Extracting Phonetic Posterior-Based Features for Detecting Multiple Sclerosis From Speech

Gábor Gosztolya, Veronika Svindt, Judit Bóna, and Ildikó Hoffmann

***Abstract*— Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system which, in addition to affecting motor and cognitive functions, may also lead to specific changes in the speech of patients. Speech production, comprehension, repetition and naming tasks, as well as structural and content changes in narratives, might indicate a limitation of executive functions. In this study we present a speech-based machine learning technique to distinguish speakers with relapsing-remitting subtype MS and healthy controls (HC). We exploit the fact that MS might cause a motor speech disorder similar to dysarthria, which, with our hypothesis, might affect the phonetic posterior estimates supplied by a Deep Neural Network acoustic model. From our experimental results, the proposed posterior posteriorgram-based feature extraction approach is useful for detecting MS: depending on the actual speech task, we obtained Equal Error Rate values as low as 13.3%, and AUC scores up to 0.891, indicating a competitive and more consistent classification performance compared to both the x-vector and the openSMILE 'ComParE functionals' attributes. Besides this discrimination performance, the interpretable nature of the phonetic posterior features might also make our method suitable for automatic MS screening or monitoring the progression of the disease. Furthermore, by examining which specific phonetic groups are the most useful for this feature extraction process, the potential utility of the proposed phonetic features could also be utilized in the speech therapy of MS patients.**

***Index Terms*— Multiple sclerosis, deep neural networks, DNN acoustic models, phonetic posteriors.**

I. Introduction

MULTIPLE sclerosis (MS) is a chronic inflammatory disease of the central nervous system [1]. Depending on the constant presence or temporal absence of symptoms, the majority of the cases are grouped into three different clinical courses: relapsing-remitting MS (RRMS), primary progressive MS (PPMS), and secondary progressive MS (SPMS), which develops from RRMS [2]. Impairments in motor skills are a central diagnostic feature in MS. Any changes in the patient's gross and fine motor skills may suggest a deterioration in the condition. Because 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. Therefore, monitoring the changes in the speech production could be effective for detecting the onset of the progression in the MS cases.

Among other symptoms, 60-70% of the MS patients have different cognitive impairments (e.g. executive function deficits such as a working-memory limitation, decreased information processing speed, or impaired cognitive flexibility, disorders of orientation, chronic fatigue), most often associated with a deteriorating quality of life. More than a third of people with MS report temporary or persistent speech disorders [3], [4]. The most frequent language and speech symptoms are motor speech disorders (dysarthria, dysphonia), word finding difficulties, limited verbal fluency [5], sentence repetition problems, limitations of the higher-level language processes [6], [7], [8], and reduced inclination for communication [9]. Although dysarthria is diagnosed only in one-third of the patients, automatic speech analysis could detect symptoms suggestive of mild motor speech disorder prior to dysarthria [10]. With a well-structured methodology, these mild symptoms could inform us about the onset of cognitive decline.

In the speech technology community, many studies have been published which focus on the automatic processing of speech with a motor speech disorder or even dysarthria. The goal of several studies is to adapt existing automatic speech recognition (ASR) systems to best suit the transcription of dysarthric speech [11], [12], [13]. Another significant research topic is that of transforming the speech to make it more understandable, by voice conversion techniques [14], [15], while several studies deal with dysarthric speech intelligibility assessment, where the goal is to measure the precise degree of failure in the motor control of speech muscles (for example by applying ASR to the speech utterances and counting the number of misidentified words) [16], [17], [18]. In contrast with these studies, we seek to develop an appropriate feature extractor approach that might allow us to detect Multiple Sclerosis (and, if possible, measure its severity), based on the dysarthric speech properties of MS.

Manuscript received 1 August 2022; revised 4 December 2022, 28 April 2023, and 17 July 2023; accepted 22 July 2023. Date of publication 7 August 2023; date of current version 14 August 2023. This work was supported in part by the National Research, Development and Innovation (Office of the Hungarian Ministry of Innovation and Technology) under Grant K-132460 and Grant TKP2021-NVA-09 and in part by the Framework of the Artificial Intelligence National Laboratory Program under Grant RRF-2.3.1-21-2022-00004. *(Corresponding author: Gábor Gosztolya.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of the Uzsoki Hospital, and performed in line with the Declaration of Helsinki.

Gábor Gosztolya is with the ELKH-SZTE Research Group on Artificial Intelligence, 1016 Szeged, Hungary, and also with the Institute of Informatics, University of Szeged, 6720 Szeged, Hungary (e-mail: ggabor@inf.u-szeged.hu).

Veronika Svindt and Ildikó Hoffmann are with the Research Center for Linguistics, 1068 Budapest, Hungary.

Judit Bóna is with the Department of Applied Linguistics and Phonetics, ELTE Eötvös Loránd University, 1053 Budapest, Hungary.

Digital Object Identifier 10.1109/TNSRE.2023.3300532

Our study lies in the area of pathological speech processing, where the goal is to identify the subjects who suffer from a particular (cognitive of physical) disease based on their speech samples, or to estimate the severity of the disease. Such systems would be ideal for screening purposes, or for monitoring the progression of specific diseases in a cheap, automatic and contact-free manner. Particularly in the last decade, several studies had been published, focusing for example on Alzheimer's disease [19], [20], Parkinson's Disease [21], [22], aphasia [23] or depression [24], [25]. Regarding the case of Multiple Sclerosis, besides calculating manual features such as speech rate and pause duration [26] or various word timing measures in fluency tasks [27], several studies found statistically significant differences in standard attributes such as jitter, shimmer, F_0, F_1 / F_2, volume (or loudness variation) and harmonics-to-noise ratio [28], [29], [30]. Looze et al. [31] calculated temporal parameters manually as well as pitch, speech tempo and pause duration (in a semi-automatic manner, based on Praat). Surprisingly, all the studies listed above employed only statistical analyses of the attributes, the only exception being de Looze et al.: by employing automatic classification (linear discriminant analysis), they were able to identify MS patients with an AUC score of 0.70 [31].

In this study, we construct a completely automatic workflow for MS detection from the speech of the subjects. Using a standard classification technique (Support Vector Machines), we apply a neural network-based feature extraction method, exploiting the fact that deficiencies in the motor control of speech muscles (and therefore, MS-caused such deficiencies as well) affect the pronunciation of specific phonemes. With our hypothesis, similarly to the case of Parkinson's disease [32], these changes are reflected in the tendency of the phonetic posteriors of a standard deep neural network (DNN) acoustic model. As the pronunciation of (specific) phonemes becomes less precise, the corresponding posterior estimates can be expected to drop, and / or their variance can be expected to increase. A further advantage of a phonetic feature extraction method is the interpretable nature of the attributes: since they are directly linked to specific phones of the given language, they might have a straightforward application in the speech therapy of the subjects.

The basis of our approach is to calculate the so-called *phonetic posteriorgrams* for the actual speech response of the subjects. Phonetic posteriorgrams store the probability vectors for each phonetic label or state for short time segments [33], [34], [35]. Since they are *local* vectors, tied to a short time segment of a larger speech recording, they are not suitable for directly representing a complete speech utterance, or used as features for subject (or subject category) discrimination. This noted, they are frequently utilized for voice conversion [36], [37] and rating articulation quality (e.g. dysarthria) [38], [39]. They have also been employed for improving speech recognition performance under noisy conditions [40], voice disorder severity estimation [41] and query-by-example spoken term detection [33]. Arias-Vergara et al. [42] built a classification workflow to evaluate the read speech of cochlear implant users. By relying on the manual transcript of the recordings and obtaining a phonetic categorical time alignment by forced alignment, they achieved average F_1 values between 0.36 and 0.65 with an SVM.

