An Automatic Dysarthric Speech Recognition Approach using Deep Neural Networks

Jun Ren
Institute of Natural and Mathematical Sciences
Massey University
Auckland, New Zealand

Mingzhe Liu
State Key Laboratory of Geohazard Prevention and Geoenvironment Protection Chengdu
University of Technology
Chengdu, China

***Abstract*—Transcribing dysarthric speech into text is still a challenging problem for the state-of-the-art techniques or commercially available speech recognition systems. Improving the accuracy of dysarthric speech recognition, this paper adopts Deep Belief Neural Networks (DBNs) to model the distribution of dysarthric speech signal. A continuous dysarthric speech recognition system is produced, in which the DBNs are used to predict the posterior probabilities of the states in Hidden Markov Models (HMM) and the Weighted Finite State Transducers framework was utilized to build the speech decoder. Experimental results show that the proposed method provides better prediction of the probability distribution of the spectral representation of dysarthric speech that outperforms the existing methods, e.g., GMM-HMM based dysarthric speech recogniztion approaches. To the best of our knowledge, this work is the first time to build a continuous speech recognition system for dysarthric speech with deep neural network technique, which is a promising approach for improving the communication between those individuals with speech impediments and normal speakers.**

***Keywords*—Dysarthric speech recognition; deep neural networks; hidden markov models**

I. Introduction

The Automatic Speech Recognition (ASR) technique for normal speech has evolved significantly over the past few years whereas the dysarthric speech recognition has not gained enough attention [1]-[3]. Dysarthric speech is produced by individuals with speech impediments, which is usually caused by weakness, paralysis, or poor coordination of the muscles responsible for speech production. Due to the high variability in dysarthric speech signals, translating of dysarthric spoken words into readable text is still a challenging task [4].

Best results for dysarthric speech recognition were provided by isolated-word ASR models and traditional ASR algorithms, such as Gaussian Mixture Models (GMMs) [5], whereas an assistive system for the dysarthria normally requires the ability of recognition of continuous speech [6]-[9]. GMM-based approaches may have difficulties to model dysarthric speech because there is more than one pronounciation for a single phone, and some of the pronounciations are same for different phones [10]. Although some alternative method for revising the false pronounciation has been proposed [11], [12], the performance of dysarthric speech recognition still requires significant improvements so as to be used in the reality applications. Recently, some projects reported to be successful in the recognition of dysarthric speech with limited vocabularies. However, so far, a large vocabulary dysarthric speech recognition system is still unavailable.

All the traditional dysarthric speech recognition systems are generally based on the structured approaches. For example, Hidden Markov models (HMMs) are used to model the sequential structure of speech signals and GMMs are used to model the distribution of the spectral representation of a waveform. Nevertheless, there are some drawbacks of such methods applied in dysarthric speech recognition [13]: 1) The basic assumption for GMM is that the input representations are Gaussian distributed, but this is not true for dysarthric speech. 2) The HMM model assumes that the observation probability of every hidden state is independent, thus the training process ignores the context information. 3) GMM is an efficient algorithm for high dimensional data, but the model often is very complex and ultimately affects the performance on the test dataset. 4) GMM is sensitive to the model parameters, thus it needs large amount of training data to train a robust model. However, the training data for dysarthric speech is low resource, and is not sufficient to estimate the means and variances for a continuous dysarthric speech recognition system.

Due to the high complexity of dysarthric speech signal, it still is a challenge to find a precise model to recognize the latent patterns. Deep Neural Networks (DNNs) and its variants have achieved significant improvement in recognition of normal speech when used as a replacement of GMMs. In this paper we utilized the hierarchical framework proposed by Hinton [14], [15] to extract a set of distinctive features from dysarthric speech and applied the robustness of this probabilistic generative model to characterize the long-span contextual influence of dysarthric speech.

DNNs can be an efficient alternative to GMMs because they possess the following advantages [16]: 1) The estimation of the posterior probabilities of HMM states does not require detailed assumptions about the data distribution. 2) DNN allows an easy way to combine diverse features, including both discrete and continuous features. 3) DNNs uses far more of the data to constrain each parameter since the output on each training case is sensitive to a large fraction of the weights.

