in *Proc. ICASSP*, 2012, pp. 4833–4836.

[18] H. Kamper, "Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models," in *Proc. ICASSP*, 2019, pp. 6535–6539.