ASHA American Speech-Language-Hearing Association

JSLHR

Clinical Focus

Perceptual Classification of Motor Speech Disorders: The Role of Severity, Speech Task, and Listener's Expertise

Michaela Pernon,[a,b,c,d] Frédéric Assal,[a,b] Ina Kodrasi,[e] and Marina Laganaro[f]

[a] Neurology Department, Geneva University Hospitals, Switzerland [b] Faculty of Medicine, University of Geneva, Switzerland [c] Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS-Université Sorbonne Nouvelle, Paris, France [d] CRMR Wilson & Parkinson Unit, Neurology Department, Hôpital Fondation Adolphe de Rothschild, Paris, France [e] Signal Processing for Communication Group, Idiap Research Institute, Martigny, Switzerland [f] Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland

ARTICLE INFO

Article History:
Received September 29, 2021
Revision received January 25, 2022
Accepted April 20, 2022

Editor-in-Chief: Bharath Chandrasekaran
Editor: Stephanie Borrie

https://doi.org/10.1044/2022_JSLHR-21-00519

ABSTRACT

Purpose: The clinical diagnosis of motor speech disorders (MSDs) is mainly based on perceptual approaches. However, studies on perceptual classification of MSDs often indicate low classification accuracy. The aim of this study was to determine in a forced-choice dichotomous decision-making task (a) how accuracy of speech-language pathologists (SLPs) in perceptually classifying apraxia of speech (AoS) and dysarthria is impacted by speech task, severity of MSD, and listener's expertise and (b) which perceptual features they use to classify.

Method: Speech samples from 29 neurotypical speakers, 14 with hypokinetic dysarthria associated with Parkinson's disease (HD), 10 with poststroke AoS, and six with mixed dysarthria associated with amyotrophic lateral sclerosis (MD-FlSp [combining flaccid and spastic dysarthria]), were classified by 20 expert SLPs and 20 student SLPs. Speech samples were elicited in spontaneous speech, text reading, oral diadochokinetic (DDK) tasks, and a sample concatenating text reading and DDK. For each recorded speech sample, SLPs answered three dichotomic questions following a diagnostic approach, (a) neurotypical versus pathological speaker, (b) AoS versus dysarthria, and (c) MD-FlSp versus HD, and a multiple-choice question on the features their decision was based on.

Results: Overall classification accuracy was 72% with good interrater reliability, varying with SLP expertise, speech task, and MSD severity. Correct classification of speech samples was higher for speakers with dysarthria than for AoS and higher for HD than for MD-FlSp. Samples elicited with continuous speech reached the best classification rates. An average number of three perceptual features were used for correct classifications, and their type and combination differed between the three MSDs.

Conclusions: The auditory-perceptual classification of MSDs in a diagnostic approach reaches substantial performance only in expert SLPs with continuous speech samples, albeit with lower accuracy for AoS. Specific training associated with objective classification tools seems necessary to improve recognition of neurotypical speech and distinction between AoS and dysarthria.

The auditory-perceptual approach is currently the "gold standard" in clinical practice for diagnostic assessment in motor speech disorders (MSDs), as well as for the assessment of severity and changes over time in voice and speech. Accurate and differential diagnosis of MSD is essential for speech-language pathologists (SLPs) to

Correspondence to Michaela Pernon: michaela.pernon@gmail.com.
Disclosure: *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

propose an appropriate treatment plan for the patient, or to anticipate declining function in neurodegenerative etiologies, or else to facilitate medical diagnosis. However, the auditory-perceptual approach has some well-known limits, giving rise to low or variable classification accuracy and low interrater agreement. The variability of accuracy in perceptual classification has been reported to be related to several factors across studies, such as listener's expertise, the type of speech task, and the type and severity of MSD.

Here, we aim at investigating (a) how these different factors influence the accuracy of perceptual classification of MSDs including apraxia of speech (AoS) and subtypes of dysarthria and (b) which perceptual features they use to classify. In the following introduction, we review the challenging differential diagnosis of MSDs in an auditory-perceptual approach before focusing on the factors that may impact on classification accuracy in SLPs.

The Challenge of Differential Diagnosis of MSDs in the Auditory-Perceptual Approach

MSDs represent a large portion of the caseload for clinicians engaged in the management of acquired neurogenic communication disorders (Duffy, 2013; Simmons & Mayo, 1997). AoS and dysarthria are two distinct types of MSDs, resulting from different etiologies such as stroke, neurodegenerative diseases, brain injury, brain tumor, and so forth (Darley et al., 1969a, 1969b, 1975; Duffy, 2013; McNeil et al., 2009). These two main MSDs have been attributed to the impairment of different motor levels and processes of motor speech production. AoS has been associated to impaired retrieving and/or assembling of speech motor plans (Blumstein, 1990; Code, 1998; Darley et al., 1975; Van der Merwe, 2021; Ziegler, 2009), a breakdown in translating encoded phonological representations to articulated speech (McNeil et al., 1997). Dysarthria, on the other hand, is defined as a deficit of the motor control and execution of the neuromuscular commands involved in speech production (Darley et al., 1969a, 1969b, 1975; Duffy, 2013; Guenther, 2016; Van der Merwe, 2021). Dysarthria has therefore been attributed to impaired motor programming and execution in models of motor speech control (Guenther, 2016; Van der Merwe, 2021). Seven subtypes of dysarthria, determined according to the impaired underlying pathophysiological neurosubsystem, have been described and classified in the framework of the standard classification by Darley et al. (1969a, 1969b, 1975) from Mayo Clinic: flaccid dysarthria (in bulbar palsy), spastic dysarthria (in pseudobulbar palsy or bilateral upper motor neuron lesion), ataxic dysarthria (in cerebellar disorders), hypokinetic dysarthria (in parkinsonism; HD), hyperkinetic dysarthria (in dystonia and chorea), and a mixed dysarthria (combining flaccid and spastic dysarthria: MD-FlSp), described in the amyotrophic lateral sclerosis (ALS), to which two subtypes have been added more recently (Duffy, 2013; dysarthria with undefined etiology and dysarthria associated with focal unilateral upper motor neuron [UUMN] lesion).

The clinical diagnosis of MSDs and the differential diagnosis between subtypes of MSDs are mainly based on auditory-perceptual criteria (Bunton et al., 2007; Duffy, 2013; Kent, 1996) and their relationship to the underlying pathophysiology. The auditory-perceptual approach remains the most commonly used method because of its convenience in terms of time and material, offering a quick description of the patients' speech characteristics (Kent, 1996), and partly because of the lack of other valid, sensitive, and robust markers (Delvaux & Pillot-Loiseau, 2020; Simmons & Mayo, 1997). However, the differential diagnosis between MSDs based on an auditory-perceptual approach can be quite challenging for SLPs due to multiple sources of biases.

A first bias is related to the internal representations the listeners have about perceptual features. The internal representations are built through the clinical experience (number of years of practice, number of MSD patients followed, degree of familiarity with the MSD subtypes, underlying pathologies and severity of MSD, and place of practice) and type and amount of training (Kim et al., 2011). With expertise, the clinicians develop the ability to recognize complex speech pattern, which builds and enriches their internal representations. Internal representation of complex speech patterns leads SLPs to direct diagnosis, without need of a checklist approach of perceptual features (Duffy, 2013). However, the multiple parameters contributing to the construction of internal representations also involve that perceptual classification is highly subjective.

In addition, mild speech disorders may be missed with an auditory-perceptual approach as audition and auditory perception of the listeners give rise to the phenomenon of phonemic restoration. It drives the listener to replace a distorted sound by another sound, influenced by the meaning or the syntactic structure of an utterance, or also the degree of familiarity with the speech corpus (Kent, 1996).

Moreover, although the underlying pathophysiology and impaired motor speech production processes and levels differ, the different MSDs also share several signs of impaired speech. For instance, most of the core perceptual clinical symptoms of AoS (distorted articulation and slow speech rate due to lengthened intersegment durations and segments, syllabification, and additions of pauses; Ballard et al., 2016; Cunningham et al., 2016; McNeil et al., 2009) are also found in dysarthria. As for dysarthria, of the 38 perceptual features used for their characterization in the

Mayo Clinic classification, several are shared by the subtypes of dysarthria, such as imprecise consonants (Darley et al., 1969a, 1969b, 1975).

Finally, there is a lack of consensus on the core sets of diagnostic features for AoS even if there are currently several attempts to define its most sensitive and specific differential diagnostic criteria (Duffy, 2013; Duffy et al., 2021; Jonkers et al., 2017; McNeil et al., 2009; Strand et al., 2014; Utianski et al., 2018; Wambaugh et al., 2019). Indeed, Molloy and Jagoe (2019) found 34 different speech features used as diagnostic criteria for AoS across 157 included studies in their scoping review. It showed a set of 14 common features, mostly perceptual, that was shared by only seven studies. In summary, the auditory-perceptual approach in the assessment of MSDs is far from being perfect, and its diagnostic and classification accuracy are affected by several factors that will be further developed in the following sections.

Factors Influencing Perceptual Classification of MSDs

Given the limits of perceptual classification highlighted above and the overlap of perceptual speech signs between subtypes of MSDs, some studies have been conducted to assess the reliability of the Mayo Clinic classification system of dysarthria or the accuracy of the classification of MSDs (Bunton et al., 2007; Zeplin & Kent, 1996; Zyski & Weisiger, 1987) or else to validate perceptual scales for AoS including perceptually based classification (Diagnostic Instrument for Apraxia of Speech; Jonkers et al., 2017; Apraxia of Speech Rating Scale 1.0 [ASRS 1.0]: Strand et al., 2014). The results of these studies do not seem to converge, but they are actually hardly comparable due to different factors that affect the decision-making process, such as the groups of speakers (number of groups, MSD subtypes, and underlying pathologies and inclusion of a control group), the severity of MSDs, the expertise/clinical experience of the listeners, the speech tasks and corpus (number and types of speech tasks and length of speech material), or the method/design of the experiments (free-choice classification task and forced-choice classification task). In the following subsections, we will briefly review each of the factors that may impact the results of auditory-perceptual classification of speech samples.

Groups of Speakers

The two main types of MSDs have mostly been studied separately in perceptual classification studies, although they still raise questions about their specific diagnostic criteria and their differential diagnosis. As a matter of fact, most perceptual classification studies include speakers with several subtypes of dysarthria (Bunton et al., 2007; Fonville et al., 2008; Lansford et al., 2014, 2016; Van der Graaff et al., 2009; Zyski & Weisiger, 1987), whereas other studies only include speakers with AoS (Duncan et al., 2020; Josephs et al., 2013). In the few studies including speakers with dysarthria and speakers with AoS, subtypes of dysarthria were not specifically considered and constitute a single heterogeneous group of dysarthria (Hybbinette et al., 2021; Jonkers et al., 2017; Mumby et al., 2007). Furthermore, a group of neurotypical speakers was rarely included (Fonville et al., 2008; Jonkers et al., 2017; Van der Graaff et al., 2009).

Severity of MSDs

Another possible confounding factor is related to the severity of MSDs that varies across speakers and groups in most studies and was usually not controlled. Severity could actually influence the dysarthric profiles of speakers of a particular dysarthria, explaining much of the interspeaker variability. This variation has been suggested to be larger than the variation across dysarthria subtypes (Kim et al., 2011; Weismer & Kim, 2010).

Degree of Expertise

The degree of expertise of SLPs, whether considered as overall clinical experience or training with specific MSD patient subtypes, has an intuitive impact on diagnostic accuracy and differential diagnosis (Simmons & Mayo, 1997). Surprisingly, most previous perceptual classification studies analyzing the impact of expertise of SLPs found no effect (Bunton et al., 2007; Fonville et al., 2008; Lansford et al., 2016; Zyski & Weisiger, 1987). Only Verkhodanova et al. (2021) observed its influence on accuracy rates in the classification of groups of speakers with dysarthria. It is possible, however, that expertise may have a larger impact when classifying in the same study speakers with AoS and speakers with dysarthria, given the lack of international consensus on the core criteria for a diagnosis of AoS and the overlap of perceptual features across subtypes of MSDs.