In this paper we applied DBNs as the acoustical model to model the distribution of the dysarthric speech signal and compared with GMM-HMM based models. Previous studies showed that the latest Recurrent Neural Network (RNN) and its extended models have achieved significant improvement [17] in normal-speech recognition. However, they need large-scale sample data for training or otherwise it will end up with over fitting problems, whereas DBNs are relatively simple, and have a good ability in extracting latent features of dysarthric speech. It can be trained more easily with limited data than those complex models [18].

II. Materials and Feature Description

In this paper, we used TORGO database, which provides more than 8400 dysarthric utterances [19]. There are several other speech databases available for dysarthric speech recognition, but they primarily provide data about the voice recordings of isolated words. The stimuli in this database came from the sentences for the Yorkston-Beukelman assessment of intelligibility [20] and the TIMIT transcript, which ensured the balance of different phonemes. The single word stimuli in the database include repetitions of English digits, the international radio alphabets, the 20 most frequent words in the British National Corpus (BNC), and a set of words selected to demonstrate relevant phonetic contrasts [21]. Other databases such as UA-Speech database [22] or Nemours database [23] mainly contain isolated words, acoustic samples of digits, radio alphabet letters, and computer commands, which is inadequate to build a continuous dysarthric speech recognition model.

In order to compare the dysarthric speech features with normal speech, Fig. 1 demonstrated the oscillogram and pitch contours of two utterances of the same sentence by a non-dysarthria speaker and a dysarthria speaker respectively. It is evident that a dysarthria speaker often has difficulties in controlling the time to speak and the prosody of their voice is also different from a normal speaker.

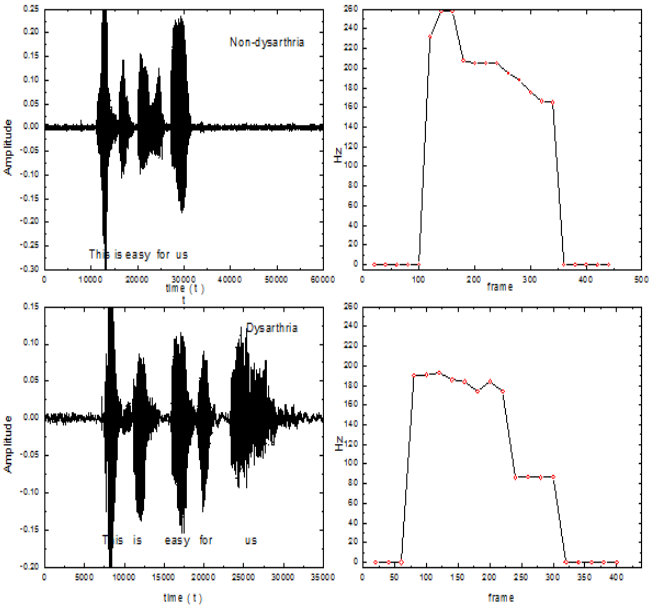

Fig. 1. Oscillogram and pitch contours of two utterances spoken by a non-dysarthria (top) and dysarthria speakers (bottom).

Traditionally, MFCCs, GMMs, and HMMs co-evolved as the ways of conducting speech recognition since the training process is computationally expensive. We adopted DBNs model in this paper instead of GMMs, which provides a more computationally intensive approach for speech recognition. The utterance signals were windowed with a 25-msec Hamming window every 10 msec. We applied vocal tract length normalization (VTLN) to the features in the feature preparing process. The GMM-HMM model was trained based on the augmented MFCC features. In order to partially overcome the conditional independence assumption of HMMs, the derivative and acceleration were also included and then a DNN was trained with the speaker adapted features.

III. Architecture of Dysarthric Speech Recognition System

The principal components of a dysarthric speech recognizer are illustrated in Fig. 2. In this paper, we mainly focus on Acoustic Modelling using DBNs.