TABLE I
THE DEMOGRAPHIC ATTRIBUTES OF THE SUBJECTS INVOLVED IN OURSTUDY

		Speaker Groups		Statistics
		MS ($n = 23$)	HC ($n = 22$)	p
Age	mean ± std range	39.00 ± 8.11 [24, 56]	39.95 ± 7.22 [28, 56]	$p = 0.685$
Gender	m / f	5 / 18	6 / 16	$p = 0.536$
Education	mean ± std range	15.05 ± 2.17 [12, 19]	16.09 ± 1.26 [12, 19]	$p = 0.100$

Regarding pathological speech processing applications, Černak et al. [43] calculated phonetic-categorical posteriorgrams and analyzed the distributions of the posterior estimates for Parkinsonian speakers and healthy controls by statistical tests. Klumpp et al. [32] also employed statistical tests to compare the posterior estimates along with the activations of the acoustic neural network for Parkinsonian and HC subjects.

The main contributions of our study are the following: (i) we calculate an utterance-level statistical feature vector from the phonetic posteriorgrams, (ii) we utilize these vectors as features in a machine learning step, predicting the subject category (whether the given subject suffers from MS or is a healthy control), (iii) we retain only specific phonetic categorical features to study their effect on classification performance. To the best of our knowledge, no study employed the first two items together (that is, constructing a completely automatic workflow of phonetic feature extraction and subject classification by machine learning) in pathological speech processing in general, nor for Multiple Sclerosis subjects in particular. Furthermore, the phonetic feature extraction method proposed does not require the presence of the time-aligned ground truth phonetic labels, therefore its application is not limited to read speech, but it can be used for free-form spontaneous speech as well. This, and the fact that the features can be broken down into phonetically interpretable subsets (i.e. contribution (iii)) might allow the feature extraction approach to be applied in the speech therapy of MS subjects in a straightforward way.

II. THE RECORDINGS USED

The tests were carried out at the Neurology Department of Uzsoki Hospital, Budapest, Hungary, and at the Research Center for Linguistics of the Eötvös Loránd Research Network, Budapest, Hungary. 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 23 MS subjects and 22 healthy controls. All 23 MS subjects belonged to the relapsed-remitting MS subtype (RRMS).

All the speakers involved in the study were native Hungarian speakers; and, mirroring the ethnic composition of Hungary, all of them were Caucasians. None of them had any hearing impairment, depression or any other known psychiatric

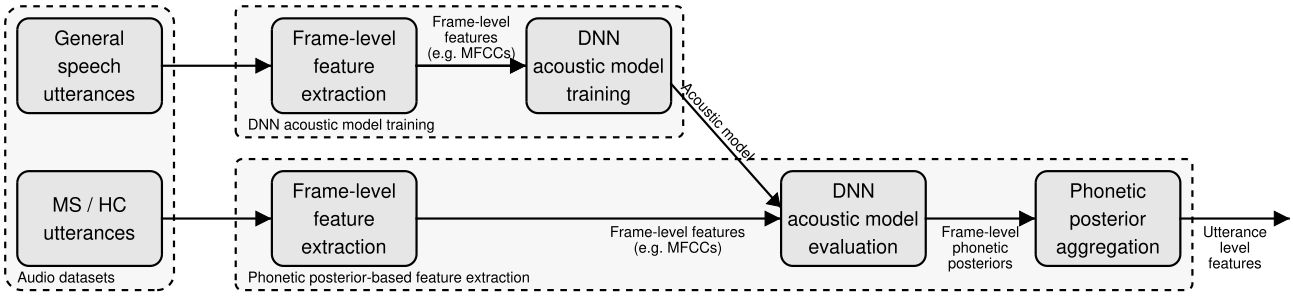

Fig. 1. The general workflow of the proposed, phonetic posterior-based feature extraction process.

condition. The demographic attributes (age in years, gender (male / female) and years of education) of the two subject groups can be seen in Table I. We used one-way ANOVA for the age and years of education attributes, and χ^2 test for gender; clearly, the MS and HC groups display no statistically significant difference ($p > 0.05$) in either of these attributes.

The linguistic protocol for collecting the speech samples from the subjects was quite extensive; in this study we use six utterances from each subject. These were, in the order they were recorded (which was the same for all subjects):

1) **Everyday:** First, the subjects were asked to talk about their everyday life (in a spontaneous manner), such as their work and how Multiple Sclerosis affects their life.
2) **Previous Day:** In this task, the subjects were asked to talk about the events of their previous day in detail.
3) **Opinion:** Next, the subjects were asked to share their opinions about vegetarianism.
4) **Narrative Recall:** Afterwards, the subjects listened to a two-minute-long historical anecdote which was unknown to them beforehand. The task of the subjects was to summarize the story heard as accurately as possible.
5) **Hobby:** Next, the subjects were asked to talk about their hobbies. This was a fairly similar task to **Everyday**.
6) **Phonetics:** In the last task, the subjects were asked to read aloud several non-words (consonant-vowel-consonant-vowel (CVCV) sequences, in which the first CVs contained a voiceless plosive [p, t, k] and one of the vowels [i:, a:, u:]).

Our aim was to have tasks which differed in the activated cognitive processes and the rate of the cognitive load required for speech production. In the two spontaneous speech tasks (everyday and hobby), speakers created personal narratives. The production quality of these two types of narratives might be useful for assessing some neurodenegerative diseases [44]. Speaking about our previous day activates the episodic memory, and requires temporal organization. This type of task proved beneficial for measuring subtle cognitive difficulties in other neurodegenerative conditions, such as mild cognitive impairment and Alzheimer disease [45], [46]. An argumentation or opinion task requires the building of complex narratives, activating social knowledge and personal experiences, and inferencing, respectively. The narrative recall task is one of the most difficult spontaneous speech tasks: it requires a set of cognitive processes, such as focused attention, working memory, temporal orientation, organization and sequencing [47]. In the phonetic task, we used voiceless consonants that are sensitive even to mild motor speech problems [48].

The recording was performed with a Sony PCM-A10 digital dictaphone through a tie clip microphone with a sampling rate of 48 kHz; later the recordings were converted to 16 kHz mono with a 16 bit resolution.

III. Phonetic Posterior-Based Feature Extraction

Our assumption was that, since motor control deficiencies of the speech muscles are characterized by a poor phonetic articulation, they can also be expected to affect the accuracy of a machine learning speech recognition system. In particular, we decided to focus on the phonetic posterior estimates provided by a DNN acoustic model; the DNN component of a HMM/DNN hybrid model, being a standard technique for automatic speech recognition [49].

Next, we will briefly describe the general concept of this deep learning acoustic model. A schematic workflow of the proposed feature extraction process is given in Fig. 1. In the following, we will use the standard term "utterance" in the sense of "a speech clip processed at once", which, in our case, is the whole audio clip (i.e. the response of the subject).

