DYSARTHRIC SPEECH RECOGNITION WITH LATTICE-FREE MMI

Enno Hermann[1,2] *and Mathew Magimai.-Doss*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École polytechnique fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

Recognising dysarthric speech is a challenging problem as it differs in many aspects from typical speech, such as speaking rate and pronunciation. In the literature the focus so far has largely been on handling these variabilities in the framework of HMM/GMM and cross-entropy based HMM/DNN systems. This paper focuses on the use of state-of-the-art sequence-discriminative training, in particular lattice-free maximum mutual information (LF-MMI), for improving dysarthric speech recognition. Through a systematic investigation on the Torgo corpus we demonstrate that LF-MMI performs well on such atypical data and compensates much better for the low speaking rates of dysarthric speakers than conventionally trained systems. This can be attributed to inherent aspects of current speech recognition training regimes, like frame subsampling and speed perturbation, which obviate the need for some techniques previously adopted specifically for dysarthric speech.

***Index Terms*—** Speech recognition, pathological speech processing, dysarthria, LF-MMI.

1. INTRODUCTION

Neurodegenerative diseases like Parkinson's or amyotrophic lateral sclerosis (ALS) not only reduce speech intelligibility, but affect the entire motor system. Assistive systems that recognise such pathological speech could therefore help carry out daily tasks, such as switching on the light or changing TV channels, that are otherwise very difficult for people with limited motor control. Although considerable progress has been made in the field of automatic speech recognition (ASR), it has been found that current commercial and open-source ASR systems still perform poorly on pathological speech data [11]. This highlights the need for further research in this area that results in tangible improvements in mainstream speech technology and thus directly improves the quality of life for people with speech disorders.

Given the scarcity of pathological speech datasets, there has been an emphasis on adapting ASR models trained on typical speech [2, 10, 13, 23]. Other works investigated transforming pathological speech to be more similar to typical speech, for example with speech enhancement methods [1] or by adjusting speech tempo [22]. Alternatively, Jiao et al. [6], Xiong et al. [22] have also employed data augmentation techniques to create additional, artifical dysarthric speech data. As many speech disorders affect the movement of articulators in the vocal tract, modelling articulatory information has been found to be beneficial [5, 19, 24].

Most previous works on dysarthric speech recognition have been in the framework of maximum likelihood trained hidden Markov model (HMM)/Gaussian mixture model (GMM) models or hybrid HMM/deep neural network (DNN) models trained with a frame-level cross entropy objective. However, as ASR is a sequence modelling problem, recent state-of-the-art systems are increasingly trained with sequence-discriminative loss functions, especially lattice-free maximum mutual information (LF-MMI) [17]. The use of such sequence-discriminative criteria has not been sufficiently explored in the context of pathological speech yet. LF-MMI has previously been applied to dysarthric speech [22], but its performance in comparison with other methods has not yet been analysed in detail.

Multiple now common techniques employed for performance or efficiency reasons in LF-MMI training and other state-of-the-art models, such as frame subsampling [21] and speed perturbation [9], potentially also give performance benefits especially on dysarthric speech. For frame subsampling, only every third frame is preserved during training and decoding for a substantial speedup. At training time, this sampling is repeated with different offsets, so that the model still sees every frame. Similarly, for dysarthric speech recognition it was suggested to increase the frame shift of dysarthric speaker during feature extraction to compensate for their lower speaking rates [4]. Speed perturbation augments the training data with multiple (usually 2) copies of itself with slightly modified speed to make models more robust to different speaking rates and to increase the amount of training data, which is crucial for neural network training on small corpora. It could thus also help with the much larger speaking rate variability found in dysarthric speech. We therefore focus our analysis on these techniques.

We evaluate our systems on the Torgo corpus of dysarthric

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).

speech [20] and unlike previous works split the evaluation between isolated and multi-word utterances to obtain more informative results. We show that time-delay neural network (TDNN) acoustic models trained with the LF-MMI objective give state-of-the-art results and especially reduce the number of insertion errors. ASR systems commonly insert many spurious words when recognising dysarthric speech [12] because it is often much slower than the speech they are typically trained on.