The raw waveform of the audio signal is converted into a sequence of acoustic vectors $x = \{x_1, \ldots, x_T\}$ during the feature extraction phase. The decoder then attempts to find the sequence of words $W = \{w_1, \ldots, w_L\}$, which is most likely to have generated x, i.e. the decoder tries to find:

$$W^* = \arg\max P(W \mid X)$$

However, this problem cannot be solved directly since it is difficult to work out the P(W|X). Therefore, the solution is to transfer it to another form through Bayesian formula:

$$W^* = \arg\max P(W \mid X) = \arg\max \frac{P(X \mid W)P(W)}{P(X)}$$

Where, P(X) is a constant for any specific phonic. $P(X|W)$ is the acoustic model; $P(W|X)$ is called posterior probability, which is simple and more straight forward to calculate, and P(W) is the class priors, which is called the language model. It is challenging to calculate the posterior probability in dysarthric speech as the disabled are not capable to pronounce phones accurately. The proposed approach applies DNN to remedy this problem.

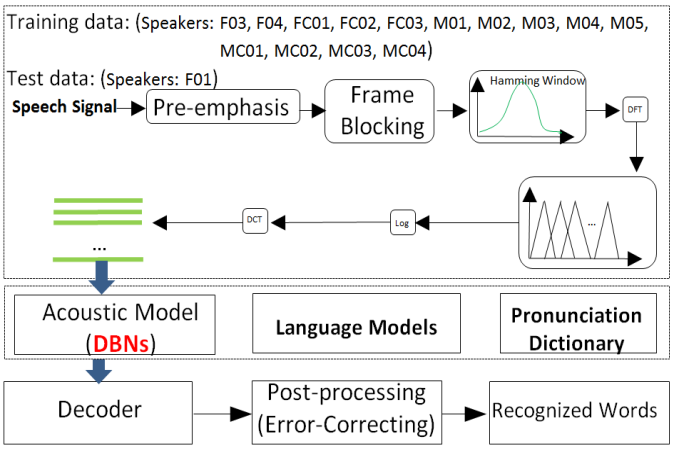

Fig. 2. Overview of the dysarhtirc speech recognition method/system.

In particular, for any given W, the corresponding acoustic model is synthesized by concatenating phone models to make words as defined by a pronunciation dictionary. The parameters of these phone models are estimated from the training data consisting of speech waveforms and their orthographic transcriptions. The language model is typically an N-gram parameters that is estimated by counting N-tuples in appropriate text corpora. In our system, we used a 2-gram language model. The Decoder operates by searching through all word sequence alternatives using pruning to remove unlikely hypotheses, which enable us to keep the search tractable. When the end of an utterance is reached, the most likely word sequence is the output. Then, the decoder is followed by a post-processing phase, a function for error-correcting, which helps to improve the accuracy further.

IV. Deep Neural Networks

DBN is a probabilistic generative model with multiple layers of stochastic hidden units above a single bottom layer of observed variables that represent a feature vector. It is a multi-layer generative model of a window of augmented speech coefficients. There are many cases of utilizing DBN into normal speech recognition and achieved significant improvement.

Over-fitting usually happens when the size of the sample data is too small comparison with the model complexity. In order to avoid over-fitting, the proposed approach uses a generative model to find out sensible features and then initialize the hidden units of the neural net with these features. Hinton [24] showed that these features can be inference using an undirected graphical model called a Restricted Boltzmann Machine (RBM). A set of RBMs can be composed of Deep belief neural networks. A typical architecture of DBNs is shown in Fig. 3.

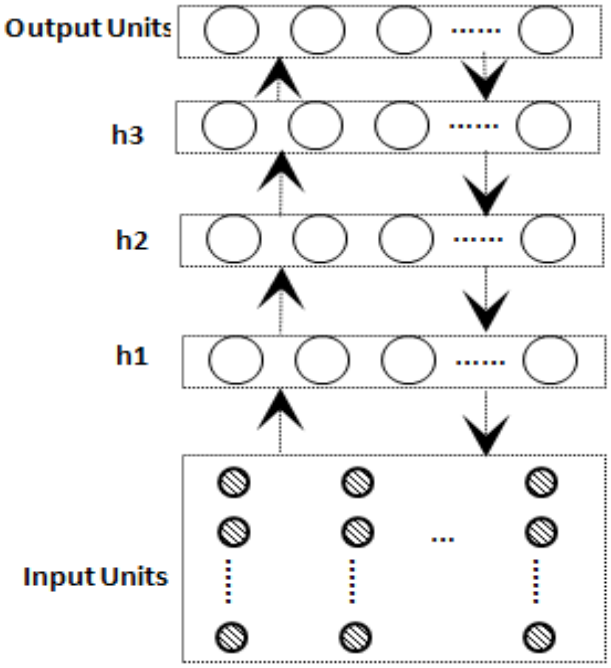

Fig. 3. Architecture of DNNs.

