Automatic Intonation Recognition for the Prosodic Assessment of Language-Impaired Children

Fabien Ringeval, Julie Demouy, György Szaszák, Mohamed Chetouani, Laurence Robel, Jean Xavier, David Cohen, and Monique Plaza

***Abstract*—This study presents a preliminary investigation into the automatic assessment of language-impaired children's (LIC) prosodic skills in one *grammatical* aspect: sentence modalities. Three types of language impairments were studied: autism disorder (AD), pervasive developmental disorder-not otherwise specified (PDD-NOS), and specific language impairment (SLI). A control group of typically developing (TD) children that was both age and gender matched with LIC was used for the analysis. All of the children were asked to imitate sentences that provided different types of intonation (e.g., descending and rising contours). An automatic system was then used to assess LIC's prosodic skills by comparing the intonation recognition scores with those obtained by the control group. The results showed that all LIC have difficulties in reproducing intonation contours because they achieved significantly lower recognition scores than TD children on almost all studied intonations ($p < 0.05$). Regarding the "*Rising*" intonation, only SLI children had high recognition scores similar to TD children, which suggests a more pronounced pragmatic impairment in AD and PDD-NOS children. The automatic approach used in this study to assess LIC's prosodic skills confirms the clinical descriptions of the subjects' communication impairments.**

***Index Terms*—Automatic intonation recognition, prosodic skills assessment, social communication impairments.**

I. Introduction

SPEECH is a complex waveform that conveys a lot of useful information for interpersonal communication and human–machine interaction. Indeed, a speaker not only produces a raw message composed of textual information when he or she speaks but also transmits a wide set of information that modulates and enhances the meaning of the produced message [1]. This additional information is conveyed in speech by prosody and can be directly (e.g., through sentence modality or word focus) or indirectly (e.g., idiosyncrasy) linked to the message. To properly communicate, knowledge of the pre-established codes that are being used is also required. Indeed, the richness of social interactions shared by two speakers through speech strongly depends on their ability to use a full range of pre-established codes. These codes link acoustic speech realization and both linguistic- and social-related meanings. The acquisition and correct use of such codes in speech thus play an essential role in the inter-subjective development and social interaction abilities of children. This crucial step of speech acquisition relies on cognition and is supposed to be functional in the early stages of a child's life [2].

A. Prosody

Prosody is defined as the supra-segmental properties of the speech signal that modulate and enhance its meaning. It aims to construct discourse through expressive language at several communication levels, i.e., *grammatical*, *pragmatic*, and *affective* prosody [3]. *Grammatical* prosody is used to signal syntactic information within sentences [4]. Stress is used to signal, for example, whether a token is being used as a noun (***con**vict*) or a verb (*con**vict***). Pitch contours signal the ends of utterances and denote whether they are, for example, questions (rising pitch) or statements (falling pitch). *Pragmatic* prosody conveys the speaker's intentions or the hierarchy of information within the utterance [3] and results in optional changes in the way an utterance is expressed [5]. Thus, it carries social information beyond that conveyed by the syntax of the sentence. *Affective* prosody serves a more global function than those served by the prior two forms. It conveys a speaker's general state of feeling [6] and includes associated changes in register when talking to different listeners (e.g., peers, young children or people of higher social status) [3].

Because prosodic deficits contribute to language, communication and social interaction disorders and lead to social isolation, the atypical prosody in individuals with communication disorders became a research topic. It appears that prosodic awareness underpins language skills, and a deficiency in prosody may affect both language development and social interaction.

Manuscript received April 17, 2010; revised August 15, 2010 and October 15, 2010; accepted October 18, 2010. Date of publication October 28, 2010; date of current version nulldate. This work was supported in part by the French Ministry of Research and Superior Teaching and by the Hubert–Curien partnership between France (EGIDE www.egide.asso.fr) and Hungary (TéT, OMFB-00364/2008). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Renato De Mori.

F. Ringeval and M. Chetouani are with the Institute of Intelligent Systems and Robotics, University Pierre and Marie Curie, 75005 Paris, France (e-mail: fabien.ringeval@free.fr; mohamed.chetouani@upmc.fr).

J. Demouy and J. Xavier are with the Department of Child and Adolescent Psychiatry, Hôpital de la Pitié-Salpêtrière, University Pierre and Marie Curie, 75013 Paris, France (e-mail: julie.demouy@yahoo.fr; jean.xavier@psl.aphp.fr).

G. Szaszák is with the Department for Telecommunication and Media Informatics, Budapest University of Technology and Economics, H-1117 Budapest, Hungary (e-mail: szaszak@tmit.bme.hu).

L. Robel is with the Department of Child and Adolescent Psychiatry, Hôpital Necker-Enfants Malades, 75015 Paris, France (e-mail: laurence.robel@free.fr).

D. Cohen and M. Plaza are with the Department of Child and Adolescent Psychiatry, Hôpital de la Pitié-Salpêtrière, University Pierre and Marie Curie, 75013 Paris, France, and also with the Institute of Intelligent Systems and Robotics, University Pierre and Marie Curie, 75005 Paris, France (e-mail: david.cohen@psl.aphp.fr; monique.plaza@psl.aphp.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2090147

B. Prosodic Disorders in Language-Impaired Children

Most children presenting speech impairments have limited social interactions, which contributes to social isolation. A developmental language disorder may be secondary to hearing loss or acquired brain injury and may occur without specific cause [7]. In this case, international classifications distinguish specific language impairment (SLI), on one hand, and language impairment symptomatic of a developmental disorder (e.g., Pervasive Developmental Disorders-PDD) on the other. The former can affect both expressive and receptive language and is defined as a "pure" language impairment [8]. The latter, PDD, is characterized by severe deficits and pervasive impairment in several areas of development such as reciprocal social interactions, communication skills and stereotyped behaviors, interests, and activities [9]. Three main disorders have been described [7]: 1) autistic disorder (AD), which manifests as early onset language impairment quite similar to that of SLI [10] and symptoms in all areas that characterize PDD; 2) Asperger's Syndrome, which does not evince language delay; and 3) pervasive developmental disorder-not otherwise specified (PDD-NOS), which is characterized by social, communicative and/or stereotypic impairments that are less severe than in AD and appear later in life.

Language-impaired children (LIC) may also show prosodic disorders: AD children often sound differently than their peers, which adds a barrier to social integration [11]. Furthermore, the prosodic communication barrier is often persistent while other language skills improve [12]. Such disorders notably affect acoustic features such as pitch, loudness, voice quality, and speech timing (i.e., rhythm).

The characteristics of the described LIC prosodic disorders are various and seem to be connected with the type of language impairment.