A. Frame-Level Feature Extraction for Audio

First, we perform some standard preprocessing steps of the audio recording, which are standard in automatic speech recognition [49]. First the spectral representation of the speech signal is calculated, which practically gives the intensity of a given frequency at a given time in the input audio signal. To summarize this information, this spectrogram is processed with a Mel-scale filter, which models human hearing as it focuses on the lower frequency range, representing this part of the spectrum in more detail. Besides processing the frequency axis, the energies are also summarized over the time axis with a sliding window with a length of typically 25ms and a step size of 10ms; this results in the concept of *frames* [49]. After this processing step, we obtain the local energies of each frequency band in 10ms steps, which gives 100 frames for each second of audio. Besides these raw filter bank energies ("FBANK"), it is also common to transform the values using a Discrete Cosine Transform (DCT) to obtain the Mel-Frequency Cepstral Coefficients or MFCCs [49]. In our actual experiments, we chose to utilize filter bank energies

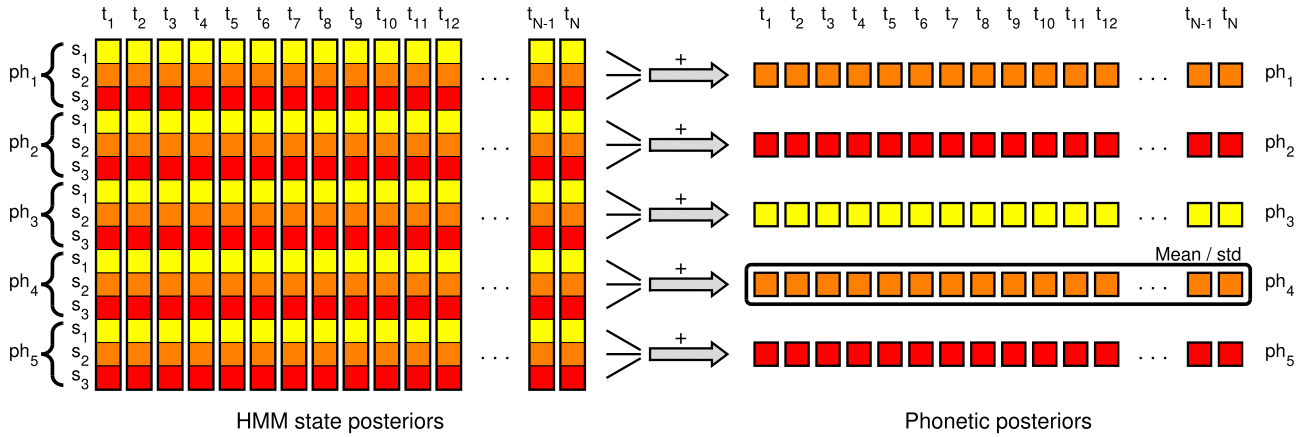

Fig. 2. The schematic diagram of the feature extraction process applied from the frame-level state posterior estimates, shown for a context-independent acoustic model. The vertical bars represent the monophone (but tri-state) posterior vectors for frame t_i, belonging to each possible phoneme ph_j. The circled region is used to calculate the means and standard deviations.

(see Section IV-A for more details), since, according to the literature, they allow a higher speech recognition performance [50]. However, we do not regard the choice of the frame-level feature set as a part of the proposed approach, as other sets could also be utilized as the input for the DNN acoustic model.

B. HMM/DNN Hybrid Models

In the next step, we employ a Deep Neural Network (DNN) to obtain the frame-level estimates of the $P(c_k|x_t)$ conditional probabilities, where x_t is the frame-level feature vector of the tth frame, and c_k is the kth phonetic class. In this step we practically obtain local (that is, frame-level) probability estimates of which phoneme was uttered at the given time, but for the sake of more efficient phonetic modelling, each phoneme is modelled by several phonetic classes (i.e. states, see Section III-C). Relying on these frame-level likelihoods, in ASR (when employing a HMM/DNN hybrid model) a Hidden Markov Model (i.e. HMM) calculates the most probable state sequence for the complete utterance; that is, for each frame it supplies the most probable class c_k (i.e. phonetic state), taking *the whole utterance* into account. This sequence is used to obtain the (word-level) transcript of the speech utterance with its time alignment (i.e. for each word its starting and ending time points within the utterance are also provided). In this study, however, we focus only on the phonetic posterior estimates, so this search step (i.e. the application of the Hidden Markov Model) is omitted.

We are aware that in the last decade, HMM/DNN hybrids have been superseded by other deep learning-based approaches in ASR as state-of-the-art. For example, one might utilize recurrent neural architectures (applying units such as Long-short term memory (LSTM, [51]) and Gated Recurrent Units (GRUs, [52]) as building blocks), or end-to-end models which have become quite popular in the past couple of years [53], [54]. Still, there are several reasons for employing the HMM/DNN model instead of applying one of these more sophisticated approaches. These include easier training, and (in the case of limited training data) a competitive or even superior performance [55], [56]. Furthermore, traditional HMM/DNN hybrids have a lower computational complexity and a smaller memory footprint, which is desirable for a pathological screening application. This is even more important when we seek to use only the acoustic model of the ASR system for feature extraction. Lastly, we would like to point out that these more sophisticated models are usually trained with a Connectionist Temporal Classification (CTC, [57]) loss. Due to this, the trained models tend to produce sparse and arbitrary posterior strike timings [58], [59], which are clearly not as straightforward to post-process either for model combination or for feature extraction as those provided by standard HMM acoustic models [59].

C. Phonetic States

The states of a HMM system are related to the phonetic set of the given language, but usually there is no direct one-to-one correspondence, as the states typically represent a finer resolution. First, one might decide to also model several acoustic phenomena like filled pauses, noises, breathing, gasps and coughs by assigning special models to them, although these vocalizations do not correspond to phones in the strict sense. Second, the phones are traditionally divided into three production states, as it is known to improve recognition performance [49]. Third, instead of working with such simple, context-independent (CI) phone labels, even better speech recognition results can be achieved by context-dependent (CD) modelling [50], where the phonetic labeling also takes the (left and right) neighbors of the actual phone into consideration. As in this HMM/DNN hybrid model, the role of the DNN acoustic model is to estimate the local (i.e. frame-level) posteriors of the HMM states, the number of the DNN outputs being the same as the number of HMM states. In this study we decided to employ context-dependent phonetic modelling, since it allows more precise phonetic posterior estimations.

D. Phonetic Posterior Aggregation

The DNN acoustic model supplies the class-conditional likelihood $P(c_k|x_t)$ values for the c_k phonetic classes

($1 \leq k \leq K$) and the x_t frame-level feature vectors (e.g. FBANKs, MFCCs), where $1 \leq t \leq T$, T being the duration of the actual utterance in frames. In our feature extraction approach, next we merge the probability estimates for each phone. That is, let ph_i denote the ith phone in the given phonetic set ($1 \leq i \leq N$) and let S_i denote the set of phonetic classes corresponding to ph_i. Then we can calculate

$$P(ph_i|x_t) = \sum_{c_k \in S_i} P(c_k|x_t) \tag{1}$$

for each phone, i.e. for $1 \leq i \leq N$. These $P(ph_i|x_t)$ values form the phonetic posteriorgrams [33], storing the local (frame-level) posterior estimates for each phone in the given language. Unfortunately, these cannot be directly employed as utterance-level features in a classification workflow, since the number of these $P(ph_i|x_t)$ aggregated estimates still depends on the duration of the utterance (i.e. T). To eliminate this dependency, next we take their mean and standard deviation values. That is,

$$\mu_i = \frac{1}{T} \sum_{t=1}^{T} P(ph_i|x_t) \tag{2}$$

and

$$\sigma_i = \sqrt{\frac{1}{T} \sum_{t=1}^{T} \Big(P(ph_i|x_t) - \mu_i\Big)^2} \tag{3}$$

are calculated. These values can then be readily used as *utterance-level* features to classify the subjects as either those having MS or as healthy controls. The scheme of this feature extraction step can be seen in Figure 2 (where, for the sake of simplicity, context-independent states are shown).