Speech Tasks

Perceptual classification may also be influenced by the type of speech elicitation tasks. In the clinical assessment procedure, SLPs use a set of speech tasks that provide information on different functional subsystems, including respiratory, phonatory, velopharyngeal, articulatory/sound level, or prosodic levels (Allison et al., 2020; Kent & Kent, 2000; Kent et al., 2000; Zeplin & Kent, 1996). In perceptual classification studies including speakers with AoS (Hybbinette et al., 2021; Strand et al., 2014) or mixing both AoS and dysarthria (Jonkers et al., 2017;

Mumby et al., 2007), the judgments were often made on large speech production samples from a complete speech and language assessment protocol. Thus, they involved several tasks for each speaker (Strand et al., 2014). Conversely, the majority of perceptual classification studies of dysarthric speakers used a single speech task or a sample of few concatenated tasks. The speech samples varied from a single sentence (Lansford et al., 2014, 2016); an extract from a text reading (Fonville et al., 2008); conversational or narrative speech samples (Bunton et al., 2007); a concatenation of text reading and spontaneous speech samples (Van der Graaff et al., 2009); or speech samples elicited with text reading, alternating motion rates (AMRs), and sustained phonation task (Zyski & Weisiger, 1987). To our knowledge, the impact of the type of speech elicitation tasks included in perceptual classification studies has not been controlled or explicitly tested, whereas effects of speech tasks have been described on the speech and voice of dysarthric speakers (Brown & Docherty, 1995; Kempler & Van Lancker, 2002; Van Lancker Sidtis et al., 2010, 2012).

Operationalized Methods and Design

Regarding the operationalized methods and the design used in perceptual classification studies, two main approaches can be devised: forced-choice classification tasks and free-choice classification tasks. In a forced-choice classification task, listeners are asked to rate specific aspects of speech (Allison et al., 2020), and they are constrained in their answer by choosing between a given number of propositions, either dichotomic (two choices) or with more choices. The number of choices varied across studies (e.g., six subtypes of dysarthria in Van der Graaff et al., 2009; 38 perceptual dimensions in Bunton et al., 2007). A free-choice classification task is a perceptual sorting task in which listeners are asked to group speech samples according to perceived similarity without operationalized speech features to guide clinician ratings (Clopper, 2008; Lansford et al., 2014, 2016). Better perceptual classification accuracy in a free-choice classification task relative to a forced-choice task has been reported on dysarthria by Zyski and Weisiger (1987). This result may be due to the fact that, in the forced-choice designs, the predefined response options could mislead the listener, especially when they are numerous.

Perceptual Classification Following a Diagnostic Approach

In the frame of forced-choice and free-choice designs, clinicians are not guided in their diagnostic reasoning, which could be helpful notably when medical history or diagnosis, and neuroimaging data are not provided. An approach that follows the clinical diagnostic reasoning and decision tree may be better suited to assess diagnosis and differential diagnosis of MSDs based on auditory-perceptual classification of speech samples. A step-by-step diagnostic approach has been suggested by Duffy (2013) as a means of establishing a diagnosis. In the presence of speech abnormality, the clinician proceeds by answering successive dichotomous questions: first, whether there is a neurological difficulty; if so, whether it is MSD or another neurological communication disorder; and in the case of MSD, if it is AoS or dysarthria. Finally, if the diagnosis is dysarthria, its subtype should be identified. It is therefore likely that better accuracy and interrater agreement is achieved via perceptual classification in a study in which the decisions follow the step-by-step clinical approach rather than a unique forced-choice decision between multiple options as proposed in most previous studies.

To investigate the ability of French-speaking SLPs to perceptually distinguish between MSDs, a forced-choice perceptual classification task was built following a diagnostic approach proceeding to the differential diagnosis step-by-step using dichotomic questions. Both neurotypical and pathological speakers with different types of MSDs, namely, AoS and dysarthria (MSD types), and among dysarthria, speakers with MD-FlSp and HD (dysarthria subtypes), were included. The potential impact of a set of relevant factors was also examined within the same study: listener's expertise, speech task (type and number), and severity of the MSD. Identifying which factors impact on the diagnosis and differential diagnosis of MSDs in the auditory-perceptual approach can help to better understand the limits of auditory-perceptual approach and lead to practical recommendations to improve it. To our knowledge, the factors of interest mentioned have never been considered simultaneously in previous studies and tested in a perceptual classification task following a diagnostic approach, nor did previous perceptual classification task involve a forced-choice experimental design mixing speech samples from AoS and dysarthria with a group of neurotypical speakers. Finally, we also aimed at determining on which perceptual features the listeners based their classification.

Method

Speakers

Speech samples were issued from 30 French-speaking adults with MSD and 29 neurotypical controls collected in the framework of a larger corpus of research on speech and MSD using the same speech elicitation protocol (Fougeron, Delvaux, et al., 2018; Laganaro et al., 2021; Pernon et al., 2020).

Table 1. Demographic and speech characteristics of groups of speakers.

Group of speakers	Pathology	*n*	Gender	M_{age} (*SD*; min–max)	Mean TotalDevScore of MonPaGe-2.0.s/32 (*SD*; min–max)	Mean perceptual score BECD/20 (*SD*; min–max)	Mean total score ASRS 1.0/64 (*SD*; min–max)
Neurotypical speakers	—	29	19 F	58 (17; 25–82)	0.38 (0.69; 0–2)	—	—
AoS speakers	Poststroke	10	6 F	52.5 (15.69; 24–72)	6.4 (3.01; 2–11)	9.1 (3.03; 5–15)	12 (13.53; 0–41)
MD-FlSp speakers	ALS	6	3 F	71.17 (4.83; 65–77)	5.17 (4.67; 1–12)	8 (4.24; 3–14)	7.5 (6.35; 2–17)
HD speakers	PD	14	3 F	73.5 (8.34; 55–83)	2 (2.22; 0–6)	5.07 (2.76; 1–10)	2.07 (1.59; 0–6)

Note. Em dashes indicate that there are no value for neurotypical speakers for these columns. BECD = Batterie d'Evaluation Clinique de la Dysarthrie; F = female; ASRS 1.0 = Apraxia of Speech Rating Scale 1.0; AoS = apraxia of speech; MD-FlSp = mixed dysarthria, combining flaccid and spastic dysarthria; ALS = amyotrophic lateral sclerosis; HD = hypokinetic dysarthria associated with Parkinson's disease; PD = Parkinson's disease.

Speakers With MSD

The 30 speakers with MSD had French as their first language or acquired before adulthood without foreign accent and were aged from 24 to 83 years old (M_{age} = 65.7, *SD* = 9.62). They were diagnosed as having mild-to-moderate MSDs at the Neurology Department of the University Hospital of Geneva between September 2018 and October 2019, with the following diagnosis: poststroke AoS (*n* = 10), HD (*n* = 14), and MD-FlSp secondary to ALS (*n* = 6).

For each speaker with MSD, the neurological diagnosis was established by neurologists at the university hospital based on standard clinical criteria, and MSD was assessed by an expert SLP based on the perceptual score of BECD[1] (Auzou & Rolland-Monnoury, 2006), the TotalDevScore of MonPaGe-2.0.s screening protocol (Laganaro et al., 2021), and the ASRS 1.0 (Strand et al., 2014). The severity scores for each subgroup of MSD are presented in Table 1. The MSD diagnosis made by the expert SLP was never incompatible with the one predicted from lesion loci or neurological diagnosis. Two of the speakers with AoS had also an associated UUMN dysarthria, and six of them were diagnosed with a concomitant, mild, nonfluent aphasia, assessed by Electronic Geneva Bedside Aphasia Scale (e-GeBAS; Chicherio et al., 2019), but with AoS being dominant. Their mean accuracy at the reading task of e-GeBAS was high (97.50%), as well as at the naming task of e-GeBAS (97.50%)

Neurotypical Speakers

Twenty-nine neurotypical French-speaking subjects aged 25–82 years old were recruited from the same linguistic region without foreign accent. They had no voice, speech, or language disorders and had good hearing (self-assessed goodness of hearing on a 10° scale: *M* = 8.17, *SD* = 1.21, min–max: 5–10).

One neurotypical speaker was excluded because of a pathological TotalDevScore (> 2) in the MonPaGe speech screening protocol. All speakers had corrected visual acuity when needed for reading tasks. Speakers' characteristics are summarized in Table 1. All speakers gave consent for participation in the study that received the approval from the local medical ethics committee of Geneva and the faculty ethics committee (Psychology Faculty, University of Geneva).

Speech Samples

The material for the perceptual classification task was composed of speech elicited with three different tasks from the MonPaGe screening protocol (Fougeron, Delvaux, et al., 2018; Fougeron et al., 2016; Laganaro et al., 2021; Pernon et al., 2020):

1. The production of two DDK tasks: a simple AMR CV "bababa": [bababa] and a complex AMR CCV "tratra": [tRatRa]. In the MonPaGe protocol, speakers are asked to repeat each sequence as fast and accurate as possible during 4 s.
2. The reading aloud of two paragraphs from the MonPaGe text module. The two paragraphs ("lundi"/"Monday" and "mercredi"/"Wednesday") are composed of 50 words and last in an average of 18.6 s (controls: *M* = 14.6 s; speakers with MSD: *M* = 22.6 s; see Appendix A).
3. The production of spontaneous speech elicited with the question: "What did you do during your last vacation or week-end?" (in French). For each participant, a sample of spontaneous speech with the same duration as the reading aloud task was extracted from the beginning of the sample.

The recordings took place in a quiet room at the University Hospital of Geneva and at the Laboratory

[1]BECD is the French acronym for "Batterie d'Evaluation Clinique de la Dysarthrie," which stands for "Clinical Assessment Test for Dysarthria."

of Psycholinguistics, Faculty of Psychology and Educational Sciences of Geneva. They were performed using a Shure SM35-XLR head-mounted microphone connected to an external sound card Focusrite Scarlet 2i4 USB. All recordings were preprocessed by removing the pause at the beginning and at the end, down-sampled to 16 kHz, and normalized to the same root-mean-square value. For spontaneous speech, the double talk has also been removed.

For the classification task, the speech samples were organized in four blocks, each including 59 samples (one per speaker, i.e., 236 samples for the four blocks) from specific speech task(s). The speakers' samples were randomly presented in each block, with blocks presented in the following fixed order:

(A) the concatenation of two paragraphs of a reading Text task "Monday" and "Wednesday" with the two DDK "bababa" and "tratratra" (referred below by "Text + DDK"; 8 s), with a 500-ms pause between the two speech tasks;
(B) the Spontaneous speech task, with a duration matched to the duration of the two paragraphs of text reading for each speaker (Task D; referred below by "Spontaneous");
(C) the concatenation of the two DDK alone: "bababa" and "tratra" (4 s each, 8 s in total; referred below by "DDK"); and
(D) the concatenation of the two paragraphs of the reading Text task "Monday" and "Wednesday" (referred below by "Text").

Listeners

A total of 40 French-speaking listeners who were either professional SLPs or students in speech-language pathology participated in the study, including (a) 20 students at the end of their first year of a master program in speech-language pathology at the Faculty of Psychology of the University of Geneva (the students have completed a course on MSDs and had received basic instructions in their differential diagnosis) and (b) 20 qualified SLPs (number of years of practice: M = 11, SD = 8, min–max: 1–40/average number of patients with MSDs followed since the beginning of SLPs' practice: M = 130, SD = 90, range: 15–350). These SLPs worked in private practice at university hospitals or in neurorehabilitation centers, all experienced in working with adult neurological patients, regularly diagnosed and treated patients with an MSD, as part of their practice.

These two groups are respectively referred to as "student SLP" and "expert SLP" below. They were from the same geographic area as the speakers. All of them self-reported having no hearing loss.

Procedure and Design

The perceptual classification of speech samples was implemented online as a forced-choice task using the Qualtrics platform and survey software (Qualtrics, 2019). The 236 speech samples were split in two subsets (Version I and Version II), each including all the four blocks of speech samples and half of the speakers, balanced in MSD's severity and medical etiologies/MSD types for pathological speakers, gender, and age for all speakers. Version I consisted of 116 speech samples (29 speakers × 4 blocks), and Version II consisted of 120 speech samples (30 speakers × 4 blocks).