The remainder of the paper is organised as follows. Section 2 describes the Torgo corpus and the ASR systems that we train on it. In Section 3 we present our results and analyse the strong performance of LF-MMI systems. Section 4 concludes the paper and summarises the main contributions.

2. EXPERIMENTAL SETUP

2.1. Dataset

We used the Torgo corpus of dysarthric speech [20], which contains about 15 hours of recordings from 15 speakers. There are 8 mostly severely dysarthric speakers (total of 6 hours of speech) and 7 control speakers (total of 9 hours of speech) with no speech disorder.

Participants were asked to do different recordings tasks. For ASR training we included only the isolated word and sentence recordings in line with previous works. We further discarded utterances that had no transcriptions or that were too short to contain any speech.

Table 1: Torgo corpus statistics.

Total utterances	16394
Total unique utterances	971
Total multi-word utterances	4161
Total unique multi-word utterances	356

Table 1 provides statistics on the utterances that we included for ASR training. It shows that the number of unique utterances is small, meaning that many are repeated within and across speakers [25]. About 75% of utterances consist of isolated words, among which are many minimal pairs, such as *rate* and *raid* without context that would disambiguate them. In fact, for 88% of isolated words there is at least one other word with a pronunciation within an edit distance of 1. The average closest edit distance is 1.16. This makes the corpus very useful for automatic assesment of speech intelligibility and similar tasks, but more challenging for ASR. Even for speakers without any speech disorders correctly recognising the minimal pairs is expected to be difficult.

2.2. Systems

2.2.1. HMM/GMM

We used the open-source Kaldi speech recognition toolkit [16] for all our experiments. We followed the typical development pipeline to train a subspace GMM (SGMM) [15] baseline model on 39-dimensional MFCC+Δ+$\Delta\Delta$ features. We used the hyperparameters and provided Kaldi recipe of España-Bonet and Fonollosa [4].[1] The code for our experiments is publicly available.[2]

We also chose to model phones independent of their position in words as suggested by Joy and Umesh [7] because of data sparsity and because the lower speaking rates lead to reduced coarticulation effects.

2.2.2. HMM/DNN

It is important to avoid excessive hyperparameter tuning on Torgo, which would easily lead to overfitting because of the little amount of data and the cross-validation approach for evaluation. Our hybrid HMM/DNN models are therefore based on the well-tuned Kaldi recipes for the 5-hour subset of the Librispeech corpus.[3]

The main system that we analyse below is a 13-layer factorised TDNN model [18] trained with the sequence-discriminative LF-MMI objective function. For comparison, we also trained a 9-layer TDNN-LSTM model with a conventional frame-wise cross-entropy (CE) objective. As is the default in Kaldi, we trained the HMM/DNN models on speed perturbed data for which the original data is augmented by perturbed versions at 0.9 and 1.1 times the original speed.

2.3. Evaluation protocol

As there are only 8 dysarthric speakers and their degree of dysarthria varies a lot, we maintain the leave-one-out cross-validation training procedure where each of the 15 speakers is evaluated separately and models are trained on the remaining 14 speakers.

Unlike previous works, we split the evaluation of isolated- and multi-word utterances by treating the two tasks separately. Otherwise the results would be less informative because of the different challenges in these two tasks. Most prior research on dysarthric speech recognition has focused on isolated words because of the lack of datasets that include continuous speech. However, we do not see this as a limitation. Speaking can require a significant effort from severely dysarthric speakers and to maximise communication efficiency they might choose to use shorter utterances. For example, the homeService corpus [14] was recorded in realistic home environments and contains simple 1–2 word commands like *"Volume up"*. Most

[1] https://github.com/cristinae/ASRdys

[2] https://github.com/idiap/torgo_asr

[3] https://github.com/kaldi-asr/kaldi/tree/master/egs/mini_librispeech

other pathological speech corpora are not recorded specifically for ASR, but for speech assessment purposes, which explains why the sentences in the Torgo corpus are often long and unnatural.

