USING ACOUSTIC DEEP NEURAL NETWORK EMBEDDINGS TO DETECT MULTIPLE SCLEROSIS FROM SPEECH

Gábor Gosztolya[1,2]*, László Tóth*[1]*, Veronika Svindt*[3]*, Judit Bóna*[4]*, Ildikó Hoffmann*[3,5]

[1]University of Szeged, Institute of Informatics, Szeged, Hungary
[2]ELRN-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
[3]Research Center for Linguistics, ELRN, Budapest, Hungary
[4]ELTE Eötvös Loránd University, Dept. of Applied Linguistics and Phonetics, Budapest, Hungary
[5]University of Szeged, Department of Linguistics, Szeged, Hungary

ABSTRACT

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system. It affects cognitive and motor functions, and the limitation of executive functions can also manifest itself in speech production. Due to this, automatic speech analysis might serve as an effective technique for assessing MS, or for monitoring the status of the patient. However, choosing the features to be extracted from the recordings is not straightforward. In the past few years, general feature extractors such as i-vectors, d-vectors and x-vectors have found their way into automatic speech analysis. In this study we show that there is no need to employ a special neural network architecture such as x-vectors to calculate effective features, but (even more) indicative features can be derived on the basis of a standard Deep Neural Network acoustic model. From our results, these features could effectively be used to distinguish MS subjects from healthy controls, as we measured AUC scores up to 0.935. We found that classification performance depended only slightly on the choice of the hidden layer used to extract our features, but the speech task performed by the subject turned out to be an important factor.

***Index Terms*—** Multiple Sclerosis, medical speech processing, Deep Neural Networks, embeddings, x-vectors

1. INTRODUCTION

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system. The main diagnostic feature of the disease is the presence of impairments in the patient's gross and fine motor skills. Since language, cognitive, and motor skills are arranged in an inseparable network in the brain, changes in one factor can induce changes in all the others. Due to this, automatic inspection of the speech production of a subject could turn out to be an effective way of examining the progression of MS in patients.

Besides other types of symptoms, roughly two-thirds of MS subjects have slight or modest cognitive impairments such as impaired cognitive flexibility, disorders of orientation, working-memory limitation or decreased information processing speed. About one-third of MS patients report temporary or persistent speech disorders [1, 2]. The most frequent language and speech symptoms include motor speech disorders (e.g. dysarthria or dysphonia), word finding difficulties, limited verbal fluency [3], limitations of the higher-level language processes [4, 5, 6, 7], and a reduced inclination for communication [8]. Although explicit dysarthria is diagnosed only in one-third of the subjects, automatic speech analysis could still be used to detect symptoms suggestive of mild motor speech disorder prior to dysarthria [9]. With a well-structured methodology, these mild symptoms could inform us about the onset of cognitive decline.

To construct an automatic speech analysis process like that, robust features have to be extracted from the speech of the subjects. In the area of medical speech processing, a standard and straightforward approach is to employ general (that is, not specific to the actual disease) feature extractors. For this, one has to keep in mind the data sparsity being typical to the field. That is, since the number of patients is very limited, and data collection is bound by the need of trained personnel (e.g. doctors to diagnose and record specific cognitive tests such as Mini Mental State Examination or Geriatric Depression Scale), the corpora recorded are usually small by Automatic Speech Recognition (ASR) standards. However, there exist quite large datasets collected for ASR purposes. A possible solution is to train some kind of statistical model on such a large external speech corpus, and employ this model to extract features for any standalone speech utterance. Perhaps the best examples for such approaches are i-vectors [10], d-vectors [11] and x-vectors [12].

This research was supported by the Hungarian Ministry of Innovation and Technology (grants no. NKFIH-1279-2/2020 and ELTE TKP2020-IKA-06), by the NRDI Office with grant K-132460 and within the framework of the Artificial Intelligence National Laboratory Program (MILAB). G. Gosztolya was also funded by the János Bolyai Scholarship of the Hungarian Academy of Sciences and by the Hungarian Ministry of Innovation and Technology New National Excellence Program ÚNKP-21-5-SZTE.

Although all three techniques were originally developed for speaker verification, they were later employed as feature extractors in other tasks as well [13, 14, 15]. From a broader perspective, these methods seek to express the difference between "general speech" (represented by the large external corpus) and the actual utterance (produced by the patient).

