Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/bbe

Original Research Article

Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge

Kamil Lahcene Kadi [a,*], *Sid Ahmed Selouani* [b], *Bachir Boudraa* [a], *Malika Boudraa* [a]

[a] *Faculty of Electronics and Computer Science, University of Sciences and Technology Houari Boumediene, Algiers, Algeria*
[b] *Department of Information Management, University of Moncton, Campus of Shippagan, Shippagan, NB E8S 1P6, Canada*

ARTICLE INFO

Article history:
Received 10 January 2015
Received in revised form
13 September 2015
Accepted 9 November 2015
Available online 27 November 2015

Keywords:
Dysarthria
Speech processing
Auditory cues
GMM
SVM
Hybrid GMM/SVM

ABSTRACT

Millions of children and adults suffer from acquired or congenital neuro-motor communication disorders that can affect their speech intelligibility. The automatically characterization of speech impairment can contribute to improve the patient's life quality, and assist experts in assessment and treatment design. In this paper, we present new approaches to improve the analysis and classification of disordered speech. First, we propose an automatic speaker recognition approach especially adapted to identify dysarthric speakers. Secondly, we suggest a method for the automatic assessment of the dysarthria severity level. For this purpose, a model simulating the external, middle and inner parts of the ear is presented. This ear model provides relevant auditory-based cues that are combined with the usual Mel-Frequency Cepstral Coefficients (MFCC) to represent atypical speech utterances. The experiments are carried out by using data of both Nemours and Torgo databases of dysarthric speech. Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs) and hybrid GMM/SVM systems are tested and compared in the context of dysarthric speaker identification and assessment. The experimental results achieve a correct speaker identification rate of 97.2% which can be considered promising for this novel approach; also the existing assessment systems are outperformed with a 93.2% correct classification rate of dysarthria severity levels.

[*] *Corresponding author at*: Faculty of Electronics and Computer Science, University of Sciences and Technology Houari Boumediene, 32 El Alia, 16111 Bab Ezzouar, Algiers, Algeria.

E-mail addresses: kkadi@usthb.dz (K.L. Kadi), selouani@umcs.ca (S.A. Selouani), bboudraa@usthb.dz (B. Boudraa), mboudraa@usthb.dz (M. Boudraa).

http://dx.doi.org/10.1016/j.bbe.2015.11.004

1. Introduction

Communication is a multidimensional dynamic process that is necessary to express thoughts, emotions and needs, allowing interaction between people and their environment. Cognition, hearing, speech production and motor coordination are involved in the communication process. If one or more of these aspects is impaired, the communication is disordered [1]. A communication disorder has a large impact on the life quality; it prevents individuals from expressing needs, wants and opinions, it also reduces the capacity to express personality, exercise autonomy and often has an impact on relationships and self-esteem [2]. Therefore, it is necessary to enhance the communication quality of individuals suffering from a verbal communication disability by offering them more possibilities to interact with their environment.

Our research focuses on one of the most common speech communication disorders associated with a neurological impairment called dysarthria. In fact, millions of adults and children through the world are affected by dysarthria which induces a reduction of their speech intelligibility [2,3]. Dysarthria is a motor speech disorder resulting from disturbed muscular control of the speech mechanism, caused by damage to the central or peripheral nervous system [4]. A few causes of dysarthria include Parkinson's disease, head injury, stroke, tumors, muscular dystrophy and cerebral palsy.

Numerous tools and methods have been developed to help dysarthric speakers. Indeed, prominent works were achieved in the field of speech recognition, speech intelligibility enhancement and automatic evaluation. In the last few years, based on the significant and newer Torgo database, Rudzicz developed a dysarthric speech recognition system and improved the intelligibility of dysarthric speech [5,6]. Although, it is worthy to note that the research effort did not focus on further functionalities such as speaker recognition which is increasingly used in various security or identity management systems. It is also important to note that the automated diagnosis and assessment that can assist clinicians to support patients suffering from speech impairments has not received enough attention and research efforts in this field are very sporadic.

This paper presents two original techniques of dysarthric speech processing, the first one is based on a biometric approach for automatic recognition of dysarthric speakers, and the second technique is designed to automatically assess the dysarthric speech with respect to the severity level. Both techniques use relevant features based on Distinctive Auditory-based Cues and Mel-Frequency Cepstral Coefficients (MFCC). The classification performance of the Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs) are compared in various experimental conditions. The experiments have been carried out by using two important world-class databases of dysarthric speech, namely, the Nemours database of dysarthric speech [7] and the Torgo database of acoustic and articulatory speech [8]. Both databases were used to build and evaluate the developed tools, which will ensure the availability of a large amount of data and rational diversity in recordings.

The original contributions reported in this paper can be summarized by the following three aspects:

(i) An original dysarthric speaker recognition system is proposed in a context where the individuals affected by verbal communication disorders are excluded from speech-enabled biometric solutions.
(ii) Computational models of auditory perceptual knowledge are proposed to improve the effectiveness of dysarthria assessment.
(iii) A new global assessment score is proposed for the Torgo database based on the second edition of Frenchay Dysarthria Assessment (FDA-2).

The remaining of the paper is structured as follows. Section 2 describes the acoustic analysis. Sections 3 and 4 present the automatic dysarthric speaker recognition system, and the technique of dysarthric speech assessment, respectively. In Section 5, the experiments and their outcomes are presented and discussed. Section 6 contains our concluding comments.

2. Acoustic analysis

The extraction of reliable parameters to represent the speech utterance waveform is an important issue in pattern recognition. This extraction process aims at extracting the relevant information contained in the speech signal while excluding the irrelevant part. Several features can be used as input parameters in speaker recognition and disorder characterization systems. Among these features, we can cite Linear Predictive Coding (LPC) coefficients, MFCC, short-time spectral envelope, short-time energy, zero crossing rate. It is important to mention that numerous studies showed that the use of human hearing properties can provide a potentially useful front-end speech representation [9].

Our acoustical analysis method consists of consolidating different sources of information about the speech signal that could be missing if just one type of features (e.g. MFCC) is used to represent the speech utterance waveform. The first information source consists of the conventional MFCCs, while the second source consists of acoustic cues derived from hearing phenomena studies.

2.1. Cepstral acoustic features

Short-term MFCC is an appropriate parameterization approach for dysarthric speech. The MFCCs have been often used in the signal processing field applied for speech disorder, like the recognition of dysarthric speech or speech disorders classification [10]. The Mel-scale introduced by Davis and Mermelstein is a mapping from a linear to a nonlinear frequency scale based on human auditory perception [11]. The Mel-scale approximation is:

$$Mel(f) = 2595\log_{10}\left(1+\frac{f}{700}\right) \quad (1)$$

where f represents the linear frequency scale.

To compute the MFCCs, a discrete cosine transform is applied to the outputs of M critical band-pass filters. These filters are triangular and ranged on the Mel-frequency scale that is linear below 1 kHz and logarithmic above. 15–24 filters can be used; in this work, approximately 2 filters per octave were used. The MFCCs are defined as:

$$MFCC_n = \sum_{m=1}^{M} X_m \cos\left(\frac{\pi n}{M}(m-0.5)\right), \quad n = 1, 2, \ldots, N \tag{2}$$

where M is the analysis order, N represents the number of Cepstral coefficients and X_m ($m = 1, 2, \ldots, M$) is the output of the mth filter applied to the log magnitude spectrum of the signal.

Several cross-validation experiments have been carried to determine the optimal frame duration. In the proposed framework, the MFCCs were calculated using a 16 ms Hamming window that contains 256 samples with an overlapping of 50%. This frame size value was found effective after carrying out experiments using frames containing 128 and 512 samples. It is important to note that this framing configuration satisfies the stationary criterion, required when processing random signals, while realizing a good trade-off between complexity and quality. The same frame length has been used to calculate the auditory-based cues presented below, which simplifies the combination of the two feature sets obtained independently.

2.2. *Auditory modeling*

All over the history of speech processing, many attempts have been performed to extract the most relevant information from speech phenomena that may be considered as inaccurate, unpredictable and versatile. Usual methods of acoustic feature extraction remain inefficient to ensure a good level of robustness, whereas humans can easily deal with uncertainty of acoustical environments and adverse conditions to understand speech. This leads many researchers to investigate the human auditory system in order to better understand this extraordinary perception capability. The ability of the auditory system to efficiently treat and interpret speech even in bad conditions, like the unintelligibility of dysarthric speech makes auditory modeling an interesting and promising approach.

