Combining Non-pathological Data of Different Language Varieties to Improve DNN-HMM Performance on Pathological Speech

Emre Yılmaz, Mario Ganzeboom, Catia Cucchiarini and Helmer Strik

CLS/CLST, Radboud University, Nijmegen, Netherlands

{e.yilmaz,m.ganzeboom,c.cucchiarini,h.strik}@let.ru.nl

Abstract

Research on automatic speech recognition (ASR) of pathological speech is particularly hindered by scarce in-domain data resources. Collecting representative pathological speech data is difficult due to the large variability caused by the nature and severity of the disorders, and the rigorous ethical and medical permission requirements. This task becomes even more challenging for languages which have fewer resources, fewer speakers and fewer patients than English, such as the mid-sized language Dutch. In this paper, we investigate the impact of combining speech data from different varieties of the Dutch language for training deep neural network (DNN)-based acoustic models. Flemish is chosen as the target variety for testing the acoustic models, since a Flemish database of pathological speech, the COPAS database, is available. We use non-pathological speech data from the northern Dutch and Flemish varieties and perform speaker-independent recognition using the DNN-HMM system trained on the combined data. The results show that this system provides improved recognition of pathological Flemish speech compared to a baseline system trained only on Flemish data. These findings open up new opportunities for developing useful ASR-based pathological speech applications for languages that are smaller in size and less resourced than English.

Index Terms: pathological speech, automatic speech recognition, Flemish, dysarthria, data merging

1. Introduction

Motor speech disorders including dysarthria caused by neuromuscular control problems [1] lead to decreased speech intelligibility and communication impairment [2]. Consequently, the life quality of dysarthric patients is negatively affected [3] and they run the risk of losing contact with friends and relatives and eventually becoming isolated from the society.

Research has shown that intensive therapy can be effective in (speech) motor rehabilitation [4–7], but various factors conspire to make intensive therapy expensive and difficult to obtain. Recent developments show that therapy can be provided without resorting to frequent face-to-face sessions with therapists by employing computer-assisted speech training systems [8]. According to the outcomes of the efficacy tests presented in [9], the user satisfaction appears to be quite high. However, most of these systems are not yet capable of automatically detecting problems at the level of individual speech sounds, which are known to have an impact on speech intelligibility [10–14].

Despite long-lasting efforts to build speaker- and text-independent ASR system for people with dysarthria (cf. Section 2), the performance of state-of-the-art systems is still much worse on this type of speech than on normal speech. One main reason is the lack of pathological speech data to train automatic speech recognition systems which can provide accurate enough recognition and speech assessment.

Training deep neural networks (DNN)-based acoustic models on large amount of pathological data to capture the within- and between-speaker variation is generally not feasible due to the limited size and structure of existing pathological speech databases. The number of recordings in dysarthric speech databases is much smaller compared to normal speech databases. Moreover, these databases contain mostly very restricted speech tasks such as reading out word and sentence lists with varying linguistic complexity.

As a remedy, combining in-domain and out-of-domain English speech data to train DNNs used for feature extraction has been proposed in [15]. In this paper, we describe another such solution to train a better DNN-hidden Markov model (HMM) system for the Dutch language which has fewer speakers and resources compared to English. We investigate combining non-dysarthric speech data from different varieties of the Dutch language to train more reliable acoustic models for a DNN-HMM ASR system. The included varieties are Northern Dutch and Flemish (Southern Dutch) which have the same phonetic alphabet and share a large amount of vocabulary. Most prominent phonetic differences between these varieties are diphthongized long vowels of Northern Dutch and articulation of several consonants. This work has been done in the framework of the CHASING project[1], in which a serious game employing ASR is being developed to provide additional speech therapy to dysarthric patients.

The rest of the paper is organized as follows. Section 2 reports on previous relevant work on ASR of dysarthric speech. Section 3 explains the rationale behind the selection of speech corpora. Section 4 summarizes the fundamentals of DNN-based ASR and details the DNN training scheme applied in this paper. The experimental setup is described in Section 5 and the recognition results are presented in Section 6. Section 7 concludes the paper.

2. Related work

Several researchers have investigated ASR performance on pathological speech. In a very recent work, Lee et al. [16] has reported the ASR performance on Cantonese aphasic speech and disordered voice. A generic DNN-HMM system provided significant improvements on disordered voice and minor improvements on aphasic speech compared to a GMM-HMM system. Takashima et al. [17] proposed a new feature extraction scheme using convolutional bottleneck networks for dysarthric speech recognition. They tested the proposed approach on a

[1] http://hstrik.ruhosting.nl/chasing/

small test set consisting of 3 repetitions of 216 words by a single male speaker with an articulation disorder and reported some gains over a system using MFCC features.