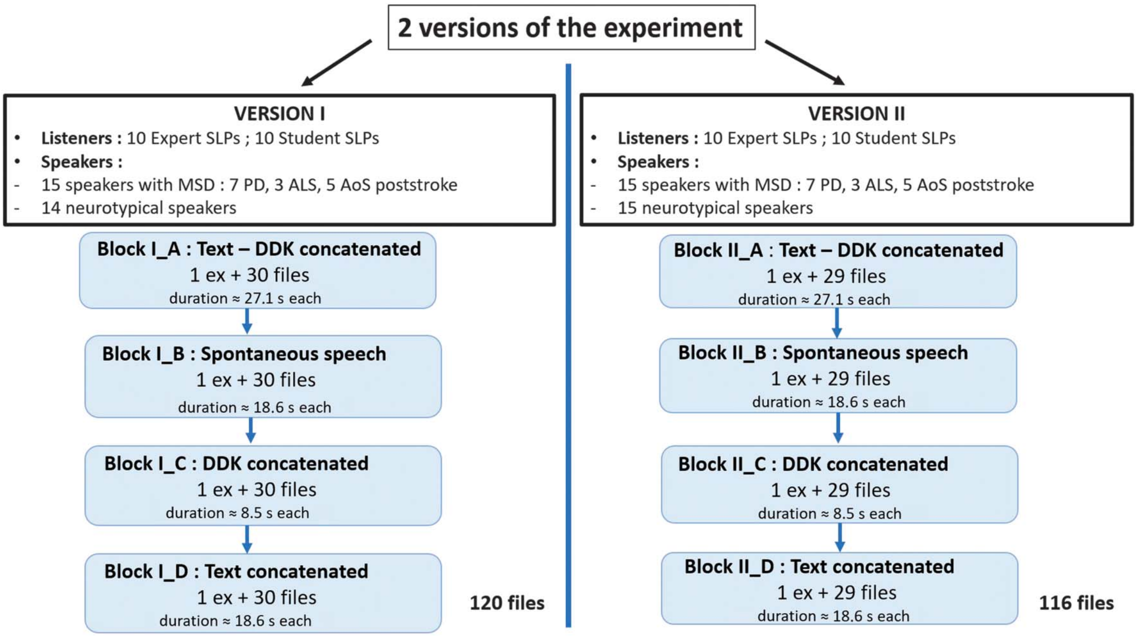
Each version was rated by half of the listeners (10 student SLPs and 10 expert SLPs). Mean duration of the procedure was around 1 hr for a whole version. The listeners could split the classification task in four parts, respecting the order defined for the four blocks (see Figure 1).

For each speech sample, listeners had to do the following (see Figure 2):

1. Determine if it corresponded to a neurotypical speaker or a pathological speaker. If the decision was "pathological speech," they had to:
2. Indicate the perceptual features on which they based their decision (only for Blocks A, B, and D in order to keep continuous part of speech in the perceptual analysis as recommended by Auzou & Rolland-Monnoury, 2006, for the perceptual evaluation). The listeners had to select the feature(s) among the eight following features: voice quality, articulation, nasality, prosody/intonation, speech rate/speech fluency, respiration, intelligibility, and naturalness of speech. For each feature, a definition was provided during the experiment (see Appendix B).
3. Decide the type of MSDs: AoS or dysarthria.
4. In case of a classification of the speech sample as dysarthria, they had to further determine the underlying pathology of the subtype of dysarthria: HD or MD-FlSp.

The questions were presented successively, similar to a clinical and diagnostic questioning (Duffy, 2013). Thus, each listener had to answer one to four questions per speech sample, depending on their first decision of neurotypical versus pathological speech sample. Before the task, listeners were instructed on which underlying medical etiologies, MSD subtypes, types of speech tasks, and speech samples they would have to rate. They also were aware that there were samples from neurotypical speakers. Neither the numbers of samples per MSD and dysarthria types nor the number, of healthy controls were disclosed. A training was conducted before each block using an

Figure 1. Repartition of the listeners, speakers, and speech samples between the two versions of the perceptual classification task. SLPs = speech-language pathologists; MSD = motor speech disorder; PD = Parkinson's disease; ALS = amyotrophic lateral sclerosis; AoS = apraxia of speech; DDK = diadochokinetic.



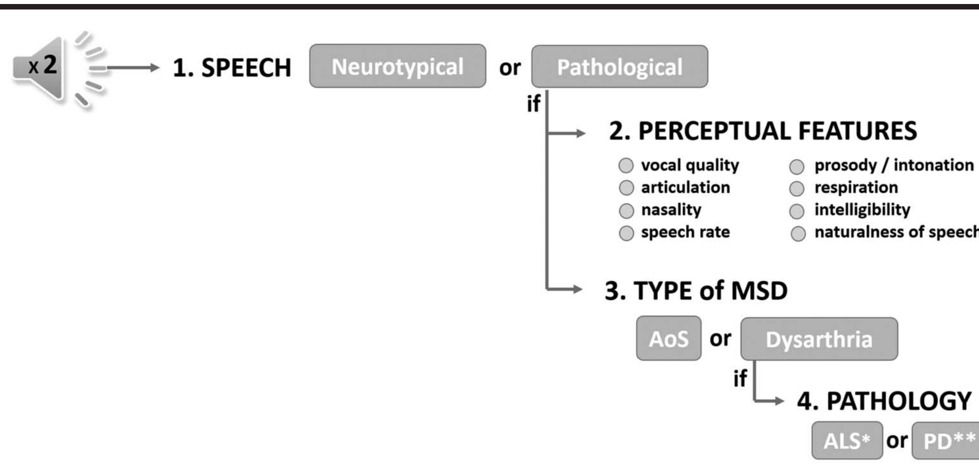
example file from the other version of the experiment. Each audio sample could be played twice at maximum.

Statistical Analyses

Statistical analyses were performed using the R software (R Core Team, 2019), with the base package, lmerTest (Kuznetsova et al., 2017), Lme4 (Bates et al., 2015), and Irr (Gamer et al., 2019) for the interrater reliability analyses.

The main analyses were run on classification accuracy using the glmer function for the generalized mixed models. We ran three glmer models with classification accuracy as the dependent variable. Each model refers to the results of the three successive dichotomic rating questions (Questions 1, 3, and 4 in Figure 1), namely, Model (i): accuracy in the classification between pathological and neurotypical speaker (Question 1); Model (ii): accuracy in the classification of MSD subtypes: AoS versus dysarthria for the subset of speakers judged as pathological

Figure 2. Rating procedure for the successive questions in the perceptual classification task. * = giving rise to the subtype of dysarthria: MD-FlSp; ** = giving rise to the subtype of dysarthria: HD. MD-FlSp = mixed dysarthria, combining flaccid and spastic dysarthria; HD = hypokinetic dysarthria associated with Parkinson's disease; AoS = apraxia; ALS = amyotrophic lateral sclerosis; PD = Parkinson's disease.



(Question 3) and Model (iii): accuracy in the classification of dysarthria: MD-FlSp dysarthria versus HD for the subset of speech samples judged as dysarthric (Question 4). Model selection was performed by including in the model all fixed factors of interest and which are known to have a potential influence on the dependent variable (groups of speakers, groups of listeners, [speech] tasks, TotalDevScore MonPaGe centered, age, and gender) and random intercepts for speakers and listeners. Since the model did not converge, random effects have been suppressed at first, namely, listeners, and then less relevant fixed factors, namely, age and gender of the speakers, were removed hierarchically from the model. The fixed effects built into the model were finally groups of speakers, groups of listeners, (speech) tasks, and the MonPaGe total deviance score of the speaker (TotalDevScore_MonPaGe), for which values were centered and the random effect was "speaker." Since contrasts were explored by turning over the intercept variable of the model to target all relevant comparisons, the resultant multitesting bias was corrected using the Bonferroni method (Bonferroni, 1936). Therefore, the significance threshold was divided by the number of models.

Interrater reliability was assessed using the intraclass correlation coefficient (ICC). ICC estimates and their 95% confident intervals were computed based on the mean correct rating per question across the four speech tasks for each speaker and per listener, separately for each group (student SLPs and expert SLPs) for common speakers, reflecting Version I or Version II of the experiment ("type": average/multiple raters, k = 10), absolute agreement ("definition"), in a two-way model ("model"). For qualitative interpretation of the ICC values, guidelines by Koo and Li (2016) were applied. According to these recommendations, ICC values of less than .50 are indicative of poor reliability, values between .50 and .75 indicate moderate reliability, values between .75 and .90 indicate good reliability, and values greater than .90 indicate excellent reliability.

Results of Question 2 on the "perceptual features" were analyzed in relationship with the classification of each MSD speech sample in Questions 3 and 4. This means that only perceptual features associated with correctly classified speech samples were analyzed. The first analysis was run on the number of perceptual features identified by the listeners for each correctly classified speech sample in the subsequent Questions 3 and 4. Two lmer models were run, Models (iv) and (v), one for the first classification of MSD (AoS vs. dysarthria) and one for the classification of dysarthria (HD vs. MD-FlSp). The "number of perceptual features" was the dependent variable with the same fixed effects and random structure as for the previous analyses. Then, the distribution of the types of perceptual features used by listeners was analyzed relative to the groups of speakers, the groups of listeners, and the tasks using Pearson's chi-square test for homogeneity (Agresti & Gottard, 2007).

Results

Classification Accuracy on Pathological Versus Neurotypical Speakers

The overall classification accuracy across the two groups of listeners on Question 1 was 72%. As shown in Table 2, accuracy was higher for speakers with MSD than for neurotypical speakers; expert SLPs had slightly higher scores than student SLPs. Both results were confirmed by the statistical analyses as presented in Table 3. In addition, classification was better on the Text and Spontaneous speech tasks, relative to the DDK task and, to a lesser extent, relative to the Text + DDK task. There was also a main effect of the TotalDevScore of MonPaGe, with better classification for more severe MSDs (see Table 3).

Classification Accuracy on AoS Versus Dysarthria

The classification accuracy across the two groups of MSD listeners on Question 3 was 74%. As shown in Table 2, accuracy was higher for speakers with dysarthria than with AoS, and expert SLPs had higher scores than student SLPs. Both results were confirmed by the statistical analyses as presented in Table 4. In addition, classification was better on the Text and Spontaneous speech tasks than on the DDK task. The accuracy for the Text + DDK task was also significantly different from the other speech tasks: higher than the DDK task, lower than the Spontaneous speech tasks (see Table 4).

Classification Accuracy on MD-FlSp Versus HD

The overall classification accuracy on Question 4 was 68%. As shown in Table 2, accuracy was higher for speakers with HD than with speakers with MD-FlSp, and expert SLPs had higher scores than student SLPs. Both results were confirmed by statistical analyses as presented in Table 5. In addition, classification was better on the Text, Spontaneous, and Text + DDK tasks than on the DDK task. There was also a main effect of the TotalDevScore of MonPaGe with better classification for more severe MSD (see Table 5).

Error Analysis

The classification matrix of all misclassified speech samples for each group of speakers and classification question is presented in Table 6. It can be observed that

Table 2. Mean accuracy rates (%) of the perceptual classification task for the three dichotomic rating questions (Questions 1, 3, and 4), with detailed results for groups of speakers, groups of listeners, and speech tasks.