Usually, the aim of parameterization approach based on auditory modeling is to examine the response of the basilar membrane and the auditory nerve to various sounds. An advanced processing that simulates the auditory cortex can also be performed. A computational model was proposed by Flanagan to estimate the basilar membrane movement. This model has been found useful for reporting the subjective auditory behavior and the acousto-mechanical process of the ear [12]. Other well-known models, such as the Cochlea Model [13], the representation of Mean Synchrony Auditory [14], and the Ensemble Interval Histogram processing [15], have been the basis of many contemporary approaches. In all these models, a band-pass filter bank is used to simulate cochlear filtering.

In recent years, there have been new interests in the improvement of front-end processing to compute robust features inspired by auditory modeling [16]. In the area of speech recognition, the psychoacoustics and auditory physiology based processing becomes the main component of robust feature extraction methods such as Gammatone Features (GFCC by Schluter et al., 2007) [17] and Power-Normalized Cepstral Coefficients (PNCC by Kim and Stern, 2012) [18].

The auditory model used throughout our experiments simulated the external, middle and inner parts of the ear. The next subsections will present all parts of the auditory model which was first proposed by Caelen [19] and adapted to be used as a front-end module in speech recognition systems by Selouani [20]. To model the various adaptive ossicle motions at the external and middle ear, a band-pass filter was used. A non-linear filter bank stimulates the basilar membrane (BM) that acts out the inner ear. The electro-mechanical transduction of hair cells and afferent fibers from which the encoding signal is generated at the synaptic endings are also considered by the BM model. Along the several organs implicated in perception and hearing, different regions are responsive to sounds with different spectral properties, due to the dissimilarity of the anatomy and the physiology. Thus, every part along the BM has a given resonance frequency for a certain input sound [19].

2.2.1. *Mid-external ear model*

A band-pass filter simulates the external and middle parts of the ear, which can be defined as:

$$s'(k) = s(k) - s(k-1) + \propto_1 s'(k-1) - \propto_2 s'(k-2), \quad 4k = 1, \ldots, K \tag{3}$$

where $s(k)$ is the speech wave, $s'(k)$ is the filtered output signal, K is the number of frame samples, $\propto_1$ and $\propto_2$ are coefficients that depend on the central frequency of the filter, its Q-factor and sampling frequency Fs. Values of 1500 Hz and 1.5 are used as central frequency and Q-factor respectively [21].

2.2.2. *Basilar membrane model*

The BM is modeled by 24 overlapping filters which represent the cochlear filter bank. The vibration of a specific part in the BM is simulated by the frequency response of a certain filter for an auditory stimulus at the outer ear [19]. For each filter of the bank, the output is defined as:

$$y_i(k) = \beta_{1,i} y_i(k-1) - \beta_{2,i} y_i(k-2) + G_i[s'(k) - s'(k-2)] \tag{4}$$

The transfer function is given by:

$$H_i(z) = \frac{G_i[1 - z^{-2}]}{1 - \beta_{1,i} z^{-1} + \beta_{2,i} z^{-2}} \tag{5}$$

where $y_i(k)$ is the BM response to a mid-external signal $s'(k)$. It constitutes the vibration amplitude in the position x_i of the BM. The parameters of the filter i are the coefficients $\beta_{1,i}$, $\beta_{2,i}$ and the gain G_i. The number of overlapping cochlear filters or channels is fixed to 24 and represented by N_c. Every filter covers about $\Delta x = 1.46$ mm of the BM. To simplify, just the coupling effects of the electro-mechanical transduction in hair cells and fibers was considered. The coupling parameters C_i, E_i and A_i represent the behavior of fibers and hair cells. Algorithm 1 represents a detailed algorithm that provides $y_i'(k)$, the outcome stimulus after the passage across entire constituents of the ear model. Using $y_i'(k)$, the energy of each channel is then computed.

Definitions

F_s	sampling frequency (16,000 Hz)
Δ_x	basic length unit of the basilar membrane (1.46 mm)
x	total length of the basilar membrane (35 mm)

N_c	total number of channels (24)
i	channel index
k	time index
K	number of total samples by frame
F_i	central frequency of channel i
$H_i, r_{i,j}, v$ and u	temporary calculation functions
E_i	direct coupling function
A_i	inverse coupling function
C_i	coupling parameter
Q_i	Q-factor
G_i	gain of the filter
$\beta_{1,i}$ and $\beta_{2,i}$	filter coefficients
$s'(k)$	filtered speech signal that passed through the mid-external ear
$y'_i(k)$	resulting stimulus

Algorithm 1. Sample by sample algorithm simulating inner ear. The aim is to compute the stimulus $y'_i(k)$.

```
Initialize $f_x = (F_s \Delta_x)^2$, $H_0 = 0$, $r_{i,j} = 0$, $E_0 = 0$.
For $i = 1$ to $N_c$ Do
  $x_i = i\Delta x;\ v = e^{(-106.5x_i)};\ F_i = 7100v - 100;\ C_i = \frac{(27v)^2}{f_x};$
  $Q_i = (-8300x + 176.3)x_i + 4;\ G_i = e^{(-80x_i)};\ u = e^{-\frac{\pi F_i}{F_s Q_i}};$
  $\beta_{1,i} = 2u\ \cos\left(\frac{2\pi F_i}{F_s}\right);\ \beta_{2,i} = u^2;$
  $E_i = \frac{1}{1+(2-E_{i-1})C_i};\ A_i = E_i C_i;$
EndDo
For $k = 1$ to $K$ Do
  For $i = 1$ to $N_c$ Do
  $H_i = (G_i(s'(k) - s'(k-2)) + \beta_{1,i} r_{i,2} - \beta_{1,i} r_{i,1})E_i + H_{i-1}A_i$
  EndDo
  For $i = N_c$ to 1 Do
  $r_{i,2} = A_i r_{i+1,3} + H_i,$ and $y'_i(k) = r_{i,3}$
  EndDo
  For $i = 1$ to $N_c$ Do
    For $j = 1$ to 2 Do
    $r_{i,j} = r_{i,j+1}$
    EndDo
  EndDo
EndDo
```

2.2.3. *Auditory distinctive features*

For the output of each channel, the absolute energy is defined as:

$$W'_i(T) = 20\log\sum_{k=1}^{K}\left|y'_i(k)\right|, \quad i = 1, 2, \ldots, N_c \tag{6}$$

T is the frame index, i indicates the channels and N_c equals 24, the total channels number. K is the frame length and thus k refers samples. To reduce the fluctuation of energy, a smoothing function is applied:

$$W_I(T) = c_0 W_i(T-1) + c_1 W'_i(T) \tag{7}$$

where $W_I(T)$ represents the smoothed energy; the coefficients c_0 and c_1 average $W_i(T-1)$ and $W'_i(T)$, such as the sum of both equals 1.

Auditory cues are computed by linear combinations of the channel outputs energies. Based on our previous studies [22], we use seven cues derived from the ear model as distinctive features: Grave/Acute (G/A), Open/Closed (O/C), Diffuse/Compact (D/C), Flat/Sharp (F/S), Mellow/Strident (M/S), Continuant/Discontinuant (C/D), and Tense/Lax (T/L). Table 1 shows the computation of the seven cues, over each frame. Based on acoustic-phonetic knowledge, we can explain, for example, the calculation of (M/S) cue. Indeed, strident phonemes are characterized by the presence of noise due to turbulence at their articulation point. Therefore, a phoneme is considered as strident if the frequency band ranging from 3800 Hz to 5300 Hz contains more energy than the band ranging from 1900 Hz to 2900 Hz. This approach has been successfully tested in the context of speech recognition in adverse conditions [23].

Table 1 – Description of the auditory-based cues.