The language models (LMs) are different for the two evaluation tasks. For isolated word recognition it is a unigram model containing all around 600 possible words, which may be preceded or followed by silence. In Section 3 we also evaluate the effect of constraining the decoding grammar so that the output is always a single word. For sentences we use a bigram LM that is trained on all the sentence data. In both cases we trained the LMs on the data of all speakers, they thus also include that of the test speaker. This is impossible to avoid because there is very high text overlap between speakers as explained in Section 2.1 and in this way we focus on improving the acoustic model (AM). Improvements on the LM side could only be obtained with LMs trained on large external corpora because the Torgo corpus is so small [25].

The language model weight for decoding in each experiment was set to the average of the best values obtained for each control speaker.[4]

3. RESULTS AND ANALYSIS

Table 2 shows the results for evaluating the baseline systems described in Section 2.2 separately on isolated-word and multi-word utterances. The word error rates (WERs) are averaged over dysarthric and control speakers for readability, but there can be substantial variation within these groups as illustrated in Figure 1.

As hypothesised, WERs on the isolated word task are high even for the control speakers because of the inherent challenge in distinguishing minimal pairs without further context. On the other hand, the sentences are recognised with only very few errors for the control group and mildly dysarthric speakers because the strong LM renders this task quite easy. Despite this advantage, WERs for moderate to severely dysarthric speakers are much higher in this case. This highlights that there is still a lot of room for improvement just on the AM side.

Table 2: WER for different systems, averaged for dysarthric (Dys) and control (Con) speakers, respectively. Every second row shows the effect of restricting the output to a single word during isolated word recognition.

	1-word LM	**Isolated**		**Sentences**	
		Dys	Con	Dys	Con
SGMM	–	56.1	19.4	41.5	4.4
	✓	47.2	18.7	–	–
CE	–	53.6	24.6	38.0	9.3
	✓	44.9	24.0	–	–
LF-MMI	–	49.2	24.0	25.9	7.9
	✓	43.0	22.0	–	–

[4]In Kaldi it is common to perform a grid search for language model weight and word insertion penalty at decoding time even on test data because differences are often small, but with the cross-validation setup on the Torgo corpus it is important to avoid tuning any parameters on a specific dysarthric speaker's data because the impact might be much larger.

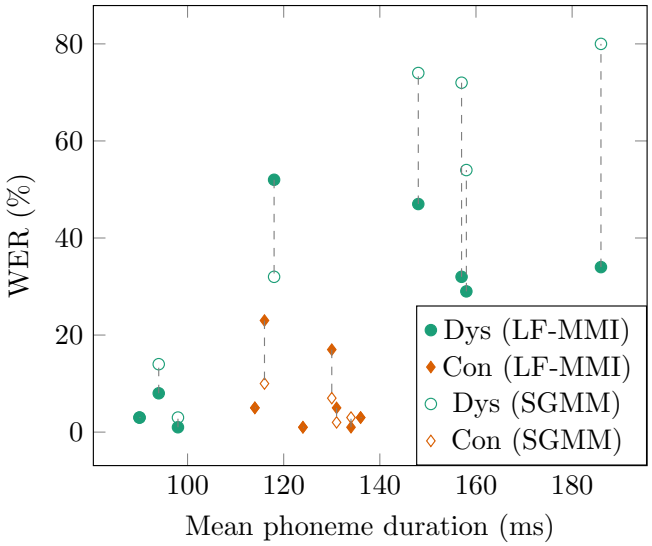

Fig. 1: Association between mean phoneme duration and WER for dysarthric (dys) and control (con) speakers. The WER results are from the LF-MMI and SGMM models on the Torgo sentence task. Mildly dysarthric speakers also achieve very low WER. Dashed lines connect results from the same speaker.

LF-MMI training always helps for dysarthric speakers except for one compared to both SGMM and CE-based models. However, the SGMM outperforms the neural network models on the control speakers, perhaps because on such a small corpus the neural networks are more sensitive to the additional variability in the training data introduced by the dysarthric speech. Indeed, if the LF-MMI system is trained for the same number of epochs and with the same hyperparameters on the control speech only, it performs much better, with WERs of 18.3% on the isolated words and 2.9% on the sentences averaged over all control speakers.