IV. Experimental Setup

A. DNN Hybrid Acoustic Model

Our Deep Neural Network acoustic models were trained on a subset of the BEA Hungarian corpus [60]; we trained the DNN on the speech of 165 subjects (60 hours of recordings). We used 40 Mel-frequency filter banks along with the raw energy as frame-level features along with Δ and $\Delta\Delta$. To improve model accuracy, we evaluated our model on a sliding window with a width of 15 frames (1845 frame-level features overall). Following this, we utilized 5 hidden layers, each consisting of 1024 ReLU neurons. Lastly, we included a softmax layer that had as many neurons as the number of states. We chose to employ the context-dependent approach for phonetic modelling [61], by which we obtained $K = 911$ phonetic states. These belonged to $N = 57$ phones, including silence, filled pauses and breathing noises. When evaluating this acoustic HMM model using a phone-level recognition (with a simple phone bigram, without any word-level information) on a test set withheld from training from the BEA corpus (roughly 3 hours and 24 minutes from 10 speakers), we measured a phonetic error rate of 25.8%.

B. Classification

In pathological speech processing tasks it is quite common to have a limited number of subjects, which correspond to the examples in the classification tasks. Under these circumstances, the Support Vector Machine (SVM) classification algorithm proved to be quite robust, so it is quite popular in this area. We employed SVMs with a linear kernel, using the libSVM implementation [62]. We utilized a cross-validation technique, where each fold consisted of 1 MS and 1 HC subject (with the exception of one fold, which consisted of 1 MS subject only) to retain the original class distribution as accurately as possible; this led to 23 folds overall. Hence, each classifier model was trained on the speech of 43 subjects (i.e. on 22 folds), and it was evaluated on the speaker(s) of the remaining fold. The C complexity parameter was set in the range $10^{-5}, 10^{-4}, \ldots, 10^{2}$.

The complexity C meta-parameter of the SVM was set by a technique called *nested cross-validation* [63]. That is, each time we trained on the data of 22 folds, we performed *another* (22-fold) cross-validation session, looking for the C meta-parameter value that led to the highest AUC score within these speakers. Afterwards, we trained an SVM model with the selected meta-parameters on the data of all speakers belonging to these 22 folds, and this model was evaluated on the remaining speaker. This way we avoided any form of peeking, which would have created a bias in our scores, had we used standard cross-validation.

C. Evaluation Metrics

We applied evaluation metrics which are commonly used in biomedical studies (e.g. [64], [65]). First, we utilized Equal Error Rate (EER), i.e. the decision threshold between the posterior estimates of the two classes (provided by the SVM classifier) was set so as to minimize the difference between sensitivity and specificity (i.e. the recalls for the two speaker categories). Since this practice in a balanced class distribution leads to quite similar accuracy, precision, recall and F-measure scores, among these we report only EER (i.e. 100%− Accuracy) along with the area under the ROC curve (AUC). As we have only two speaker categories, the AUC value of the two appears to be the same.

D. X-Vectors Baseline

As a competitive baseline, first we used the so-called *x-vector* features [66]. Originally developed for speaker recognition, x-vectors were later employed as features in a wide variety of speech analysis tasks ranging from emotion recognition [67] through sleepiness detection [68] to various pathological speech processing tasks [22], [69].

x-vector features are extracted from the speech recordings by special-structure neural networks, with two distinct parts. The lower layers operate at the *frame level*: the input of the first hidden layer are actually the frame-level feature vectors of the speech recording. After these frame-level layers, there is a special *statistics pooling* layer, which aggregates the activations of the last frame-level hidden layer for all frames of the utterance by taking their mean and standard deviation.

These values are then used as input for the next, *segment-level* layers. The last layer is the *softmax* output layer [66]. This special neural network is trained to detect speakers; therefore, it has the same number of neurons in its output layer as the number of speakers in the training dataset. During the evaluation, the activations of some segment-level layers are taken. These activations (or embeddings) capture information from the speakers over the whole audio-signal, and they are called the *x-vectors* [66], [69].

In our case, this x-vector network was trained on the very same 60-hour subset of the BEA Hungarian Database as our DNN acoustic model was. Although it is standard practice during x-vector training to add noise and reverberation both to increase training data size and to improve the noise robustness of the model, in our preliminary tests we found this procedure to be counterproductive. (This is probably because our MS recordings are of good acoustic quality.) Similar to the DNN acoustic model, we used 40 Mel-frequency filter bank energies ("FBANK") as frame-level features. We employed the Kaldi toolkit both for the x-vector model training and evaluation (i.e. feature extraction) [70].

E. The ComParE Functionals Baseline

Another common attribute set was applied as a baseline: we also utilized the 6373-sized 'ComParE functionals' feature set [71], extracted by the openSMILE tool [72]. The feature set includes energy, spectral, cepstral (MFCC) and voicing related frame-level attributes, which serve as the basis of utterance-level aggregation by specific functionals (like the mean, standard deviation, 1st and 99th percentiles, peak statistics, etc.). This method was utilized in dozens of speech processing tasks such as estimating the degree of nativeness [73] and sleepiness [74], and detecting stuttering [75].

V. Results

Table II shows the metric scores obtained for all six speech tasks investigated. For the two baseline approaches, perhaps the most obvious observation is that the scores significantly vary with the speech task: the recordings obtained from 'Previous Day' were by far the most suboptimal ones (at least, from an automatic MS-HC discrimination perspective), leading to EER scores of 40% (meaning that only 60% of the subjects were actually categorized correctly) and 35.6%, x-vector and ComParE functionals, respectively. The AUC scores of 0.725 and 0.713 (measured for the same case) are not that high either. In contrast, using the x-vector features for the 'Opinion' task led to a low (13.3%) EER score, and we measured high AUC scores (i.e. 0.879 or over) for the tasks 'Opinion' and 'Hobby', and for 'Narrative Recall' and 'Phonetics', x-vectors and ComParE functionals, respectively. For the other speech tasks, the metric scores were in between these extreme values, with EER values between 22% and 27% and AUC scores between 0.775 and 0.850.

Regarding the EER and AUC values achieved by the proposed, phonetic posterior-based features, we highlighted in **bold** those cases where our method outperformed both x-vectors and ComParE functionals (i.e. it obtained a lower EER or a higher AUC score). This was the case in roughly half of the cases: out of the six speech tasks, a lower EER was obtained in two cases, while a higher AUC score was produced for three speech tasks. Notably, scores in the worst cases were better than either for x-vectors and for ComParE functionals: we obtained an EER score of 31.1% and an AUC value of 0.739 (both for the 'Previous Day' task), while these values were 35.6% and 40% (EER), and 0.725 and 0.713 (AUC) for the two baseline feature sets.