Specific Language Impairment: Intonation has been studied very little in children with SLI [13]. Some researchers hypothesized that intonation provides reliable cues to grammatical structure by referring to the theory of phonological bootstrapping [14], which claims that prosodic processing of spoken language allows children to identify and then acquire grammatical structures as inputs. Consequently, difficulties in the processing of prosodic feature such as intonation and rhythm may generate language difficulties [15]. While some studies concluded that SLI patients do not have significant intonation deficits and that intonation is independent of both morphosyntactic and segmental phonological impairments [16]–[18], some others have shown small but significant deficits [13], [19], [20]. With regards to *intonation contours production*, Wells and Peppé [13] found that SLI children produced less congruent contours than typically developing children. The authors hypothesized that SLI children understand the pragmatic context but fail to select the corresponding contour. On the topic of *intonation imitation tasks*, the results seem contradictory. Van der Meulen *et al.* [21] and Wells and Peppé [13] found that SLI children were less able to imitate prosodic features. Several interpretations were proposed: 1) the weakness was due to the task itself rather than to a true prosodic impairment [21]; 2) a failure in working memory was more involved than prosodic skills [21]; and 3) deficits in intonation production at the phonetic level were sufficient to explain the failure to imitate prosodic features [13]. Conversely, Snow [17] reported that children with SLI showed a typical use of falling tones and Marshall *et al.* [18] did not find any difference in the ability to imitate intonation contours between SLI and typically developing children.

Pervasive Developmental Disorders: Abnormal prosody was identified as a core feature of individuals with autism [22]. The observed prosodic differences include monotonic or machine-like intonation, aberrant stress patterns, deficits in pitch and intensity control and a "concerned" voice quality. These inappropriate patterns related to communication/sociability ratings tend to persist over time even while other language skills improve [23]. Many studies have tried to define the prosodic features in Autism Spectrum Disorder (ASD) patients (for a review see [13]). With regards to *intonation contours production* and *intonation contours imitation tasks*, the results are contradictory. In a reading-aloud task, Fosnot and Jun [24] found that AD children did not distinguish questions and statements; all utterances sounded like statements. In an imitation condition task, AD children performed better. The authors concluded that AD subjects can produce intonation contours although they do not use them or understand their communicative value. They also observed a correlation between intonation imitation skills and autism severity, which suggests that the ability to reproduce intonation contours could be an index of autism severity. Paul *et al.* [3] found no difference between AD and TD children in the use of intonation to distinguish questions and statements. Peppé and McCann [25] observed a tendency for AD subjects to utter a sentence that sounds like a question when a statement was appropriate. Le Normand *et al.* [26] found that children with AD produced more words with flat contours than typically developing children. Paul *et al.* [27] documented the abilities to reproduce stress in a nonsense syllable imitation task of an ASD group that included members with high-functioning autism, Asperger's syndrome and PDD-NOS. Perceptual ratings and instrumental measures revealed small but significant differences between ASD and typical speakers.

Most studies have aimed to determine whether AD or SLI children's prosodic skills differed from those of typically developing children. They rarely sought to determine whether the prosodic skills differed between diagnostic categories. We must note that whereas AD diagnostic criteria are quite clear, PDD-NOS is mostly diagnosed by default [28]; its criteria are relatively vague, and it is statistically the largest diagnosed category [29].

Language researchers and clinicians share the challenging objective of evaluating LIC prosodic skills by using appropriate tests. They aim to determine the LIC prosodic characteristics to improve diagnosis and enhance children's social interaction abilities by adapting remediation protocols to the type of disorder. In this study, we used automated methods to assess one aspect of the *grammatical* prosodic functions: sentence modalities (cf. Section I-A).

C. Prosody Assessment Procedures

Existing prosody assessment procedures such as the American ones [3], [30], the British PROP [31], the Swedish one [20],

and the PEPS-C [32] require expert judgments to evaluate the child's prosodic skills. For example, prosody can be evaluated by recording a speech sample and agreeing on the transcribed communicative functions and prosody forms. This method, based on various protocols, requires an expert transcription. As the speech is unconstrained during the recording of the child, the sample necessarily involves various forms of prosody between the speakers, which complicates the acoustic data analysis. Thus, most of the prosodic communication levels (i.e., *grammatical*, *pragmatic* and *affective*, cf. Section I-A) are assessed using the PEPS-C with a constrained speech framework. The program delivers pictures on a laptop screen both as stimuli for expressive utterances (output) and as response choices to acoustic stimuli played by the computer (input). For the input assessment, there are only two possible responses for each proposed item to avoid undue demand on auditory memory. As mentioned by the authors, this feature creates a bias that is hopefully reduced by the relatively large number of items available for each task. For the output assessment, the examiner has to judge whether the sentences produced by the children can be matched with the prosodic stimuli of each task. Scoring options given to the tester are categorized into two or three possibilities to score the imitation such as "good/fair/poor" or "right/wrong." As the number of available items for judging the production of prosody is particularly low, this procedure does not require a high level of expertise. However, we might wonder whether the richness of prosody can be evaluated (or categorized) in such a discrete way. Alternatively, using many more evaluation items could make it difficult for the tester to choose the most relevant ones.

Some recent studies have proposed automatic systems to assess prosody production [33], speech disorders [34] or even early literacy [35] in children. Multiple challenges will be faced by such systems in characterizing the prosodic variability of LIC. Whereas acoustic characteristics extracted by many automatic speech recognition (ASR) systems are segmental (i.e., computed over a time-fixed sliding window that is typically 32 ms with an overlap ratio of 1/2), prosodic features are extracted in a supra-segmental framework (i.e., computed over various time scales). Speech prosody concerns many perceptual features (e.g., pitch, loudness, voice quality, and rhythm) that are all included in the speech waveform. Moreover, these acoustic correlates of prosody present high variability due to a set of contextual (e.g., disturbances due to the recording environment) and speaker's idiosyncratic variables (e.g., affect [36] and speaking style [37]). Acoustic, lexical, and linguistic characteristics of solicited and spontaneous children's speech were also correlated with age and gender [38].

As characterizing speech prosody is difficult, six design principles were defined in [33]: 1) *highly constraining methods* to reduce unwanted prosodic variability due to assessment procedure contextual factors; 2) a "*prosodic minimal pairs*" *design* for one task to study prosodic contrast; 3) *robust acoustic features* to ideally detect automatically the speaker's turns, pitch errors and mispronunciations; 4) *fusion of relevant features* to find the importance of each on the other in these disorders; 5) *both global and dynamical features* to catch specific contrasts of prosody; and 6) *parameter-free techniques* in which the algorithms either are based on established facts about prosody (e.g., the phrase-final lengthening phenomenon) or are developed in exploratory analyses of a separate data set whose characteristics are quite different from the main data in terms of speakers.

The system proposed by van Santen *et al.* [33] assesses prosody on *grammatical* (lexical stress and phrase boundary), *pragmatic* (focus and style), and *affective* functions. Scores are evaluated by both humans and a machine through spectral, fundamental frequency and temporal information. In almost all tasks, it was found that the automated scores correlated with the mean human judgments approximately as well as the judges' individual scores. Similar results were found with the system termed PEAKS [34] wherein speech recognition tools based on hidden Markov models (HMMs) were used to assess speech and voice disorders in subjects with conditions such as a removed larynx and cleft lip or palate. Therefore, automatic assessments of both speech and prosodic disorders are able to perform as well as human judges specifically when the system tends to include the requirements mentioned by [33].