In speech recognition, it is common to use a Gaussian-Bernoulli RBM in which the hidden units are binary but the input units are linear with Gaussian noise. For Gaussian-Bernoulli RBM's the learning procedure is very similar except that the visible activities are measured in units equal to the standard deviation of the noise [24].

V. Experiments and Evaluation

A. Experimental Conditions

The dysarthric speech used in this paper is provided by the TORGO database. Our system were trained without applying the data considered for testing, which is the leave-one-subject-out methodology and applied random cross validation for parameter tuning. For example, we used all utterances of F01 as the test set while the remaining utterances from other subjects were used for training; this process was repeated for each dysarthric speaker. Before commencing the actual experiment, we conducted a pilot study to select the optimal experimental parameters, during which several pairs of configurations were tested.

B. Training Deep Belief Networks as Acoustical Model

As it is explained in [15], RBMs can be stacked and trained in a greedy manner to form DBNs; they were applied to extract phoneme posteriors probability in our study. The training processing can generally be separated into four steps: 1) A DBN was pre-trained and layered as RBM greedily. For Gaussian-binary RBMs, we ran 100 epochs with a fixed learning rate of 0.002 while for binary-binary RBMs we used 40 epochs with a learning rate of 0.005. 2) A DNN was then created by adding a "softmax" output layer to the network. The outputs of the lower layers were fed as input of the upper layers. 3) Discriminative fine-tuning by back-propagation was done to adjust the weights and to make them better at predicting the probability distribution over the states of monophone. 4) Applying the DBN as acoustical model. The sequence of the predicted probability distribution was fed into a standard Viterbi decoder.

C. Results and Discussion

In the parameter selection process, we split F01 as the test dataset. Fig. 4 shows the performance for different amount of hidden layers are used while the number of input frames was fixed at 17 and 1024 nodes in every layer.

Fig. 5 shows word error rates decreasing monotonically with the number of hidden layers of each input across different number of frames. As can be seen adding more hidden layers improved the performance. Nevertheless, the improvements were decreased when the number of hidden layers was more than five. In addition, similar results were obtained when layer numbers was 5 for 17 frames or 21 frames.

In order to reduce the computation cost and get an acceptable result, we fixed the number of layers at 5 and used 1024 nodes in each hidden layer. Table I shows the test result of different dysarthria speakers of this setting and several variant model of GMM-HMM.

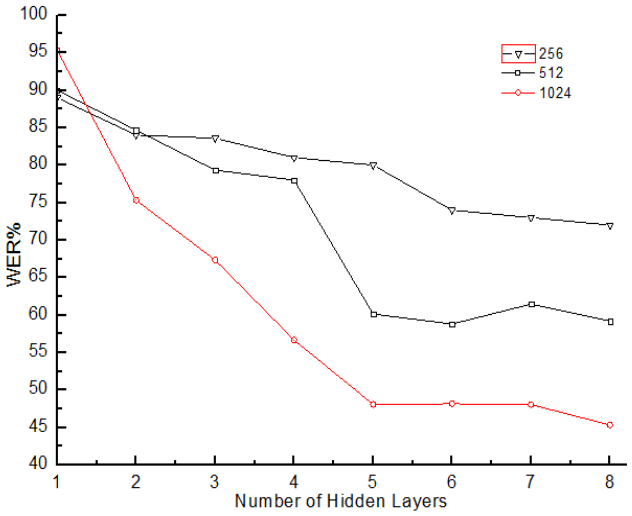

Fig. 4. Accuracy versus the number of frames fed into the DBNs across three different numbers of nodes per layer.