TABLE II
THE EQUAL ERROR RATE (EER) AND AUC SCORES OBTAINED WITH THEPROPOSED PHONETIC-BASED FEATURES

Features	Speech Task	EER	AUC
x-vectors (baseline)	Everyday	26.7%	0.775
	Previous Day	40.0%	0.725
	Opinion	13.3%	0.883
	Narrative Recall	22.2%	0.850
	Hobby	17.8%	0.881
	Phonetics	22.2%	0.775
ComParE functionals (baseline)	Everyday	22.2%	0.802
	Previous Day	35.6%	0.713
	Opinion	26.7%	0.810
	Narrative Recall	17.8%	0.905
	Hobby	26.7%	0.830
	Phonetics	17.8%	0.879
Phonetic posterior-based (proposed)	Everyday	**13.3%**	**0.872**
	Previous Day	**31.1%**	**0.739**
	Opinion	17.8%	**0.891**
	Narrative Recall	22.2%	0.824
	Hobby	22.2%	0.834
	Phonetics	22.2%	0.759

Overall, the proposed, phonetic posterior-based features led to a more balanced, more robust performance across the six speech tasks: while with x-vectors, the EER values lay in the range $[13.3\%, 40.0\%]$ and the AUC scores also lay in a large interval (i.e. $[0.725, 0.883]$), these ranges being notably smaller with the proposed approach (EER scores in the range $[13.3\%, 31.1\%]$ and AUC values in the range $[0.739, 0.891]$). This is reflected in the mean, median and standard deviation values as well (see Table III): the phonetic posterior-based features led to a higher mean EER and to lower mean and median AUC scores, while the standard deviations were lower for both metrics. This means that, although the actual speech task affects the MS classification performance to some extent, the phonetic posterior-based features proved to be more robust in this aspect than the x-vector and the ComParE functionals features. This finding, along with the (slightly) better mean and median values, the more compact feature vector and the interpretable nature of our attributes confirm the feasibility of the proposed feature extraction approach for detecting Multiple Sclerosis.

VI. Investigating the Phonetic Subsets

Besides performing machine learning experiments, we also wanted to investigate the potential improvement of speech therapy of MS patients. Therefore, we sought to determine the actual phones which prove to be useful as the basis for useful features. This is why, besides using *all the phones* in our phonemic set, next we investigate the usefulness of

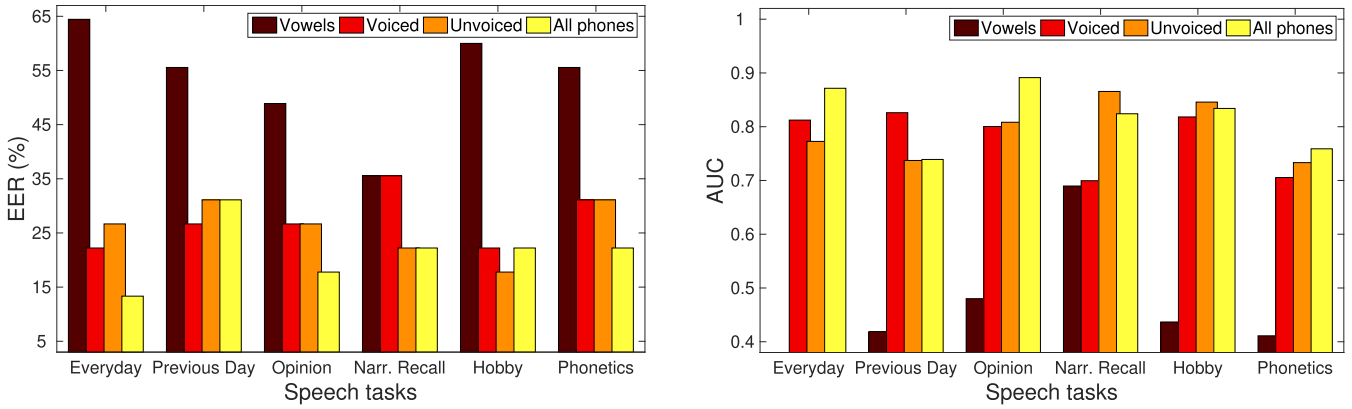

Fig. 3. The Equal Error Rate (left) and Area Under Curve (right) values for the six speech tasks for four phonetic categories (vowels, voiced consonants, unvoiced consonants, and all phones).

TABLE III
MEAN, MEDIAN AND STANDARD DEVIATION VALUES OF THE EQUALERROR RATE (EER) AND AUC SCORES, AGGREGATED FOR THE SIX SPEECHTASKS

		EER	AUC
x-vectors	Mean	23.7%	0.815
	Median	22.2%	0.813
	Standard dev.	9.2%	0.066
ComParE functionals	Mean	24.4%	0.823
	Median	24.4%	0.820
	Standard dev.	6.7%	0.067
Phonetic posteriors-based	Mean	21.5%	0.820
	Median	22.2%	0.829
	Standard dev.	5.9%	0.060

specific phonetic *groups*. This practically means that, after the phonetic posterior aggregation step of our feature extraction workflow (see Section III-D), we keep only the attributes (i.e. mean and standard deviation values) corresponding to specific phonemes in our feature vectors. The rest of our machine learning steps (e.g. nested cross-validation classification, SVM hyperparameter selection, evaluation) is carried out exactly as before, i.e. like that described in Section IV. This also means that the EER and AUC values obtained are directly comparable to those reported so far (i.e. those in Table II); as a matter of fact, we will use them as reference values.

To construct our phonetic groups, we followed the phonetic categories based on the work of Megyesi [76]. See Table IV for the broader categories of Hungarian consonants in SAMPA notation. First, we used the following categories:

1) **Vowels** This category consisted of the phones [O, a:, E, e:, i, o, 2, u, y],
2) **Voiced consonants** This category consisted of the phones [b, d, d', g, dz, j', v, z, z', h, m, n, n', l, r, j],
3) **Unvoiced consonants** This category consisted of the phones [p, t, t', k, ts, c', f, s, s'].

The results obtained with these phonetic categories can be seen in Figure 3. For reference, we also included the values corresponding to using all phonetic posterior-based features (the 'All phones' case), i.e. those reported in the bottom part of Table II. Relying on the vowel-based features led to a very low discrimination performance: the Equal Error Rate values were as high as 64.5%, while the AUC scores turned out to be lower than 0.500. The sole exception to this was the 'Narrative Recall' speech task, but even in this case, the metric scores were the worst among those measured.

Regarding the cases of using only the voiced or only the unvoiced consonants, the results were actually much better (and more consistent as well): we obtained EER scores between 17.8% and 35.6%, and AUC scores between 0.700 and 0.866. The values, in some cases, even exceeded the reference values (i.e. those measured with using all the phones), but the difference is probably not statistically significant. This, in our view, indicates that even a subset of the (already compact) phonetic posterior-based features allows an efficient discrimination of the MS and HC subjects, and there are useful attributes that correspond both to voiced and to unvoiced phones.

Next, we shall investigate another categorization of the (Hungarian) phones; i.e., we will use the following categories:

1) **Nasals** This category consisted of the phones [m, n, n'],
2) **Fricatives** This category consisted of the phones [f, s, s', v, z, z', h],
3) **Plosives** This category consisted of the phones [p, t, t', k, b, d, d', g],
4) **Affricates** This category consisted of the phones [ts, c', dz, j'].