D. Aims of This Study

Our main objective was to propose an automatic procedure to assess LIC prosodic skills. This procedure must differentiate LIC patients from TD children using prosodic impairment, which is a known clinical characteristic of LIC (cf. Section I-B). It should also overcome the difficulties created by categorizing the evaluations and by human judging bias (cf. Section I-C). The motives of these needs were twofold: 1) the acoustic correlates of prosody are perceptually much too complex to be fully categorized into items by humans; and 2) these features cannot be reliably judged by humans who have subjective opinions [39] in as much as inter-judge variability is also problematic. Indeed, biases and inconsistencies in perceptual judgment were documented [40], and the relevant features for characterizing prosody in speech were defined [41], [42]. However, despite progress in extracting a wide set of prosodic features, there is no clear consensus today about the most efficient features.

In the present study, we focused on the French language and on one aspect of the prosodic *grammatical* functions: sentence modalities (cf. Section I-A). As the correspondences between "prosody" and "sentence-type" are language specific, the intonation itself was classified in the present work. We aimed to compare the performances among different children's groups (e.g., TD, AD, PDD-NOS and SLI) in a proposed intonation imitation task by using automated approaches.

Imitation tasks are commonly achieved by LIC patients even with autism [43]. In a patient, this ability can be used to test the prosodic field without any limitations due to their language disability. Imitation tasks introduce bias in the data because the produced speech is not natural and spontaneous. Consequently, the intonation contours that were reproduced by subjects may not correspond with the original ones. However, all subjects were confronted with the same task of a single protocol of data recording (cf. Section V-B). Moreover, the prosodic patterns that served to characterize the intonation contours were collected from TD children (cf. Section III-D). In other words, the bias introduced by TD children in the proposed task was included in the system's configuration. In this paper, any significant devia-

tion from this bias will be considered to be related to *grammatical* prosodic skill impairments, i.e., intonation contours imitation deficiencies.

The methodological novelty brought by this study lies in the combination of static and dynamic approaches to automatically characterize the intonation contours. The static approach corresponds to a typical state-of-the-art system: statistical measures were computed on pitch and energy features, and a decision was made on a sentence. The dynamic approach was based on hidden Markov models wherein a given intonation contour is described by a set of prosodic states [44].

The following section presents previous works that accomplished intonation contours recognition. Systems that were used in this study are described in Section III. The recruitment and the clinical evaluation of the subjects are presented in Section IV. The material used for the experiments is given in Section V. Results are provided in Section VI while Section VII is devoted to a discussion, and Section VIII contains our conclusions.

II. Related Works in Intonation Recognition

The automatic characterization of prosody was intensively studied during the last decade for several purposes such as emotion, speaker, and speech recognition [45]–[47] and infant-directed speech, question, dysfluency, and certainty detection [48]–[51]. The performance achieved by these systems is clearly degraded when they deal with spontaneous speech or certain specific voice cases (e.g., due to the age of a child [52] or a pathology [53]). The approaches used for automatically processing prosody must deal with three key questions: 1) the *time scale* to define the extraction locus of features (e.g., speaker turn and specific acoustic or phonetic containers such as voiced segments or vowels) [54]; 2) the *set of prosodic descriptors* used for characterizing prosody (e.g., low-level descriptors or language models); and 3) the choice of a *recognition scheme* for automatic decisions on the *a priori* classes of the prosodic features. Fusion techniques were proposed to face this apparent complexity [55], [56]. A fusion can be achieved on the three key points mentioned above, e.g., unit-based (vowel/consonant) fusion [57], features-based (acoustic/prosodic) fusion [58], and classifier-based fusion [59].

Methods that are used to characterize the intonation should be based on pitch features because the categories they must identify are defined by the pitch contour. However, systems found in the literature have shown that the inclusion of other types of information such as energy and duration is necessary to achieve good performance [60], [61]. Furthermore, detection of motherese, i.e., the specific language characterized by high pitch values and variability that is used by a mother when speaking to her child, requires others types of features than those derived from pitch to reach satisfactory recognition scores [59].

Narayanan *et al.* proposed a system that used features derived from the Rise-Fall-Connection (RFC) model of pitch with an n-gram prosodic language model for four-way pitch accent labeling [60]. RFC analysis considers a prosodic event as being comprised of two parts: a rise component followed by a fall component. Each component is described by two parameters: amplitude and duration. In addition, the peak value of pitch for the event and its position within the utterance is recorded in the RFC model. A recognition score of 56.4% was achieved by this system on the Boston University Radio News Corpus (BURNC), which includes 3 hours of read speech (radio quality) produced by six adults.

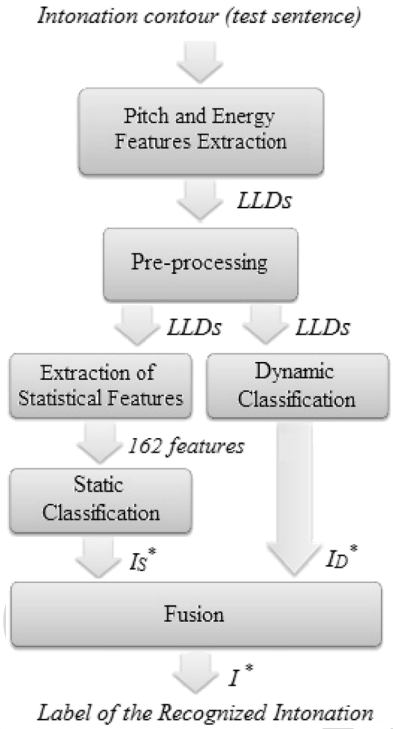


Fig. 1. Scheme of the intonation recognition system.

Rosenberg *et al.* compared the discriminative usefulness of units such as vowels, syllables, and word levels in the analysis of acoustic indicators of pitch accent [61]. Features were derived from pitch, energy, and duration through a set of statistical measures (e.g., max, min, mean, and standard deviation) and normalized to speakers by a z-score. By using logistic regression models, word level was found to provide the best score on the BURNC corpus with a recognition rate of 82.9%.

In a system proposed by Szaszák *et al.* [44], an HMM-based classifier was developed with the aim of evaluating intonation production in a speech training application for hearing impaired children. This system was used to classify five intonation classes and was compared to subjective test results. The automatic classifier provided a recognition rate of 51.9%, whereas humans achieved 69.4%. A part of this work was reused in this study as a so-called "dynamic pitch contour classifier" (cf. Section III-B).

III. Intonation Contours Recognition

The processing stream proposed in this study includes steps of prosodic information extraction and classification (Fig. 1). However, even if the data collection phase is realized up-stream (cf. Section V-B), the methods used for characterizing the intonation correspond to a recognition system. As the intonation contours analyzed in this study were provided by the imitation of prerecorded sentences, the *speaker turn* unit was used as a data input for the recognition system. This unit refers to the moment where a child imitates one sentence. Therefore, this study does not deal with read or spontaneous speech but rather with

constrained speech where spontaneity may be found according to the child.