Cue	Description
(G/A)	Measures the difference of energy between low frequencies (50–400 Hz) and high frequencies (3800–6000 Hz): $(W_1 + \cdots + W_5) - (W_{20} + \cdots + W_{24})$
(O/C)	A phoneme is considered closed if the energy of low frequencies (230–350 Hz) is greater than that of the middle frequencies (600–800 Hz). Hence, the O/C cue is calculated by: $W_8 + W_9 - W_3 - W_4$
(D/C)	Compactness reflects the prominence of the central formant region (800–1050 Hz) compared with the surrounding regions (300–700 Hz) and (1450–2550 Hz): $W_{10} + W_{11} - (W_4 + \cdots + W_8 + W_{13} + \cdots + W_{17})/5$
(F/S)	A phoneme is considered sharp if the energy in (2200–3300 Hz) is more important than the energy in (1900–2900 Hz): $W_{17} + W_{18} + W_{19} - W_{11} - W_{12} - W_{13}$
(M/S)	Strident phonemes are characterized by a presence of noise because of a turbulence at their articulation point which leads to more energy in (3800–5300 Hz) than in (1900–2900 Hz): $W_{21} + W_{22} + W_{23} - W_{16} - W_{17} - W_{18}$
(C/D)	Quantifies the variation of the spectrum magnitude by comparing the energy of current and preceding frames. $\sum_{i=1}^{N_c} W_i(T) - W_a(T) - W_i(T-1) + W_a(T-1)$ $W_i(T)$ is the energy of channel i $W_a T)$ is the energy average over all channels of current frame T.
(T/L)	Measures the difference of energy between middle frequencies (900–2000 Hz) and relative high frequencies (2650–5000 Hz): $(W_{11} + \cdots + W_{16}) - (W_{18} + \cdots + W_{23})$

Figs. 1 and 2 give examples of the auditory-based feature cues computed from dysarthric and non-dysarthric speech for the pronounced phrase: "The sin is sitting the who", taken from the Nemours corpus.

Visible differences can be noticed in the figures representing auditory-based features of a dysarthric and non-dysarthric speaker. These differences are statistically analyzed using a boxplot diagram, which displays the distribution of the auditory features based on their minimum, first quartile, median, third quartile and maximum. Fig. 3 represents the distribution of the auditory-based features that were calculated from the sentences cited above.

The median is represented by the central mark on each box, the edges are the 25th and 75th percentiles, and the whiskers stretch out to the extreme auditory feature point. As shown in Fig. 3, the discrimination can easily be done between the sentences pronounced by the dysarthric and the

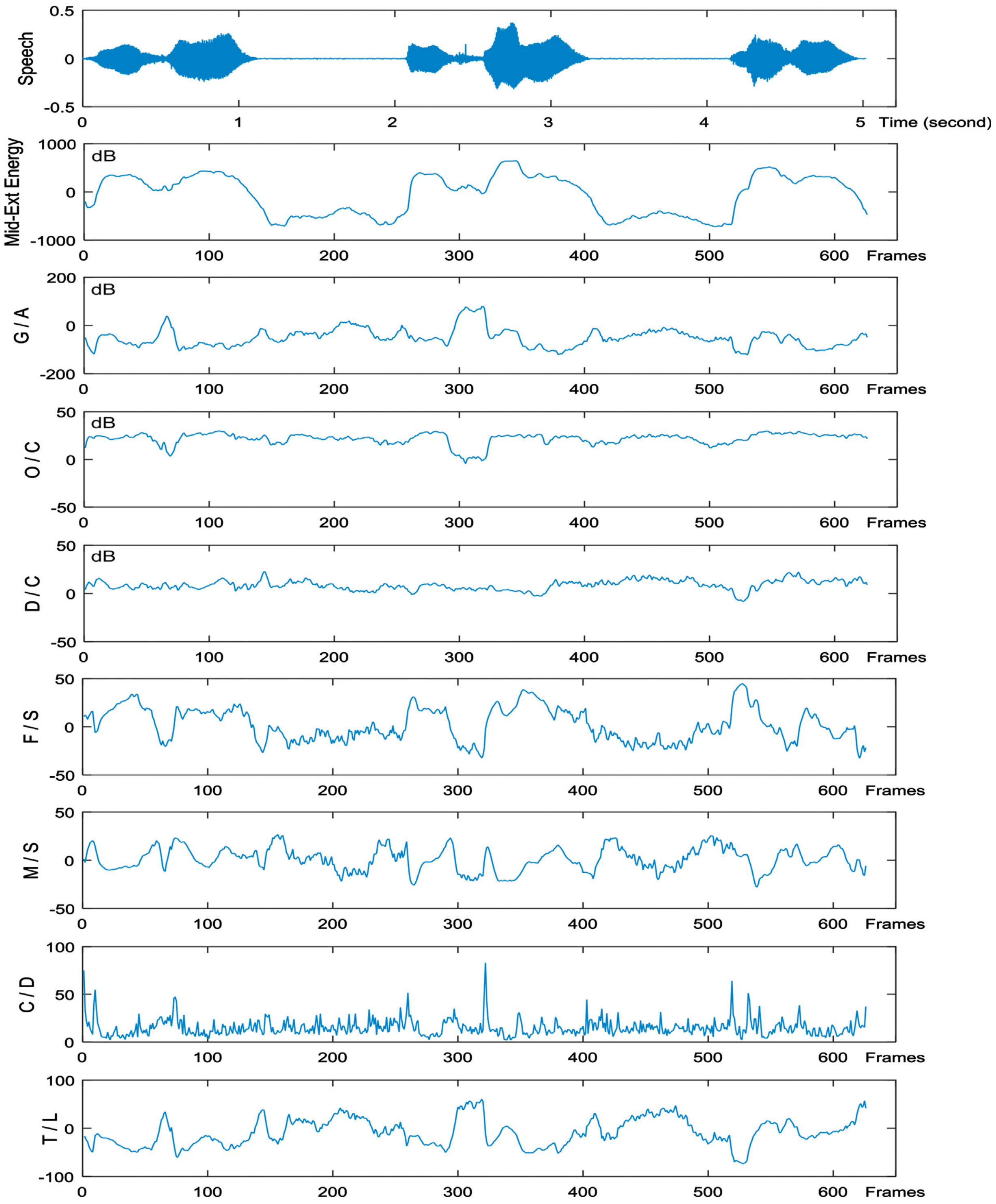

Fig. 1 – Auditory-based features cues of the following phrase pronounced by a dysarthric speaker: "The sin is sitting the who". G/A: Grave/Acute; O/C: Open/Closed; D/C: Diffuse/Compact; F/S: Flat/Sharp; M/S: Mellow/Strident; C/D: Continuant/Discontinuant; T/L: Tense/Lax.

non-dysarthric speaker for Grave/Acute, Open/Closed, Flat/Sharp, Mellow/Strident and Tense/Lax. However, the difference is not obvious for Continuant/Discontinuant and Diffuse/Compact features.

In addition to the boxplot representation, a one-way analysis of variance (ANOVA) was carried out with each of the eight features as the dependent variable. The results in Table 2 show that the dysarthric/non-dysarthric factor had an important number of significant effects on auditory-based features. These include seven features (Mid-Ext Energy, G/A, O/C, F/S, M/S, C/D, T/L). This corresponds to the observations made in Figs. 1 and 2, and on the boxplot diagrams.