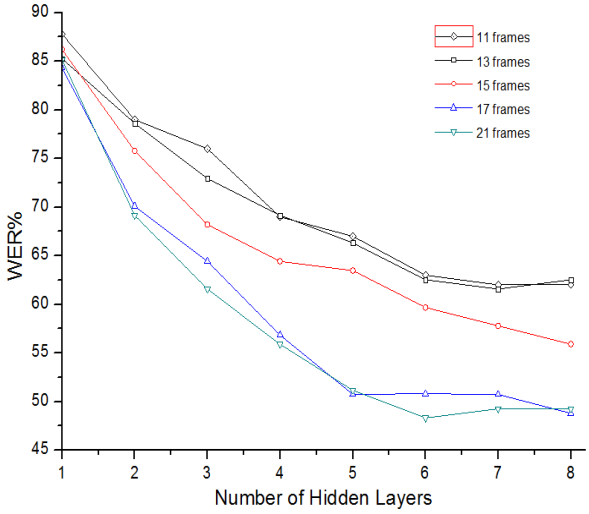

Fig. 5. Automatic speech recognition accuracy measured against number of hidden layers in acoustic model (DBNs).

TABLE I. WER OF DYSARTHRIA SPEAKERS ACROSS SEVERAL DIFFERENT MODELS

	Severely				**Moderate-to-Severely**	**Moderately**	**Very mild**	
	F01	**M01**	**M02**	**M04**	**M05**	**F03**	**M03**	**F02**
Baseline(monophone)	74.28	75.64	72.91	71.98	70.12	70.08	69.33	68.76
Triphone(tri1)	71.76	73.45	76.02	74.12	70.04	69.93	67.46	68.53
Triphone(tri2a)	67.12	65.2	68.31	67.52	64.88	64.2	62	61.78
Triphone(tri2b)	74.4	76.54	75.22	73.27	72.3	74.56	73.25	70.36
MMI(tri2b)	83.53	78.36	81.35	79.58	76.86	74.13	71.42	72.05
MMI(LDA+MLLT) (tri2b)	81.37	86.9	85.09	83.46	80.28	79.82	76.45	72.64
MPE(tri2b)	77.81	75.77	76.62	75.66	72.66	72.12	69.3	70.23
LDA+MLLT+SAT (tri3b)	58.09	56.69	57.68	60.04	57.39	58.36	55.69	56.03
MMI(tri3b)	59.59	57.47	59.4	58.63	57.3	56.9	56.7	56.4
MMI+fMMI (tri3b)	54.68	56.64	54.36	55.65	54.6	53.96	53.47	53.67
LDA+MLLT+SAT+ SAT(tri3b)	54.14	52.9	56.62	58.43	55.63	54.98	53.72	53.96
DBN	48.56	49.32	45.59	47.36	46.68	46.92	46.56	45.9

Results indicate that the trained model tends to perform better when the test speakers have higher intelligibility scores, and most individuals with moderate-to-severe and severe dysarthria tend to generate relatively higher ASR recognition accuracy rates. However, there are also some exceptions. For example, one of the severe dysarthria speakers, M01, got the lowest recognition accuracy. The reason for this may be due to different variability of different speakers, when they speak at a different time. All the test results got from DBNs were less than 50%, which is better than the initial results reported in [10].

This study has shown that the use of DBNs makes the recognizer more robust against the data variation of speech signal produced by different degrees of severity of dysarthric speech. The sentence error rates of this experiment are still a little bit high; we are exploring to find a better neural network algorithm to build a specific language model and correct the insert or addition errors of dysarthric speech in order to improve the word and sentence error rate further.

VI. CONCLUSION AND FUTURE WORK

This paper showed that the incorporation of the DNN model is useful in obtaining high probable phonemes with dysarthric acoustics. Overall, the results achieved here are superior to similar work discussed in Section 1. However, given the limited number of dysarthric speech samples provided in the database, the results can be considered preliminary; more work with additional data sets would be required to make more conclusive claims.

In future, further studies will be necessary to explore the performance of some different kinds of advanced Deep Neural Networks algorithms, applying different input features streams to improve the system's performance on a larger dataset.

ACKNOWLEDGMENT

The first author was supported by the Chinese Scholarship Council, and the database used in this paper is provided by Linguistic Data Consortium (LDC) via the data scholarship. The second author is supported by the Youth Innovation Research Team of Sichuan Province (2015TD0020) and CDUT (KYTD201404).