During the features extraction step, both pitch and energy features, i.e., low-level descriptors (LLDs), were extracted from the speech by using the Snack toolkit [62]. The fundamental frequency was calculated by the ESPS method with a frame rate of 10 ms. Pre-processing steps included an anti-octave jump filter to reduce pitch estimation errors. Furthermore, pitch was linearly extrapolated on unvoiced segments (no longer than 250 ms, empirically) and smoothed by an 11-point averaging filter. Energy was also smoothed with the same filter. Pitch and energy features were then normalized to reduce inter-speaker and recording-condition variability. Fundamental frequency values were divided by the average value of all voiced frames, and energy was normalized to 0 dB. Finally, both first-order and second-order derivates (Δ and $\Delta\Delta$) were computed from the pitch and energy features so that a given intonation contour was described by six prosodic LLDs, as a basis for the following characterization steps.

Intonation contours were then separately characterized by both static and dynamic approaches (cf. Fig. 1). Before the classification step, the static approach requires the extraction of LLD statistical measures, whereas the dynamic approach is optimized to directly process the prosodic LLDs. As these two approaches were processing prosody in distinctive ways, we assumed that they were providing complementary descriptions of the intonation contours. Output probabilities returned by each system were thus fused to get a final label of the recognized intonation. A ten-fold cross-validation scheme was used for the experiments to reduce the influence of data splitting in both the learning and testing phases [63]. The folds were stratified, i.e., intonation contours were equally distributed in the learning data sets to insure that misrepresented intonation contours were not disadvantaged during the experiments.

A. Static Classification of the Intonation Contour

This approach is a typical system for classifying prosodic information by making an intonation decision on a sentence using LLD statistical measures concatenated into a super-vector. Prosodic features, e.g., pitch, energy and their derivates (Δ and $\Delta\Delta$), were characterized by a set of 27 statistical measures (Table I) such that 162 features in total composed the super-vector that was used to describe the intonation in the static approach. The set of statistical measures included not only traditional ones such as maximum, minimum, the four first statistical moments, and quartiles but also perturbation-related coefficients (e.g., jitter and shimmer), RFC derived features (e.g., the relative positions of the minimum and maximum values) and features issued from question detection systems (e.g., the proportion/mean of rising/descending values) [49].

The ability of these features to discriminate and characterize the intonation contours was evaluated by the RELIEF-F algorithm [64] in a ten-fold cross-validation framework. RELIEF-F was based on the computation of both *a priori* and *a posteriori* entropy of the features according to the intonation contours. This algorithm was used to initialize a sequential forward selection (SFS) approach for the classification step. Ranked features were sequentially inserted in the prosodic features super-vector,

TABLE I
SET OF STATISTICAL MEASURES USED FOR STATIC MODELING OF PROSODY

Measure	Description
Max	Value of the maximum
RPmax	Relative position of the maximum
Min	Value of the minimum
RPmin	Relative position of the minimum
RP_AD	Absolute difference between RPmax and RPmin
Range_n	Range divided by RP_AD
Mean	Mean value
STD	Standard deviation value
Skewness	Third statistical moment
Kurtosis	Fourth statistical moment
Q1	Value of the first quartile
Median	Median value
Q3	Value of the third quartile
IQR	Inter Quartile Range
IQR_STD_AD	Absolute difference between IQR and STD
Jitter / Shimmer	Coefficient of pitch / energy perturbation
Slope	First coefficient of the regression slope
OnV	Onset value (start value)
TaV	Target value (middle value)
OfV	Onset value (end value)
TaVOnV_AD	Absolute difference between TaV and OnV
OfVOnV_AD	Absolute difference between OfV and OnV
OfVTaV_AD	Absolute difference between OfV and TaV
%↑	Proportion of rising values
%↓	Proportion of descending values
μ↑	Mean of rising values
μ↓	Mean of descending values

and we only kept those that created an improvement in the classification task. This procedure has permitted us to identify the relevant prosodic features for intonation contour characterization. However, the classification task was done 162 times, i.e., the number of extracted features in total. A k-nearest-neighbors algorithm was used to classify the features (k was set to three); the k-nn classifier estimates the maximum likelihood on *a posteriori* probabilities of recognizing an intonation contour I_n ($n = 1, 2, \dots N$ intonation classes) on a tested sentence S by searching the k_n labels (issued from a learning phase) that contain the closest set of prosodic features to those issued from the tested sentence S. The recognized intonation I_S^* was obtained by an argmax function on the estimates of the *a posteriori* probabilities $p_{\text{stat}}(I_n|S)$ (1) [63]:

$$p_{\text{stat}}(I_n|S) = \frac{k_n}{k}$$
$$I_S^* = \underset{n \in 1:N}{\operatorname{argmax}} \left[p_{\text{stat}}(I_n|S) \right] \quad (1)$$

B. Dynamic Classification of the Intonation Contour

The dynamic pitch contour classifier used hidden Markov models (HMMs) to characterize the intonation contours by using prosodic LLDs provided by the feature extraction steps. This system was analogous to an ASR system; however, the features were based on pitch and energy, and the prosodic contours were thus modeled instead of phoneme spectra or

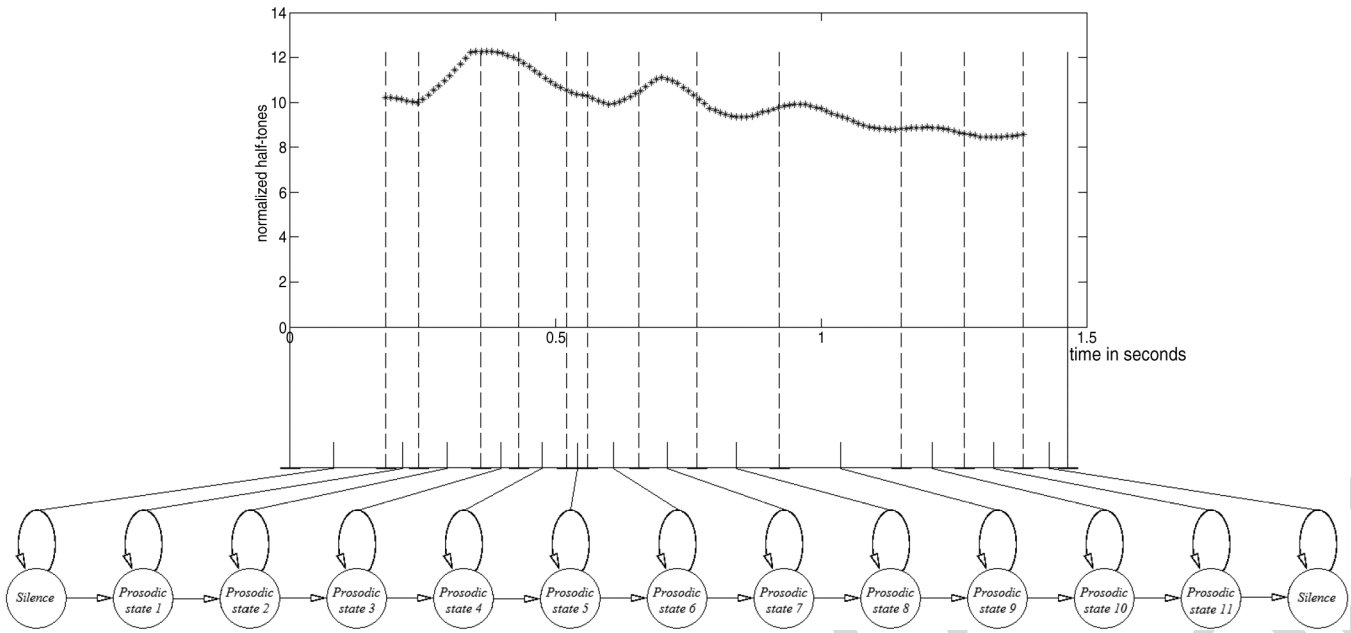


Fig. 2. Principle of HMM prosodic modeling of pitch values extracted from a sentence.

