INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Multi-Stage DNN Training for Automatic Recognition of Dysarthric 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

Incorporating automatic speech recognition (ASR) in individualized speech training applications is becoming more viable thanks to the improved generalization capabilities of neural network-based acoustic models. The main problem in developing applications for dysarthric speech is the relative in-domain data scarcity. Collecting representative amounts of dysarthric speech data is difficult due to rigorous ethical and medical permission requirements, problems in accessing patients who are generally vulnerable and often subject to altering health conditions and, last but not least, the high variability in speech resulting from different pathological conditions. Developing such applications is even more challenging for languages which in general have fewer resources, fewer speakers and, consequently, also fewer patients than English, as in the case of a mid-sized language like Dutch. In this paper, we investigate a multi-stage deep neural network (DNN) training scheme aimed at obtaining better modeling of dysarthric speech by using only a small amount of in-domain training data. The results show that the system employing the proposed training scheme considerably improves the recognition of Dutch dysarthric speech compared to a baseline system with single-stage training only on a large amount of normal speech or a small amount of in-domain data.

Index Terms: Pathological speech, automatic speech recognition, deep neural networks, dysarthria

1. Introduction

Speech disorders caused by neuromuscular control problems [1] like dysarthria can reduce speech intelligibility and cause communication impairment [2]. This can negatively affect the life quality of dysarthric patients [3] who run the risk of losing social contact and eventually becoming isolated from society. Recent research has shown that intensive therapy can be effective in (speech) motor rehabilitation [4–7]. Conventional speech therapy provided by a speech therapist is costly. Recent developments show that therapy can be provided by employing computer-assisted speech training systems [8]. According to the outcomes of the efficacy tests presented in [9], user satisfaction towards such a system 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]. Our goal is to develop more robust acoustic models for pathological speech and incorporate automatic speech recognition (ASR) technology to detect these problems.

Despite long-lasting efforts to build speaker- and text-independent ASR systems for people with dysarthria, the performance of state-of-the-art systems is still considerably lower on this type of speech than on normal speech. Past ASR experiments on dysarthric speech mostly included GMM-HMM systems [15–20]. More recently Lee et al. [21] reported 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. [22] proposed a new feature extraction scheme using convolutional bottleneck networks for dysarthric speech recognition.

Training robust deep neural networks (DNN)-based acoustic models to capture the within- and between-speaker variation in dysarthric speech 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 that in normal speech databases. Moreover, these databases mostly contain very restricted speech tasks such as reading out word and sentence lists with varying linguistic complexity.

To remedy the data scarcity problem, [23] combined in-domain and out-of-domain English speech data to train DNNs for improved feature extraction. In previous work [24], we described a similar solution to train a better DNN-hidden Markov model (HMM) system for the Dutch language, a language that has fewer speakers and resources compared to English. In particular, we investigated combining non-dysarthric speech data from different varieties of the Dutch language to train more reliable acoustic models for a DNN-HMM ASR system. This work was conducted 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. In this research we employed a 6-hour Dutch dysarthric speech database that had been collected in a previous project (EST) [25]. The serious game developed in the CHASING project also serves as a useful data collection tool for pathological speech research. The dysarthric speech material recently collected during the CHASING field studies, which we refer to as the CHASING01 speech database, is used for testing, while the EST database is employed for training purposes.

In the present work, we apply a multi-stage DNN training procedure using a large amount of out-of-domain and a small amount of in-domain data. A two-stage version of this training procedure has been applied to multilingual training of DNNs which is commonly used to obtain acoustic models for under-resourced languages [26, 27]. In these studies, considerable improvements have been reported on both low- and high-resourced languages thanks to the hidden layers trained on multiple languages.

In the first stage of the training, we train models on normal Dutch and Flemish speech, which has been shown to provide improved recognition of dysarthric speech compared to training on only one variety [24]. The background model obtained in the first stage is retrained on normal and adult Dutch speech only for language adaptation and the EST dysarthric speech database is used for domain adaptation in subsequent training stages. The final models are then applied to the recently collected dysarthric speech data from the CHASING01 database.

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

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-303

The rest of the paper is organized as follows. Section 2 details the DNN training scheme applied in this paper. Section 3 explains the selection of various speech corpora for the proposed training scheme. The experimental setup is described in Section 4 and the recognition results are presented in Section 5. Section 6 concludes the paper.

2. Multi-stage DNN training

The DNN training applied in this paper is organized in multiple steps. In the first step, a background DNN is trained on large quantities of normal speech data. The amount of training data used during the initial training phase can be increased by including speech data from different speaker groups such as normally speaking elderly people and children. In the following step, the layers of this DNN are retrained using only speech data that resembles the target speech, e.g. using Dutch dysarthric speech and/or Dutch elderly speech. The aim of the second step is to tune the DNN on dysarthric speech as this is the type of speech to be recognized.