		Accuracy (%)		
Question 1		**Pathological speech samples**	**Neurotypical speech samples**	**Total**
Total judgments		85.63	57.11	71.61
Groups of listeners	Expert SLP	84.00	61.47	72.92
	Student SLP	87.25	52.76	70.30
Speech tasks	Text + DDK	87.00	52.76	70.17
	Spontaneous	86.83	64.14	75.68
	DDK	84.17	44.66	64.75
	Text	84.50	66.90	75.85
Question 3		**AoS speech**	**Dysarthria speech samples**	**Total**
Total judgments		58.98	83.24	74.19
Group of listeners	Expert SLP	69.09	87.64	80.56
	Student SLP	48.20	78.89	67.71
Speech tasks	Text + DDK	56.12	80.67	71.46
	Spontaneous	74.24	83.54	80.00
	DDK	41.34	81.82	67.27
	Text	63.01	87.41	78.21
Question 4		**MD-FlSp speech**	**HD speech samples**	**Total**
Total judgments		60.12	71.53	67.85
Groups of listeners	Expert SLP	64.41	75.27	71.74
	Student SLP	55.35	67.46	63.58
Speech tasks	Text + DDK	64.37	74.43	71.10
	Spontaneous	59.55	77.09	71.27
	DDK	50.00	61.20	57.85
	Text	65.85	73.81	71.20

Note. SLP = speech-language pathologist; DDK = diadochokinetic; AoS = apraxia of speech; MD-FlSp = mixed dysarthria, combining flaccid and spastic dysarthria; HD = hypokinetic dysarthria associated with Parkinson's disease.

the speech samples of neurotypical speakers were most often misclassified as belonging to the class of speakers with HD or speakers with AoS. The speech samples of speakers with AoS were confused in a large proportion with speakers with MD-FlSp. The speech samples of speakers with MD-FlSp were mainly misclassified as those of speakers with HD.

Interrater Reliability

All results of the interrater reliability analyses are presented in Table 7. The overall mean interrater ICC across the three questions was good (.78). The ICCs were better for Question 1 than for Questions 3 and 4. Globally, expert SLPs had higher ICC values than student SLPs,

Table 3. Results of the generalized linear mixed models estimating differences between groups of listeners, groups of speakers, and speech tasks, regarding the accuracy of the classification of pathological and neurotypical speakers.

Model (i): glmer(accuracy ~ groups of speakers + groups of listeners + speech tasks + TotalDevScore MonPaGe centered + (1|speaker), data = data_Neurotypical_Pathological, family = "binomial")					
Comparisons		**β**	***SE***	***z***	***p***
Groups of speakers	Neurotypical–pathological	1.29	0.48	2.67	< .01*
Groups of listeners	Expert SLP–student SLP	0.19	0.08	–2.44	.014*
Speech tasks	Text + DDK–DDK	–0.37	0.11	–3.46	< .001*
	Text + DDK–Spontaneous	0.40	0.11	3.64	< .001*
	Text + DDK–Text	0.42	0.11	3.75	< .001*
	DDK–Spontaneous	0.78	0.11	7.02	< .001*
	DDK–Text	0.79	0.11	7.13	< .001*
	Spontaneous–Text	0.01	0.11	0.34	.908
TotalDevScore MonPaGe centered		0.26	0.08	3.12	< .01*

Note. Three models were necessary to perform all the comparisons for tasks; according to the Bonferroni correction, the significance threshold equals .017, indicated by the asterisk (*). *SE* = standard error; SLP = speech-language pathologist; DDK = diadochokinetic.

Table 4. Results of the generalized linear mixed models estimating differences between groups of speakers, groups of listeners, and speech tasks, regarding the accuracy of the classification of speakers with apraxia of speech (AoS) and speakers with dysarthria.

Model (ii): glmer(accuracy ~ groups of speakers + groups of listeners + speech tasks + TotalDevScore MonPaGe centered + (1|speaker), data = data_AoS_Dysarthric, family = "binomial")					
Comparisons		**β**	***SE***	***z***	***p***
Groups of speakers	AoS–dysarthria	1.38	0.34	4.07	< .001*
Groups of listeners	Expert SLP–student SLP	0.84	0.12	7.24	< .001*
Speech tasks	Text + DDK–DDK	–0.23	0.15	–1.53	.126
	Text + DDK–Spontaneous	0.56	0.16	3.47	< .01*
	Text + DDK–Text	0.34	0.16	2.06	.039
	DDK–Spontaneous	0.79	0.16	4.90	
	DDK–Text	0.57	0.16	3.47	< .001*
	Spontaneous–Text	–0.22	0.17	–1.28	.201
TotalDevScore MonPaGe centered		0.02	0.05	0.50	.614

Note. Three models were necessary to perform all the comparisons for tasks; according to the Bonferroni correction, the significance threshold equals .017, indicated by the asterisk (*). *SE* = standard error; SLP = speech-language pathologist; DDK = diadochokinetic.

except for Question 4. These results must also be evaluated considering a wide range in most of the analyses, notably for Questions 3 and 4, which considerably lowers the qualitative interpretation of the values.

Perceptual Features

The average number of perceptual features used by listeners was around 3, as shown in Table 8. As confirmed by statistical analysis, no difference was found between speech samples correctly classified as AoS or dysarthria as presented in Table 9 and between speech samples correctly classified as HD dysarthria or MD-FlSp dysarthria as shown in Table 10.

The main effect of centered TotalDevScore of MonPaGe reached significance for the two classifications with the number of used perceptual features being larger for more severe MSD as presented in Tables 9 and 10. The main effect of the groups of listeners was revealed only for the classification of speech samples as AoS and dysarthria: The average number of perceptual features given by expert SLPs was significantly higher than the student SLPs (see Table 9).

The distribution of perceptual features was significantly different between the classification of speech sample as speakers with AoS versus speakers with dysarthria, $\chi^2(7,\ N = 1{,}501) = 259.62,\ p < .001$. The three most frequent perceptual signs for AoS were "speech rate/fluency," "articulation," and "naturalness of speech." For the speakers with dysarthria, they were "voice quality," "speech rate/fluency," and "articulation" (see Table 8). The difference between the distributions of perceptual features used by expert SLPs and student SLPs reaches significance, $\chi^2(7,\ N = 1{,}501) = 15.57,\ p = .029$. The analyses also revealed a significant difference between tasks, $\chi^2(14,\ N = 1{,}501) = 36.36,\ p < .001$, linked to differences between

Table 5. Results of the generalized linear mixed models estimating differences between groups of speakers, groups of listeners, and speech tasks, regarding the accuracy of the classification of speakers with MD-FlSp and speakers with HD.

Model (iii): glmer(accuracy ~ groups of speakers + groups of listeners + speech tasks + TotalDevScore MonPaGe centered + (1|speaker), data = data_MD-FlSp_HD, family = "binomial")					
Comparisons		**β**	***SE***	***z***	***p***
Groups of speakers	MD-FlSp–HD	1.13	0.43	2.62	< .01*
Groups of listeners	Expert SLP–student SLP	0.43	0.15	2.89	< .01*
Speech tasks	Text + DDK–DDK	–0.68	0.20	–3.34	< .001*
	Text + DDK–Spontaneous	–0.02	0.21	–0.09	.931
	Text + DDK–Text	–0.05	0.21	–0.27	.791
	DDK–Spontaneous	0.66	0.20	3.25	< .01*
	DDK–Text	0.62	0.21	3.03	< .01*
	Spontaneous–Text	–0.04	0.21	–0.18	.857
TotalDevScore MonPaGe centered		0.16	0.06	2.59	< .01*

Note. Three models were necessary to perform all the comparisons for tasks, according to the Bonferroni correction, the significance threshold equals 0.017, indicated by the asterisk (*). MD-FlSp = mixed dysarthria, combining flaccid and spastic dysarthria; HD = hypokinetic dysarthria associated with Parkinson's disease; *SE* = standard error; SLP = speech-language pathologist; DDK = diadochokinetic.

Table 6. Classification matrix: percentages of misclassified speech samples for each group of speakers and classification question.

Variable	Neurotypical speech samples (n = 2,320) % (n misclassified)	AoS speech sample (n = 800) % (n misclassified)	HD speech samples (n = 1,120) % (n misclassified)	MD-FlSp speech samples (n = 480) % (n misclassified)
Judged as neurotypical (Question 1)	—	3 (24)	23.04 (258)	13.33 (64)
Judged as AoS (Question 3)	16.77 (389)	—	13.93 (156)	16.67 (80)
Judged as HD (Question 4)	17.20 (399)	16.5 (132)	—	27.92 (134)
Judged as MD-FlSp (Question 4)	8.92 (207)	21.75 (174)	17.95 (201)	—

Note. AoS = apraxia of speech; HD = hypokinetic dysarthria associated with Parkinson's disease; MD-FlSp = mixed dysarthria, combining flaccid and spastic dysarthria.

the task Text − DDK with the other tasks, Text + DDK − Spontaneous: $\chi^2(7, N = 1{,}501) = 30.11$, $p < .001$; Text + DDK − Text: $\chi^2(7, N = 1{,}501) = 18.22$, $p < .011$; Text − Spontaneous: $\chi^2(7, N = 1{,}501) = 7.06$, $p = .423$, especially for the feature "intelligibility," which rate was lowest for Text − DDK task, and the feature "articulation," which rate was highest for Text + DDK task (see Table 8).

The distribution of perceptual features was significantly different between speakers with MD-FlSp and speakers with HD, $\chi^2(7, N = 782) = 51.16$, $p < .001$. The three most frequent perceptual signs for speakers with MD-FlSp were "speech rate/fluency," "voice quality," and "articulation" in quite equally rates, and the three most frequent perceptual signs for speakers with HD were "voice quality" (which rate was higher), "speech rate/fluency," and "articulation" (see Table 8). The statistical analyses revealed no difference between groups of listeners, $\chi^2(7, N = 782) = 9.88$, $p = .196$, and also no difference between tasks, $\chi^2(14, N = 782) = 18.38$, $p = .190$.

Discussion

This study aimed to investigate the factors affecting the accuracy of SLP listeners to perceptually classify speech samples from neurotypical and pathological MSD speakers (AoS and dysarthria) in a forced-choice classification task in the framework of a diagnostic approach (Duffy, 2013). The potential effects of the following factors were assessed simultaneously in the same study: listener's expertise, speech task eliciting speech samples, and severity of MSD.

Before discussing the specific results on the factors of interest in further details, we will briefly compare the overall classification accuracy obtained in this study to results from previous studies. Then, the results will be discussed at each step of the classification procedure, and finally, the number and profiles of perceptual features used by listeners to correctly classify will be interpreted.

Substantial Accuracy Rates of Perceptual Classifications in a Diagnostic Approach

The overall rates of correct classification of neurotypical versus MSD speech samples were substantial (72%) but not perfect, and the interrater reliability was globally good. At each step of the classification procedure, significant differences were observed between groups of speakers and between groups of listeners. First, the MSD speakers were better identified than neurotypical speakers; second, higher accuracy rates were obtained for speakers with dysarthria than for speakers with AoS; and third, speakers with HD were better classified than speakers with MD-FlSp. Classification accuracy and interrater reliability were higher overall for expert SLPs than for student SLPs. Finally, an effect of the speech task was observed with better classification accuracy for spontaneous speech and text reading samples relative to DDK alone or Text + DDK.

Table 7. Interrater reliability (intraclass correlation [95% confidence interval])[a] by question for each group of speech-language pathologists (SLPs).

Question	Expert SLPs	Student SLPs	All listeners	Qualitative interpretation[b]
Q1	.91 [.82, .97]	.90 [.78, .96]	.91 [.80, .93]	Excellent
Q3	.80 [.50, .96]	.59 [.17, .91]	.70 [.35, .94]	Moderate
Q4	.73 [.35, .94]	.75 [.12, .99]	.74 [.18, .96]	Moderate
All questions	.81 [.56, .94]	.75 [.36, .95]	.78 [.44, .94]	Good

[a]Two-way model, multiple raters (average), absolute agreement. [b]Koo and Li (2016).

Table 8. Mean number and distributions (%) of perceptual features used by listeners (Question 2) for the classification of pathological speakers correctly judged according to groups of speakers, groups of listeners, and speech tasks.

