Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech

H. Christensen[1], *M. B. Aniol*[2], *P. Bell*[2], *P. Green*[1], *T. Hain*[1], *S. King*[2], *P. Swietojanski*[2]

[1]Department of Computer Science, University of Sheffield, Sheffield, S1 4DP, UK
{h.christensen,p.green,t.hain}@dcs.shef.ac.uk
[2]Centre for Speech Technology Research, University of Edinburgh, Edinburgh, EH8 9AB, UK
{m.b.aniol,peter.bell,simon.king, p.swietojanski}@ed.ac.uk

Abstract

Recently there has been increasing interest in ways of using out-of-domain (OOD) data to improve automatic speech recognition performance in domains where only limited data is available. This paper focuses on one such domain, namely that of disordered speech for which only very small databases exist, but where *normal* speech can be considered OOD. Standard approaches for handling small data domains use adaptation from OOD models into the target domain, but here we investigate an alternative approach with its focus on the feature extraction stage: OOD data is used to train feature-generating deep belief neural networks. Using AMI meeting and TED talk datasets, we investigate various tandem-based speaker independent systems as well as maximum a posteriori adapted speaker dependent systems. Results on the UAspeech isolated word task of disordered speech are very promising with our overall best system (using a combination of AMI and TED data) giving a correctness of 62.5%; an increase of 15% on previously best published results based on conventional model adaptation. We show that the relative benefit of using OOD data varies considerably from speaker to speaker and is only loosely correlated with the severity of a speaker's impairments.

Index Terms: Speech recognition, Tandem features, Deep belief neural network, Disordered speech

1. Introduction

Large vocabulary automatic speech recognition (ASR) research has in recent years been driven partly by increasingly bigger datasets. However, there are many domains in which only small amounts of *in-domain* data is available for training purposes; either because they represent challenging acoustic environments, where recordings are difficult to obtain, or they represent rarely occurring speaking styles, such as highly emotional speech. Research into how to increase performance of ASR systems through the use of readily available large *out-of-domain* (ODD) datasets is therefore receiving a lot of interest.

This paper is concerned with one such small data domain, namely the recognition of *disordered* speech such as is often needed when working in the area of assistive technology for people with severe physical impairments. Their underlying neuro-motor conditions tend to co-occur with speech articulatory motor control problems, which causes speech disorders: this condition is known as *dysarthria*.

People with disordered speech will often be able to communicate with family, friends and carers with little or no problems, whilst at the same time being close to unintelligible to un-familiar listeners [1]. ASR systems that have only been trained on *normal* speech, can in this respect be regarded as 'un-familiar' listeners and the resulting poor performances renders off-the-shelf systems unusable for all but speakers with the most mild impairments [2]. As a results, ASR systems must be designed specifically for the target domain of dysarthric speech if not for the individual speaker. At the same time this domain is and will remain inherently 'small' in terms of data because of a lack of dysarthric speakers, and because each speaker can find it upsetting and difficult to produce a substantial amount of speech.

In this paper we investigate ways of boosting the recognition of dysarthric speech by treating *normal* speech as OOD.

1.1. Previous work and contributions of presented work

The standard way of making use of OOD knowledge has taken the form of using models trained on OOD data and performing speaker adaptation like maximum a posteriori (MAP) [3] and maximum likelihood linear regression (MLLR) [4] to a target domain. Several studies have investigated adaptation techniques with disordered speech [5, 6]. In [7] we investigated the use of MAP adaptation from a normal speech speaker independent model onto the dysarthric domain; likewise, [8] also explored the use of OOD models with MAP with some success.

The alternative approach we present here uses the OOD training data at the feature extraction stage to improve the quality of the features by training deep belief neural networks (DNNs) [9] for extracting features for tandem-based systems [10]. We will explore both the standard tandem features and the newly introduced 'multi-level adaptive network' (MLAN) features, which add a further neural network layer to the tandem features [11]. Recent work has shown the promise of these techniques for multiple cross-domain scenarios, such as cross-language ASR [12, 13] and cross-domain ASR [11, 14]. Aniol [15] showed some promising results for disordered speech.