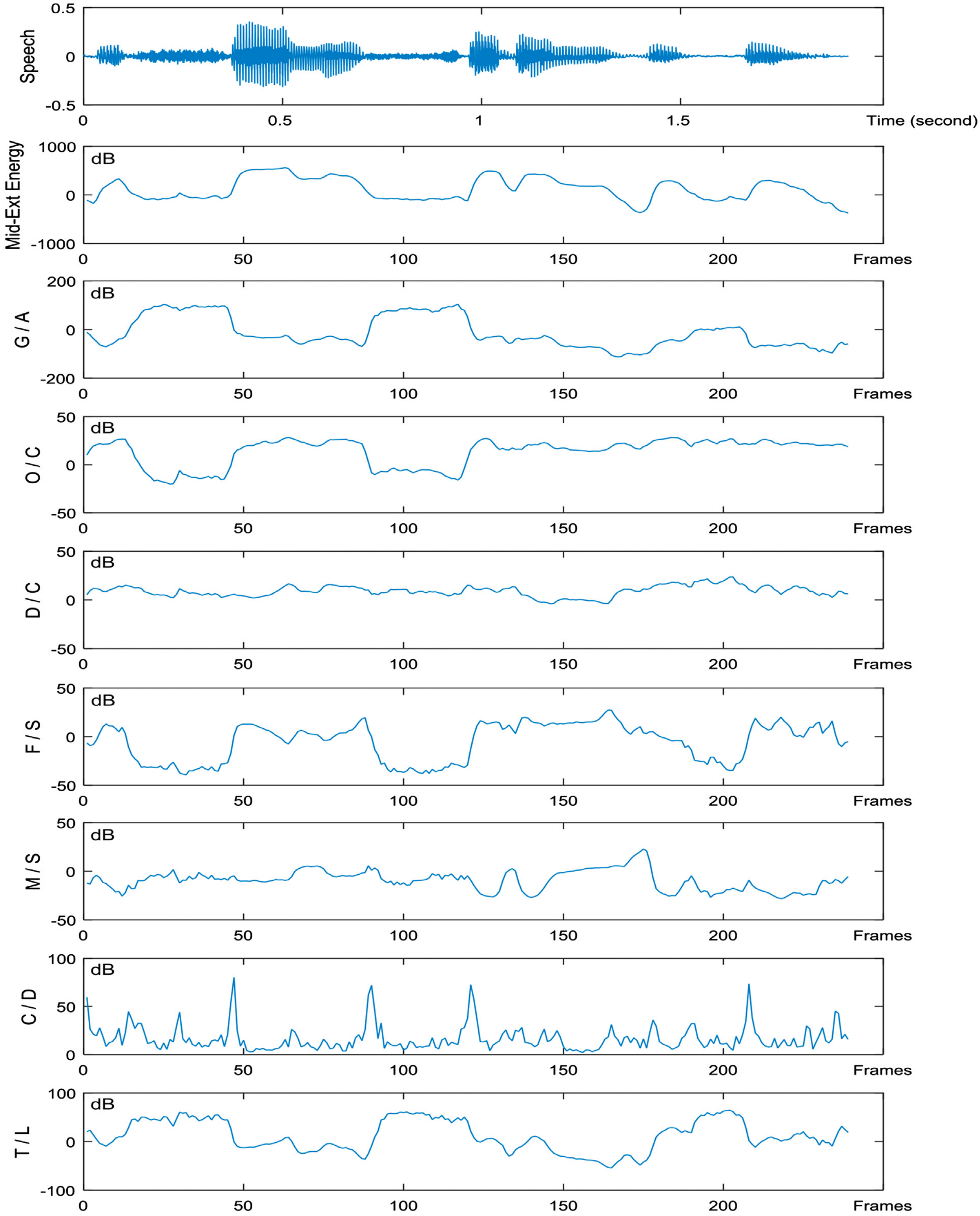

Fig. 2 – Auditory-based features cues of the following phrase pronounced by a non-dysarthric speaker: "The sin is sitting the who". G/A: Grave/Acute; O/C: Open/Closed; D/C: Diffuse/Compact; F/S: Flat/Sharp; M/S: Mellow/Strident; C/D: Continuant/ Discontinuant; T/L: Tense/Lax.

3. Automatic dysarthric speaker recognition

Automatic speaker recognition systems can recognize a person's identity that is embedded in her or his voice [24]. In this section we describe a new speaker recognition system designed to include individuals who suffer from speech disorders into voice-based biometric systems.

This approach can be used in many applications such as remote access, monitoring, voice matching for forensics, or for

Table 2 – Statistical significance (*p* values) is based on one-way ANOVAs, with dysarthric versus non-dysarthric as the independent variable. Significance (in boldface) was reached at values of $p < 0.05$.

Features	Mid-ext energy	G/A	O/C	D/C	F/S	M/S	C/D	T/L
p-value	**$2.02e^{-9}$**	**$6.08e^{-25}$**	**$5.32e^{-31}$**	0.410	**$9.81e^{-8}$**	**$1.23e^{-30}$**	**0.047**	**$3.21e^{-18}$**

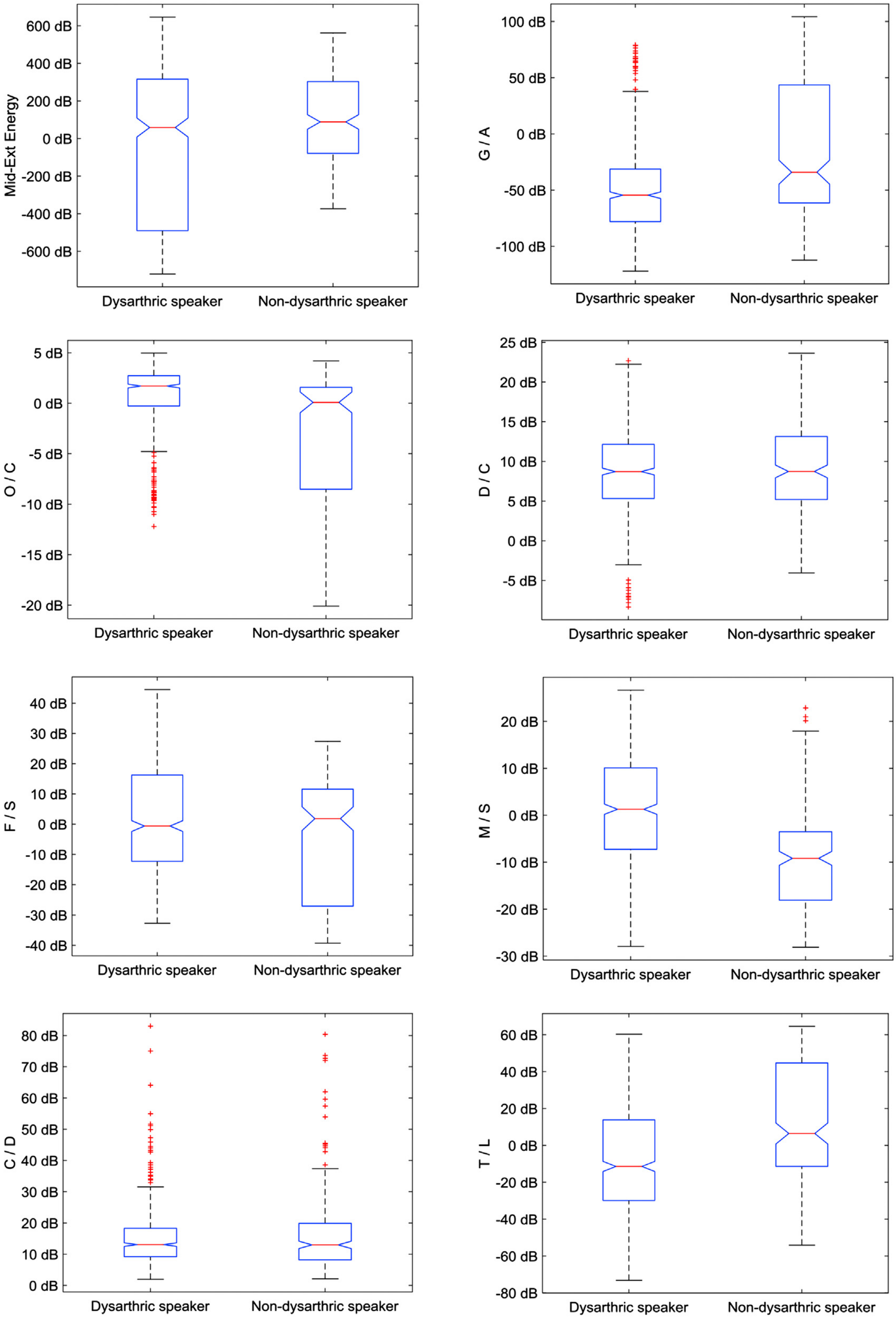

Fig. 3 – Boxplot of each auditory-based features (dysarthric/non-dysarthric speaker).

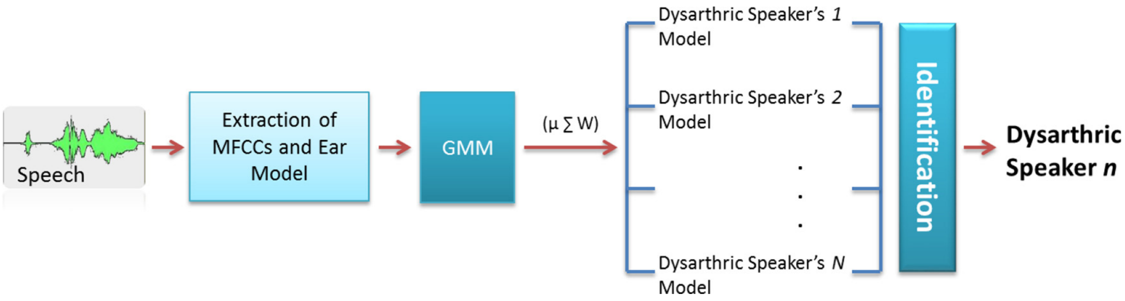
Speech
Extraction of MFCCs and Ear Model
GMM
(μ ∑ W)
Dysarthric Speaker's 1 Model
Dysarthric Speaker's 2 Model
Dysarthric Speaker's N Model
Identification
Dysarthric Speaker n

Fig. 4 – Simplified diagram of the dysarthric speaker identification process using GMM.