			Perceptual features							
			Distribution (%)							
Speech samples correctly judged AoS and dysarthric speakers		***n* (*SD*)**	**Voice quality**	**Articulation**	**Nasal resonance**	**Prosody intonation**	**Speech rate fluency**	**Breathing**	**Intelligibility**	**Naturalness of speech**
Total of judgments		3.28 (1.64)	15.85	18.08	7.85	12.57	20.44	7.51	4.78	12.92
Groups of speakers	AoS speakers	3.53 (1.56)	9.74	21.18	3.75	13.39	25.02	5.24	5.79	15.88
	Dysarthric speakers	3.12 (1.67)	20.04	15.96	10.67	12.01	17.30	9.06	4.08	10.88
Groups of listeners	Expert SLP	3.38 (1.71)	15.24	18.96	8.07	13.07	20.01	7.17	3.96	13.53
	Student SLP	3.17 (1.56)	16.52	17.12	7.62	12.02	20.92	7.87	5.69	12.24
Speech tasks	Text + DDK	3.25 (1.57)	16.33	19.58	7.72	11.79	21.05	8.43	2.89	12.21
	Spontaneous	3.26 (1.65)	15.42	17.42	8.68	12.40	19.61	7.03	6.44	12.99
	Text	3.33 (1.71)	15.78	17.16	7.07	13.62	20.69	7.01	5.04	13.62

			Perceptual features							
			Distribution (%)							
Speech samples correctly judged MD-FlSp and HD speakers		***n* (*SD*)**	**Voice quality**	**Articulation**	**Nasal resonance**	**Prosody intonation**	**Speech rate fluency**	**Breathing**	**Intelligibility**	**Naturalness of speech**
Total of judgments		3.26 (1.69)	20.58	14.85	11.51	12.49	16.26	9.51	4.08	10.72
Groups of speakers	MD-FlSp speakers	3.92 (2.06)	16.21	16.01	11.46	11.96	16.90	8.60	6.72	12.15
	HD speakers	2.93 (1.37)	23.47	14.08	11.54	12.84	15.84	10.10	2.35	9.78
Groups of listeners	Expert SLP	3.32 (1.76)	21.32	15.15	11.70	12.20	15.58	9.26	3.23	11.56
	Student SLP	3.18 (1.61)	19.69	14.48	11.27	12.84	17.09	9.80	5.12	9.71
Speech tasks	Text + DDK	3.20 (1.65)	20.43	17.46	11.52	10.45	16.75	10.10	3.09	10.21
	Spontaneous	3.22 (1.68)	20.44	13.63	12.36	12.93	15.24	9.82	5.08	10.51
	Text	3.35 (1.75)	20.88	13.48	10.62	14.08	16.83	8.59	4.06	11.46

Note. AoS = apraxia of speech; SLP = speech-language pathologist; DDK = diadochokinetic; MD-FlSp = mixed dysarthria, combining flaccid and spastic dysarthria; HD = hypokinetic dysarthria associated with Parkinson's disease.

Table 9. Results of the linear mixed models estimating differences between groups of speakers, groups of listeners, and speech tasks, regarding the number of perceptual features used for the classification of speakers with apraxia of speech (AoS) and speakers with dysarthria judged as pathological.

Model (iv): lmer(Number of perceptual features ~ groups of speakers + groups of listeners + speech tasks + TotalDevScore MonPaGe centered + (1|speaker), data = data_perceptual_features_MSD, REML = FALSE)					
Group	**Sum of squares**	**Mean square**	**Num/den *df***	***F***	***p***
Groups of speakers (AoS–dysarthria)	0.19	0.187	1, 29.11	0.11	.75
Groups of listeners (expert SLP–student SLP)	15.57	15.57	1, 1472.02	8.85	< .01*
Speech tasks (Text + DDK–Spontaneous–Text)	0.002	0.001	2, 1474.03	0.0005	.999
TotalDevScore MonPaGe centered	47.31	47.31	1, 29.40	26.88	< .001*

Note. SLP = speech-language pathologist; DDK = diadochokinetic. *p value reached significance when less than .05.

All accuracy rates (ranging between 58% and 86%) of the present forced perceptual classification task were higher than in most of previous perceptual classification studies (Bunton et al., 2007; Fonville et al., 2008; Van der Graaff et al., 2009; Zyski & Weisiger, 1987). As mentioned in the introduction, accuracy rates depend on different parameters, notably on the experimental design, on the type of questions, and on the number of groups of speakers, explaining variability across studies. Indeed, depending on the type of dysarthria, some groups of speakers received very low accuracy in previous studies (1% for flaccid dysarthric speakers and 55% for hypokinetic dysarthric speakers in Zyski & Weisiger, 1987). The overall accuracy rates for all listener groups (between 19% and 56%) were considered as not reliable enough for clinical purposes by Zyski and Weisiger (1987). Fonville et al. (2008) found an average accuracy rate of 35% for judgments done on six groups of speakers, when Van der Graaff et al. (2009) found high accuracy rates for the identification of dysarthria that dropped to about 40% when it came to choose the subtype of dysarthria. The lower rates in the previous studies mentioned may also be explained by the number of speaker groups, ranging from four to seven. The overall better performance of the listeners in the present classification task and the good interrater reliability are likely linked to the design of the experiment based on a forced-choice classification task in a diagnostic approach including successive dichotomic questions guiding the step-by-step reasoning. However, even if higher than those of previous perceptual classification studies, the observed classification accuracies are still far from perfect, as an important proportion of the speech samples are misclassified.

The Expertise of Listeners Matters

Among the analyzed factors, the expertise of the listener (of the SLPs) affected classification accuracy at all classification steps. The better accuracy in expert SLP relative to student SLP on all questions observed in this study is in contradiction to null results reported in previous studies (Bunton et al., 2007; Lansford et al., 2016; Zyski & Weisiger, 1987). By contrast, the present findings are in line with the recent results of perceptual classification of speakers with Parkinson's disease by untrained and trained (student SLP) listeners in the study by Verkhodanova et al. (2021) but also with SLPs' self-assessment of their clinical progress in the survey by Simmons and Mayo (1997). Indeed, Simmons and Mayo reported that the classification of dysarthria subtypes was at least perceived to be progressively easier for those who had been working with MSDs longer. In fact, one of the most common responses to the question of what factors make differential diagnosis of dysarthria difficult to perform was a lack of experience with the specific disorder.

Table 10. Results of the linear mixed models estimating differences between groups of speakers, groups of listeners, and speech tasks, regarding the number of perceptual features used for the classification of speakers with MD-FlSp and speakers with HD judged as dysarthric.

Model (v): lmer(Number of perceptual features ~ groups of speakers + groups of listeners + speech tasks + TotalDevScore MonPaGe centered + (1|speaker), data = data_perceptual_features_Dysarthria, REML = FALSE)					
	Sum of squares	**Mean square**	**Num/den *df***	***F***	***p***
Groups of speakers (MD-FlSp–HD)	1.21	1.21	1, 19.45	0.71	.411
Groups of listeners (expert SLP–student SLP)	3.45	3.45	1, 763.16	2.01	.157
Speech tasks (Text + DDK–Spontaneous–Text)	0.30	0.15	2, 763.99	0.09	.918
TotalDevScore MonPaGe centered	25.55	25.55	1, 19.27	14.87	< .01*

Note. MD-FlSp = mixed dysarthria, combining flaccid and spastic dysarthria; HD = hypokinetic dysarthria associated with Parkinson's disease; SLP = speech-language pathologist; DDK = diadochokinetic. *p value reached significance when less than .05.

Here, the overall accuracy differences between the two groups of listeners are related to student SLPs misclassifying neurotypical speakers as pathological and speakers with an AoS as dysarthria. This was confirmed by the results of the analysis of the interrater reliability, higher for expert SLPs than for student SLPs.

The 20 student SLPs of this study had completed a course on MSDs during their first year of a master program in speech-language pathology and had received basic training in both dysarthria and AoS and their differential diagnosis. These hours of training appear to be insufficient to perform at the same level as expert SLPs trained in clinical settings (Kearns & Simmons, 1988). It should be noted, however, that even on the group of expert SLPs, the parameters of the number of year of practice and the average number of patients with MSDs followed since the beginning of SLPs' practice that may have an impact on the results were not controlled statistically in this study.

Effect of Speech Task

A specific issue addressed in this study that was not explored in most previous perceptual classification tasks was the impact of speech tasks (type and number) on the judgments done by SLPs. The Text Reading Aloud task and Spontaneous speech task clearly had the highest classification accuracy rates, a result that has also been reported in a recent study by Verkhodanova et al. (2021), who examined the effects of speech elicited in an interview and in a reading task. By contrast, speech elicited with a DDK task did not seem to help classification neither in isolation nor concatenated with a continuous task, as it actually dropped classification accuracy by 10%–20%.

The observation that the Spontaneous and Text speech tasks were found to be the most sensitive tasks to detect pathological speech and to distinguish between MSDs indicates that they allow to most fully express the profile of MSDs, even with a short sample. Such continuous speech tasks probably contain sufficient information on prosody, speech rate, articulatory precision, voice quality, and so forth, which is the reason why they are also recommended for perceptual analyses of dysarthria subtypes and rating of the global features of "intelligibility" or "naturalness of speech" in clinical practice (Auzou & Rolland-Monnoury, 2006).

It is surprising that speech elicited with DDK tasks did not contribute to the decision of the listeners in each classification question of this experiment, as DDK has been shown to help the differential diagnosis between neurotypical speakers and speakers with MD-FlSp with ALS (Rong & Heidrick, 2021) and between AoS and dysarthria (Duffy, 2013; Jonkers et al., 2017; Ogar et al., 2006). However, for this latter diagnosis, the main contribution is often linked to the difference in performance between sequential motion rate DDK and AMR DDK based on quantitative analyses on speech rate (Duffy, 2013; Kent et al., 2022; Strand et al., 2014). Here, the low contribution of DDK to correct classification may therefore be related, on the one hand, to the fact that it only included AMRs' sequences and, on the other hand, to the fact that the diagnostic value of such task probably relies on objective/acoustic measures of speech rate. Finally, the reduced contribution of DDK to correct classification is in line with Weismer's view of DDK tasks (2006) considering that the "oromotor, nonverbal tasks" as DDK (also called "pseudospeech tasks," "quasispeech tasks," "paraspeech tasks," "speech tasks," "nonspeech oral tasks," "speechlike tasks," or else categorized in "[speechlike] maximum performance tests"; Bunton, 2008; Kent, 2015; Kent et al., 1987; Ziegler et al., 2019) may not be useful in clinical diagnosis, not representative and informative on the speech production to the same extent as conversational/spontaneous speech, because of a specific sensitivity of the motor control processes to the task. This viewpoint considering DDK task as speechlike task is also shared by Kent et al. (1987) and Ziegler (2002, 2003) and has been proven empirically by Ziegler et al. (2019), who showed that DDK and other maximum performance tests did not contribute to the diagnosis of dysarthria. The results for the Text + DDK in our experiment, achieving lower accuracy rates than Text speech task and Spontaneous speech task, would also support this view.

Differences Between Groups of Speakers

Speech and Voice Normality Representations Among SLPs

Surprisingly, on the first classification question, speech samples from MSD speakers were overall better identified than those from neurotypical speakers (86% vs. 58%). This result may be related to a bias toward classifying speech samples as pathological in forced-choice tasks. Nevertheless, it is in contrast to other results of previous perceptual classification studies that included a group of neurotypical speakers (Fonville et al., 2008; Van der Graaff et al., 2009), reporting results in the opposite direction (lower accuracy rates for pathological than for neurotypical speakers). However, as already discussed above, the overall classification accuracy rates of the other perceptual classification studies were much lower than in this study (between 19% and 56% in Zyski & Weisiger, 1987, around 35% in Fonville et al., 2008), which was also related to the number of speaker groups to be classified. Misclassification analyses showed that neurotypical speakers were here mainly confused with speakers with HD and speakers with AoS. For the speakers with HD, it could be related to the mild degree of severity of this group; for the speakers with AoS, the lack of expertise of student SLPs could, in large part, explain it.