cepstra. The dynamic description of intonation requires a determination of both the location and the duration of the intonation units that represent different states in the prosodic contours (Fig. 2). Statistical distributions of the LLDs were estimated by Gaussian mixture models (GMMs) as mixtures of up to eight Gaussian components. Observation vectors (*prosodic states* in Fig. 2) were six-dimensional, i.e., equal to the number of LLDs. Because some sentences were conveying intonation with much shorter duration than others, both a fixed and a varying number of states was used according to sentence duration to set the HMMs for the experiments. A fixed number of 11-state models patterned by eight Gaussian mixtures were found to yield the best recognition performance in empirical optimization for Hungarian. In this case, the same configuration was applied to French because the intonations we wished to characterize were identical to those studied in [44]. Additionally, a silence model was used to set the HMM's configuration states for the beginning and the ending of a sentence. The recognized intonation I_D^* was obtained by an argmax function on the *a posteriori* probabilities $p_{\mathrm{dyn}}(I_n|S)$ (2)

$$p_{\mathrm{dyn}}(I_n|S) = \frac{p(S|I_n)p(I_n)}{p(S)}$$
$$I_D^* = \underset{n\in 1:N}{\operatorname{argmax}}\,[p_{\mathrm{dyn}}(I_n|S)] \qquad (2)$$

The estimation of $p_{\mathrm{dyn}}(I_n|S)$ was decomposed in the same manner as in speech recognition; according to Bayes' rule, $p(S|I_n)$ specifies the prosodic probability of observations extracted from a tested sentence S, where $p(I_n)$ is the probability associated with the intonation contours and $p(S)$ is the probability associated with the sentences.

C. Fusion of the Classifiers

Because the static and dynamic classifiers provide different information by using distinct processes to characterize the intonation, a combination of the two should improve recognition performance. Although many sophisticated decision techniques do exist to fuse them [55], [56], we used a weighted sum of the *a posteriori* probabilities:

$$I^* = \underset{n\in 1:N}{\operatorname{argmax}}\,(\alpha * p_{\mathrm{stat}}(I_n|S) + (1-\alpha) * p_{\mathrm{dyn}}(I_n|S)). \quad (3)$$

This approach is suitable because it provides the contribution of each classifier used in the fusion. In (3), the label I^* of the final recognized intonation contour is attributed to a sentence S by weighting the *a posteriori* probabilities provided by both static and dynamic based classifiers by a factor $\alpha(0 \leq \alpha \leq 1)$. To assess the similarity between these two classifiers, we calculated the Q statistic [50]:

$$Q_{\mathrm{stat,dyn}} = \frac{N^{00}N^{11} - N^{01}N^{10}}{N^{00}N^{11} + N^{01}N^{10}} \qquad (4)$$

where N^{00} is the number of times both classifiers are wrong, N^{11} is the number of times both classifiers are correct, N^{01} is the number of times when the first classifier is correct and the second is wrong and N^{10} is the number of times when the first classifier is wrong and the second classifier is correct. The Q statistic takes values between $[-1; 1]$ and the closer the value is to 0, the more dissimilar the classifiers are. For example, $Q_{\mathrm{stat,dyn}} = 0$ represents total dissimilarity between the two classifiers. The Q statistic was used to evaluate how complementarity the audio and visual information is for dysfluency detection in a child's spontaneous speech [50].

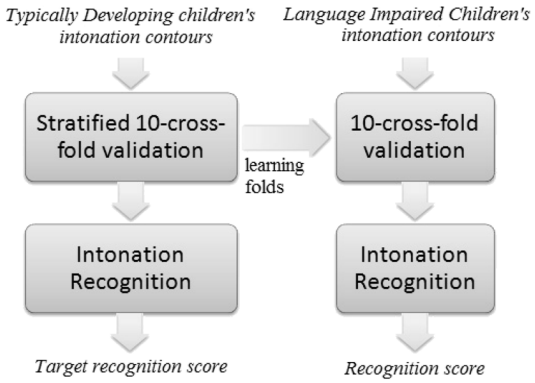


Fig. 3. Strategies for intonation contours recognition.

TABLE II
SOCIODEMOGRAPHIC AND CLINICAL CHARACTERISTICS OF SUBJECTS

Characteristic	AD	PDD-NOS	SLI
Age in years	$9.8_{3.5}$	$9.8_{2.2}$	$9.2_{3.9}$
Male – Female	10 – 2	9 – 1	10 – 3
ADI-R scores			
Social impairment	$21.1_{5.8}$	$12.7_{7.8}$	Not-relevant
Communication	$19.3_{5.2}$	$8.5_{6.4}$	Not-relevant
Repetitive interest	$6.4_{2.4}$	$2.0_{1.6}$	Not-relevant
Total	$50.7_{12.8}$	$25.7_{15.4}$	Not-relevant
CARS scores	$33.2_{15.4}$	$22.3_{5.4}$	Not-relevant

Statistics are given in the following style: $[\text{Mean}]_{(\text{standard}-\text{deviation})}$; AD: autism disorder; PDD-NOS: pervasive developmental disorder-not otherwise specified; SLI: specific language impairment; SD: standard deviation; ADI-R: autism diagnostic interview-revised [66]; CARS: child autism rating scale [67].

D. Recognition Strategies

Recognition systems were first used on the control group data to define the target scores for the intonation contours. To achieve this goal, TD children's sentences were stratified according to the intonation in a ten-fold cross-validated fashion and the *a posteriori* probabilities provided by both static and dynamic intonation classifiers were fused according to (3). LIC prosodic abilities were then analyzed by testing the intonation contours whereas those produced by the control group were learned by the recognition system (Fig. 3).

The TD children's recognition scheme was thus cross-validated with those of LIC: testing folds of each LIC group were all processed with the ten learning folds that were used to classify the TD children's intonation contours. Each testing fold provided by data from the LIC was thus processed ten times. For comparison, the relevant features set that was obtained for TD children by the static classifier was used to classify the LIC intonation contours. However, the optimal weights for fusion of both static and dynamic classifiers were estimated for each group separately, i.e., TD, AD, PDD-NOS, and SLI.