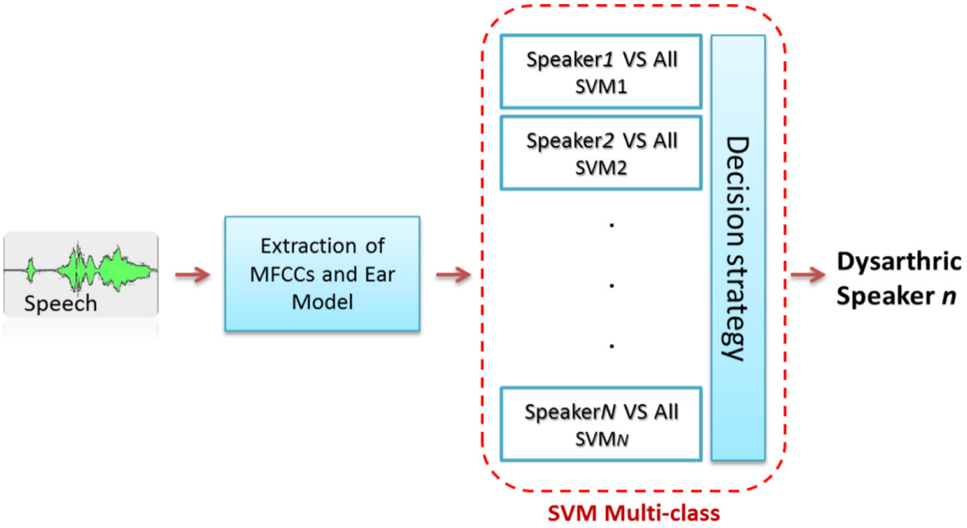

Fig. 5 – Simplified diagram of the dysarthric speaker identification process using Multiclass-SVM.

medical tools to identify patients. It can have important usages in combination with other modalities in biometry applications such as fingerprint, face and iris identification.

The design of automatic recognition system tries to mimic the human ability to identify speaker from speech. The human auditory system extracts specific cues for speaker-dependent perception. These perceptual cues have different levels, ranging from high levels such as pronunciation, prosody and semantic to low levels such as acoustic and phonetic of speech. The differences between speakers come mainly from variation in physiological characteristics of speech production organs, and some of acquired or learned pronunciation reflexes [24].

Thereby, signal-based low-level cues are used as front-end of the proposed dysarthric speaker identification system (Figs. 4 and 5). The task aims at determining an unknown dysarthric speaker's identity by comparing the voice model against *N* templates (1:*N*). The voice print is represented by the MFCCs parameters and the auditory cues while the GMM, SVM and hybrid GMM/SVM methods are applied to modeling and classification.

4. Automatic assessment of dysarthric speech

The intelligibility of dysarthric speech can be ranged from near-normal to unintelligible, depending on the severity of the disease. Usually, to measure the disorder's severity or a rehabilitation progress, a large set of tests are used to assess the level of intelligibility. The automatic methods of assessment can help clinicians to monitor the dysarthria disease [25].

Several approaches of automatic assessment of dysarthric speech have been developed:

- In [26], feed-forwards artificial neural networks (ANNs) and Support Vector Machines (SVMs) with phonological features have been used to design discriminative models for dysarthric speech.
- A combination of the statistical GMM and soft-computing technique of ANNs was used along with MFCCs and speech rhythm metrics in [27], achieving 86% accuracy over four degrees of severity levels.
- In [28], a classifier based on Mahalanobis distance and discriminant analysis was developed for dysarthria severity classification using acoustic features, where 95% accuracy was achieved over two levels of severity.
- Linear discriminant analysis (LDA) was combined with a SVM automatic classifier using prosodic features as front-end. This method achieved a classification rate of 93% over four severity levels of dysarthria in [25].
- Recently, an automatic intelligibility assessment system which performs a binary classification by capturing atypical variation in dysarthric speech by using LDA and SVM classifiers was proposed [29]. The experiments were carried out on the Torgo database.

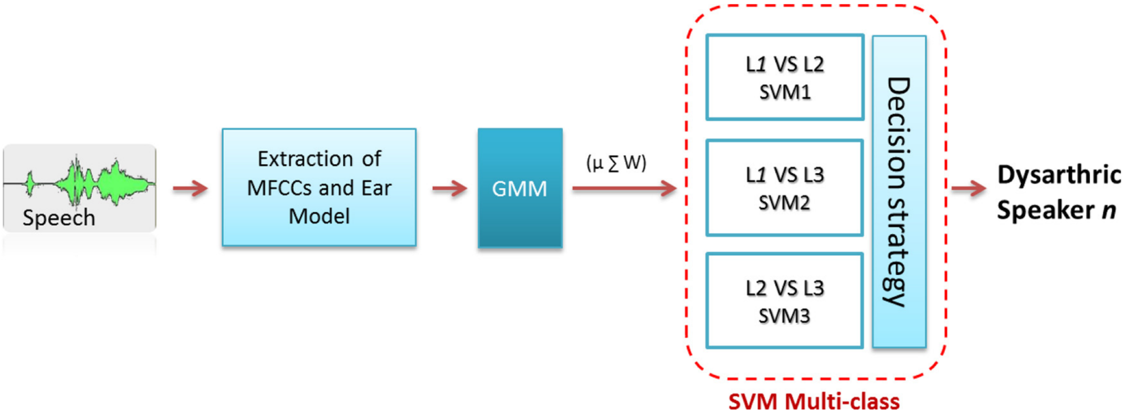

Fig. 6 – Simplified diagram of the dysarthric speech assessment process using GMM/SVM.

In our approach, the front-end processing already computed for the task of dysarthric speaker identification (MFCCs + Auditory cues) is used to assess the severity level of dysarthria through the different classifiers. The task consists of classifying three severity levels of dysarthria using both the Nemours dysarthric speech database and the Torgo database of acoustic and articulatory speech (Fig. 6). We have normalized both databases in order to minimize the combining data conflict.

The evaluation of the dysarthria severity levels is based on the *Frenchay Dysarthria Assessment* [30]. Files containing the detailed FDA of the dysarthric speakers are available for both Nemours and Torgo database. The FDA was first published in 1983, following a research which identified the pattern and character of speech production and oromotor movements involved with proven neurological diseases. The test protocol should assist with diagnosis, guide treatment and have well validity and reliability. The FDA second edition (FDA-2) has been amended to integrate recent knowledge about motor speech disorders and their contribution to neurological diagnosis. For the second edition (2008), some items included in the first edition (1983) have been left out as they were considered inadequate or redundant for the aim of management and diagnosis [31]. To determine the assessment global score of each dysarthric speaker in this work, we rely on the research in [7]. For the Torgo database, which is more recent than Nemours, we consider the amendments in the FDA-2 besides the FDA-1 assessment included for each of the eight dysarthric speakers.

5. Experiments

5.1. *Speech material*

The availability of suitable databases remains a critical issue that disrupted the field of speech analysis and processing of verbal communication disorders. In this field, the main constraint is related to recording conditions and to availability and approval of patients and health agencies. The few available databases have been designed by an experimented staff following very restrained conditions. Most of data even exist for a few languages, usually North American English, and with lower amount of recordings in non-American English, Spanish, Korean or German [32].

To train and develop our automatic dysarthric speech systems, we expect avoiding the lack of data and the lack of diversity in the recordings by combining two significant dysarthric speech databases, namely, the Nemours database of dysarthric speech and the Torgo database of acoustic and articulatory speech from speakers with dysarthria.

It is important to control the various differences across databases in order to minimize the data combination effects on the experiments. For the Nemours database, the recording sessions were conducted in a small sound dampened room, while for the Torgo database an acoustic noise reduction was performed. For both databases, the Pulse Code Modulation (PCM) was used for speech coding, and the speech signal was sampled at 16 kHz rate with 16 bits sample resolution. The Resource Interchange File Format (RIFF) was used for audio files. Furthermore, we have performed normalization and a silence removal at the pre-processing stage on all recording materials.

5.1.1. *Nemours database*

Nemours is one of the few available databases of dysarthric speech. It includes recordings of 11 American patients who suffer from dysarthria with different degrees of severity. The dysarthric speakers produced 814 short and nonsensical sentences. Furthermore, the database contains two commonly used passages, the connected-speech paragraphs: ‘‘Grandfather’’ and ‘‘Rainbow’’ pronounced by each dysarthric patient. Every sentence in the database has the following form: ‘‘The X is Ying the Z’’, generated by randomly selecting X and Z, without substitution, from an ensemble of 74 monosyllabic nouns and selecting Y from an ensemble of 37 disyllabic verbs without substitution. This process produced 37 sentences, from which 37 additional sentences were generated by swapping the nouns X and Z. Thus, through the entire set of 74 sentences, each patient pronounced twice each noun and verb. The complete database has been marked at the word level, and also at the phoneme level for the sentences of 10 patients. Finally, a speech pathologist who conducted the recording session produced the whole speech corpus as the healthy control (HC).