Despite the obvious similarities between such cross-language studies and the *normal* vs. *dysarthric* framework proposed here, there are notable difference which makes the application to this new domain non-trivial and worth investigating. The degree and type of inter- and intra-speaker variabilities which occur for non-impaired speakers (even if speaking in different languages). Dysarthric speakers will typically only have a reduced phone set they can utilise, and there is often a large variation between each instance of a word. Other factors not present for multi-language and multi-accent domains are the effect of tiredness and general physical wellness.

In the remainder of the paper we describe our experimental setup (Section 2) and results (Section 3) addressing the question of which features and data work best for the OOD pre-training framework and to what degree this depends on the speaker.

2. Experimental setup

The underlying methodology of this study compares a range of different systems of increasing complexity, each of which use OOD data in a different way. Each system has been individually optimised and performance is compared using percent correct on the UAspeech isolated word[1] task. The UAspeech database [16] was chosen as it is one of the largest databases available for English dysarthric speech and with 15 speakers has a relatively large variation of severity of speech impairment. For the OOD data we have chosen to work with two different OOD datasets: the 'TED talk' [17] and the AMI meeting room datasets [18] and their corresponding pre-trained feature extraction front-ends. Further details about the data can be found in Section 2.1.

Although previous work outlined above has shown MLAN feature-based systems to outperform tandem-based systems in OOD frameworks, it is unclear to what degree this ports to the normal vs. disordered scenario, and we therefore chose to include both types of feature generation framework in the study. For comparison, we have also investigated the effect of speaker adaptation (using MAP) and alongside this, how standard *speaker dependent* (SD) systems fare with the OOD and in-domain data. Finally, a number of baseline systems were incorporated in the study based on ordinary PLP-based *speaker independent* (SI) systems. More details about the individual features and training strategies are given in sections 2.2 and 2.3. Section 2.4 provides information on decoding and scoring.

2.1. Data

2.1.1. In-domain dataset: UAspeech

The UAspeech database contains synchronised audio and visual streams from 15 speakers (4 female and 11 male). The dysarthric speakers were asked to repeat single words from 5 groups: 10 digits, 29 Nato alphabet letters, 19 command words ('delete', 'enter' etc.), 100 common words ('the','will' etc.), and 300 uncommon words chosen to be phonetically rich and complimentary to the remaining words ('Copenhagen','chambermaid' etc.). In total, each speaker has produced around 70 minutes of speech. Full details of the corpora can be found in [16].

The speakers all have a type dysarthric speech, and accompanying the database are percent intelligibly scores as obtained from listening tests with unfamiliar listeners. These range from 4% to 95%. Following previously published work using the UAspeech for ASR (e.g. [5]) the data was divided into training and test data with a 2:1 split, using blocks 1 and 3 for training and block 2 for testing.

2.1.2. Out-of-domain datasets: TED and AMI

The 'TED talk' dataset [17] consists of a series of lectures comprising a total of 138 hours of training data. Most lectures have a single American English native speaker speaking in a well-rehearsed, planned fashion which - although not read - bears strong similarity to data types such as broadcast news. The recordings are all from close-talking microphones on headsets and of high quality. In contrast the AMI dataset (126.8 hours) [18] consists of meeting room headset microphone recordings with multiple speakers per session. The speech is conversational of nature and there is a relatively large variety in accent, (although all speakers can be considered fluent in English).

2.2. Feature extraction

Two different tandem-based feature extraction frameworks have been investigated: a standard tandem-based feature generator and an MLAN-based generator.

The term 'tandem features' refers to feature vectors compromised of a conventional feature vector - in our case a 13-dimensional PLP vector with added first and second order derivatives - augmented with features extracted from a pre-trained DNN [10]. Recently an extension to the original tandem features was proposed, Multi-level Adaptive Networks (MLAN) [11] where tandem features are passed through a further neural network trained on phone-level labels, before being augmented with the original PLP features. For each dataset (AMI, TED and UAspeech), both tandem and MLAN features have been extracted.