These techniques differ in their approach for capturing the distribution of this "general speech". i-vectors employ a Gaussian Mixture Model (the so-called Universal Background Model or UBM) to model the distribution of the frame-level features (e.g. MFCCs). However, based on their performance, i-vectors cannot be considered state-of-the-art any more, as they have been surpassed by Deep Neural Network (DNN) based x-vectors. x-vectors utilize a DNN with a special architecture to allow the pooling of frame-level information (processed by the lower, frame-level layers) into utterance-level (handled by the higher layers). The feature extraction step consists of evaluating the fully trained network on the actual utterance, and returning the activations of one of the utterance-level layers (*embedding*) [12]. d-vectors can be viewed as an intermediate solution, where the neural network is trained on the speaker IDs on the frame level, and the utterance-level features are obtained as the activations of the last hidden layer averaged out for the whole recording [11].

Notice that all these techniques employ some specific steps such as training an UBM for i-vectors, or training a separate (and perhaps special) neural network. On the other hand, HMM/DNN hybrid acoustic models (trained for ASR purposes) are available quite commonly. Furthermore, in the past decade the research community has developed techniques to efficiently train such acoustic models. Therefore, using such a DNN acoustic model for feature extraction in a medical task might be beneficial, as long as the performance achievable is competitive. In this study we propose a technique to extract effective features from the activations of such a DNN acoustic model, and show that this approach can outperform the scores of i-vectors and x-vectors.

2. ACOUSTIC DNN EMBEDDING FEATURES

For the above reasons, we will employ a standard feed-forward Deep Neural Acoustic model to extract features. Therefore, the first step of the proposed workflow is to train such a model (where necessary, as there are several such models available). Of course, for this step we need a (larger) external corpus that has annotated and time-aligned phonetic labels. In our view, however, this is not a limitation from a practical point of view, as such datasets are quite easy to obtain. The result of this step will be a (frame-level) HMM/DNN hybrid acoustic model.

In the second step, this DNN model is evaluated on the utterances produced by the subjects. Instead of the output layer (providing the posterior estimates of the context-dependent phonetic states), we focus on the hidden layers, and the activation values of these layers will be noted. Since these vectors are still present at the frame level, we aggregate them over the whole utterance in the third step. For aggregation, we propose four approaches: mean, standard deviation, skewness and kurtosis. We might also concatenate the results of these techniques. The result of this step are the utterance-level feature vectors, being the size of 1-4× the number of neurons in the given hidden layer. These values can be used directly as features in the classification step.

3. THE RECORDINGS USED

All the tests were carried out at the Neurology Department of Uzsoki Hospital Budapest and at the Research Institute for Linguistics of the Eötvös Loránd Research Network in Budapest. The study was approved by the Ethics Committee of the Uzsoki Hospital, and it was conducted in accordance with the Declaration of Helsinki. In the current study we use the recordings of 22 MS subjects (8 males and 14 females) and 19 healthy controls (5 males and 14 females). From the 22 MS subjects, 15 belonged to the relapsing-remitting (RRMS), 3 to the secondary-progressive (SPMS), and 4 to the primary-progressive (PPMS) type; however, in our experiments we did not treat these MS subtypes separately.

We collected the speech samples using the following protocol. The subjects were first asked to talk about their **previous day**. Afterwards, they listened to a two-minute-long historical anecdote that was unknown to them beforehand. The task of the subjects was to summarize the story heard as accurately as possible (**narrative recall**). In the last task, the subjects were asked to read aloud several specific non-words (CVCV sequences), in which the first CVs contained a voiceless plosive [p, t, k] and one of the vowels [i:, a:, u:]) (**phonetics**). This way, we obtained three recordings from each subject, corresponding to the three different tasks, which were also different in nature: two of them were spontaneous speech tasks, differing in the type of recall from memory, while the third one was a simple reading task. We used a Sony PCM-A10 digital dictaphone with a tie clip microphone; the recordings were converted to 16 kHz mono with a 16 bit resolution.

4. EXPERIMENTAL SETUP

4.1. DNN Hybrid Acoustic Model

Our Deep Neural Network acoustic models were trained on a subset of the BEA Hungarian corpus [16], on the speech of 116 subjects (44 hours). We used only spontaneous speech from this corpus; the special vocalizations, being typical for spontaneous speech (e.g. filled pauses, breathing sounds, laughter and gasps) were all marked in the transcriptions, and were included in the phonetic set as special labels.

We used 40 Mel-frequency filter banks along with raw energy as frame-level features along with Δ and $\Delta\Delta$, and

Table 1. The AUC scores obtained for the different speaker tasks, feature subsets and DNN layer embeddings.