Notice that all the categories include both voiced and unvoiced phones. The results obtained with the phonetic features derived from these phonetic categories can be seen in Figure 4. Relying on the attributes derived from the 'Fricatives' phonetic category led to the least effective discrimination performance: for two speech tasks ('Everyday' and 'Narrative Recall'), the EER and AUC values obtained actually fell below the level attainable by random guessing. These attributes were also clearly worse than most of the other phonetic categories or the 'All phones' case used for reference (speech tasks 'Opinion' and 'Hobby'). Surprisingly, though, for the 'Phonetics' speech task we obtained an EER value matching the 'All phones' case (22.2%) and the third-best AUC value of all the investigated phonetic groups in this experiment (AUC = 0.870). This might be due to the fixed content of this speech

TABLE IV

THE PHONETIC CATEGORIES OF THE HUNGARIAN CONSONANTS,FOLLOWING THE WORK OF MEGYESI [76]

	Labial / labiodental		Dental / alveolar		Palatal		Velar		Glottal
	unvoiced	voiced	unvoiced	voiced	unvoiced	voiced	unvoiced	voiced	unvoiced
Plosives	p	b	t	d	t'	d'	k	g	
Affricates			ts	dz	c'	j'			
Fricatives	f	v	s	z	s'	z'			h
Nasals		m		n		n'			
Laterals				l					
Tremulants				r					
Glides						j			

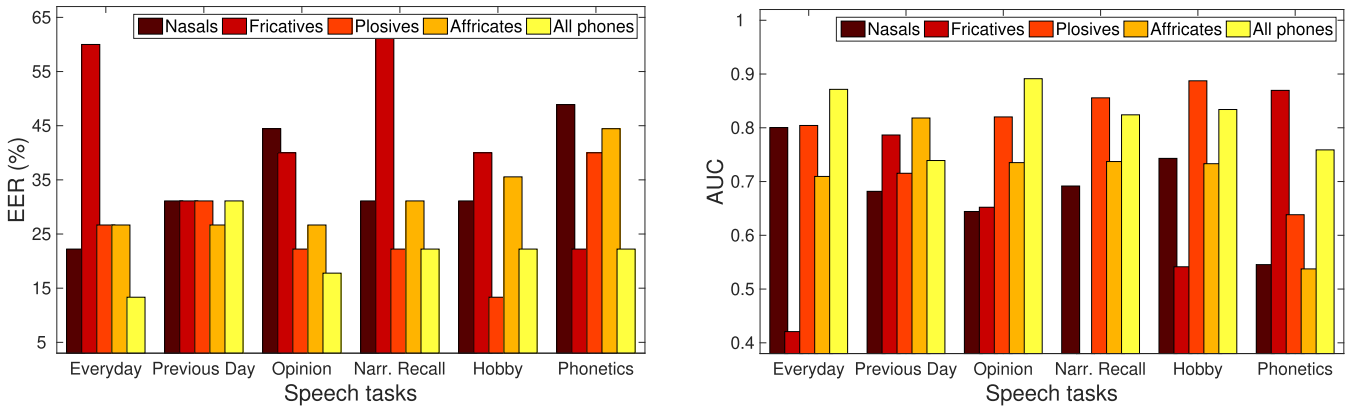

Fig. 4. The Equal Error Rate (left) and Area Under Curve (right) values for the six speech tasks with five phonetic categories (nasals, fricatives, plosives, affricates and all phones).

TABLE V

MEAN, MEDIAN AND STANDARD DEVIATION VALUES OF THE EqualERROR RATE (EER) AND AUC SCORES, AGGREGATED FOR THE SIX SPEECH TASKS(LEFT), AND FOR THE FIVE SPONTANEOUS SPEECH TASKS (RIGHT)

	All Six Speech Tasks						Spontaneous Speech Tasks					
	EER			AUC			EER			AUC		
	Mean	Median	Std.	Mean	Median	Std.	Mean	Median	Std.	Mean	Median	Std.
All phones	21.5%	22.2%	5.9%	0.820	0.829	0.060	21.3%	22.2%	6.6%	**0.832**	0.834	0.059
Vowels	53.3%	55.6%	10.1%	0.455	0.428	0.131	52.9%	55.6%	11.3%	0.463	0.437	0.145
Voiced	27.4%	26.7%	5.2%	0.777	0.806	0.058	26.7%	26.7%	5.4%	**0.791**	0.812	**0.052**
Unvoiced	25.9%	26.7%	5.2%	0.794	0.791	0.056	**24.9%**	26.7%	**5.1%**	**0.806**	**0.808**	**0.052**
Nasals	34.8%	31.1%	9.9%	0.685	0.687	0.087	**32.0%**	31.1%	**8.0%**	**0.712**	0.692	**0.061**
Fricatives	43.0%	40.0%	16.4%	0.607	0.597	0.199	47.1%	40.0%	**14.4%**	0.555	0.542	**0.170**
Plosives	25.9%	24.4%	9.1%	0.787	0.812	0.093	**23.1%**	**22.2%**	**6.6%**	**0.817**	0.820	**0.065**
Affricates	31.9%	28.9%	7.1%	0.712	0.734	0.093	**29.3%**	**26.7%**	**4.0%**	**0.747**	0.735	**0.042**

task because, unlike the other five speech tasks (which were spontaneous ones), this one consisted of reading the same fixed text.

From the other phonetic groups, nasals led to mediocre accuracy scores, as we obtained EER scores between 22.2% and 48.9%, and AUC values between 0.546 and 0.800 with the corresponding attributes. Still, this performance was much more uniform than it was in the 'Fricatives' case, especially for the five spontaneous speech tasks. This also holds for plosives and affricates: leaving the 'Phonetics' speech task aside, we obtained EER values between 13.3% and 31.1%, and between 26.7% and 35.6%, plosives and affricates, respectively. Similarly, the AUC scores fell between 0.715 and 0.887, and between 0.709 and 0.818, plosives and affricates, respectively. Among the two, plosives yielded better scores for five speech tasks out of six, making it the best phonetic group tested: this led to the best metric values as well (EER = 13.3% and AUC = 0.887 for the 'Hobby' speech task).

A. Robustness Across the Speech Tasks

The left hand side of Table V shows the mean, median and standard deviation values of the two metrics for all the phonetic groups examined. These values support our previous findings: while relying on the vowels led to an unacceptable classification performance, voiced consonants and unvoiced consonants proved to be more (and equally) useful. Although the mean and median values for these two phonetic categories both fell a bit below those obtained using all the phones (an absolute EER difference of 4.4% – 5.9% and an AUC difference of 0.023 – 0.043), these still led to a relatively good classification performance. The standard deviation of the scores (calculated over the six speech tasks) was also

remarkably similar, indicating that the proposed features are quite robust as well.

Regarding the other four phonetic categories, these values reinforce our findings that the fricatives did not allow an efficient MS discrimination performance: the mean and median EER values were quite close to 50%, while the AUC scores lay close to 0.500. This was also accompanied by quite high standard deviation values, indicating a lack of robustness. Using nasals or affricates turned out to be much better, but clearly plosives gave the best performance. However, the robustness of these attribute subgroups was notably below the 'All phones' case: while the latter led to standard deviation values of 5.9% and 0.060, EER and AUC, respectively, with the latter four phonetic groups we measured standard deviations between 7.1% and 9.9%, and around 0.09. This indicates that the utility of these phonetic groups varied more with the speech tasks than with all the phonetic posterior-derived features.

B. Robustness Across the Spontaneous Speech Tasks

When inspecting Figure 4, we found that the spontaneous speech tasks led to quite similar performance scores. Owing to this, next we calculated the mean, median and standard deviation scores for the five spontaneous speech tasks (i.e. 'Everyday', 'Previous Day', 'Opinion', 'Narrative Recall' and 'Hobby') (see the right hand side of Table V; better values are shown as **bold**). When using all the phones, the values did not differ significantly from the case of relying on all six speech tasks. Relying on the vowels or the fricatives also led to a similar performance for all six speech tasks: while having an unusable MS discrimination performance, the standard deviation values even rose a bit. For the voiced and unvoiced consonant categories, however, the mean and median metric scores improved, while standard deviation values fell for both metrics. Indeed, from Fig. 3 we can see that the 'Phonetics' speech task (the only one which was not a spontaneous one) led to the highest (or, for the voiced consonants, to the second-highest) EER and the lowest AUC scores for these two phonetic categories.