All TED and UAspeech networks share the same architecture with 4 layers and 1024 hidden units in each with the same phone set modelled at the output. As was found in [12], with appropriate regularisation, good results can be obtained even for as little as 1 hour of training data. PCA was applied to all output posteriors in order to de-correlate and to reduce dimensionality from 45 to 30. The nets were trained on globally-normalised PLP features with added energy and first and second order derivatives. For further details on how the TED networks are trained, please refer to [19]. The AMI networks were trained on filterbank outputs and the AMI features are stacked bottleneck features as described in [18].

Because of the difference in style of data as well as their associated feature extraction networks, we expect the AMI and the TED OOD feature generators to be complementary to each other to some degree. This can be illustrated by looking at cross-recognition results: applying the TED test sets to the corresponding TED models gives a word error rate (WER) of 24.9%, whereas when applying the TED test set to the AMI models a WER of 30.7% is obtained [20].

2.3. Acoustic modelling

All Hidden Markov Models (HMMs) were trained using the maximum likelihood (ML) criterion. State-clustered, triphones having Gaussian mixture models with standard mixing-up to 16 components per state were used. Both the tandem and the MLAN features used are 69-dimensional and the HMM systems were trained starting from a monophone system and subsequently doing triphone training. Systems based on single-pass-retraining were also tested but overall very little difference was found between the two different training strategies. All final tri-phone systems were optimised with respect to the number of states with most systems achieving the best performance around 1300-1500 states for MLAN based systems and around 500 for tandem based systems.

2.4. Decoding

The UAspeech task is single word recognition and it was decided to follow the decoding strategy deployed in [7]. A uniform language model was used, with a word grammar network containing silence models at the start and end, and all possible

[1] Isolated word recognition is an appropriate task for dysarthric speech as it reflects 'command-and-control' applications, which are particularly relevant for this group of people.

System	In-domain	Out-of-domain	
	UAspeech	AMI	Ted
PLP ML-SD	50.9	-	-
PLP ML-SI	50.6	22.4	
+SD-MAP	54.1	40.1	
Tandem ML-SD	55.8	55.9	54.4
Tandem ML-SI	56.0	57.5	55.0
+SD-MAP	57.9	61.8	60.8

Table 1: *Word accuracy rates for UAspeech and AMI based PLP and Tandem system. All systems are tested on the UAspeech test set. See text for system name descriptions.*

test words in parallel. The dictionary contains 256 entries (the number of different words in the test set) with an average of 1.66 pronunciations per word.

3. Results

Table 1 shows all the main PLP and tandem-based results in percent correct as averaged over all speakers, when tested on the UAspeech test set using models containing only UAspeech (i.e., in-domain data only) and OOD (AMI and TED).

Before discussing the benefits of using OOD features generated from pre-trained DNNs, it is interesting to look at the UAspeech-only results in comparison to previously published work. The table shows the 'PLP'-based results, which are here for reference and were first published in [7] – these are the speaker dependent (*PLP ML-SD*), speaker independent (*PLP ML-SI*) and speaker adapted (*PLP ML-SI+SD-MAP*) systems, with the *PLP ML-SI* system being the previously highest scoring system with 54.1%. New for the current work are the tandem-based results, which all collectively improve on the previous results with between 9 and 12% relatively. Results for a similar SD tandem system is reported in [15] with an overall correctness of 52.3% in comparison to the 55.8% correctness achieved for the *Tandem ML-SD* in this study.

3.1. Effect of OOD feature generators

The OOD PLP and tandem results are also shown in Table 1. As explained in the introduction, using features extracted from DNNs pre-trained on ODD is an alternative to the conventional method of doing adaptation from the OOD models to the target domain. Results of both approaches are given in Table1: for the AMI data, the tandem-based system show an improvement in comparison with the PLP-based MAP system, 61.8% vs. 40.1%, a relative improvement of over 54%!. In general, comparing the OOD-based results to the UAspeech-only baseline results in Table 1 shows an increase in performance for all systems *except* the *TED Tandem ML-SI* system which has a lower correctness than the corresponding *Tandem ML-SI* system (55.0% vs. 56.0%). For the normal ML systems, the improvements range from 2.7% to 7.3% relatively; for the MAP versions of these systems larger relative improvements are seen - up to 7.9%.