We have investigated multiple parameters that may influence the accuracy of the final model, such as the number of retrained layers and the learning rate. Moreover, various types of speech data have been used to explore their impact on the modeling accuracy of the final DNN model. Normal speech has been used due to its abundance compared to other deviant speech types. Since the majority of dysarthric speakers are older than 50, elderly speech data is also relevant in this scenario. Finally, normal speech data from a related variety of Dutch, namely Flemish, is included to obtain the background model. Using speech data from different language varieties led to a mild improvement in recognition accuracy in a previous study [24]. Since both varieties share the phonetic alphabet, we learn several hidden layers and a softmax layer on both varieties with the aim of learning more reliable hidden layers. The following section continues this paper by describing the speech corpora that have been used during the experiments.

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. There have been multiple Dutch-Flemish speech data collection efforts [28, 29] which facilitate the integration of both Dutch and Flemish data in the present research. For training purposes, we used the CGN corpus [28], which contains representative collections of contemporary standard Dutch as spoken by adults in the Netherlands and Flanders. Considering that the high median age in our database of dysarthric speech is 66.5 years, we have also included elderly speech data from the JASMIN corpus [29] to the Dutch normal speech in the training phase.

The EST Dutch dysarthric speech database [25] contains dysarthric speech from ten patients with Parkinson's Disease (PD), four patients who have had a Cerebral Vascular Accident (CVA), one patient who suffered Traumatic Brain Injury (TBI) and one patient having dysarthria due to a birth defect. Based on the meta-information, the age of the speakers is in the range of 34 to 75 years with a median of 66.5 years. The level of dysarthria varies from mild to moderate. The dysarthric speech collection for this database was achieved in several experimental contexts. The speech tasks presented to the patients in these contexts consist of numerous word and sentence lists with varying linguistic complexity. The database includes 12 Semantically Unpredictable Sentences (SUSs) with 6- and 13-word declarative sentences, 12 6-word interrogative sentences, 13 Plomp and Mimpen sentences, 5 short texts, 30 sentences with /t/, /p/ and /k/ in initial position and unstressed syllable, 15 sentences with /a/, /e/ and /o/ in unstressed syllables, production of 3 individual vowels /a/, /e/ and /o/, 15 bisyllabic words with /t/, /p/ and /k/ in initial position and unstressed syllable and 25 words with alternating vowel-consonant composition (CVC, CVCVCC, etc.).

As mentioned above, for testing purposes we use the CHASING01 dysarthric speech database that was recently collected in the first stage of the CHASING project. This database contains speech of 5 patients who participated in speech training experiments and were tested at 6 different times during the treatment. For each set of audio files, the following material was collected: 12 SUSs, 30 /p/, /t/, /k/ sentences in which the first syllable of the last word is unstressed and starts with /p/, /t/ or /k/, 15 vowel sentences with the vowels /a/,/e/ and /o/ in stressed syllables, appeltaarttekst (*apple cake recipe*) in 5 parts. Utterances that deviated from the reference text due to pronunciation errors (e.g. restarts, repeats, hesitations, etc.) were removed. After this subselection, the utterances from 3 male patients remained and were included in the test set. These speakers are 67, 62 and 59 years old, two of them having PD and the third having had a CVA.

4. Experimental Setup

4.1. Database details

The CGN components with read speech, spontaneous conversations, interviews and discussions were used for acoustic model training. The duration of the normal Flemish (FL) and northern Dutch (NL) speech data used for training is 186.5 and 255 hours, respectively. The combined training data (Nor. FL+NL) contains 441.5 hours in total. The total duration of the elderly speech recordings in the JASMIN database (Eld. NL) is 10 hours and 10 minutes.

The EST Dutch dysarthric speech database (Dys. NL) contains 6 hours and 16 minutes of dysarthric speech material from 16 speakers [25]. The speech segments with pronunciation errors (e.g. restarts, repeats, hesitations, etc.) were excluded from the training set to maintain integrity of the results on ASR performance evaluation. Additionally, the segments including a single word and pseudoword were also excluded, since the sentence reading tasks are more relevant in our project context. The total duration of the dysarthric speech data eventually selected for training is 4 hours and 47 minutes.

The CHASING01 speech database, which was used for testing, contains 721 utterances (6231 words) with corresponding manual transcriptions that match the reference text. The total duration of this speech data is 55 minutes.