	Previous Day			**Narrative Recall**			**Phonetics**		
Feature Subset	Layer 3	Layer 4	Layer 5	Layer 3	Layer 4	Layer 5	Layer 3	Layer 4	Layer 5
Mean	0.639	0.651	0.663	0.787	**0.785**	**0.766**	**0.921**	**0.931**	**0.928**
Std.	**0.754**	**0.775**	**0.801**	**0.818**	0.737	**0.749**	0.899	0.904	**0.926**
Skewness	0.665	0.632	0.734	0.696	0.708	0.742	0.883	**0.923**	**0.938**
Kurtosis	0.646	0.636	0.763	0.718	0.718	**0.768**	0.847	0.907	**0.926**
Mean + std.	0.723	0.711	0.794	0.775	0.754	0.723	**0.938**	**0.933**	0.926
All	0.562	0.641	0.751	0.785	0.785	0.723	0.921	**0.935**	0.928

evaluated our model on a sliding window of 15 frames (1845 frame-level features overall). We utilized 5 hidden layers, each consisting of 1024 ReLU neurons, and a softmax layer that had as many neurons as the number of states. We utilized context-dependent (CD) phonetic mapping. For the 57 phones, we employed the standard tree-based clustering method for state tying [17] with the Kullback-Leibler divergence-based criterion [18], which led to 911 tied states.

4.2. Acoustic DNN Embedding Feature Extraction

We used the activations of the three upper layers (i.e. layers 3, 4 and 5) of the DNN acoustic model, which this led to a 1024-sized vector for each frame. For utterance-level aggregation, we experimented with the arithmetic mean, standard deviation, skewness and kurtosis of the frame-level values; furthermore, we evaluated concatenated versions of the mean and standard deviation, and all four combination methods (which led to 2048 and 4096 attributes, respectively). We standardized the values before utilizing them in the classification step (i.e. we converted them to have zero mean and unit variance).

4.3. Classification and Evaluation

We employed Support Vector Machines to predict whether the speakers belonged to the MS or to the HC group. We used the libSVM implementation [19] with a linear kernel (nu-SVR method); the C complexity parameter was set in the range 10^{-5}, . . ., 10^{1}. Due to the small number of examples, we chose to perform cross-validation (CV); one fold always consisted of the features of one control subject and one having MS. To avoid any form of peeking, we employed *nested cross-validation* [20]; that is, each time we trained our model on the data of 21 folds, *another* (21-fold) cross-validation session was performed, in order to find the C meta-parameter value that gave the highest AUC score within these speakers. Afterwards, we trained an SVM model with the selected C value on the data of all the speakers that belonged to these 21 folds, and this model was evaluated on the (one or two) speakers of the remaining fold. In our first experiment, we focused on the AUC value of the predictions.

5. RESULTS

Table 1 shows the AUC scores obtained for the different speaker tasks, hidden layers and utterance-level aggregation approaches. Among the four standard aggregation methods (mean, standard deviation, skewness and kurtosis), the best values and those being close are shown in **bold**. In general, it seems that taking the activations of the last hidden layer (i.e. Layer 5) lead to higher AUC scores than those corresponding to the lower layer. Since it is well-known that the higher layers of a neural network are usually more task-dependent than the lower ones, this observation can perhaps be interpreted by the fact that Layer 5 captures information which is more related to the phonetic content of the recording. On the other hand, the AUC values corresponding to the lower layers are not remarkably lower; actually, the highest score of 0.818 for the task of Narrative Recall was achieved by using Layer 3.

Regarding the efficiency of the four aggregation approaches, taking the mean and the standard deviation of the activations were, without a doubt, the most successful technique. Although in two times out of the nine tested ones, skewness and kurtosis both led to one of the best values, this occurred in five and six cases, mean and standard deviation, respectively. The most robust case was perhaps taking the standard deviation of the Layer 5 activations. We also experimented with combining (concatenating) these feature vectors (i.e. using early fusion); in these cases, **bold** means that the combined model could surpass all incorporated individual feature sets. Although this approach led to improvements in a few cases, the increase in the AUC vales was never outstanding, indicating that this combination was not really useful.

The largest difference was clearly due to the speaker task utilized. The two spontaneous speech tasks led to quite similar AUC scores: 0.639. . .0.801 and 0.696. . .0.818, Previous Day and Narrative Recall, respectively. Compared to these scores, the AUC values measured for the Phonetics task are remarkably better: they fall in the range 0.847. . .0.938. It is worth noting that, when using the activations of Layer 5, all AUC values were 0.926 or above. The success of this task, in our opinion, shows that it is worth making the subjects utter such phonetic combinations to detect Multiple Sclerosis.