Similarly, for nasals, plosives and affricates we measured lower mean and median EER scores, higher AUC values, and significantly lower standard deviations for the spontaneous speech tasks alone. Specifically, for plosives, the (mean and median) EER values were the same or only 1.8% higher, while the AUC scores were only about 0.02 lower than in the reference case, with practically the same standard deviations. This, in our opinion, means that they give a more robust performance on the spontaneous speech tasks than when other types of speech tasks (in our case, a special reading task) are also included.

VII. Conclusion and Limitations

In this study, we investigated the automatic processing of the speech of Multiple Sclerosis subjects and healthy controls. To automatically distinguish the two speaker groups, we developed a feature extraction method based on the phonetic posterior estimates of the acoustic part of a Hidden Markov Model / Deep Neural Network (HMM/DNN) hybrid model (phonetic posteriorgrams). We performed our experiments on the recordings of 45 subjects, and our recording protocol involved six different speech tasks. Based on our experimental results, the proposed method leads to a slightly better classification performance than with x-vectors or ComParE functionals (applied as competitive baselines), and it is more robust: the classification performance turned out to be less sensitive to the actual speech task. Besides this, although the size of the feature vector extracted from each utterance depends on the phonetic set of the given language, it is still extremely compact: in our case we worked with only 82 features for each recording.

In the next part of our study we exploited the interpretable nature of the proposed, phonetic posterior-based features. Since all the attributes are directly linked to a phone of the phonetic set, we examined the utility of specific phonetic subsets by discarding the attributes corresponding to other phones. By repeating this experiment for several phonetic subsets, we found that the 'Plosives' category led to a subject classification performance close to using all phones, which increased further when we only retained the spontaneous speech tasks. This also means that, by focusing on the posterior estimates of the plosive phones, our proposed method might be applicable in the speech therapy of MS patients as well.

Regarding the limitations of our study, having around 50 subjects as samples is quite a small dataset from a machine learning point of view, but it is fairly common and accepted in pathological speech processing studies. Another concern might be the special nature of the speech tasks, but since we obtained similar classification performance scores for the five spontaneous speech tasks, the proposed method appears to be quite insensitive to the actual instructions. It is also unclear whether using more sophisticated phonetic posterior estimation methods such as LSTMs or GRUs would improve the performance of the subsequent MS identification step, or just the other way around: even simpler context-independent phonetic states could prove to be adequate. We plan to investigate this in the near future.

Another possible limitation might come from the strong connection between the proposed features and the phonetic set of the given language, in the case where we apply the proposed feature extraction approach to subjects speaking in another language such as English, Spanish or Chinese. This would probably require repeating the phonetic subset selection experiments for the phonetic groups of the target language, or using some multilingual phonetic set when calculating the phonetic posterior estimates [35], [77]. This said, in our opinion, the competitive MS discrimination performance, the robustness to speech tasks, the compact size of the feature set, and the interpretable nature of the attributes (allowing a potential application in speech therapy) all demonstrate the utility of the proposed feature extraction method.

References

[1] I. Szirmai, *Neurologia*. Budapest, Hungary: Medicina, 2006.
[2] F. D. Lublin et al., "Defining the clinical course of multiple sclerosis: The 2013 revisions," *Neurology*, vol. 14, no. 3, pp. 278–286, 2014.

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

[4] S. Renauld, L. Mohamed-Saïd, and J. Macoir, "Language disorders in multiple sclerosis: A systematic review," *Multiple Sclerosis Rel. Disorders*, vol. 10, pp. 103–111, Nov. 2016.

[5] A. Delgado-Álvarez et al., "Cognitive processes underlying verbal fluency in multiple sclerosis," *Frontiers Neurol.*, vol. 11, Jan. 2021, Art. no. 629183.

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

[7] B. Yamout et al., "Vocal symptoms and acoustic changes in relation to the expanded disability status scale, duration and stage of disease in patients with multiple sclerosis," *Eur. Arch. Oto-Rhino-Laryngol.*, vol. 266, no. 11, pp. 1759–1765, Nov. 2009.

[8] G. Noffs et al., "What speech can tell us: A systematic review of dysarthria characteristics in multiple sclerosis," *Autoimmunity Rev.*, vol. 17, no. 12, pp. 1202–1209, Dec. 2018.

[9] F. J. F. Gerald, B. E. Murdoch, and H. J. Chenery, "Multiple sclerosis: Associated speech and language disorders," *Austral. J. Human Commun. Disorders*, vol. 15, no. 2, pp. 15–35, Dec. 1987.

[10] D. Mulfari, G. Meoni, M. Marini, and L. Fanucci, "Machine learning assistive application for users with speech disorders," *Appl. Soft Comput.*, vol. 103, May 2021, Art. no. 107147.

[11] H. Tolba and A. S. El-Torgoman, "Towards the improvement of automatic recognition of dysarthric speech," in *Proc. ICCSIT*, Beijing, China, Aug. 2009, pp. 277–281.

[12] M. Kim, Y. Kim, J. Yoo, J. Wang, and H. Kim, "Regularized speaker adaptation of KL-HMM for dysarthric speech recognition," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 9, pp. 1581–1591, Sep. 2017.

[13] J. Yu et al., "Development of the CUHK dysarthric speech recognition system for the UA speech corpus," in *Proc. Interspeech*, Sep. 2018, pp. 2938–2942.

[14] S. Chandrakala and N. Rajeswari, "Representation learning based speech assistive system for persons with dysarthria," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 9, pp. 1510–1517, Sep. 2017.

[15] C.-Y. Chen, W.-Z. Zheng, S.-S. Wang, Y. Tsao, P.-C. Li, and Y.-H. Lai, "Enhancing intelligibility of dysarthric speech using gated convolutional-based voice conversion system," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 4686–4690.

[16] G. Van Nuffelen, C. Middag, M. De Bodt, and J. Martens, "Speech technology-based assessment of phoneme intelligibility in dysarthria," *Int. J. Lang. Commun. Disorders*, vol. 44, no. 5, pp. 716–730, Jan. 2009.

[17] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, "Automatic intelligibility classification of sentence-level pathological speech," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 132–144, Jan. 2015.

[18] J. Fritsch and M. Magimai-Doss, "Utterance verification-based dysarthric speech intelligibility assessment using phonetic posterior features," *IEEE Signal Process. Lett.*, vol. 28, pp. 224–228, 2021.

[19] K. Lopez-De-Ipiña et al., "On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature," *Cognit. Comput.*, vol. 7, no. 1, pp. 44–55, 2015.

[20] A. Ablimit, K. Scholz, and T. Schultz, "Deep learning approaches for detecting Alzheimer's dementia from conversational speech of ILSE study," in *Proc. Interspeech*, Incheon, (South) Korea, Sep. 2022, pp. 3348–3352.

[21] A. Pompili et al., "Assessment of Parkinson's disease medication state through automatic speech analysis," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 4591–4595.

[22] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect Parkinson's disease from speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1155–1159.

[23] G. Chatzoudis, M. Plitsis, S. Stamouli, A. Dimou, N. Katsamanis, and V. Katsouros, "Zero-shot cross-lingual aphasia detection using automatic speech recognition," in *Proc. Interspeech*, Sep. 2022, pp. 2178–2182.

[24] Z. Zhao et al., "Hybrid network feature extraction for depression assessment from speech," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 4956–4960.

[25] A. Z. Jenei, G. Kiss, and D. Sztahó, "Detection of speech related disorders by pre-trained embedding models extracted biomarkers," in *Proc. SPECOM*, Gurugram, India, 2022, pp. 279–289.