4.2. Implementation Details

The recognition experiments were performed using the Kaldi ASR toolkit [30]. A standard feature extraction scheme was 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 was trained on the 39-dimensional MFCC features including the deltas and delta-deltas. We also trained a GMM-HMM system on the LDA-MLLT features, followed by training models with speaker adaptive training using FMLLR features.

Table 1: *Word error rates in % obtained on the test set for different number of retrained layers (# of Retr. Lay.) and retraining initial learning rate (Retr. Init. LR)*

Training	Retraining	# of Retr. Lay.	Retr. Init. LR	WER (%)
Nor. NL	-	-	-	21.3
Dys. NL	-	-	-	17.3
Nor. NL	Dys. NL	all	0.008	12.1
Nor. NL	Dys. NL	5	0.008	12.6
Nor. NL	Dys. NL	4	0.008	12.8
Nor. NL	Dys. NL	3	0.008	12.0
Nor. NL	Dys. NL	2	0.008	12.4
Nor. NL	Dys. NL	1	0.008	**11.0**
Nor. NL	Dys. NL	softmax	0.008	11.9
Nor. NL	Dys. NL	all	0.0008	11.6
Nor. NL	Dys. NL	5	0.0008	11.8
Nor. NL	Dys. NL	4	0.0008	11.8
Nor. NL	Dys. NL	3	0.0008	11.8
Nor. NL	Dys. NL	2	0.0008	12.0
Nor. NL	Dys. NL	1	0.0008	12.2
Nor. NL	Dys. NL	softmax	0.0008	13.6

This system was used to obtain the state alignments required for DNN training.

The DNNs with 6 hidden layers and 2048 sigmoid hidden units at each hidden layer were trained on the 40-dimensional log-mel filterbank features with the deltas and delta-deltas. The DNN training was done by mini-batch Stochastic Gradient Descent with an initial learning rate of 0.008 and a minibatch size of 256. The default initial learning rate of 0.008 was used in the first training stage. The time context size was 11 frames achieved by concatenating ±5 frames. A trigram language model trained on the target transcriptions of the sentence tasks was used during recognition of the sentence tasks.

5. Results and Discussion

We performed several ASR experiments using the speech data described in Section 4.1. Firstly, we explored the impact of the number of retrained layers and initial learning rate on the recognition accuracy in a two-stage training setting. The Word Error Rates (WER) obtained on the CHASING01 test set after having trained models on the normal speech database (Nor. NL) and retrained on EST's dysarthric speech database (Dys. NL) are presented in Table 1. The lowest WER is marked in bold. Two recognizers trained on the Nor. NL database and Dys. NL separately provide a baseline WER of 21.3% and 17.3%, respectively. WERs yielded by the recognizers with two-stage training are considerably lower than those of the baseline systems and vary between 11.0%-13.6%. The recognition accuracy is obtained by only retraining the softmax and the last hidden layer with an initial learning rate that is the same as the initial learning rate used in the first stage. The results for different numbers of retrained layers do not follow a pattern, hence, it is difficult to formulate a superior retraining strategy. However, we can conclude that retraining only the softmax layer with a relatively low learning rate results in a reduced recognition accuracy of 13.6% with respect to retraining more layers with the same learning rate or retraining with a higher learning rate. It is important to mention that the amount of in-domain data used in the retraining stage will have an impact on the choice of the number of retrained layers.

Table 2: *Word error rates in % obtained on the test set for different number of retrained layers (# of Retr. Lay.) and retraining initial learning rate (Retr. Init. LR)*

Training	Retraining	# of Retr. Lay.	Retr. Init. LR	WER (%)
Nor. NL+FL	Dys. NL	all	0.008	12.8
Nor. NL+FL	Dys. NL	5	0.008	12.9
Nor. NL+FL	Dys. NL	4	0.008	12.6
Nor. NL+FL	Dys. NL	3	0.008	12.5
Nor. NL+FL	Dys. NL	2	0.008	12.5
Nor. NL+FL	Dys. NL	1	0.008	12.2
Nor. NL+FL	Dys. NL	softmax	0.008	**11.3**
Nor. NL+FL	Dys. NL	all	0.0008	12.0
Nor. NL+FL	Dys. NL	5	0.0008	11.9
Nor. NL+FL	Dys. NL	4	0.0008	12.0
Nor. NL+FL	Dys. NL	3	0.0008	12.3
Nor. NL+FL	Dys. NL	2	0.0008	12.3
Nor. NL+FL	Dys. NL	1	0.0008	12.0
Nor. NL+FL	Dys. NL	softmax	0.0008	12.2