The large improvements of LF-MMI models on the sentences are because they make much fewer insertion errors, indicating that they are better equipped to handle very low speaking rates. Figure 1 shows how speaking rate and WER are correlated. We approximate speaking rate information by computing mean phoneme durations from forced alignments of the training data with the methodology of Xiong et al. [22]. It can be seen that dysarthric speakers have the lowest speaking rates and also the highest WERs. There are another three mildly dysarthric speakers that have normal or even slightly shorter phoneme durations that the ASR system recognises

very well.

For the sake of completeness, we also evaluated all speech grouped together and with the methodology of España-Bonet and Fonollosa [4]. We substantially outperform their best results obtained with hybrid HMM/DNN systems that were the previous state of the art on this corpus.

In the following sections we will analyse the performance of LF-MMI in more detail.

3.1. Constrained language model

Every second row in Table 2 also shows the results of forcing the decoder to output only a single word for the isolated-word utterances. This consistently improves results across speakers, in particular for the most severely dysarthric ones because their very low speaking rate otherwise leads to a large number of insertion errors. This suggests that the WER on the sentence task where the number of words is not known a priori could also be reduced by appropriately tuning the word insertion penalty during decoding for each speaker or utterance. However, this penalty would need to be set in an unsupervised manner by automatically estimating speaking rates.

3.2. Speed perturbation

As mentioned in Section 2, the training data for the hybrid HMM/DNN systems was augmented with two speed-perturbed copies. To test the effects of this we trained LF-MMI models on the original data only, but increasing the number of epochs by a factor of 3 to compensate for the lower amount of training data. Results, shown in Table 3, are overall still better than SGMMs and cross-entropy models, maintaining a big reduction in insertion errors as indicated by the sentence results. This suggests that this reduction can at least in part be attributed to the sequence-discriminative objective function.

However, the performance on control speech is better when no speed perturbation is applied. It is then on par with the SGMM results, but still worse than training on speed-perturbed control speech only as observed above. This is perhaps because applying further distortions to dysarthric speech makes the training data too variable to perform well on unimpaired speech.

Table 3: LF-MMI systems trained without speed perturbation still outperform SGMMs. The isolated word results use the constrained LM.

	Speed perturbation	**Isolated**		**Sentences**	
		Dys	Con	Dys	Con
SGMM	–	47.2	18.7	41.5	4.4
LF-MMI	✓	43.0	22.0	25.9	7.9
	–	46.4	21.4	30.2	4.2

3.3. Frame shift

Previous work [4] proposed to apply a frame shift of 15 ms to the dysarthric data while maintaining the usual 10 ms for the control speech to compensate for the lower speaking rates of dysarthric speakers. However, the good performance of the LF-MMI systems suggests that it might not be necessary in these models. Our results in Table 4 confirm that a constant frame shift of 10 ms for the entire data does not reduce performance on dysarthric speech. This is useful because the constant frame shift does not require prior knowledge about the speaker.

Table 4: Applying a 15 ms frame shift to dysarthric and 10 ms to control speakers compared with a constant 10 ms shift throughout. The isolated word results use the constrained LM.

	Frame shift	**Isolated**		**Sentences**	
		Dys	Con	Dys	Con
LF-MMI	15/10 ms	43.0	22.0	25.9	7.9
	10 ms	42.9	22.5	25.9	8.1

4. CONCLUSIONS

We applied LF-MMI training to dysarthric speech and demonstrated that it also yields strong results on such a small and atypical dataset. Our results are a new state of the art on the Torgo corpus that can serve as strong baselines for further research. When analysing these improvements we found that especially insertion errors are reduced, which are otherwise very frequent due to the low speaking rates of dysarthric speakers. Contributing factors to this are the frame subsampling of LF-MMI, data augmentation with speed perturbed speech and the sequence-discriminative objective function itself. Further analysis is required to determine the importance of each of these factors. While hybrid HMM/DNN systems reduce the number of errors on dysarthric speech, we observed that they do not work as well for control speakers as systems trained only on control speech or a traditional HMM/GMM system. This calls for further research into improving speech recognition for everyone.