IV. RECRUITMENT AND CLINICAL EVALUATIONS OF SUBJECTS

A. Subjects

Thirty-five monolingual French-speaking subjects aged 6 to 18 years old were recruited in two university departments of child and adolescent psychiatry located in Paris, France (Université Pierre et Marie Curie/Pitié-Salpêtière Hospital and Université René Descartes/Necker Hospital). They consulted for patients with PDD and SLI, which were diagnosed as AD, PDD-NOS, or SLI according to the DSM-IV criteria [8]. Socio-demographic and clinical characteristics of the subjects are summarized in Table II.

To investigate whether prosodic skills differed from those of TD children, a monolingual control group ($n = 73$) matched for chronological age (mean age $= 9.8$ years; standard deviation $=$ 3.3 years) with a ratio of 2 TD to 1 LIC child was recruited in elementary, secondary, and high schools. None of the TD subjects had a history of speech, language, hearing, or general learning problems.

AD and PDD-NOS groups were assigned from patients' scores on the Autism Diagnostic Interview-Revised [66] and the Child Autism Rating Scale [67]. The psychiatric assessments and parental interviews were conducted by four child-psychiatrists specialized in autism. Of note, all PDD-NOS also fulfilled diagnostic criteria for Multiple Complex Developmental Disorder [68], [69], a research diagnosis used to limit PDD-NOS heterogeneity and improve its stability overtime [70]. SLI subjects were administered a formal diagnosis of SLI by speech pathologists and child psychiatrists specialized in language impairments. They all fulfilled criteria for Mixed Phonologic–Syntactic Disorder according to Rapin and Allen's classification of Developmental Dysphasia [9]. This syndrome includes poor articulation skills, ungrammatical utterances and comprehension skills better than language production although inadequate overall for their age. All LIC subjects received a psychometric assessment for which they obtained Performance Intellectual Quotient scores above 70, which meant that none of the subjects showed mental retardation.

B. Basic Language Skills of Pathologic Subjects

To compare basic language skills between pathological groups, all subjects were administered an oral language assessment using three tasks from the ELO Battery [71]: 1) *Receptive Vocabulary*; 2) *Expressive Vocabulary*; and 3) *Word Repetition*. ELO is dedicated to children 3–11 years old. Although many subjects of our study were older than 11, their oral language difficulties did not allow the use of other tests because of an important floor-effect. Consequently, we adjusted the scoring system and determined the severity levels. We determined for each subject the corresponding age for each score and calculated the discrepancy between "verbal age" and "chronological age." The difference was converted into severity levels using a five-level Likert-scale with 0 standing for the expected level at that chronological age, 1 standing for a 1-year deviation from the expected level at that chronological age, 2 for 2-years deviation, 3 for 3-years deviation, and 4 standing for 4 or more years of deviation.

Receptive Vocabulary: This task containing 20 items requires word comprehension. The examiner gives the patient a picture booklet and tells him or her: "*Show me the picture in which there is a* ...." The subject has to select from among four pictures the one corresponding to the uttered word. Each correct identification gives one point, and the maximum score is 20.

Expressive Vocabulary: This task containing 50 items calls for the naming of pictures. The examiner gives the patient a

TABLE III
BASIC LANGUAGE SKILLS OF PATHOLOGIC SUBJECTS

Task from ELO [71]	AD	PDD-NOS	SLI
Receptive Vocabulary	$2.4_{1.6}$	$1.9_{1.5}$	$1.9_{1.0}$
Expressive Vocabulary	$2.0_{1.8}$	$1.2_{1.8}$	$1.4_{1.1}$
Word Repetition	$2.9_{1.5}$	$2.7_{1.4}$	$3.5_{0.7}$

Statistics are given in the following style: $[\text{Mean}]_{(\text{standard}-\text{deviation})}$; AD: autism disorder; PDD-NOS: pervasive developmental disorder-not otherwise specified; SLI: specific language impairment.

booklet comprised of object pictures and asks him or her "*What is this?*" followed by "*What is he/she doing*?" for the final ten pictures, which show actions. Each correct answer gives one point and the maximum score for objects is 20 for children from 3 to 6, 32 for children from 6 to 8, and 50 for children over 9.

Word Repetition: This task is comprised of 2 series of 16 words and requires verbal encoding and decoding. The first series contains disyllabic words with few consonant groups. The second contains longer words with many consonant groups, which allows the observation of any phonological disorders. The examiner says "*Now, you are going to repeat exactly what I say. Listen carefully, I won't repeat*." Then, the patient repeats the 32 words, and the maximum score is 32.

As expected given clinical performance skills in oral communication, no significant differences were found in vocabulary tasks depending on the groups' mean severity levels (Table III): $p = 0.5$ for the *receptive* task and $p = 0.4$ for the *expressive* task. All three groups showed an equivalent delay of 1 to 2 years relative to their chronological ages. The three groups were similarly impaired in the *word repetition* task, which requires phonological skills. The average delay was 3 years relative to their chronological ages ($p = 0.8$).

V. DATABASE DESIGN

A. Speech Materials

Our main goal was to compare the children's abilities to reproduce different types of intonation contours. In order to facilitate reproducibility and to avoid undue cognitive demand, the sentences were phonetically easy and relatively short. According to French prosody, 26 sentences representing different modalities (Table IV) and four types of intonations (Fig. 4) were defined for the imitation task. Sentences were recorded by means of the *Wavesurfer* speech analysis tool [72]. This tool was also used to validate that the intonation contour of the sentences matched the patterns of each intonation category (Fig. 4) The reader will have to be careful with the English translations of the sentences given in Table IV as they may provide different intonation contours due to French prosodic dependencies.

B. Recording the Sentences

Children were recorded in their usual environment, i.e., the clinic for LIC and elementary school/high school for the control group. A middle quality microphone (*Logitech USB Desktop*) plugged to a laptop running *Audacity* software was used for the recordings. In order to limit the perception of the intonation groups among the subjects, sentences were randomly played