The present findings on misclassification of neurotypical speakers in the context of higher overall classification accuracy further question the knowledge/representation of normality and variability of the speech and voice of neurotypical speakers by the listeners, with unstable internal standards (Bunton et al., 2007; Delvaux & Pillot-Loiseau, 2020). Here, the neurotypical speakers originated from different areas of French-speaking Switzerland and could have slightly different regional accents. Using a free-classification perceptual task of speakers with different regional American English accents, Clopper and Pisoni (2007) showed that the regional/linguistic experience of the listeners differentially affected perceptual similarity ratings of speakers. Besides regional accent, other factors may have contributed to the misclassification of speech samples from neurotypical speakers. Delvaux and Pillot-Loiseau (2020) showed that multiple other factors could affect voice and speech in neurotypical speakers, such as smoking habits (Gilbert & Weismer, 1974) or chronic diseases, associated medications, musical and theatre practice, profession, or idiosyncrasies such as glottal fry. These sources of voice and speech variations, possibly perceived as pathological signs, could be considered to have had a potential influence on the accuracy of the ratings. A closer check of the misclassified neurotypical speakers showed that two of them were among the oldest speakers of this group. The impact of aging on voice and speech, combined with interspeaker variability that becomes more important with age (Eichhorn et al., 2018; Fougeron, d'Alessandro, & Lancia, 2018; Fougeron et al., 2021; Ramig et al., 2001), could partly explain these results. This indicates that signs of aging in voice and speech can be confused with signs of neurological disorders, especially by student SLPs, also suggesting that they should be trained to recognize the characteristics of "neurotypicality" in aging voice and speech.

Uncertain Perceptual Diagnosis of AoS

The accuracy rates of the classification of speakers with dysarthria (83%) were higher than those of MSD speakers with AoS (59%). The error analysis showed that speakers with AoS were mostly confused with MD-FlSp. This could be partly related to some shared perceptual signs, such as slowed speech rate and the presence of distorted articulation (Darley et al., 1969a, 1969b, 1975; Duffy, 2013; McNeil et al., 2009; Molloy & Jagoe, 2019).

AoS is rarely pure, often mixed with aphasia and/or dysarthria (Duffy, 2013; McNeil et al., 2009), generating misclassifications and more variability in ratings, especially for severe patients. Here, only speakers with dominant AoS were included in this group; only two out of the 10 speakers with dominant AoS had associated UUMN dysarthria. The perceptual features of UUMN dysarthria in these two speakers, such as changes in voice quality (Duffy, 2013), could have led to a diagnosis of dysarthria. For these speakers, it is also possible that the most salient MSD has not been selected by the listeners. On the other hand, six of them had concomitant mild nonfluent aphasia. For these speakers, the main sign was latencies for lexical/phonological access in spontaneous speech, not easy to distinguish from additions of pauses associated with AoS (Ballard et al., 2016; Cunningham et al., 2016; McNeil et al., 2009).

The concomitant presentations of AoS with aphasia and/or dysarthria and their potential variability are likely to make the adherence to standard criteria challenging (Molloy & Jagoe, 2019). The lack of clear consensus for diagnostic criteria of AoS and its differential diagnosis with dysarthria (Allison et al., 2020; Molloy & Jagoe, 2019) are therefore coherent with the low accuracy rate in the identification of speech samples from speakers with AoS, as it does not allow decisions based on clear features. Altogether, this probably leads to less stable/ less defined internal representations of AoS features for SLPs, thus explaining the lower classification accuracy for AoS.

However, an additional issue in the classification of AoS versus dysarthria is the modality of the perceptual task, which was auditory only for this study. In some previous reliability studies including AoS (Jonkers et al., 2017; Mumby et al., 2007), listeners performed perceptual classifications from video recordings that received substantial interrater reliability. Recent reviews by Molloy and Jagoe (2019) and Allison et al. (2020) highlighted the importance of visual information such as the presence of "groping" in the diagnostic process of AoS (see also Bailey et al., 2019). This missing visual cue could partly explain the lower accuracy rates for this MSD type in our auditory-perceptual classification task.

Prototypical and Less Prototypical Dysarthria Subtypes of the Mayo Clinic Classification System

Finally, within the classification of the group of speakers with dysarthria, the "mixed" subtype of dysarthria combining spastic and flaccid components associated with ALS could partly explain the lower accuracy rate of classification of this group of speakers relative to HD. Zyski and Weisiger (1987) also reported that HD was consistently identified with greater accuracy than other subtypes of dysarthria ranging from 55% to 73% by all groups of listeners, and the group of speakers with MD-FlSp was the most difficult subtype to classify in the study by Fonville et al. (2008; 14% of accuracy). The latter argued that participants classified speech samples according to the component, flaccid or spastic, that were perceptually dominant, missing the mixed pattern. The “mixed” subtype appears to be more variable and less prototypical—partially confused with the other group of dysarthric speakers, namely, those with HD. The higher accuracy in classifying HD may be related to its specific and consistent pathognomonic perceptual features. Borrie et al. (2012), for instance, suggested that the most consistent and regular features are, as in hypokinetic dysarthria, the more “learnable and amenable to perceptual training” (p. 295).

The mild degree of severity of our group of speakers with HD, reflecting early stages of the disease, probably also played a role in our results by facilitating the judgments of listeners. Indeed, the HD associated with Parkinson’s disease is known to become more variable with changes of speech and voice over the course of the disease at later stages (Forrest et al., 1989; Ho et al., 1998; Moya-Galé & Levy, 2019; Skodda et al., 2013). As shown by the error analysis, the accuracy rates of classification of speech samples from speakers with HD have, however, been lowered, as they were confused in a larger proportion with neurotypical speakers. In fact, four of them had very mild dysarthria and were probably undetected by listeners. They were indeed among the 10 worst misclassified speakers.

As already mentioned for speakers with AoS, visual cues are also lacking for the dysarthric speakers in this experiment. Indeed, crucial information is available in the context of the clinical physical examination, such as, for example, the reduced range of orofacial movements specific to speakers with HD or the possible presence of lingual fasciculations in speakers with MD-FlSp, reflecting lower motor neuron involvement. More broadly, in clinical contexts, the auditory-perceptual approach is completed with the case history of the patients (and possibly with neuroimaging data) and contributing to the differential diagnosis. This missing information should be considered as a limit of auditory-perceptual classification relative to clinical practice and could also explain why some speakers are misclassified in this study.

Overall, another potential source of influence in the misclassification of the speakers with MSDs could be the effects of current or past speech therapy. This parameter was not controlled for in this study. Through the impact of possible compensatory strategies, it could have modified the perceptual characteristics of the speech of MSD speakers, such as those of articulatory accuracy or speech rate (Martens et al., 2015; Mendoza Ramos et al., 2021).

Main Perceptual Features Characterizing MSD Subtypes in Line With Theorical Data

All the listeners of this study used, on average, three perceptual features out of the eight features proposed in the multiple-choice questions to classify MSD types and dysarthria subtypes. The limited set of features was included to take into account the observations by Kent (1996) that multiple dimensions, in addition to requiring more time and being correlated, could result in more erroneous answers.

Although the same set of eight features was proposed for all speech samples, the distribution of perceptual features identified by SLPs for speakers with AoS and dysarthria is significantly different. Qualitatively, the three most frequent perceptual features associated with correct classification of speech samples from speakers with AoS are “speech rate/fluency” (20%), “articulation” (18%), and “naturalness of speech” (13%), which is congruent with the literature on AoS. In fact, the two main reported perceptual symptoms of AoS are “distorted articulation” and “slow speech rate due to lengthened intersegment durations and segments,” “syllabification,” and “additions of pauses” (Ballard et al., 2016; Cunningham et al., 2016; McNeil et al., 1997). In contrast, the rates of the perceptual features of “breathing” and “nasal resonance,” respectively 9% and 11%, even if low, are higher for the correct classification of dysarthric speech samples.

Focusing on the correct classification of the two subgroups of speakers with dysarthria, the three most frequent perceptual signs selected by SLPs are “voice quality,” “speech rate/fluency,” and “articulation.” The “voice quality” feature is in line with the potentially impaired laryngeal function/level in dysarthria (Darley et al., 1969a, 1969b, 1975). Here, the profiles of speakers with MD-FlSp and HD differ, the “voice quality” being the most frequent sign used to characterize HD, followed by “speech rate/fluency” and, at last, “articulation.” They also reflected the pathognomonic perceptual features described for hypokinetic dysarthria in HD (Darley et al., 1969a, 1969b, 1975; Duffy, 2013; Ho et al., 1998; Moya-Galé & Levy, 2019), with the main clusters and perceptual features being reduced loudness (hypophonia), less variability of pitch, hoarse and breathy voice, momentary rushes of speech, variable speech rate, imprecise articulation, and dysfluent speech production (Skodda et al., 2013). For MD-FlSp, these three most frequent perceptual features received equivalent rates (“speech rate/fluency,” “voice quality,” and “articulation”) and are consistent with the main perceptual signs reported for the mixed flaccid–spastic dysarthria by Darley et al. (1969a, 1969b, 1975), Duffy (2013), or Tomik and Guiloff (2010), namely, strained, breathy, and hoarse voice and excessive prolonged speech segments resulting in a slow speech rate, imprecise consonants, and hypernasality.

The observation that the number of features identified was larger for more severe MSD is consistent with the occurrence of other perceptual features with the severity of the MSD, for both HD at later stages of the neurodegenerative disease (Ho et al., 1998; Skodda et al., 2013) and MD-FlSp, presenting in its course more marked features of each component, flaccid and spastic (Tomik & Guiloff, 2010; Yunusova et al., 2019).

Conclusions and Future Directions

The present forced-choice perceptual classification study of speakers with AoS and dysarthria and neurotypical speakers achieved substantial overall accuracy rates, but with important variations depending on group of speakers, listener's expertise, and speech task. The overall classification accuracy (72%) is far from perfect, but nevertheless much higher than in previous perceptual classification task. While extrapolating the findings is limited by the presence and choice between two dysarthria types, this could lead to a recommendation for the use of a diagnostic approach for the perceptual classification of MSDs for SLPs, who are then guided step-by-step in their reasoning for the differential diagnosis. It indicates that, at least in expert SLPs, the perceptual expertise largely contributes to the diagnosis, yet leaving a part of incertitude, which should be repeated on a more proportionate and larger sample size and complemented with alternative approaches and additional information.

In future studies, it may be informative to focus on the comparison of correctly classified with misclassified speech samples with fine-grained acoustic analyses to better understand the reasons of perceptual misclassification. Current technological means allow for the use of combined acoustic, physiological, or kinematic measures in clinical settings to supplement perceptual analyses (Duffy, 2016). The support of automatic classification could also be a promising tool for clinicians by providing diagnostic markers to better distinguish AoS from dysarthria (Kodrasi et al., 2020a, 2020b) and to detect mild and very mild speech and voice changes in MSDs (Tracy et al., 2020).

Author Contributions

Michaela Pernon: Conceptualization (Equal), Data curation (Lead), Formal Analysis (Lead), Investigation (Lead), Methodology (Equal), Resources (Lead), Software (Lead), Visualization (Lead), Writing – original draft (Lead), Writing – review & editing (Equal). **Frédéric Assal:** Conceptualization (Supporting), Funding acquisition (Lead), Investigation (Supporting), Project administration (Lead), Supervision (Lead), Validation (Lead), Writing – review & editing (Supporting). **Ina Kodrasi:** Formal Analysis (Supporting), Methodology (Supporting), Writing – review & editing (Supporting). **Marina Laganaro:** Conceptualization (Equal), Data curation (Supporting), Formal Analysis (Supporting), Funding acquisition (Lead), Investigation (Supporting), Methodology (Equal), Project administration (Lead), Resources (Supporting), Software (Supporting), Supervision (Lead), Validation (Lead), Visualization (Supporting), Writing – original draft (Supporting), Writing – review & editing (Equal).

Acknowledgments

This research was supported by Swiss National Science Foundation Grant CRSII5_173711 for the Sinergia "MoSpeeDi" project. The authors gratefully acknowledge their colleagues from the MonPaGe and MoSpeeDi projects, more particularly Cécile Fougeron, Maryll Fournet, Nathalie Lévêque, Véronique Delvaux, Rolland Trouville, Daria D'Alessandro, Grégoire Python, Monica Lancheros, Marion Bourqui, and Pauline Pellet Cheneval. The authors also would like to thank all the patients, clinicians, and graduate students who participated in this study, as well as Eric Ménétré and Gilles Reymond for their helpful comments and advice.