[26] V. Svindt, J. Bóna, and I. Hoffmann, "Changes in temporal features of speech in secondary progressive multiple sclerosis (SPMS)—Case studies," *Clin. Linguistics Phonetics*, vol. 34, no. 4, pp. 339–356, Apr. 2020.

[27] E. Barois, Y. Sagawa, S. Yilmaz, E. Magnin, and P. Decavel, "What (more) can verbal fluency tell us about multiple sclerosis?" *Ann. Phys. Rehabil. Med.*, vol. 64, no. 2, Mar. 2021, Art. no. 101394.

[28] J. Rusz et al., "Characteristics of motor speech phenotypes in multiple sclerosis," *Multiple Sclerosis Rel. Disorders*, vol. 19, pp. 62–69, Jan. 2018.

[29] P. Vizza et al., "Methodologies of speech analysis for neurodegenerative diseases evaluation," *Int. J. Med. Informat.*, vol. 122, pp. 45–54, Feb. 2019.

[30] G. Noffs et al., "Acoustic speech analytics are predictive of cerebellar dysfunction in multiple sclerosis," *Cerebellum*, vol. 19, no. 5, pp. 691–700, Oct. 2020.

[31] C. De Looze et al., "Effects of cognitive impairment on prosodic parameters of speech production planning in multiple sclerosis," *J. Neuropsychol.*, vol. 13, no. 1, pp. 22–45, Mar. 2019.

[32] P. Klumpp et al., "The phonetic footprint of Parkinson's disease," *Comput. Speech Lang.*, vol. 72, Mar. 2022, Art. no. 101321.

[33] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. ASRU*, 2009, pp. 421–426.

[34] K. Kintzley, A. Jansen, and H. Hermansky, "Event selection from phone posteriorgrams using matched filters," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 1905–1908.

[35] J. C. Vásquez-Correa, P. Klumpp, J. R. Orozco-Arroyave, and E. Nöth, "PhoNet: A tool based on gated recurrent neural networks to extract phonological posteriors from speech," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 549–553.

[36] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[37] W.-Z. Zheng, J.-Y. Han, C.-K. Lee, Y.-Y. Lin, S.-H. Chang, and Y.-H. Lai, "Phonetic posteriorgram-based voice conversion system to improve speech intelligibility of dysarthric patients," *Comput. Methods Programs Biomed.*, vol. 215, Mar. 2022, Art. no. 106602.

[38] Y. Jiao, V. Berisha, and J. Liss, "Interpretable phonological features for clinical applications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5045–5049.

[39] B. M. Halpern, J. Fritsch, E. Hermann, R. van Son, O. Scharenborg, and M. Magimai-Doss, "An objective evaluation framework for pathological speech synthesis," in *Proc. 14th ITG Conf.*, Sep. 2021, pp. 1–5.

[40] M. Wöllmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6822–6826.

[41] Y. Liu, T. Lee, T. Law, and K. Y. Lee, "Acoustical assessment of voice disorder with continuous speech using ASR posterior features," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 6, pp. 1047–1059, Jun. 2019.

[42] T. Arias-Vergara, J. R. Orozco-Arroyave, M. Cernak, S. Gollwitzer, M. Schuster, and E. Nöth, "Phone-attribute posteriors to evaluate the speech of cochlear implant users," in *Proc. Interspeech*, Sep. 2019, pp. 3108–3112.

[43] M. Cernak, J. R. Orozco-Arroyave, F. Rudzicz, H. Christensen, J. C. Vásquez-Correa, and E. Nöth, "Characterisation of voice quality of Parkinson's disease using differential phonological posterior features," *Comput. Speech Lang.*, vol. 46, pp. 196–208, Nov. 2017.

[44] A. Abraham, R. Schubotz, and D. Y. von Cramon, "Thinking about the future versus the past in personal and non-personal contexts," *Brain Res.*, vol. 1233, pp. 106–119, Oct. 2008.

[45] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features," *Comput. Speech Lang.*, vol. 53, pp. 181–197, Jan. 2019.

[46] G. Gosztolya et al., "Cross-lingual detection of mild cognitive impairment based on temporal parameters of spontaneous speech," *Comput. Speech Lang.*, vol. 69, Sep. 2021, Art. no. 101215.

[47] R. A. Mar, "The neuropsychology of narrative: Story comprehension, story production and their interrelation," *Neuropsychologia*, vol. 42, no. 10, pp. 1414–1434, Jan. 2004.

[48] J. Bóna, V. Svindt, and I. Hoffmann, "Voice onset time of Hungarian voiceless plosives in multiple sclerosis," in *Proc. ISSP*, Dec. 2020, pp. 202–205.

[49] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[50] G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[51] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[52] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.

[53] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12449–12460.

[54] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," in *Proc. Interspeech*, Aug. 2021, pp. 1509–1513.

[55] M. Panzner and P. Cimiano, "Comparing hidden Markov models and long short term memory neural networks for learning action representations," in *Proc. MOD*, Volterra, Italy, 2016, pp. 94–105.

[56] M. Schmitt, N. Cummins, and B. W. Schuller, "Continuous emotion recognition in speech—Do we need recurrence?" in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 2808–2812.

[57] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, 2006, pp. 369–376.

[58] H. Sak et al., "Learning acoustic frame labeling for speech recognition with recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4280–4284.

[59] G. Kurata and K. Audhkhasi, "Guiding CTC posterior spike timings for improved posterior fusion and knowledge distillation," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 1616–1620.

[60] 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 *Proc. TSD*, Brno, Czech Republic, Sep. 2014, pp. 424–431.

[61] J. Odell, "The use of context in large vocabulary speech recognition," Ph.D. dissertation, Oueen's College, Univ. Cambridge, 1995. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=53b50cb6dea86c52729cab4068b7ceeabf2e888d

[62] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.

[63] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, Jul. 2010.

[64] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, J. Villalba, J. R. Orozco-Arroyave, and N. Dehak, "Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's disease," *Appl. Soft Comput.*, vol. 62, pp. 649–666, Jan. 2018.

[65] J. Fritsch, S. Wankerl, and E. Nöth, "Automatic diagnosis of Alzheimer's disease using neural network language models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5841–5845.

[66] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5329–5333.

[67] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 7169–7173.

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

[69] J. V. Egas-López, G. Kiss, D. Sztahó, and G. Gosztolya, "Automatic assessment of the degree of clinical depression from speech using x-vectors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 8502–8506.

[70] D. Povey et al., "The Kaldi speech recognition toolkit," in *Proc. ASRU*, Big Island, HI, USA, Dec. 2011, pp. 1–4.

[71] B. Schuller et al., "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 2001–2005.

[72] F. Eyben, M. Wöllmer, and B. Schuller, "OpensMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, Florence, Italy, Oct. 2010, pp. 1459–1462.

[73] B. Schuller et al., "The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition," in *Proc. Interspeech*, Sep. 2015, pp. 478–482.

[74] B. W. Schuller et al., "The INTERSPEECH 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 2378–2382.

[75] B. Schuller et al., "The ACM multimedia 2022 computational paralinguistics challenge: Vocalisations, stuttering, activity, & mosquitoes," in *Proc. 30th ACM Int. Conf. Multimedia*, Lisbon, Portugal, Oct. 2022, pp. 7120–7124.

[76] B. Megyesi. (2003). *The Hungarian Language: A Short Descriptive Grammar*. [Online]. Available: http://stp.ling.uu.se/~bea/publ/megyesi-hungarian.pdf

[77] X. Li et al., "Universal phone recognition with a multilingual allophone system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 8249–8253.