TABLE IV
SPEECH MATERIAL FOR THE INTONATION IMITATION TASK

Intonation	Modality	Sentence
Descending *statements*	Declarative, affirmative	"David a mangé un croissant." *"David ate a croissant."*
		"Je viens d'arriver de l'école." *"I'm coming from the school."*
	Declarative, negative	"Cette maison ne me plaît pas du tout." *"This house does not appeal to me at all."*
		"Il n'est pas encore l'heure." *"It's not yet time."*
	Declarative, dubitative	"Je ne suis pas sûr de pouvoir le faire." *"I'm not about to do so."*
		"Il me semble qu'il ne soit pas encore prêt." *"It seems to me he is not yet ready"*
	Exclamatory, emphatic	"C'est Rémy qui va être content." *"That's Rémy whom will be happy."*
		"C'est ainsi que vont les choses." *"So things are going."*
Falling *questions* / *statements*	Interrogative	"Où se tient-il ?" *"Where is he standing?"*
		"Comment vas-tu ?" *"How are you?"*
	Interrogative, reduced order	"Pouvez-vous passer à mon bureau ?" *"Can you go to my office?"*
		"Pourriez-vous nous accorder un instant ?" *"Can you give us a moment?"*
	Exclamatory	"Oh non, je ne te le donnerais pas." *"Oh no, I won't give it to you."*
		"Comme je suis content !" *"Because I'm happy!"*
	Imperative, order/ counseling	"Ne l'abîme pas !" *"Do not ruin it!"*
		"Dis-moi la vérité !" *"Tell me the truth!"*
Floating *statements*	Declarative	"Anna viendra avec toi." *"Anna will come with you."*
		"Je suis très content que tu sois venu." *"I am very glad you came."*
	Exclamatory	"J'aime les crêpes au chocolat." *"I like pancakes with chocolate."*
		"Il n'aime pas le sucre en poudre." *"He does not like powdered sugar."*
Rising *questions*	Interrogative, short questions	"Qui ?" / *"Who?"*
		"Un croissant ?" / *"A croissant?"*
		"Pardon ?" / *"Pardon?"*
		"A l'intérieur ?" / *"On the inside?"*
		"Ah bon ?" / *"Really?"*
		"Quoi ?" / *"What?"*

with an order that was fixed prior to the recordings. During the imitation task, subjects were asked to repeat exactly the sentences they had heard even if they did not catch one or several words. If the prosodic contours of the sentences were too exaggeratedly reproduced or the children showed difficulties, then the sentences were replayed a couple of times.

To ensure that clean speech was analyzed in this study, the recorded data were carefully controlled. Indeed, the reproduced sentences must as much as possible not include false-starts, repetitions, noises from the environment or speech not related to the task. All of these perturbations were found in the recordings. As they might influence the decision taken on the sentences when characterizing their intonation, sentences reproduced by the children were thus manually segmented and post-processed. Noisy sentences were only kept when they presented false-starts or repetitions that could be suppressed without changing the intonation contour of the sentence. All others noisy sentences were rejected so that from a total of 2813 recorded sentences,

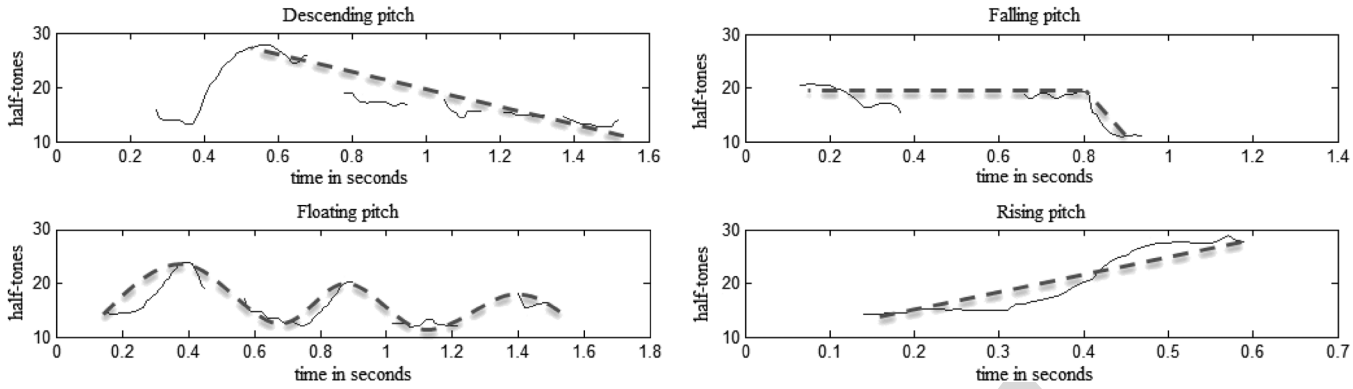


Fig. 4. Groups of intonation according to the prosodic contour: (a) "Descending pitch," (b) "Falling pitch," (c) "Floating pitch" and (d) "Rising pitch." (a): "That's Rémy whom will be content.," (b): "As I'm happy!," (c): "Anna will come with you.," (d): "Really?" Estimated pitch values are shown as solid lines while the prosodic prototypes are shown as dashed lines.

TABLE V
QUANTITY OF ANALYZED SENTENCES

Intonation	REF	TD	AD	PDD-NOS	SLI
Descending	8	580	95	71	103
Falling	8	578	94	76	104
Floating	4	291	48	40	52
Rising	6	432	70	60	78
All	26	1881	307	247	337

REF: speech material; TD: typically developing; AD: autism disorder; PDD: pervasive developmental disorders not-otherwise specified; SLI: specific language impairment.

2772 sentences equivalent to 1 hour of speech in total were kept for analysis (Table V).

VI. RESULTS

Experiments conducted to study the children's prosodic abilities in the proposed intonation imitation task were divided into two main steps. The first step was composed of a duration analysis of the reproduced sentences by means of statistical measures such as mean and standard deviation values. In the second step, we used the classification approaches described in Section III to automatically characterize the intonation. The recognition scores of TD children are seen as targets to which we can compare the LIC. Any significant deviation from the mean TD children's score will be thus considered to be relevant to *grammatical* prosodic skill impairments, i.e., intonation contours imitation deficiencies. A non-parametric method was used to make a statistical comparison between children's groups, i.e., a p-value was estimated by the Kruskal–Wallis method. The p-value corresponds to the probability that the compared data have issued from the same population; $p < 0.05$ is commonly used as an alternative hypothesis where there is less than 5% of chance that the data have issued from an identical population.

A. Typically Developing Children

Sentence Duration: Results showed that the patterns of sentence duration were conserved for all intonation groups when the sentences were reproduced by TD children ($p > 0.05$). Consequently, the TD children's imitations of the intonation contours have conserved the duration patterns of the original sentences (Table VI).

TABLE VI
SENTENCE DURATION STATISTICS OF TYPICALLY DEVELOPING CHILDREN

Intonation	REF	TD
Descending	$1.7_{0.3}$	$1.7_{0.6}$
Falling	$1.2_{0.3}$	$1.3_{1.4}$
Floating	$1.6_{0.2}$	$1.6_{0.4}$
Rising	$0.7_{0.2}$	$0.5_{0.2}$

Statistics for sentence duration (in s,) are given in the following style: $[\text{Mean}]_{(\text{standard}-\text{deviation})}$; REF: reference sentences; TD: typically developing.

TABLE VII
STATIC, DYNAMIC AND FUSION INTONATION RECOGNITION PERFORMANCES FOR TYPICALLY DEVELOPING CHILDREN

Intonation	Static	Dynamic	Fusion	$Q_{stat,dyn}$
Descending	61	55	64	0.1688
Falling	55	48	55	0.3830
Floating	49	71	72	0.6754
Rising	93	95	95	0.2716
All	67	64	70	0.4166

Performances are given as percentage of recognition from a stratified ten-fold cross-validation based approach.