5.1.2. *Torgo database*

Torgo database was developed by the departments of Computer Science and Speech Language Pathology at the University of Toronto, in collaboration with the Holland–Bloorview Kids Rehabilitation Hospital in Toronto. It includes about 23 h of English speech data collected between 2008 and 2010 from 8 dysarthric speakers (3 females and 5 males) with different level of intelligibility, and 7 non-dysarthric speakers (3 females and 4 males) as a control group. The acoustic data were recorded through two different microphones; the first is an array microphone with 8 recording elements placed at a distance of 61 cm facing the speaker, and the second microphone is a head-mounted electret. The sampling rate was 44.1 kHz for the first microphone and 22.1 kHz for the second. A down-sampling at 16 kHz was performed on the acoustic signals by the database designers [33]. Therefore, all the speech signals provided in the Torgo database are sampled at 16 kHz, which also corresponds to the sampling frequency of the Nemours speech database.

It is important to note that the use of the 16 kHz sampling frequency is very common in speech processing and particularly in speech recognition applications. This value insures that the relevant information contained in the speech is present. Indeed, the practical bandwidth of a (wide-band) speech signal is about 8 kHz, meaning that the percentage of the speech signal energy beyond 8 kHz is negligibly small. If we sample this speech signal at a rate of 16 kHz (twice of the bandwidth), or 16,000 samples per second, the aliasing effect will be very small [34].

Moreover, the database contains 3D recording of articulatory movements performed with a 3D electro-magnetic articulograph (EMA) system, allowing detailed studies on the nature and direction of dysarthric speech activity [8].

The Torgo database data is structured by speaker and by session. In this work, to make the automatic processing more useful, we reorganized the database by creating a new directory for each speaker. We sorted out all the audio recordings of the dysarthric speakers according to the type of the speech text consistency as follows: non-words, short words and restricted sentences. Thus, the dysarthric speaker's directory contains one folder for each of these script types plus another folder for the unrestricted sentences speech, after renaming all the speech wave files.

5.2. *Subjects*

The speakers from the Nemours database are 11 young adult males affected by different category of dysarthria in consequence of either cerebral palsy (CP) or head trauma (HT), and one male adult as a control speaker. Seven of the patients have CP, among them, two have athetoid CP (one quadriplegic), three have spastic CP with quadriplegia, and two have a combination of athetoid and spactic along quadriplegia. A two letters code is attributed to each dysartric speaker: BB, BK, BV, FB, JF, KS, LL, MH, RK, RL, and SC. Knowing that the perceptual data and the speech assessment did disregarded the too mild case and the too severe case, corresponding to the patient FB and KS respectively [35].

The Torgo database of dysarthric articulation consists of aligned acoustics from speakers with either cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS) that are two of the more recurrent causes of speech disability [36]. Each of the eight dysarthric speaker has recorded approximately 500 utterances, while every speaker of the control group has recorded about 1200 utterances. All participants have been attributed a code that begin with 'F' for female speakers and with 'M' for male speakers, then the letter 'C' follows the gender code for speakers of the control group. The last two digits designate the order in which that participant was enrolled. Thus, the 8 dysarthric speakers considered over our experiments are: F01, F03, F04, M01, M02, M03, M04, and M05 [8,37].

5.3. *Front-end processing*

The experimental dataset contains a total of 1330 sentences. It results from 70 sentences recorded by each dysarthric speaker from the both databases (11 from the Nemours database and 8 from the Torgo database).

In order to detect clean speech segments and therefore optimize the treatments, a silence removal and segmentation of audio streams that contain speech is applied on all audio recordings involved in the experiments. The approach is based on signal energy and spectral centroid characteristics, using a threshold method to detect the speech segments [38].

For each sentence uttered by each patient, twenty features are computed: 8 auditory distinctive cues (G/A, O/C, D/C, F/S, M/S, C/D, T/L and mid-external-energy) and 12 coefficients of the usual MFCC. These features are used as front-end processing of both automatic systems developed in this work: the dysarthric speaker identification system and the dysarthric speech assessment system.

5.4. *Speaker recognition of dysarthric speakers*

The task consists of developing a text-independent GMM, SVM and GMM/SVM based speaker identification systems, especially adapted to speakers suffering from dysarthria.

GMM is a generative technique used to estimate the probability density function (PDF) from a dataset, and SVM is a binary classifier that models the optimal boundary to discriminate between two classes. Divers hybrid frameworks combining GMMs and SVMs have been developed [39].

The entire dataset is partitioned into two subsets: a training subset that contains 70% of the sentences recorded by each patient, and a test subset which includes the remaining 30% of the sentences. The training subset is used to train and build the classifiers that recognize each of the 19 dysarthric speakers; the test subset is then used for evaluating the capability of the systems to recognize each patient.

The recognition rate is assumed as the performance evaluation metric of the system. It is calculated by the ratio of the right identifications to the total number of identification trials [40].

5.4.1. *GMM classifier*

The efficiency of Gaussian mixture models for modeling the speaker identity results from their ability to model arbitrary densities by capturing the speaker-dependent spectral forms [41].

Training: For each class C_i that represents one of the dysarthric speakers, the training is initiated to get a model including the parameters of every Gaussian distribution m of the class, which are: an average vector $\mu_{i,m}$, a covariance matrix $\Sigma_{i,m}$, and a weight for each Gaussian $w_{i,m}$. These characteristics are computed after performing a sufficient amount of iterations to insure the convergence of the expectation-maximization (EM) algorithm [42]. One model is generated for representing each dysarthric speaker identity.

Recognition: Each processed signal X is represented by an input acoustical vector x. The size d of this vector is the number of acoustical parameters extracted from the signal. The PDF of each feature vector for a given class C_i is estimated as follow:

$$p(x\backslash C_i) = \sum_{m=1}^{M} w_{i,m} \cdot \frac{1}{\sqrt{(2\pi)^d |\Sigma_{i,m}|}} \cdot e^{A_{i,m}} \tag{8}$$

$$A_{i,m} = \left(-\frac{1}{2}(x - \mu_{i,m})^T \cdot \frac{1}{\Sigma_{i,m}} \cdot (x - \mu_{i,m}) \right)$$

where M is the dimension of the Gaussian distribution.

The algorithm computes the likelihood that the signal X corresponds to the class C_i. The prior probability of each class is assumed to be the same. Therefore, the maximum of the probability density function indicates the closest match.

Table 3 provides the performance evaluation rate of the tested systems and the optimal system parameters.

It is important to note that a t-test statistical hypothesis testing was successfully performed for the results of all our experiments.

The results show that the combination of MFCCs and Auditory distinctive features as a front-end processing performs better than the baseline MFCC system over most GMM model orders. Indeed, the hybrid MFCCs–Auditory cues method achieves a 97.2% correct identification rate for an order model of 64 Gaussians.

5.4.2. *SVM classifier*

The SVM was proposed by Vapnik as a novel machine learning approach, through introducing the kernel function [43]. This approach projects a data to a new high-dimensional space allowing linear separation. The classifier establishes a linear hyperplan separator that maximizes the margin between two data groups.

The kernel function is the key constituent of the SVM classifier. The learning ability and the generalization capability depend on the choice of this function. In our experiments, a radial basis function (RBF) is applied as the kernel function. It is formulated as:

$$k(x, y) = \exp\left(-\frac{||x - y^2||}{2 \times \sigma^2} \right) \tag{9}$$

where the σ parameter represents the Gaussian width that is tunable through experiments. Then, we use n-fold cross validation to get the box constraint parameter C and the RBF-σ.

Table 3 – Identification performance for different acoustic features and several GMM model orders.

Model order	MFCCs	Auditory cues	MFCCs + auditory cues
M = 4	93.4	81.9	91.3
M = 8	93.1	89.3	93.4
M = 16	94.8	92.7	95.1
M = 32	95.3	93.7	95.3
M = 64	**95.8**	93.9	**97.2**
M = 128	95.6	**94**	97.1

Bold values indicate the best performance for each column.
Bold underline value indicates the global best performance.

A multiclass-SVM using the "one-against-all" approach is set to distinguish the nineteen dysarthric speaker identities. This method consists of building one binary SVM per identity, which is trained to discriminate the observations of one patient from the observations of all remaining patients. The necessary number of binary classifiers (SVMs) is 19.