Table 2. Comparison of the i-vector, x-vector and acoustic DNN embedding feature extraction approaches. The best and close-to-best values for a given metric and task are shown as **bold**. (Acc. = classification accuracy, Prec. = precision, Rec. = recall, Spec. = specificity)

Speaker task	Feature set	Acc.	Prec.	Rec.	Spec.	F_1	AUC
Previous Day	i-vectors	73.2%	76.2%	72.7%	73.7%	74.4	0.763
	x-vectors	63.4%	66.7%	63.6%	63.2%	65.1	0.718
	Acoustic DNN embeddings	**78.0%**	**81.0%**	**77.3%**	**78.9%**	**79.1**	**0.801**
Narrative Recall	i-vectors	**73.2%**	**76.2%**	**72.7%**	**73.7%**	**74.4**	0.727
	x-vectors	**73.2%**	**76.2%**	**72.7%**	**73.7%**	**74.4**	**0.754**
	Acoustic DNN embeddings	68.3%	71.4%	68.2%	68.4%	69.8	**0.749**
Phonetics	i-vectors	73.2%	76.2%	72.7%	73.7%	74.4	0.761
	x-vectors	68.3%	71.4%	68.2%	68.4%	69.8	0.727
	Acoustic DNN embeddings	**90.2%**	**90.9%**	**90.9%**	**89.4%**	**90.9**	**0.926**

5.1. Comparison with i-vectors and x-vectors

Although the obtained AUC scores were quite high in some cases, the real potential of a method can be judged only by comparison. For this, we chose i-vectors and x-vectors, calculated by the Kaldi framework [21], and were trained on the same subset of the BEA corpus which we used to train the DNN acoustic model. From the tested variations in Table 1, we chose the "Std." case with taking the activations of Layer 5, as we regarded this to be the most robust configuration.

Following our preliminary tests, we used 20 MFCC features and their Δs for i-vectors as frame-level attributes, while x-vectors were trained on FBANKs. In this experiment we used several evaluation metrics: besides AUC, we calculated classification accuracy, precision, recall, specificity (practically recall for the healthy control category) and F_1. These scores were taken besides Equal Error Rate (EER).

Table 2 shows the achieved metric values; the best values for each speaker task (and those being close to it) are again shown as **bold**. It can clearly be seen that the proposed acoustic DNN embedding feature extraction approach turned out to be, in general, superior to both i-vectors and x-vectors. For two out of the three speaker tasks, this technique led to the best values for all calculated evaluation metrics, and the difference was especially large for task Phonetics. For the Narrative Recall task, the AUC score was better than that of i-vectors, and was on par with x-vectors. For some reason, however, this performance could not manifest in really high accuracy and F_1 values, although the difference was usually moderate (4.5. . .5.3%, absolute). It should also be noted that this was not the best case for the proposed approach, but it was also outperformed by the Layer 5 case for this Narrative Recall speaker task.

6. CONCLUSIONS AND DISCUSSION

In this study we focused on the detection of Multiple Sclerosis from the speech of the subject. We noted that it is standard practice in the medical speech processing area to utilize general techniques (such as i-vectors and x-vectors) for feature extraction from the utterances, as these methods allow one to use general, large speech corpora which are independent of the actual domain (e.g. MS). However, these methods usually have to be trained just for this aim, their training targets might significantly differ from the domain of the application (e.g. training for speaker recognition), and they might have a special DNN structure (just as x-vectors do).

Due to these reasons, we proposed to utilize a standard HMM/DNN hybrid acoustic model in the feature extraction step. These, besides usually having a traditional neural network architecture (for example, we used a simple feed-forward one, without even being a time-delay neural network), also have the advantage of being commonly available. We tried out using the activations of several hidden layers, and (inspired by d-vectors) experimented with four approaches to aggregate the frame-level values into utterance level. We achieved higher AUC scores than traditional i-vectors and x-vectors for two of the three speaker tasks tested, while for the third case we measured competitive scores. For the particular speaker task "Phonetics", all our metrics were around 90%, while the AUC score appeared to be 0.926.

Of course, the metric values of the proposed approach were significantly affected by the speaker task, while it was not true for i-vectors and x-vectors (or only to a much more limited extent). We attribute this to the nature of the feature sets. That is, the main purpose of i-vectors and x-vectors is to reflect the identity of the speakers. On the other hand, the embeddings of a DNN acoustic model (and especially those if its higher layers) are more related to the phonetic content of an utterance. Although definite dysarthria is present only at roughly one-third of the MS patients, it might be the case that similar articulation symptoms might have manifested for the other MS subjects as well. In our hypothesis, this was detected by the proposed feature extraction method, allowing it to significantly outperform the other techniques tested.