3.2. How to best use OOD

Comparing the tandem and the MLAN-based systems gives some insight into the best ways of using the OOD data.

Table 2 introduce the MLAN results, which are all better than the tandem systems in Table 1 to which they are comparable. In terms of which OOD data and feature generation to use, for the best tandem-based systems we observe that UAspeech (57.9%) < UAspeech+TED (60.8%) < UAspeech+AMI (61.8%) and correspondingly for the best MLAN-based system we get that UAspeech+TED (61.3%) < UAspeech+AMI (61.8%) < UAspeech+AMI+TED (62.5%). These conclusions are based on the SD-MAP systems. However, the picture is less clear from the ML-SI systems where TED is the worst for the tandem features, but the second best choice of OOD for the MLAN-based system.

System	Out-of-domain		
	AMI	TED	AMI+TED
MLAN ML-SD	57.8	57.1	58.1
MLAN ML-SI	58.1	58.6	60.1
+SD-MAP	61.8	61.3	62.5

Table 2: *Word accuracy rates for MLAN-based systems. All systems are tested with the UAspeech test set. See text for system name descriptions.*

For the AMI dataset, there is only a small difference between the tandem and the MLAN system, but for TED, the MLAN system is better than the tandem system (55.0% vs. 58.6%). When looking at the MAP adapted system, the picture is again less clear with all systems having performances between 60.8% and 61.8%. The only exception is the overall best performing system, which is the *AMI+TED MLAN SI+SD-MAP* with a correctness of 62.5%. This is a relative increase of 15.5% compared to the previously best published result of 54.1% [7].

3.3. Inter-speaker variabilities

In [7] we observed a large variation from speaker to speaker as to which system was the best for them. For the systems presented here, the best system for any of the 15 UAspeech speakers is always one of the MAP adapted systems. However, which data and feature set is best varies with 3, 3, 1, 2, and 7 speakers favouring the *AMI tandem ML-SI+SD-MAP*, the *AMI MLAN ML-SI+SD-MAP*, the *TED tandem ML-SI+SD-MAP*, the *TED MLAN ML-SI+SD-MAP* and the *AMI+TED ML-SI+SD-MAP* systems respectively.

We also observe that the benefit of using OOD data varies considerably from speaker to speaker and is only loosely correlated with the severity of a speaker's impairments. For each speaker we compared the performance of the OOD systems with the corresponding UAspeech-only system. We found that although there appears to be an overall decreasing trend where the less severely dysarthric speakers see smaller added benefit from OOD, there are clearly some deviations from this. For example, two speakers with 6 and 7% intelligence respectively, obtain vastly different improvements from the OOD systems: where the former sees very little improvement (7.3%) in performance and the 7% speaker improves with (31.5%).

In [21], we investigate further how the speaker specific variations observed at the phone level posterior probabilities output from the DNNs can be used to learn more accurate, speaker-specific transcriptions.

4. Discussion and conclusions

The work presented here is motivated by our interest in improving the performance of automatic recognition for dysarthric speech - a domain in which only relatively small amounts of data is available. We address the issue by investigating ways of using OOD data (i.e. normal speech) to boost feature generation and thereby the acoustic modelling of dysarthric speech. Tandem and MLAN feature generating front-ends using DNNs have been pre-trained on the TED talk and AMI meeting datasets and tested on the UAspeech isolated word task of dysarthric speech. We have demonstrated a large improvement on previously published results, with an increase of up to 15% for our best system for a MAP adapted MLAN system pre-trained on AMI and TED data. For individual speakers (each with very varying speech impairments and degrees of intelligibility) there is some variability in terms of which OOD and feature type would provide them with the best performing system. For future work we intend to explore ways of improving the training strategies for both the pre-training and the in-domain HMM training stage to better reflect speech impairment characteristics specific to the individual speaker.