References

Agresti, A., & Gottard, A. (2007). Independence in multi-way contingency tables: S. N. Roy's breakthroughs and later developments. *Journal of Statistical Planning and Inference, 137*(11), 3216–3226. https://doi.org/10.1016/j.jspi.2007.03.006

Allison, K. M., Cordella, C., Iuzzini-Seigel, J., & Green, J. R. (2020). Differential diagnosis of apraxia of speech in children and adults: A scoping review. *Journal of Speech, Language, and Hearing Research, 63*(9), 2952–2994. https://doi.org/10.1044/2020_JSLHR-20-00061

Auzou, P., & Rolland-Monnoury, V. (2006). *BECD: Batterie d'Evaluation Clinique de la Dysarthrie* [Clinical Assessment Test for Dysarthria]. Ortho Édition.

Bailey, K., Kearney, E., Roberts, E., & Yunusova, Y. (2019). Towards movement-based outcome measures for apraxia of speech: A systematic review. *Aphasiology, 33*(8), 943–969. https://doi.org/10.1080/02687038.2018.1510111

Ballard, K. J., Azizi, L., Duffy, J. R., McNeil, M. R., Halaki, M., O'Dwyer, N., Layfield, C., Scholl, D. I., Vogel, A. P., & Robin, D. A. (2016). A predictive model for diagnosing stroke-related apraxia of speech. *Neuropsychologia, 81,* 129–139. https://doi.org/10.1016/j.neuropsychologia.2015.12.010

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Blumstein, S. E. (1990). Phonological deficits in aphasia: Theoretical perspectives. In A. Caramazza (Ed.), *Cognitive neuropsychology and neurolinguistics: Advances in models of cognitive function and impairment* (pp. 33–53). Erlbaum.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità [Statistical class theory and calculus of probability]. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8,* 3–62.

Borrie, S. A., McAuliffe, M. J., & Liss, J. M. (2012). Perceptual learning of dysarthric speech: A review of experimental studies. *Journal of Speech, Language, and Hearing Research, 55*(1), 290–305. https://doi.org/10.1044/1092-4388(2011/10-0349)

Brown, A., & Docherty, G. J. (1995). Phonetic variation in dysarthric speech as a function of sampling task. *International Journal of Language & Communication Disorders, 30*(1), 17–35. https://doi.org/10.3109/13682829509031320

Bunton, K. (2008). Speech versus nonspeech: Different tasks, different neural organization. *Seminars in Speech and Language, 29*(4), 267–275. https://doi.org/10.1055/s-0028-1103390

Bunton, K., Kent, R. D., Duffy, J. R., Rosenbek, J. C., & Kent, J. F. (2007). Listener agreement for auditory-perceptual ratings of dysarthria. *Journal of Speech, Language, and Hearing Research, 50*(6), 1481–1495. https://doi.org/10.1044/1092-4388(2007/102)

Chicherio, C., Genoud-Prachex, T., Assal, F., & Laganaro, M. (2019). *E-GeBAS: Electronic Geneva Bedside Aphasia Scale* [Computer program]. https://www.unige.ch/fapse/logotools/fr/adultes

Clopper, C. G. (2008). Auditory free classification: Methods and analysis. *Behavior Research Methods, 40*(2), 575–581. https://doi.org/10.3758/BRM.40.2.575

Clopper, C. G., & Pisoni, D. B. (2007). Free classification of regional dialects of American English. *Journal of Phonetics, 35*(3), 421–438. https://doi.org/10.1016/j.wocn.2006.06.001

Code, C. (1998). Models, theories and heuristics in apraxia of speech. *Clinical Linguistics & Phonetics, 12*(1), 47–65. https://doi.org/10.3109/02699209808985212

Cunningham, K. T., Haley, K. L., & Jacks, A. (2016). Speech sound distortions in aphasia and apraxia of speech: Reliability and diagnostic significance. *Aphasiology, 30*(4), 396–413. https://doi.org/10.1080/02687038.2015.1065470

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969a). Clusters of deviant speech dimensions in the dysarthrias. *Journal of Speech and Hearing Research, 12*(3), 462–496. https://doi.org/10.1044/jshr.1203.462

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969b). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research, 12*(2), 246–269. https://doi.org/10.1044/jshr.1202.246

Darley, F. L., Aronson, A. E., & Brown, J. R. (1975). *Motor speech disorders*. W. B. Saunders.

Delvaux, V., & Pillot-Loiseau, C. (2020). Perceptual judgment of voice quality in nondysphonic French speakers: Effect of task-, speaker- and listener-related variables. *Journal of Voice, 34*(5), 682–693. https://doi.org/10.1016/j.jvoice.2019.02.013

Duffy, J. R. (2013). *Motor speech disorders: Substrates, differential diagnosis, and management* (3rd ed.). Elsevier.

Duffy, J. R. (2016). Motor speech disorders: Where will we be in 10 years? *Seminars in Speech and Language, 37*(3), 219–224. https://doi.org/10.1055/s-0036-1584154

Duffy, J. R., Utianski, R. L., & Josephs, K. A. (2021). Primary progressive apraxia of speech: From recognition to diagnosis and care. *Aphasiology, 35*(4), 560–591. https://doi.org/10.1080/02687038.2020.1787732

Duncan, E. S., Donovan, N. J., & Sajjadi, S. A. (2020). Clinical assessment of characteristics of apraxia of speech in primary progressive aphasia. *American Journal of Speech-Language Pathology, 29*(1S), 485–497. https://doi.org/10.1044/2019_AJSLP-CAC48-18-0225

Eichhorn, J. T., Kent, R. D., Austin, D., & Vorperian, H. K. (2018). Effects of aging on vocal fundamental frequency and vowel formants in men and women. *Journal of Voice, 32*(5), 644.e1–644.e9. https://doi.org/10.1016/j.jvoice.2017.08.003

Fonville, S., van der Worp, H. B., Maat, P., Aldenhoven, M., Algra, A., & van Gijn, J. (2008). Accuracy and inter-observer variation in the classification of dysarthria from speech recordings. *Journal of Neurology, 255*(10), 1545–1548. https://doi.org/10.1007/s00415-008-0978-4

Forrest, K., Weismer, G., & Turner, G. S. (1989). Kinematic, acoustic, and perceptual analyses of connected speech produced by Parkinsonian and normal geriatric adults. *The Journal of the Acoustical Society of America, 85*(6), 2608–2622. https://doi.org/10.1121/1.397755

Fougeron, C., d'Alessandro, D., & Lancia, L. (2018). Reduced coarticulation and aging. *The Journal of the Acoustical Society of America, 144*(3), 1905–1905. https://doi.org/10.1121/1.5068348

Fougeron, C., Delvaux, V., Ménard, L., & Laganaro, M. (2018). The MonPaGe_HA database for the documentation of spoken French throughout adulthood. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 4301–4306). European Language Resources Association.

Fougeron, C., Delvaux, V., Pernon, M., Lévêque, N., Borel, S., Pellet, P., Bagou, O., Trouville, R., Ménard, L., Catalano Chiuvé, S., Lopez, U., Kocjancic-Antolik, T., & Laganaro, M. (2016). MonPaGe: Un protocole informatisé d'évaluation de la parole pathologique en langue française [MonPaGe: A computerized protocol for the evaluation of pathological speech in French]. In N. Joyeux & S. Topouzkhanian (Eds.), *Actes du colloque UNADREO «Orthophonie et technologies innovantes»* (pp. 291–303). Ortho Edition.

Fougeron, C., Guitard-Ivent, F., & Delvaux, V. (2021). Multidimensional variation in adult speech as a function of age. *Language, 6*(4), 176. https://doi.org/10.3390/languages6040176

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various coefficients of interrater reliability and agreement (Version 0.84)*. https://cran.r-project.org/web/packages/irr/index.html

Gilbert, H. R., & Weismer, G. G. (1974). The effects of smoking on the speaking fundamental frequency of adult women. *Journal of Psycholinguistic Research, 3*(3), 225–231. https://doi.org/10.1007/BF01069239

Guenther, F. H. (2016). *Neural control of speech*. MIT Press. https://doi.org/10.7551/mitpress/10471.001.0001

Ho, A. K., Iansek, R., Marigliani, C., Bradshaw, J. L., & Gates, S. (1998). Speech impairment in a large sample of patients with Parkinson's disease. *Behavioural Neurology, 11*(3), 131–137. https://doi.org/10.1155/1999/327643

Hybbinette, H., Östberg, P., & Schalling, E. (2021). Intra- and interjudge reliability of the apraxia of speech rating scale in early stroke patients. *Journal of Communication Disorders, 89,* 106076. https://doi.org/10.1016/j.jcomdis.2020.106076

Jonkers, R., Feiken, J., & Stuive, I. (2017). Diagnosing apraxia of speech on the basis of eight distinctive signs. *Canadian Journal of Speech-Language Pathology and Audiology, 41*(3), 303–319.

Josephs, K. A., Duffy, J. R., Strand, E. A., Machulda, M. M., Senjem, M. L., Lowe, V. J., Jack, C. R., & Whitwell, J. L. (2013). Syndromes dominated by apraxia of speech show distinct characteristics from agrammatic PPA. *Neurology, 81*(4), 337–345. https://doi.org/10.1212/WNL.0b013e31829c5ed5

Kearns, K. P., & Simmons, N. N. (1988). Interobserver reliability and perceptual ratings: More than meets the ear. *Journal of Speech and Hearing Research, 31*(1), 131–136. https://doi.org/10.1044/jshr.3101.131

Kempler, D., & Van Lancker, D. (2002). Effect of speech task on intelligibility in dysarthria: A case study of Parkinson's disease. *Brain and Language, 80*(3), 449–464. https://doi.org/10.1006/brln.2001.2602

Kent, R. D. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology, 5*(3), 7–23. https://doi.org/10.1044/1058-0360.0503.07

Kent, R. D. (2015). Nonspeech oral movements and oral motor disorders: A narrative review. *American Journal of Speech-Language Pathology, 24*(4), 763–789. https://doi.org/10.1044/2015_AJSLP-14-0179

Kent, R. D., & Kent, J. F. (2000). Task-based profiles of the dysarthrias. *Folia Phoniatrica et Logopaedica, 52*(1–3), 48–53. https://doi.org/10.1159/000021512

Kent, R. D., Kent, J. F., & Rosenbek, J. C. (1987). Maximum performance tests of speech production. *Journal of Speech and Hearing Disorders, 52*(4), 367–387. https://doi.org/10.1044/jshd.5204.367

Kent, R. D., Kent, J. F., Weismer, G., & Duffy, J. R. (2000). What dysarthrias can tell us about the neural control of speech. *Journal of Phonetics, 28*(3), 273–302. https://doi.org/10.1006/jpho.2000.0122

Kent, R. D., Kim, Y., & Chen, L. M. (2022). Oral and laryngeal diadochokinesis across the life span: A scoping review of methods, reference data, and clinical applications. *Journal of Speech, Language, and Hearing Research, 65*(2), 574–623.