Shahamiri and Salim [18] proposed an artificial neural network-based system trained on digit utterances from nine non-dysarthric and 13 dysarthric individuals affected by Cerebral Palsy (CP). They reported word recognition accuracies of 74.7%, 67.4% and 51.7% on mild (66-99% speech intelligibility), moderate (33-66% speech intelligibility) and high (less than 33% speech intelligibility) dysarthric speaker as an independent test set. Christensen et al. [19] trained their models solely on 18 hours of speech of 15 dysarthric speakers due to CP leaving one speaker out as test set. The different degrees of severity were reported through classes and percent intelligibility scores from listening tests with unfamiliar listeners: Very low 2-17%, Low 28-43%, Mid 58-62%, and High 86-95%. There were 4, 3, 3, and 5 speakers in every class, respectively. Recognition results for Very low ranged from 4.1-12.9%, Low 7.0-22.2%, Mid 30.3-50.2% , and High 46.6-68.5%. This shows that there is overlap between classes and that the recognition results do not always exactly match the intelligibility scores given by listeners.

Rudzicz [20] compared the performance of a speaker-dependent and a speaker-adaptive GMM-HMM systems on the Nemours database [21]. Their system was trained on the WSJ corpus. The test set consisted of speech from 11 dysarthric speakers due to CP or head trauma. Every speaker recorded 74 nonsense sentences. The recognizer provided recognition rates below 10% on the speech of 4 speakers with severe dysarthria. For moderately and mildly dysarthric speakers, recognition accuracy was between 11-30% and 31-60% respectively. Kyeong Seong et al. [22] proposed a weighted finite state transducer (WSFT)-based ASR correction technique applied to a recognition system trained also on the WSJ corpus. They reported an average accuracy of 47.1% when recognizing the speech of 10 dysarthric speakers from the same dysarthric database. Similar work had been proposed by Caballero-Morales and Cox [23] previously.

Mengistu and Rudzicz [24] combined dysarthric data of eight dysarthric speakers with that of seven normal speakers, leaving one out as test set and obtained an average increase by 13.0% in comparison to models trained on non-dysarthric speech only. They also noted that context-dependent HMMs showed little improvement over context-independent ones. In one of the earliest work on Dutch pathological speech by Sanders et al. [25], a pilot study was presented on ASR of Dutch dysarthric speech data obtained from two speakers with a birth defect and a cerebrovascular accident. Both speakers were classified as mild dysarthric by a speech pathologist.

From the previous descriptions, it appears that it is difficult to fully compare results between these publications due to the differences in types of speech materials, types of dysarthria, reported severity, and dataset used for training and testing. Additionally, dysarthric speech is highly variable in nature, not only due to its various etiologies and degrees of severity, but also because of possible individually deviating speech characteristics. This may negatively influence the capability of speaker-independent systems to generalize over multiple dysarthric speakers.

Possible improvements may come from recent advances in DNN-based acoustic model yielding impressive results in the field of non-dysarthric speech recognition [26]. These results show promising increases in the speaker-independent recognition accuracies when compared to those obtained with traditional GMM-HMMs. Therefore, we investigate how the DNN-HMM system trained on normal speech perform on the recognition of dysarthric speech with a focus on the amount of available training data from different varieties of the Dutch language.

3. Speech corpora selection

Given the limited availability of dysarthric speech data, we investigate to what extent already existing databases of Dutch normal speech can be employed to train DNNs and optimize their performance on dysarthric speech. The ASR technology to be developed in the CHASING project is primarily intended for dysarthric patients who live in the Netherlands and speak the Northern Dutch variety. However, we thought it would be interesting to also investigate the usability of speech databases of the Southern variety of Dutch spoken in Flanders and also known as Flemish. First, because these two varieties are mutually intelligible and their phonetic alphabets are arguably similar, apart from some well described phonological and phonetic differences. Second, because using Flemish speech would open up the possibility of adapting the game that is now been developed for patients in the Netherlands for use by patients in Flanders.