In Table 2, we present a similar set of results by varying the content of data used in the initial training phase. Speech from a related Dutch language variety, Flemish, was used. Background models were trained on both the Northern and Flemish varieties of Dutch (Nor. NL+FL) instead of the Northern variety only, as was done in the previous paragraph. The goal of these experiments was to investigate the impact of adding speech from a related language variety to the training procedure on the modeling accuracy of the final models tuned on the Dys. NL data. The best recognition accuracy was provided by the system obtained with retraining of the softmax layer with a relatively high initial learning rate. That system has a WER of 11.3% which is comparable with, but not better than the previous best performing system presented in Table 1. Using in-domain speech data for training, the performance gain reported in previous experiments [24] cannot be obtained in this scenario.

Finally, the impact of retraining the background acoustic model on the EST Dysarthric data (Dys. NL) and speech data from Dutch elderly (Eld. NL) is shown in Table 3. The presented WER results vary between 13.9%-15.5% for different training parameters. From these results, it can clearly be seen that the performance of the final acoustic models deteriorate when elderly speech is used for retraining in all scenarios. The impact of the mismatch between the elderly and dysarthric elderly speech, e.g. reduced speaking rate and articulation skills, appears to be more salient than the increase in the amount of retraining data on the recognition accuracy.

To summarize, we can conclude that using in-domain data in the described two-stage training scheme improves the recognition performance significantly whereas merging different speech types in a single-stage training scheme provides only minor improvements [31]. Adding relevant types of data, i.e., the Flemish Dutch variety during background model training and using elderly speech data for retraining, does not improve the recognition accuracy of the final models.

6. Conclusions

In this paper, we applied a multi-stage DNN training scheme to obtain robust acoustic models in the framework of a serious game to be used as an individualized speech therapy tool. These models are applied to Dutch dysarthric speech, which is

Table 3: *Word error rates in % obtained on the test set for different number of retrained layers (# of Retr. Lay.) and retraining initial learning rate (Retr. Init. LR)*

Training	Retraining	# of Retr. Lay.	Retr. Init. LR	WER (%)
Nor. NL	Dys.+Eld. NL	all	0.008	15.1
Nor. NL	Dys.+Eld. NL	5	0.008	15.1
Nor. NL	Dys.+Eld. NL	4	0.008	15.5
Nor. NL	Dys.+Eld. NL	3	0.008	15.1
Nor. NL	Dys.+Eld. NL	2	0.008	14.7
Nor. NL	Dys.+Eld. NL	1	0.008	14.7
Nor. NL	Dys.+Eld. NL	softmax	0.008	**13.9**
Nor. NL	Dys.+Eld. NL	all	0.0008	14.8
Nor. NL	Dys.+Eld. NL	5	0.0008	15.1
Nor. NL	Dys.+Eld. NL	4	0.0008	15.3
Nor. NL	Dys.+Eld. NL	3	0.0008	14.7
Nor. NL	Dys.+Eld. NL	2	0.0008	15.1
Nor. NL	Dys.+Eld. NL	1	0.0008	15.0
Nor. NL	Dys.+Eld. NL	softmax	0.0008	15.2

more challenging to recognize than normal speech due to its increased variation. The data recently collected through the game could be used for testing, while the dysarthric data already available from the EST database were used for training. The applied multi-stage training approach aims to learn a background model trained on more general data in the initial stage. That model was then retrained on in-domain data in the second stage to get a domain-specific model.

We performed several ASR experiments by varying two training parameters, namely the number of retrained layers and the initial learning rate used in the second stage. The results have shown that this kind of training provides large improvements in recognition accuracy compared to baseline systems trained either on normal speech or on dysarthric speech. Moreover, we investigated the inclusion of various speech types such as normal speech from a related language variety for background model training and elderly speech for retraining in further recognition experiments. The recognition results suggest that adding normal speech data from a language variety does not bring improvement compared to a recognizer trained on only normal speech from the target language. Adding elderly data reduced the recognition performance compared to retraining only on dysarthric speech, most likely due to the increased mismatch between the training and target speech.

7. Acknowledgements

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

8. 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*, Sept. 2016, pp. 2503–2507.

[15] 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.

[16] 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.

[17] 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.

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

[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] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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*, Sept. 2016, pp. 218–222.

[25] 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.

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

[27] 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.

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

[29] 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.

[30] 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.

[31] M. Ganzeboom, E. Yılmaz, C. Cucchiarini, and H. Strik, “On the development of an ASR-based multimedia game for speech therapy: Preliminary results,” in *International Workshop on Multimedia for Personal Health and Health Care (MM Health)*, Oct. 2016, pp. 3–8.