Kim, Y., Kent, R. D., & Weismer, G. (2011). An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria. *Journal of Speech, Language, and Hearing Research, 54*(2), 417–429. https://doi.org/10.1044/1092-4388(2010/10-0020)

Kodrasi, I., Pernon, M., Laganaro, M., & Bourlard, H. (2020a). *Automatic and perceptual discrimination between dysarthria, apraxia of speech, and neurotypical speech*. arXiv preprint arXiv:2011.07542. https://doi.org/10.48550/arXiv.2011.07542

Kodrasi, I., Pernon, M., Laganaro, M., & Bourlard, H. (2020b). Automatic discrimination of apraxia of speech and dysarthria using a minimalistic set of handcrafted features. In *Proceedings of the Annual Conference of the International Speech Communication Association, Shanghai, China*.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTestPackage: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/JSS.V082.I13

Laganaro, M., Fougeron, C., Pernon, M., Levêque, N., Borel, S., Fournet, M., Catalano Chiuvé, S., Lopez, U., Trouville, R., Ménard, L., Burkhard, P. R., Assal, F., & Delvaux, V. (2021). Sensitivity and specificity of an acoustic and perceptual-based tool for assessing motor speech disorders in French: The MonPaGe screening protocol. *Clinical Linguistics & Phonetics, 35*(11), 1060–1075. https://doi.org/10.1080/02699206.2020.1865460

Lansford, K. L., Berisha, V., & Utianski, R. L. (2016). Modeling listener perception of speaker similarity in dysarthria. *The Journal of the Acoustical Society of America, 139*(6), EL209–EL215. https://doi.org/10.1121/1.4954384

Lansford, K. L., Liss, J. M., & Norton, R. E. (2014). Free-classification of perceptually similar speakers with dysarthria. *Journal of Speech, Language, and Hearing Research, 57*(6), 2051–2064. https://doi.org/10.1044/2014_JSLHR-S-13-0177

Martens, H., Van Nuffelen, G., Dekens, T., Hernández-Díaz Huici, M., Kairuz Hernández-Díaz, H. A., De Letter, M., & De Bodt, M. (2015). The effect of intensive speech rate and intonation therapy on intelligibility in Parkinson's disease. *Journal of Communication Disorders, 58,* 91–105. https://doi.org/10.1016/j.jcomdis.2015.10.004

McNeil, M. R., Robin, D. A., & Schmidt, R. A. (1997). Apraxia of speech: Definition, differentiation and treatment. In M. R. McNeil (Ed.), *Clinical management of sensorimotor speech disorders* (pp. 311–344). Thieme Medical Publishers.

McNeil, M. R., Robin, D. A., & Schmidt, R. A. (2009). Apraxia of speech. In M. R. McNeil (Ed.), *Clinical management of sensorimotor speech disorders* (pp. 249–268). Thieme.

Mendoza Ramos, V., Paulyn, C., Van den Steen, L., Hernandez-Diaz Huici, M. E., De Bodt, M., & Van Nuffelen, G. (2021). Effect of boost articulation therapy (BArT) on intelligibility in adults with dysarthria. *International Journal of Language & Communication Disorders, 56*(2), 271–282. https://doi.org/10.1111/1460-6984.12595

Molloy, J., & Jagoe, C. (2019). Use of diverse diagnostic criteria for acquired apraxia of speech: A scoping review. *International Journal of Language & Communication Disorders, 54*(6), 875–893. https://doi.org/10.1111/1460-6984.12494

Moya-Galé, G., & Levy, E. S. (2019). Parkinson's disease-associated dysarthria: Prevalence, impact and management strategies. *Research and Reviews in Parkinsonism, 9,* 9–16. https://doi.org/10.2147/JPRLS.S168090

Mumby, K., Bowen, A., & Hesketh, A. (2007). Apraxia of speech: How reliable are speech and language therapists' diagnoses? *Clinical Rehabilitation, 21*(8), 760–767. https://doi.org/10.1177/0269215507077285

Ogar, J., Willock, S., Baldo, J., Wilkins, D., Ludy, C., & Dronkers, N. (2006). Clinical and anatomical correlates of apraxia of speech. *Brain and Language, 97*(3), 343–350. https://doi.org/10.1016/j.bandl.2006.01.008

Pernon, M., Lévêque, N., Delvaux, V., Assal, F., Borel, S., Fougeron, C., Trouville, R., & Laganaro, M. (2020). MonPaGe, un outil de screening francophone informatisé d'évaluation perceptive et acoustique des troubles moteurs de la parole (dysarthries, apraxie de la parole) [MonPaGe, a French-speaking computerized screening tool for the perceptual and acoustic evaluation of motor speech disorders (dysarthria, apraxia of speech)]. *Rééducation orthophonique, 281,* 169–197.

Qualtrics. (2019). [Computer software]. https://www.qualtrics.com

Ramig, L. O., Gray, S., Baker, K., Corbin-Lewis, K., Buder, E., Luschei, E., Coon, H., & Smith, M. (2001). The aging voice: A review, treatment data and familial and genetic perspectives. *Folia Phoniatrica et Logopaedica, 53*(5), 252–265. https://doi.org/10.1159/000052680

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rong, P., & Heidrick, L. (2021). Spatiotemporal control of articulation during speech and speechlike tasks in amyotrophic lateral sclerosis. *American Journal of Speech-Language Pathology, 30*(3S), 1382–1399. https://doi.org/10.1044/2020_AJSLP-20-00136

Simmons, K. C., & Mayo, R. (1997). The use of the Mayo Clinic system for differential diagnosis of dysarthria. *Journal of Communication Disorders, 30*(2), 117–132. https://doi.org/10.1016/S0021-9924(96)00058-5

Skodda, S., Grönheit, W., Mancinelli, N., & Schlegel, U. (2013). Progression of voice and speech impairment in the course of Parkinson's disease: A longitudinal study. *Parkinson's Disease, 2013*. https://doi.org/10.1155/2013/389195

Strand, E. A., Duffy, J. R., Clark, H. M., & Josephs, K. (2014). The apraxia of speech rating scale: A tool for diagnosis and description of apraxia of speech. *Journal of Communication Disorders, 51,* 43–50. https://doi.org/10.1016/j.jcomdis.2014.06.008

Tomik, B., & Guiloff, R. J. (2010). Dysarthria in amyotrophic lateral sclerosis: A review. *Amyotrophic Lateral Sclerosis, 11*(1–2), 4–15. https://doi.org/10.3109/17482960802379004

Tracy, J. M., Özkanca, Y., Atkins, D. C., & Hosseini Ghomi, R. H. (2020). Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease. *Journal of Biomedical Informatics, 104,* 103362. https://doi.org/10.1016/j.jbi.2019.103362

Utianski, R. L., Duffy, J. R., Clark, H. M., Strand, E. A., Botha, H., Schwarz, C. G., Machulda, M. M., Senjem, M. L., Spychalla, A. J., Jack, C. R., Jr., Petersen, R. C., Lowe, V. J., Whitwell, J. L., & Josephs, K. A. (2018). Prosodic and phonetic subtypes of primary progressive apraxia of speech. *Brain and Language, 184,* 54–65. https://doi.org/org/10.1016/j.bandl.2018.06.004

Van der Graaff, M., Kuiper, T., Zwinderman, A., Van de Warrenburg, B., Poels, P., Offeringa, A., Van der Kooi, A., Speelman, H., & De Visser, M. (2009). Clinical identification of dysarthria types among neurologists, residents in neurology and speech therapists. *European Neurology, 61*(5), 295–300. https://doi.org/10.1159/000206855

Van Der Merwe, A. (2021). New perspectives on speech motor planning and programming in the context of the four-level model and its implications for understanding the pathophysiology underlying apraxia of speech and other motor speech disorders. *Aphasiology, 35*(4), 397–423. https://doi.org/10.1080/02687038.2020.1765306

Van Lancker Sidtis, D., Cameron, K., & Sidtis, J. J. (2012). Dramatic effects of speech task on motor and linguistic planning in severely dysfluent parkinsonian speech. *Clinical Linguistics & Phonetics, 26*(8), 695–711. https://doi.org/10.3109/02699206.2012.696307

Van Lancker Sidtis, D., Rogers, T., Godier, V., Tagliati, M., & Sidtis, J. J. (2010). Voice and fluency changes as a function of speech task and deep brain stimulation. *Journal of Speech, Language, and Hearing Research, 53*(5), 1167–1177. https://doi.org/10.1044/1092-4388(2010/09-0154)

Verkhodanova, V., Coler, M., Jonkers, R., & Lowie, W. (2021). How expertise and language familiarity influence perception of speech of people with Parkinson's disease. *Clinical Linguistics & Phonetics,* 1–18. https://doi.org/10.1080/02699206.2021.2003433

Wambaugh, J. L., Bailey, D. J., Mauszycki, S. C., & Bunker, L. D. (2019). Interrater reliability and concurrent validity for the Apraxia of Speech Rating Scale 3.0: Application with persons with acquired apraxia of speech and aphasia. *American Journal of Speech-Language Pathology, 28*(2S), 895–904. https://doi.org/10.1044/2018_AJSLP-MSC18-18-0099

Weismer, G. (2006). Philosophy of research in motor speech disorders. *Clinical Linguistics & Phonetics, 20*(5), 315–349. https://doi.org/10.1080/02699200400024806

Weismer, G., & Kim, Y.-J. (2010). Classification and taxonomy of motor speech disorders: What are the issues? In B. Maassen & P. H. H. M. van Lieshout (Eds.), *Speech motor control: New developments in basic and applied research* (pp. 229–241). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199235797.003.0013

Yunusova, Y., Plowman, E. K., Green, J. R., Barnett, C., & Bede, P. (2019). Clinical measures of bulbar dysfunction in ALS. *Frontiers in Neurology, 10,* 106. https://doi.org/10.3389/fneur.2019.00106

Zeplin, J., & Kent, R. D. (1996). Reliability of auditory-perceptual scaling of dysarthria. In D. Robin, K. Yorkson, & D. R. Buekelman (Eds.), *Disorders of motor speech: Recent advances in assessment, treatment, and clinical characterization* (pp. 145–154). Brookes.

Ziegler, W. (2002). Task-related factors in oral motor control: Speech and oral diadochokinesis in dysarthria and apraxia of speech. *Brain and Language, 80*(3), 556–575. https://doi.org/10.1006/brln.2001.2614

Ziegler, W. (2003). Review. *Aphasiology, 17*(1), 3–36. https://doi.org/10.1080/729254892

Ziegler, W. (2009). Modelling the architecture of phonetic plans: Evidence from apraxia of speech. *Language and Cognitive Processes, 24*(5), 631–661. https://doi.org/10.1080/01690960802327989

Ziegler, W., Schölderle, T., Brendel, B., Amsellem, J., & Staiger, A. (2019). Higher-faster-farther: Maximum performance tests in the assessment of neurogenic speech impairment. *Folia Phoniatrica et Logopaedica, 71*(5–6), 261–274. https://doi.org/10.1159/000495784

Zyski, B. J., & Weisiger, B. E. (1987). Identification of dysarthria types based on perceptual analysis. *Journal of Communication Disorders, 20*(5), 367–378. https://doi.org/10.1016/0021-9924(87)90025-6

Appendix A

Content of the Paragraphs of the Text Module of MonPaGe Screening Tool

* 1st paragraph: «Lundi, le chat, le loup et Papa vont à Bali. Les copains sont tout contents.»
Translation: "Monday, the cat, the wolf and Daddy go to Bali. The friends are very happy."
* 2d paragraph: «Mercredi, Papy dit: "Toi, le chat! Tu es doux, tu es chou, tu n'as pas de poux! Mais pas ce loup: il a une cape rouge et je n'aime pas ce gars-là!"»
Translation: "On Wednesday, Grandpa said: "You, the cat! You're sweet, you're cute, you don't have lice! But not that wolf: he has a red cape and I don't like that guy!""

Appendix B

Definitions of the Perceptual Criteria Given to the Judges During the Experiment

Voice quality: Overall impression of alterations in voice quality relating to one or more parameters (pitch, intensity, timbre), more or less marked.

Articulation: Overall impression of an alteration in articulation precision, more or less severe, affecting the production of consonants or vowels.

Nasal resonance: More or less marked disturbance of nasal resonance, which may be in the direction of hyponasality or hypernasality.

Prosody/intonation: More or less marked deficit of the suprasegmental characteristics of speech, which may be manifested by an inadequacy of the modulations of pitch and/or vocal intensity, diminished or excessive, by a difficulty in producing prosodic patterns (assertion, questioning, exclamation).

Speech rate/fluency: More or less marked alteration of speech rate and fluency, characterized by abnormally slow or fast speech rate, paroxysmal fluctuations and speech rate accelerations, initiation impairments, inappropriate or frequent pauses, and dysfluencies (repetitions, palilalia, prolongations, and blocks).

Breathing: Disturbance of breathing during speech production, more or less marked. Impairments in respiratory control may result in forced inhalations and exhalations that interrupt the flow of speech, inspiratory noises (breath and stridor) or shortened breath groups.

Naturalness of speech: Speech is perceived as abnormal due to the impairment of one or more parameters (articulation, vocal quality, speech fluency, nasal resonance, and breathing), which may be discrete to severe.