Fortunately, there have been multiple Dutch-Flemish speech data collection efforts [27, 28] which facilitate the integration of both Dutch and Flemish data in the research reported in this paper. For training purposes, we used the CGN corpus [27], which contains representative collections of contemporary standard Dutch as spoken by adults in the Netherlands and Flanders. The components with read speech, spontaneous conversations, interviews and discussions are used for training the acoustic models in the present experiments.

For testing purposes, we decided to use the largest collection of pathological speech that is available for the Dutch language, the Flemish COPAS database [29]. In the meantime, a database of Northern Dutch dysarthric speech has been compiled [30]. In the course of the CHASING project, this collection will be augmented with additional material and then used for further experiments to optimize ASR back-end used in the developed serious game.

The COPAS corpus contains recordings from 122 Flemish normal speakers and 197 Flemish speakers with speech disorders such as dysarthria, cleft, voice disorders, laryngectomy and glossectomy. The dysarthric speech component contains recordings from 75 Flemish patients affected by Parkinson's disease, traumatic brain injury, cerebrovascular accident and multiple sclerosis who exhibit dysarthria at different levels of severity.

The word reading tasks used in this paper are taken from the Dutch Intelligibility Assessment (DIA) [31] material which contains 35 versions of 50 consonant-vowel-consonant (CVC) words organized in 3 subgroups. Moreover, all sentence reading tasks with annotations, namely 2 isolated sentence reading tasks, 11 text passages with reading level difficulty of AVI 7 and 8 and Text Marloes, are also included in the test data.

4. Training DNNs for Dysarthric Speech

4.1. Fundamentals of DNN-based ASR

A single artificial neuron, which is the basic element of the DNN structure, receives N input values $\mathbf{v} = [v_0, v_1, ..., v_{N-1}]$ with weights $\mathbf{w} = [w_0, w_1, ..., w_{N-1}]$ and an offset value b. To compute the neuron output y, a non-linear function f is applied

Table 1: Word error rates in % obtained on the word and sentence COPAS test sets

Acoustic models	Training Data	WordDys	WordNor	SentDys	SentNor
GMM+MFCC	FL	77.2	56.1	38.2	13.0
GMM+MFCC	FL+NL	78.7	61.0	37.4	14.7
GMM+LDA-MLLT	FL	74.4	50.9	37.4	11.3
GMM+LDA-MLLT	FL+NL	74.9	55.0	37.0	12.5
DNN+FBANK	FL	65.0	37.9	28.1	4.7
DNN+FBANK (w/o retraining on FL)	FL+NL	64.9	38.4	26.8	4.7
DNN+FBANK (with retraining on FL)	FL+NL	**63.7**	**36.2**	**26.3**	**4.4**

the weighted sum z of all outputs of the previous layer and the offset, i.e., $y = f(z) = f(\mathbf{w}^T\mathbf{v}+b)$. A DNN consists of L layers of M artificial neurons and the output of the $(l-1)^{\text{th}}$ layer with M_{l-1} neurons is the input of the l^{th} layer with M_l neurons which is formulated as $\mathbf{v}_l = f(\mathbf{z}_l) = f(\mathbf{W}_l\mathbf{v}_{l-1} + \mathbf{b_l})$ where the dimensions of $\mathbf{v}_l$, $\mathbf{W}_l$, $\mathbf{v}_{l-1}$ and $\mathbf{b_l}$ are M_l, $(M_l \times M_{l-1})$, M_{l-1} and M_l respectively. M_0 is the number of neurons in the input layer which is equal to the dimension of the speech features. The non-linear activation function f maps an M_{l-1} vector to an M_{l-1} vector. The activation function applied at the output layer is the softmax function in order to get output values in the range $[0, 1]$ for the HMM state posterior probabilities

$$\mathbf{v}_{L+1} = P(q_i|\mathbf{o}) = \frac{e^{z_L^i}}{\sum_m^{M_{L+1}} e^{z_L^m}}, \quad (1)$$

where M_{L+1} is equal to the number of HMM states.

The DNN-HMM training is achieved in three main stages [32, 33]. Firstly, a GMM-HMM setup is trained to obtain the structure of the DNN-HMM model, initial HMM transition probabilities and training labels of the DNNs. Then, the pre-training algorithm described in [34] is applied to obtain a robust initialization for the DNN model. Finally, the back-propagation algorithm [35] is applied to train the DNN that will be used as the emission distribution of the HMM states.