5. Acknowledgements

The research leading to these results was supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). The authors wish to Arnab Ghoshal of University of Edinburgh for his advice and expertise.

6. References

[1] K. T. Mengistu and F. Rudzicz, “Comparing humans and automatic speech recognition systems in recognizing dysarthric speech,” in *Proceedings of the Canadian Conference on Artificial Intelligence*, St. John’s Canada, May 25–27 2011.

[2] L. Ferrier, H. Shane, H. Ballard, T. Carpenter, and A. Benoit, “Dysarthric speakers intelligibility and speech characteristics in relation to computer speech recognition.” *Augmentative and Alternative Communication*, vol. 11, pp. 165–174, 1995.

[3] J. Gauvain and C.-H. Lee, “MAP estimation of continuous density hmm: theory and applications,” in *Proceeding HLT’91 Proceedings of the workshop on Speech and Natural Language*, 1992.

[4] C. J. Legetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models,” . *Computer Speech and Language*, pp. 171–185, 1995.

[5] H. V. Sharma and M. Hasegawa-Johnson, “State transition interpolation and map adaptation for hmm-based dysarthric speech recognition,” in *HLT/NAACL Workshop on Speech and Language Processing for Assistive Technology (SLPAT)*, 2010, pp. 72–79.

[6] K. T. Mengistu and F. Rudzicz, “Adapting acoustic and lexical models to dysarthric speech,” in *Proceedings of ICASSP’11*, 2011, pp. 4924–4927.

[7] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, “A comparative study of adaptive, automatic recognition of disordered speech,” in *Proc Interspeech 2012*, Portland, Oregon, US, Sep 2012.

[8] H. V. Sharma and M. Hasegawa-Johnson, “Acoustic model adaptation using in-domain background models for dysarthric speech recognition,” *Computer Speech and Language*, 2012.

[9] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[10] H. Hermansky, D. Ellis, and S. Sharma, “Tandem connectionist feature extraction for conventional HMM systems,” in *Proceedings of ICASSP’00*, Istanbul, Turkey, June 2000, pp. 1635–1630.

[11] P. Bell, M. Gales, P. Lanchantin, X. Liu, Y. Long, S. Renals, P. Swietojanski, and P. Woodland, “Transcription of multi-genre media archives using out-of-domain data,” in *Proceesing of IEEE Workshop on Spoken Language Technology*, Miami, US, Dec 2012.

[12] P. Swietojanski, A. Ghoshal, and S. Renals, “Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR,” in *Workshop on Spoken Language Technology*, Miami, US, December 2012.

[13] Y. Qian, J. Xu, D. Povery, and J. Liu, “Strategies for using MLP based features with limited target-language training data,” in *Proceedings of ASRU*, 2011, pp. 354–358.

[14] A. Stolcke, F. Grezl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, “Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons.” in *Proc. IEEE ICASSP*, vol. 1, Toulouse, France, 2006, pp. 321–324.

[15] M. Aniol, “Tandem features for dysarthic speech recognition,” Master’s thesis, Edinburgh University, United Kingdom, 2012.

[16] 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, 2008, pp. 22–26.

[17] M. Cettolo, C. Girardi, , and M. Federico, “Wit3: Web inventory of transcribed and translated talks,” in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[18] T. Hain, L. Burget, J. Dines, P. N. Garner, F. Grezl, A. el Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, “Transcribing meetings with the amida systems,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 2, pp. 486–498, 2011.

[19] E. Hasler, P. Bell, A. Ghoshal, B. Haddow, F. M. P. Koehn, S. Renals, and P. Swietojanski, “The UEDIN systems for the IWSLT 2012 evaluation,” in *Proceedings of IWSLT2012*, December 2012.

[20] P. Bell, P. Swietojanski, and S. Renals, “Multi-level adaptive networks in tandem and hybrid ASR systems,” in *ICASSP’13*, 2013.

[21] H. Christensen, P. Green, and T. Hain, “Learning speaker-specific pronunciations of disorderes speech,” in *Interspeech’13*, 2013.