Table 4 shows the performance of the SVM system using the different front-end processing:

For the SVM identification method, the best rate is obtained with the MFCCs features.

5.4.3. *Effects of time normalization on the performance of GMM and SVM classifiers*

The GMM achieves a high correct classification rate compared to SVM. This difference can probably be explained by the ability of GMM and statistical-based methods in general, to deal with variable time lengths of input data. SVM cannot effectively process the utterances having different and changing durations provided by both Nemours and Torgo databases. Hence, the use of SVM requires defining the same time length for all speech utterances. Such normalization process is inherent to the SVM input data preparation to ensure the homogeneity of the front-end processing.

5.4.4. *GMM/SVM classifier*

To benefit from both the data description ability of the GMM and the high classification performance of the SVM, we combine the two systems described above in order to use the Gaussian distributions as a parametric base of the SVMs classifiers.

Table 4 – One-against-all SVM identification performance for different acoustic features.

	MFCCs	Auditory cues	MFCCs + auditory cues
Multiclass-SVM	**84.7**	70.1	84.2

Bold values indicates the global best performance.

Table 5 – Hybrid GMM/SVM identification performance for different acoustic features and different GMM model orders.

Model order	MFCCs	Auditory cues	MFCCs + auditory cues
M = 8	77.1	53.7	90
M = 16	77.6	55.8	86.3
M = 32	**82.6**	**57.1**	**91.1**
M = 64	81	55.7	90.5

Bold values indicate the best performance for each column.
Bold underline value indicates the global best performance.

Table 6 – Severity levels of dysarthric speakers from the Nemours database.

Patients	KS	SC	BV	BK	RK	RL	JF	LL	BB	MH	FB
Severity (%)	–	49.5	42.5	41.8	32.4	26.7	21.5	15.6	10.3	7.9	7.1

To reach a better efficiency, only the mean parameters of the GMM models are involved in the SVM front-end processing [44]. Table 5 provides the performance evaluation rate of the hybrid GMM/SVM and the optimal system parameters.

The best performance achieved with the GMM/SVM identification system is 91.1% using the combined MFCCs–Auditory cues as front-end.

5.5. *Assessment process*

For both Nemours and Torgo databases, an assessment of the motor functions for each dysarthric speaker was carried out according to the standardized Frenchay Dysarthria Assessment (FDA, Enderby, 1983) by a speech pathologist. The test is divided into 8 sections: reflex, respiration, lips, jaw, palate, laryngeal, tongue and intelligibility. Besides this, the FDA contains a section on influencing factors such as sensation and rate. Through the eight sections, a global set of 28 perceptual dimensions are assessed using a 9-point scale rate.

To build our automatic system that assesses the dysarthria severity, a single numerical value is needed to represent the global FDA-score of each patient. Based on [7], the severity levels of the dysarthric speakers of Nemours are showed in Table 6.

We have to mention that these scores are not available for the dysarthric speakers of the Torgo database. Only a file containing the 9-point scale rate of the 28 perceptual dimensions are provided for each patient. Therefore, a new global FDA-score is proposed for the eight dysarthric participants of the Torgo database, based on the perceptual dimension scores on the basis of the recent FDA-2 protocol [31].

Table 7 presents the FDA-rating scale used by clinicians to assess dysarthric speaker's capacity, on a range of behaviors related to speech function (Section). The ability of each parameter is rated from normal-function to no-function using the 9-point scale, which contains 5 descriptors (a, b, c, d and e) + ½ point. The new methodology for the estimation of global FDA-scores proposed in this study is based on the following two-step process:

– In the first step, for each section, an underlying-score is estimated by averaging the perceptual dimension rates that belong to the respective section. According to the FDA-2 protocol, we removed the jaw tests from the evaluation because the dysarthric patients are rarely affected by a jaw impairment, thus the information did not serve the assessment [31].
– In the second step, a global rate is calculated in percentage (%) using the seven underlying-scores' (of the first step) average. The FDA-2 specifies that laryngeal function and lip movements contribute substantially to lack of the intelligibility, and thereby it can be suitable to focus on these speech functions more than the other areas [31]. On this basis, we have assigned more weights for the laryngeal and lip sections in the estimation of the new global FDA-score.

Table 8 shows the final results of this new score estimation methodology for the Torgo database's patients.

The FDA score expresses the intelligibility level and it is corresponding inversely to the dysarthria severity level. All dysarthric speakers can be divided into three subgroups, based on their assessment. The subgroups are: 'mild L1' that includes the patients F04, F03, M03, FB, MH, BB and LL; 'severe L2' containing the participants M05, F01, JF, RL, RK, BK and BV; and the more 'severe L3' which includes the speakers M02, M01, M04, SC and KS.

The training subset is composed of 70% of the complete dataset and it is used to train the different classifiers for discriminating the L1, L2 and L3 severity levels. Then, the evaluation of the system capacity to classify the sentences into the corresponding severity levels is performed by using the remainder 30% of the dataset that represents the test subset. The performance evaluation metric of the assessment system is obtained as follows:

$$\textit{Performance rate} = \left(\frac{\textit{number of correct severity classifications}}{\textit{total number of severity classification trials}}\right)\times 100 \qquad (10)$$

Table 7 – Structure of the FDA-rating scale.

Section	Perceptual dimension
Reflex	Cough Swallow Dribble/drool
Resp.	At rest In speech
Lips	At rest Spread Seal Alternate In speech
Jaw	At rest In speech
Palate	Fluids Maintenance In speech
Laryngeal	Time Pitch Volume In speech
Tongue	At rest Protrusion Elevation Lateral Alternate In speech
Intel.	Words Sentences Conversation

Table 8 – The new proposed Frenchay Dysarthria Assessment (FDA) scores of dysarthric speakers from the Torgo database.

Patients	M04	M01	M02	F01	M05	M03	F03	F04
Score rate (%)	44.35	49.79	49.79	55.87	57.96	94.29	96.67	96.67

5.5.1. *GMM classifier*

To represent each of the three severity levels, a model including a Gaussian distribution is generated. Table 9 shows the assessment rates and the optimal system parameters.

The GMM method using MFCCs and Auditory distinctive features achieved a 93.2% correct classification rate, over 3 levels of dysarthria severity; L1, L2 and L3.

5.5.2. *SVM classifier*

A multiclass-SVM using the "one-against-one" method is performed to discriminate the severity levels of dysarthria. Binary classifiers are set to distinguish classes C_i and C_j, knowing that: $0 < i \le I$ and $0 < j < i$, where I is the number of groups. $(I(I-1))/2$ binary SVMs are needed to classify I groups [45].

The multiclass-SVM includes three binary SVMs. A "majority voting" method (best candidate) is used as the decision strategy of the system. Table 10 presents the performance of the SVM assessment system using the different speech features.

Table 9 – Assessment performance for different acoustic features and several GMM model orders.

Model order	MFCCs	Auditory cues	MFCCs + Auditory cues
M = 4	70.4	71.3	68.4
M = 8	76.8	65.7	72.6
M = 16	83.9	75.2	83.6
M = 32	88.9	83.6	87.1
M = 64	90.5	86.5	90.7
M = 128	**92.3**	87.1	**93.2**
M = 256	92.2	**88.6**	92.9

Bold values indicate the best performance for each column.
Bold underline value indicates the global best performance.

Table 10 – One-against-one SVM assessment performance for different acoustic features.

	MFCCs	Auditory cues	MFCCs + auditory cues
Multiclass-SVM	**76.6**	66.4	75.3

Bold values indicates the global best performance.

Table 11 – Hybrid GMM/SVM assessment performance for different acoustic features and different GMM model orders.

Model order	MFCCs	Auditory cues	MFCCs + Auditory cues
M = 8	76.6	62.9	**78.8**
M = 16	**75.5**	61.8	74.2
M = 32	73.2	**63.1**	74.2
M = 64	74.5	62.4	72.6

Bold values indicate the best performance for each column.
Bold underline value indicates the global best performance.

The best performance of 76.6% is achieved by the SVM-based system using the MFCCs.

5.5.3. *GMM/SVM classifier*

We evaluated a hybrid GMM/SVM system with different GMM model orders and the "one-against-one" multiclass-SVM approach. Table 11 shows the performance results.

By combining the MFCCs and Auditory cues, and using eight Gaussians, the hybrid GMM/SVM classifier achieved a 78.8% correct classification rate.

6. Conclusion

For both dysarthric speaker identification and dysarthria severity level assessment, the proposed combination of MFCCs and Auditory-based cues achieves a satisfactory performance. This combination outperforms the MFCCs in the case of GMM and hybrid GMM/SVM classifiers.