4.2. Tuning DNNs on Flemish Speech

The DNN training applied in this paper is organized in two steps. In the first step, the DNN training is performed on the combined speech data containing Flemish and Northern Dutch normal speech. Both varieties sharing the phonetic alphabet, we learn several hidden layers and an output layer on both varieties with the aim of learning more reliable hidden layers. The amount of training data used during the initial training phase can be increased by including more speech data from different speech types such as elderly and children speech. In the scope of this work, we only consider using normal speech to analyze the impact of the data merging on the recognition performance.

In the second step, the softmax layer of this DNN is re-trained only on the Flemish data. Retraining the softmax layer achieves the fine-tuning of the DNN targets on the target Flemish speech. This two-step training approach resembles the multilingual DNN training scheme for cross-lingual knowledge transfer which is commonly used for obtaining acoustic models for under-resourced languages, e.g. [36, 37]. In these studies, considerable improvements have been reported on both low- and high-resourced languages thanks to the hidden layers trained on multiple languages.

5. Experimental Setup

5.1. Databases

The CGN components with read speech, spontaneous conversations, interviews and discussions are used for acoustic model training. The duration of the normal Flemish (FL) and northern Dutch (NL) speech data used training is 186.5 and 255 hours respectively. The combined training data (FL+NL) contains 441.5 hours in total.

For testing the acoustic models, we have classified the speech material in the COPAS database based on the type of the speaker (normal vs. pathological) and speech material (word vs. sentence) resulting in 4 test sets. The speech segments in which the speaker do not utter the target word are discarded to be able to evaluate the recognizer errors only. There are 687 different words and 212 different sentences in the test data. The test set containing the word tasks uttered by normal speakers (WordNor) and speakers with disorders (WordDys) consists of 6154 and 8648 utterances with a total duration of 1.5 and 2 hours, respectively. The test set containing the sentence tasks uttered by normal speakers (SentNor) and speakers with disorders (SentDys) consists of 1918 (15,149) and 1034 (8287) sentences (words) with a total duration of 1.5 and 1 hour, respectively.

5.2. Implementation Details

The recognition experiments are performed using the Kaldi ASR toolkit [38]. A standard feature extraction scheme is used by applying Hamming windowing with a frame length of 25 ms and frame shift of 10 ms. A conventional context dependent GMM-HMM system with 40k Gaussians and 5925 triphone states is trained on the 39-dimensional MFCC features including the deltas and delta-deltas. This system is used to obtain the state alignments required for DNN training. We also trained a GMM-HMM system on the LDA-MLLT features as a second baseline system.

The DNNs with 6 hidden layers and 2048 sigmoid hidden units at each hidden layer are trained on the 40-dimensional log-mel filterbank features with the deltas and delta-deltas. The DNN training is done by mini-batch Stochastic Gradient Descent with an initial learning rate of 0.008 and a minibatch size of 256. The time context size is 11 frames achieved by concatenating $\pm$5 frames. A unigram (trigram) language model trained on the target transcriptions of the word (sentence) tasks is incorporated in the recognition of the word (sentence) tasks.

6. Results and Discussion

We have performed ASR experiments using the speech data described in Section 5.1. The recognition results obtained on

the word and sentence tasks uttered by normal and pathological speakers from the COPAS database are presented in the columns of Table 1. The lowest WER for each column is marked in bold. The recognition performance obtained on the sentence readings task is more relevant compared to isolated word recognition in the context of the developed CHASING serious game. We report results on both word and sentence task results for completeness.

The conventional GMM-HMM trained on FL data using the MFCC features provides a WER of 38.2% on the dysarthric sentence utterances and a WER of 77.2% on the dysarthric word utterances. This large gap between in recognition accuracy obtained on sentence and word recognition task is due to the very challenging nature of recognizing phonetically similar, monosyllabic words and pseudowords in isolation. The GMM-HMM system trained on FL+NL data reduces the normal speech recognition from 13.0% to 14.7% in sentence reading tasks and from 56.1% to 61.0% in word reading tasks, while increasing dysarthric sentence recognition accuracy from 38.2% to 37.4%. The performance reduction in normal speech is comprehensible, since adding Northern Dutch data increases the mismatch between the training and testing conditions. Training GMM-HMM on the combined data does not always improve dysarthric speech recognition with 0.8% decrease on sentence tasks and 1.5% increase on word tasks in the WER.

Compared to MFCC features, using LDA-MLLT-transformed features considerably reduces the WERs obtained on normal speech as expected. On the other hand, the gains obtained on pathological speech by using these features are minimal. This is due to the fact that there is a significant mismatch between the type of speech on which the transformation is learned and applied to in the case of pathological speech.