7. REFERENCES

[1] K. Laakso, K. Brunnegård, L. Hartelius, and E. Ahlsén, "Assessing high-level language in individuals with multiple sclerosis: A pilot study," *Clinical Linguistics & Phonetics*, vol. 14, no. 5, pp. 329–349, 2000.

[2] S. Renauld, L. Mohamed-Saïd, and J. Macoir, "Language disorders in multiple sclerosis: A systematic review," *Multiple Sclerosis and Related Disorders*, vol. 10, no. Nov, pp. 103–111, 2016.

[3] A. Delgado-Álvarez, J.A. Matias-Guiu, C. Delgado-Alonso, L. Hernández-Lorenzo, A. Cortés-Martínez, L. Vidorreta, P. Montero-Escribano, V. Pytel, and J. Matias-Guiu, "Cognitive processes underlying verbal fluency in multiple sclerosis," *Frontiers in Neurology*, vol. 11, 2021.

[4] F.L. Darley, J.R. Brown, and N.P. Goldstein, "Dysarthria in multiple sclerosis," *Journal of Speech and Hearing Research*, vol. 15, no. 2, pp. 229–245, 1972.

[5] L. Hartelius, B. Runmarker, and O. Andersen, "Prevalence and characteristics of dysarthria in a multiple-sclerosis incidence cohort: Relation to neurological data," *Folia Phoniatrica et Logopaedica*, vol. 52, no. 4, pp. 160–177, 2000.

[6] B. Yamout, N. Fuleihan, T. Hajj, A. Sibai, O. Sabra, H. Rifai, and A.-L. Hamdan, "Vocal symptoms and acoustic changes in relation to the expanded disability status scale, duration and stage of disease in patients with multiple sclerosis," *European Archives of Oto-Rhino-Laryngology*, vol. 266, no. Nov, pp. 1759–1765, 2009.

[7] G. Noffs, T. Perera, S.C. Kolbe, C.J. Shanahan, F.M.C. Boonstra, A. Evans, H. Butzkueven, A. van der Walt, and A.P. Vogel, "What speech can tell us: A systematic review of dysarthria characteristics in Multiple Sclerosis," *Autoimmunity Reviews*, vol. 17, no. 12, pp. 1202–1209, 2018.

[8] F.J. Fitz Gerald, B.E. Murdoch, and H.J. Chenery, "Multiple sclerosis: Associated speech and language disorders," *Australian Journal of Human Communication Disorders*, vol. 15, no. 2, pp. 15–35, 1987.

[9] D. Mulfari, G. Meoni, M. Marini, and L. Fanucci, "Machine learning assistive application for users with speech disorders," *Applied Soft Computing*, vol. 103, no. May, 2021.

[10] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, V. Hubeika, and F. Castaldo, "Support Vector Machines and Joint Factor Analysis for speaker verification," in *Proceedings of ICASSP*, 2009, pp. 4237–4240.

[11] E. Variani, X. Lei, E. McDermott, I.L. Moreno, and J. G-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proceedings of ICASSP*, 2014, pp. 4080–4084.

[12] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker verification," in *Proceedings of ICASSP*, 2018, pp. 5329–5333.

[13] J. Grzybowska and S. Kacprzak, "Speaker age classification and regression using i-vectors," in *Proceedings of Interspeech*, 2016, pp. 1402–1406.

[14] M. Huckvale, A. Beke, and M. Ikushima, "Prediction of sleepiness ratings from voice by man and machine," in *Proceedings of Interspeech*, 2020, pp. 4571–4575.

[15] J.V. Egas-López, M. Vetráb, L. Tóth, and G. Gosztolya, "Identifying conflict escalation and primates by using ensemble x-vectors and Fisher vector features," in *Proceedings of Interspeech*, 2021, pp. 476–480.

[16] T. Neuberger, D. Gyarmathy, T.E. Gráczi, V. Horváth, M. Gósy, and A. Beke, "Development of a large spontaneous speech database of agglutinative Hungarian language," in *Proceedings of TSD*, 2014, pp. 424–431.

[17] J.J. Odell, *The Use of Context in Large Vocabulary Speech Recognition*, Ph.D. thesis, University of Cambridge, 1995.

[18] G. Gosztolya, T. Grósz, L. Tóth, and D. Imseng, "Building context-dependent DNN acousitc models using Kullback-Leibler divergence-based state tying," in *Proceedings of ICASSP*, 2015, pp. 4570–4574.

[19] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[20] G.C. Cawley and N.L.C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2079–2107, 2010.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proceedings of ASRU*, 2011.