Our first achievement is in the field of dysarthric speaker recognition where the 64-mixture GMM-based classifier using a combination of conventional MFCCs and auditory-based distinctive features achieves the best identification rate of 97.2%. This result can be considered as very promising if it is compared with the state-of-the-art systems dealing with the recognition of speakers suffering from verbal communication disorders.

The second achievement of this study is the performance of our severity level classifier that also outperforms the state-of-art in the field of automatic assessment and diagnosis of dysarthric speech. Indeed, the combination of MFCCs and the auditory-based cues used as input of the GMM classifier achieves a correct classification rate of 93.2%. It is important to note that our experimental findings are obtained by using datasets of dysarthric speech involving two major linguistic resources in the market: the Nemours and Torgo databases.

Finally, our third achievement is the new global FDA-score that we proposed for the eight dysarthric participants of the Torgo database, based on the perceptual dimension scores on the basis of the recent FDA-2 protocol. Indeed, this new protocol introduced improvements that include the emphasizing of few speech characteristics, such as laryngeal function and lip movements in order to deal with impairments related to the lack of intelligibility, and the removal of some others that have been considered redundant or unreliable to assessment and treatment, such as the jaw evaluation.

The proposed systems are expected to be useful for the accessibility of dysarthric speakers to biometric systems and for the automatic assessment and/or diagnosis of dysarthric patients.

Acknowledgment

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the University

of Moncton (NB, Canada) and the University of Sciences and Technology Houari Boumediene (Algiers, Algeria).

REFERENCES

[1] Melf RS. Communication disorders. Available from http://emedicine.medscape.com/article.

[2] Roth C. Dysarthria. In: Caplan B, Deluca J, Kreutzer JS, editors. Encyclopedia of clinical neuropsychology. Springer; 2011. p. 905–8.

[3] American Speech-Language-Hearing Association. Available from: http://www.asha.org/.

[4] Enderby P. Disorders of communication: dysarthria. In: Tselis AC, Booss J, editors. Handbook of clinical neurology. Elsevier; 2013.

[5] Rudzicz F. Using articulatory likelihoods in the recognition of dysarthric speech. J Speech Commun 2012;54:430–44.

[6] Rudzicz F. Adjusting dysarthric speech signals to be more intelligible. J Speech Commun 2013;27:1163–77.

[7] Menendez-Pidal X, Polikoff JB, Peters SM, Leonzio JE, Bunnell HT. The Nemours database of dysarthric speech. Fourth international conference on spoken language; 1996. pp. 1962–5.

[8] Rudzicz F, Namasivayam AK, Wolff T. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Lang Resour Eval 2012;46(4):523–41.

[9] O'Shaughnessy D. Speech communication: human and machine. IEEE Press; 2001.

[10] Shahamiri SR, Salim SSB. Artificial neural networks as speech recognizers for dysarthric speech: identifying the best-performing set of MFCC parameters and studying a speaker-independent approach. J Adv Eng Inform 2014;28 (1):102–10.

[11] Davis S, Mermelstein P. Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 1980;28(4):357–66.

[12] Flanagan JL. Models for approximating basilar membrane displacement. Bell System Technol J 1960;39:1163–91.

[13] Lyon RF. A computational model of filtering, detection, and compression in the cochlea. IEEE Int Conf Acoust Speech Signal Process. 1982. pp. 1282–5.

[14] Seneff S. A joint synchrony/mean-rate model of auditory speech processing. J Phonet 1988;16:55–76.

[15] Ghitza O. Auditory models and human performance in tasks related to speech coding and speech recognition. IEEE Trans Speech Audio Proc SAP 1994;2:115–32.

[16] Stern R, Morgan N. Hearing is believing: biologically inspired methods for robust automatic speech recognition. IEEE Signal Process Mag 2012;29(6):34–43.

[17] Schluter R, Bezrukov L, Wagner H, Ney H. Gammatone features and feature combination for large vocabulary speech recognition. Acoustics, speech and signal processing, ICASSP 2007, IEEE international conference, vol. 4. 2007. pp. 649–52.

[18] Kim C, Stern RM. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. Acoustics, speech and signal processing (ICASSP), 2012 IEEE international conference. 2012. pp. 4101–4.

[19] Caelen J. Space/time data-information in the ARIAL project ear model. Speech Commun 1985;4:457–67.

[20] Selouani SA. Speech processing and soft computing. New York: Springer; 2011.

[21] Selouani SA, O'Shaughnessy D, Caelen J. Incorporating phonetic knowledge into an evolutionary subspace approach for robust speech recognition. Int J Comput Appl 2007;29:143–54.

[22] Selouani SA, Caelen J. Recognition of Arabic phonetic features using neural networks and knowledge-based system: a comparative study. Int J Artif Intell Tools 1999;8 (1):73–103.

[23] Selouani SA, Tolba H, O'Shaughnessy D. Auditory-based acoustic distinctive features and spectral cues for robust automatic speech recognition in Low-SNR car environments. Conference of the North American Chapter of the association for computational linguistics on human language technology: HLT-NAACL; 2003. pp. 91–3.

[24] Mary L. Extraction and representation of prosody for speaker, speech and language recognition. Springer briefs in speech technology; 2012 [Chap 1].

[25] Kadi KL, Selouani SA, Boudraa B, Boudraa M. Automated diagnosis and assessment of dysarthric speech using relevant prosodic features. In: Yang G, Ao S, Gelman L, editors. Transactions on engineering technologies. Springer; 2013.

[26] Rudzicz F. Phonological features in discriminative classification of dysarthric speech. ICASSP; 2009.

[27] Selouani SA, Dahmani H, Amami R, Hamam H. Using speech rhythm knowledge to improve dysarthric speech recognition. Int J Speech Technol 2012;15(1):57–64.

[28] Paja MS, Falk TH. Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech. Interspeech; 2012.

[29] Kim J, Kumar N, Tsiartas A, Li M, Narayanan S. Automatic intelligibility classification of sentence-level pathological speech. Comput Speech Lang 2014;29(1):132–44.

[30] Enderby PM. Frenchay dysarthria assessment. PRO-ED; 1983.

[31] Enderby PM, Palmer R. Frenchay dysarthria assessment. Second Ed (FDA-2). PRO-ED; 2008.

[32] Baghai-Ravary L, Beet SW. Automatic speech signal analysis for clinical diagnosis and assessment of speech disorders. Springer briefs in electrical and computer engineering; 2013.

[33] LDC. Catalog number LDC2012S02. Linguistic Data Consortium; 2012, Available from: http://catalog.ldc.upenn.edu/docs/LDC2012S02/README.txt.

[34] Deng L, O'Shaughnessy D. Speech processing: a dynamic and optimization-oriented approach. Marcel Dekker, Inc; 2003.

[35] Polikoff JB, Bunnel HT. The Nemours database of dysarthric speech: a perceptual analysis. 14th international congress of phonetic sciences; 1999. pp. 783–6.

[36] Kent RD, Rosen K. Motor control perspectives on motor speech disorders. In: Maassen B, Kent RD, Peters H, Van LP, Hulstijn W, editors. Speech motor control in normal and disordered speech. Oxford University Press; 2004. p. 285–311. chap 12.

[37] Rudzicz F. Articulatory knowledge in the recognition of dysarthric speech. IEEE Trans Audio Speech Lang Process 2011;19(4).

[38] Giannakopoulos T, Pikrakis A. Introduction to audio analysis. Academic Press, Elsevier; 2014.

[39] Campbell WM, Karam ZN. A framework for discriminative SVM/GMM systems for language recognition. Interspeech. 2009. pp. 2195–8.

[40] Abd El-Samie FE. Information security for automatic speaker identification. Springer briefs in speech technology. Springer; 2011.

[41] Reynolds DA, Rose RC. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans Speech Audio Process 1995;3(1):72–83.

[42] Dempster AP, Laird NM, Rubin DB. Maximum-likelihood from incomplete data via the EM algorithm. J Acoust Soc Am 1977;39(1):1–38.

[43] Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw 1999;10(5):988–99.

[44] liu M, Xie Y, Yao Z, Dai B. A new hybrid GMM/SVM for speaker verification. the 18th international conference on pattern recognition. IEEE; 2006.

[45] Fleury A, Vacher M, Noury N. SVM-based multimodal classification of activities of daily living in health smart homes: sensors, algorithms, and first experimental results. IEEE Trans Inform Technol Biomed 2010;14(2): 274–83.