The DNNs trained on FBANK features provide considerable improvements on all speech types and reading tasks. These improvements are as large as 8.9% on dysarthric sentence utterances and 9.9% on dysarthric word utterances. Training the DNN-HMMs on FL+NL data improves the performance on dysarthric sentence reading tasks with an absolute improvement of 1.3% without retraining the softmax layer on Flemish data. The same system does not improve the recognition accuracy of dysarthric word reading tasks. After the final step of applying softmax layer retraining for tuning the DNN targets to Flemish phones, we can get an improved recognition performance in all cases compared to the baseline DNN system trained only on Flemish data. To be specific, the WER obtained on dysarthric sentence reading task decreases from 28.1% to 26.3%, while the WER obtained on dysarthric word reading task reduces from 65.0% to 63.7%.

From these results, it can be seen that training DNN-HMM systems on training data containing speech from different varieties of a language can improve the recognition performance at moderate levels. Despite the large gap between the performance on pathological and normal speech, the presented speaker-independent recognition results obtained on different types of pathological speech at different severity levels are encouraging. Building text- and speaker-independent ASR systems that can be used in clinical applications appears to be within reach, even for languages with more limited resources than English.

7. Conclusions

In this paper, we have investigated combining speech data from different varieties of a mid-sized language for training a DNN-HMM system. The DNN-based acoustic models were trained on normal Flemish and Northern Dutch speech and speaker-independent recognition experiments were performed on pathological Flemish speech. The results have shown that the proposed data merging technique in this context can improve the recognition of pathological speech, especially after a second training phase in which the DNN targets are tuned to the phones of the specific variety involved in the testing setup, Flemish in this case. These results are promising for future work aiming to develop useful ASR-based pathological speech applications for languages that are smaller in size and less resourced than English.

8. Acknowledgements

This research is funded by the NWO research grant with Ref. no. 314-99-101 (CHASING).

9. References

[1] J. R. Duffy, *Motor speech disorders: substrates, differential diagnosis and management*. St. Louis: Mosby, 1995.

[2] R. D. Kent and Y. J. Kim, "Toward an acoustic topology of motor speech disorders," *Clin Linguist Phon*, vol. 17, no. 6, pp. 427–445, 2003.

[3] M. Walshe and N. Miller, "Living with acquired dysarthria: the speaker's perspective," *Disability and rehabilitation*, vol. 33, no. 3, pp. 195–215, 2011.

[4] L. O. Ramig, S. Sapir, C. Fox, and S. Countryman, "Changes in vocal loudness following intensive voice treatment (LSVT) in individuals with Parkinsons disease: A comparison with untreated patients and normal age-matched controls," *Movement Disorders*, vol. 16, pp. 79–83, 2001.

[5] S. K. Bhogal, R. Teasell, and M. Speechley, "Intensity of aphasia therapy, impact on recovery," *Stroke*, vol. 34, no. 4, pp. 987–993, 2003.

[6] G. Kwakkel, "Impact of intensity of practice after stroke: issues for consideration," *Disability and Rehabilitation*, vol. 28, no. (13-14), pp. 823–830, 2006.

[7] M. Rijntjes, K. Haevernick, A. Barzel, H. van den Bussche, G. Ketels, and C. Weiller, "Repeat therapy for chronic motor stroke: a pilot study for feasibility and efficacy," *Neurorehabilitation and Neural Repair*, vol. 23, pp. 275–280, 2009.

[8] L. J. Beijer and A. C. M. Rietveld, "Potentials of telehealth devices for speech therapy in Parkinson's disease, diagnostics and rehabilitation of Parkinson's disease," *InTech*, pp. 379–402, 2011.

[9] L. J. Beijer, A. C. M. Rietveld, M. B. Ruiter, and A. C. Geurts, "Preparing an E-learning-based Speech Therapy (EST) efficacy study: Identifying suitable outcome measures to detect within-subject changes of speech intelligibility in dysarthric speakers," *Clinical Linguistics and Phonetics*, vol. 28, no. 12, pp. 927–950, 2014.

[10] M. S. De Bodt, H. M. Hernandez-Diaz, and P. H. Van De Heyning, "Intelligibility as a linear combination of dimensions in dysarthric speech," *Journal of Communication Disorders*, vol. 35, no. 3, pp. 283–292, 2002.