In future work we plan to focus on additional ways for making ASR systems more invariant to speaking rate variability. For example, segmental training was recently found to be an effective way to normalise segment durations [3]. Does this also apply in the case of dysarthric speech? We will also cross-evaluate models that were trained on the UA-Speech [8] and homeService [14] corpora. They both contain only 1–2 word utterances as well and can therefore be compared with our isolated word recognition task.

References

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

[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*, pages 3642–3645, 2013.

[3] S. P. Dubagunta and M. Magimai.-Doss. Segment-level Training of ANNs Based on Acoustic Confidence Measures for Hybrid HMM/ANN Speech Recognition. In *Proc. ICASSP*, pages 6435–6439, 2019.

[4] C. España-Bonet and J. A. R. Fonollosa. Automatic Speech Recognition with Deep Neural Networks for Impaired Speech. In *Proc. IberSpeech*, pages 97–107, 2016.

[5] S. Hahm, D. Heitzman, and J. Wang. Recognizing Dysarthric Speech due to Amyotrophic Lateral Sclerosis with Across-Speaker Articulatory Normalization. In *Proc. Workshop on Speech and Language Processing for Assistive Technologies*, pages 47–54, 2015.

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

[7] N. M. Joy and S. Umesh. Improving acoustic models in TORGO dysarthric speech database. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(3):637–645, 2018.

[8] 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*, pages 1741–1744, 2008.

[9] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur. Audio Augmentation for Speech Recognition. In *Proc. Interspeech*, pages 3586–3589, 2015.

[10] K. T. Mengistu and F. Rudzicz. Adapting Acoustic and Lexical Models to Dysarthric Speech. In *Proc. ICASSP*, pages 4924–4927, 2011.

[11] M. Moore, H. Venkateswara, and S. Panchanathan. Whistle-blowing ASRs: Evaluating the Need for More Inclusive Speech Recognition Systems. In *Proc. Interspeech*, pages 466–470, 2018.

[12] M. Moore, M. Saxon, H. Venkateswara, V. Berisha, and S. Panchanathan. Say What? A Dataset for Exploring the Error Patterns That Two ASR Engines Make. In *Proc. Interspeech*, pages 2528–2532, 2019.

[13] M. B. Mustafa, S. S. Salim, M. N. Al-Qatab, and B. E. Siong. Severity-Based Adaptation with Limited Data for ASR to Aid Dysarthric Speakers. *PLoS ONE*, 9(1):1–11, 2014.

[14] M. Nicolao, H. Christensen, S. Cunningham, P. Green, and T. Hain. A Framework for Collecting Realistic Recordings of Dysarthric Speech - the homeService Corpus. In *Proc. LREC*, 2016.

[15] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas. The subspace Gaussian mixture model - A structured model for speech recognition. *Computer Speech & Language*, 25:404–439, 2010.

[16] 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.

[17] 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*, pages 2751–2755, 2016.

[18] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur. Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. In *Proc. Interspeech*, pages 3743–3747, 2018.

[19] F. Rudzicz. Articulatory Knowledge in the Recognition of Dysarthric Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):947–960, 2011.

[20] F. Rudzicz, A. K. Namasivayam, and T. Wolff. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources & Evaluation*, 46(4):523–541, 2012.

[21] H. Sak, A. Senior, K. Rao, and F. Beaufays. Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition. In *Proc. Interspeech*, pages 1468–1472, 2015.

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

[23] E. Yılmaz, M. Ganzeboom, C. Cucchiarini, and H. Strik. Combining Non-pathological Data of Different Language Varieties to Improve DNN-HMM Performance on Pathological Speech. In *Proc. Interspeech*, pages 218–222, 2016.

[24] E. Yılmaz, V. Mitra, C. Bartels, and H. Franco. Articulatory Features for ASR of Pathological Speech. In *Proc. Interspeech*, pages 2958–2962, 2018.

[25] Z. Yue, F. Xiong, H. Christensen, and J. Barker. Exploring Appropriate Acoustic and Language Modelling Choices for Continuous Dysarthric Speech Recognition. In *Proc. ICASSP (to appear)*, 2020.