REFERENCES

[1] A.L. Maas, P. Qi, Z. Xie, A. Y. Hannun, C.T. Lengerich, D. Jurafsky, A.Y. Ng. "Building DNN Acoustic Models for Large Vocabulary Speech Recognition." arXiv preprint arXiv:1406.7806, 2014 Jun 30.

[2] P. Raghavendra, E. Rosengren, and S. Hunnicutt. "An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems." Augmentative and Alternative Communication, 17(4), pp. 265-275, 2001.

[3] F. Rudzicz, G. Hirst, and P.V. Lieshout. "Vocal tract representation in the recognition of cerebral palsied speech." Journal of Speech, Language, and Hearing Research, 55(4), pp. 1190-1207, 2012 Aug 1.

[4] V. Young, and A. Mihailidis. "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review." Assistive Technology, 22(2), pp. 99-112, 2010.

[5] S.R. Shahamiri, and S. Salim. "A multi-views multi-learners approach towards dysarthric speech recognition using multi-nets artificial neural networks." IEEE Transactions on Neural Systems and Rehabilitation Engineering, 22(5), pp. 1053-1063, 2014.

[6] J.F. Gemmeke, et al., Self-taught assistive vocal interfaces: An overview of the ALADIN project. 2013.

[7] S. Fager, D.R. Beukelman, M. Fried-Oken, T. Jakobs, & J. Baker. Access interface strategies. Assistive Technology, 24(1), pp. 25-33, 2012.

[8] D. Beukelman, S. Fager, and A. Nordness. "Communication support for people with ALS." Neurology Research International, 2011 Apr 14.

[9] S. Fager, L. Bardach, S. Russell, and J. Higginbotham. "Access to augmentative and alternative communication: New technologies and clinical decision-making." Journal of Pediatric Rehabilitation Medicine: An Inderdisciplinary Approach, 5.1, 2012 Jan 1.

[10] K.T. Mengistu, and F. Rudzicz. "Adapting acoustic and lexical models to dysarthric speech." 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4924-4927, 2011 May 22.

[11] H. V. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," Computer Speech and Language, vol. 27, no. 6, pp. 1147–1162, 2013.

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

[13] S. Oue, R. Marxer, and F. Rudzicz. Automatic dysfluency detection in dysarthric speech using deep belief networks. In 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT). p.60, 2015.

[14] G.E. Hinton, S. Osindero, and Y.W. Teh, A fast learning algorithm for deep belief nets. Neural computation, pp. 1527-1554, 18(7), 2006.

[15] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle, Greedy layer-wise training of deep networks. Advances in neural information processing systems, 19: p. 153, 2007.

[16] A.R. Mohamed, G.E. Dahl, and G. Hinton, Acoustic modeling using deep belief networks. IEEE Transactions on Audio, Speech, and Language Processing, 20(1) pp. 14-22, 2012.

[17] G.E. 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*, 20.1, pp. 30-42, 2012.

[18] A. Hannun, C. Case, et al. "Deep speech: Scaling up end-to-end speech recognition." arXiv preprint: 1412.5567, 2014.

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

[20] K.M. Yorkston, and D.R. Beukelman, Assessment of intelligibility of dysarthric speech. Pro-ed, 1984.

[21] R.D. Kent, G. Weismer, J.F. Kent, and J.C. Rosenbek, Toward phonetic intelligibility testing in dysarthria. Journal of Speech and Hearing Disorders, 54(4): p. 482-499, 1989.

[22] H. Kim, M. Hasegawa-Johnson, A. Perlman, Dysarthric speech database for universal access research. Interspeech. pp. 1741-1744, 2008.

[23] X. Menendez-Pidal, J.B. Polikoff, S.M. Peters et al. The nemours database of dysarthric speech, InSpoken Language, 1996. ICSLP 96. Proceedings, Fourth International Conference on 1996 Oct 3 pp. 1962-1965. 1996.

[24] G.E. Hinton. "A practical guide to training restricted boltzmann machines." Neural Networks: Tricks of the Trade. Springer Berlin Heidelberg, pp. 599-619, 2012.