[11] Y. Yunusova, G. Weismer, R. D. Kent, and N. M. Rusche, "Breath-group intelligibility in dysarthria: characteristics and underlying correlates," *J Speech Lang Hear Res.*, vol. 48, no. 6, pp. 1294–1310, 2005.

[12] G. Van Nuffelen, C. Middag, M. De Bodt, and J.-P. Martens, "Speech technology-based assessment of phoneme intelligibility in dysarthria," *International Journal of Language & Communication Disorders*, vol. 44, no. 5, pp. 716–730, 2009.

[13] D. V. Popovici and C. Buică-Belciu, "Professional challenges in computer-assisted speech therapy," *Procedia - Social and Behavioral Sciences*, vol. 33, pp. 518 – 522, 2012.

[14] M. Ganzeboom, M. Bakker, C. Cucchiarini, and H. Strik, "Intelligibility of disordered speech: Global and detailed scores," in *Proc. INTERSPEECH, To appear*, Sept. 2016.

[15] 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.

[16] T. Lee, Y. Liu, P.-W. Huang, J.-T. Chien, W. K. Lam, Y. T. Yeung, T. K. T. Law, K. Y. Lee, A. P.-H. Kong, and S.-P. Law, "Automatic speech recognition for acoustical analysis and assessment of cantonese pathological voice and speech," in *Proc. ICASSP*, 2016, pp. 6475–6479.

[17] Y. Takashima, T. Nakashika, T. Takiguchi, and Y. Ariki, "Feature extraction using pre-trained convolutive bottleneck nets for dysarthric speech recognition," in *Proc. EUSIPCO*, 2015, pp. 1426–1430.

[18] S. R. Shahamiri and S. S. B. Salim, "Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach," *Advanced Engineering Informatics*, vol. 28, pp. 102–110, 2014.

[19] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech." in *INTERSPEECH*, 2012, pp. 1776–1779.

[20] F. Rudzicz, "Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech," in *Proc. of the 9th International ACM SIGACCESS Conference on Computers and Accessibility*, 2007, pp. 255–256.

[21] X. Menéndez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The Nemours database of dysarthric speech," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 1996, pp. 1962–1966.

[22] W. Seong, J. Park, and H. Kim, "Dysarthric speech recognition error correction using weighted finite state transducers based on context-dependent pronunciation variation," in *Computers Helping People with Special Needs*, ser. Lecture Notes in Computer Science, 2012, vol. 7383, pp. 475–482.

[23] S.-O. Caballero-Morales and S. J. Cox, "Modelling errors in automatic speech recognition for dysarthric speakers," *EURASIP J. Adv. Signal Process*, pp. 1–14, Jan. 2009.

[24] K. T. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *Proc. ICASSP*, may 2011, pp. 4924–4927.

[25] E. Sanders, M. B. Ruiter, L. J. Beijer, and H. Strik, "Automatic recognition of Dutch dysarthric speech: a pilot study," in *Proc. INTERSPEECH*, 2002, pp. 661–664.

[26] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[27] N. Oostdijk, "The spoken Dutch corpus: Overview and first evaluation," in *Proc. LREC*, 2000, pp. 886–894.

[28] C. Cucchiarini, J. Driesen, H. Van hamme, and E. Sanders, "Recording speech of children, non-natives and elderly people for HLT applications: the JASMIN-CGN Corpus," in *Proc. LREC*, May 2008, pp. 1445–1450.

[29] C. Middag, "Automatic analysis of pathological speech," Ph.D. dissertation, Ghent University, Belgium, 2012.

[30] E. Yılmaz, M. Ganzeboom, L. Beijer, C. Cucchiarini, and H. Strik, "A Dutch dysarthric speech database for individualized speech therapy research," in *Proc. LREC*, 2016, pp. 792–795.

[31] M. De Bodt, C. Guns, and G. Van Nuffelen, "NSVO: handleiding," Vlaamse Vereniging voor Logopedie: Herentals, Tech. Rep., 2006.

[32] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[33] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Springer-Verlag London, 2015.

[34] G. Hinton, "A practical guide to training restricted Boltzmann machines," Department of Computer Science, University of Toronto, Tech. Rep. UTML TR 2010003, 2010.

[35] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural Networks, 1989. IJCNN., International Joint Conference on*, 1989, pp. 593–605 vol.1.

[36] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proc. SLT*, Dec 2012, pp. 246–251.

[37] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. ICASSP*, May 2013, pp. 7304–7308.

[38] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, Dec. 2011.