Intonation Recognition: Recognition scores on TD children's intonation contours are given in Table VII. For comparison, we calculated the performance of a naïve classifier, which always attributes the label of the most represented intonation, e.g., "*Descending*," to a given sentence. The Q statistics (cf. Section III-C) were computed for each intonation to evaluate the similarity between classifiers during the classification task.

The naïve recognition rate of the four intonations studied in this paper was 31%. The proposed system raises this to 70%, i.e., more than twice the chance score, for 73 TD subjects aged 6 to 18. This recognition rate is equal to the average value of scores that were obtained by other authors on the same type of task, i.e., intonation contours recognition, but on adult speech data and for only six speakers [60], [61]. Indeed, the age effect on the performance of speech processing systems has been

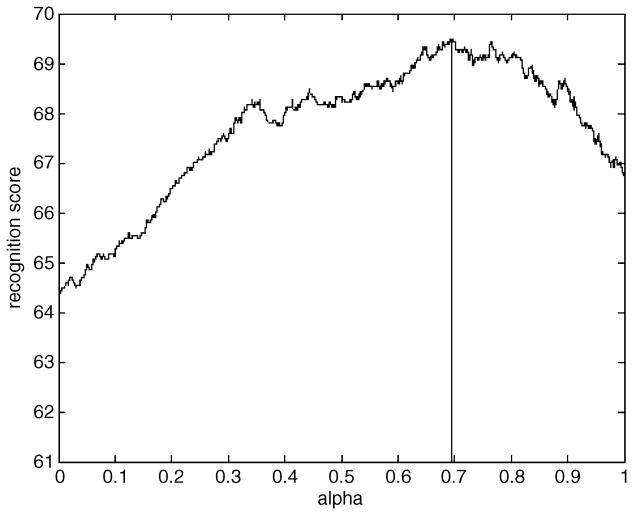


Fig. 5. Fusion recognition scores as function of weight alpha attributed to both static ($\alpha = 1$) and dynamic classifier ($\alpha = 0$).

shown to be a serious disturbing factor especially when dealing with young children [52]. Surprisingly, the static and dynamic classifiers were similar for the "*Floating*" intonation even when the dynamic recognition score was clearly higher than the static one (Table VII). However, because this intonation contains the smallest set of sentences (cf. Table IV), a small dissimilarity between classifiers was sufficient to improve the recognition performance. The concept of exploiting the complementarity of the classifiers used to characterize the intonation contours (cf. Section III-C) was validated as some contours were better recognized by either the static or dynamic approach. Whereas both "*Rising*" and "*Floating*" intonations were very well recognized by the system, "*Descending*" and "*Falling*" intonations provided the lowest recognition performances. The low recognition score of the "*Falling*" intonation may be explained by the fact that this intonation was represented by sentences that contained too many ambiguous modalities (e.g., question/order/counseling etc.) compared with the others.

The best recognition scores provided by the fusion of the two classifiers were principally conveyed by the static approach rather than by the dynamic one (Fig. 5).

As the "*Floating*" intonation had a descending trend, it was confused with the "*Descending*" and "*Falling*" intonations but never with "*Rising*" (Table VIII). The "*Rising*" intonation appeared to be very specific because it was very well-recognized and was only confused with "*Falling*." Confusions with respect to the "*Falling*" intonation group were numerous as shown by the scores, and were principally conveyed by both the "*Descending*" and "*Floating*" intonations.

The set of relevant prosodic features that was provided by the SFS method, which was used for the static-based intonation classification (cf. Section III-A), is mostly constituted of both Δ and $\Delta\Delta$ derivates (Table IX): 26 of the 27 relevant features were issued from these measures. Features extracted from pitch are more numerous than those from energy, which may be due to the fact that we exclusively focused on the pitch contour when recording the sentences (cf. Section V-A). About half of

TABLE VIII
CONFUSION MATRIX OF THE INTONATION RECOGNITION FOR TYPICALLY DEVELOPING CHILDREN

Intonation	Descending	Falling	Floating	Rising
Descending	377	58	151	2
Falling	104	320	116	46
Floating	50	33	212	0
Rising	2	16	4	416

Tested intonations are given in rows while recognized ones (I^*) are given in columns. Diagonal values from top-left to bottom-right thus correspond to sentences that were correctly recognized by the system while all others are miscategorized.

TABLE IX
RELEVANT PROSODIC FEATURES SET IDENTIFIED BY STATIC RECOGNITION

Pitch	Energy
R – RPmax	Δ – IQR
Δ – Q1	Δ – Shimmer
Δ – Q3	Δ – Slope
Δ – Jitter	Δ – TaV
Δ – Slope	Δ – TaVOnV_AD
Δ – OfVTaV_AD	ΔΔ – RPmax
ΔΔ – RPmin	ΔΔ – RPmin
ΔΔ – RP_AD	ΔΔ – Q3
ΔΔ – STD	ΔΔ – OnV
ΔΔ – Q1	ΔΔ – TaV
ΔΔ – Median	ΔΔ – OfVOnV_AD
ΔΔ – Q3	
ΔΔ – IQR	
ΔΔ – Jitter	
ΔΔ – OnV	
ΔΔ – OfVOnV_AD	

R: raw data (i.e., static descriptor), Δ: first-order derivate, ΔΔ: second-order derivate (Δ, and ΔΔ are both dynamic descriptor).

the features set include measures issued from typical question detection systems, i.e., values or differences between values at onset/target/offset and relative positions of extrema in the sentence. The others are composed of traditional statistical measures of prosody (e.g., quartiles, slope, and standard deviation values). All 27 relevant features provided by the SFS method during static classification were statistically significant for characterizing the four types of intonation contours ($p < 0.05$).

B. Language-Impaired Children

Sentence Duration: All intonations that were reproduced by LIC appeared to be strongly different from those of TD children when comparing sentence duration ($p < 0.05$): the duration was lengthened by 30% for the three first intonations and by more than 60% for the "*Rising*" contour (Table X). Moreover, the group composed of SLI children produced significantly longer sentences than all other groups of children except for the case of "Rising" intonation.

Intonation Recognition: The contributions from the two classification approaches that were used to characterize the intonation contours were similar among all pathologic groups but different from that for TD children: static, $\alpha = 0.1$; dynamic, $1 - \alpha = 0.9$ (Fig. 6). The dynamic approach was thus found

TABLE X
SENTENCE DURATION STATISTICS OF THE GROUPS

Intonation	REF	TD	AD	PDD-NOS	SLI
Descending	$1.7_{0.3}$	$1.7_{0.6}$	$2.2_{0.9}$ ***T,S**	$2.2_{0.8}$ ***T,S**	$2.4_{0.9}$ ***T,A,P**
Falling	$1.2_{0.3}$	$1.3_{1.4}$	$1.6_{0.6}$ ***T,S**	$1.7_{0.8}$ ***T,S**	$1.8_{0.8}$ ***T,A,P**
Floating	$1.6_{0.2}$	$1.6_{0.4}$	$2.1_{0.7}$ ***T,S**	$2.1_{0.5}$ ***T,S**	$2.4_{1.0}$ ***T,A,P**
Rising	$0.7_{0.2}$	$0.5_{0.2}$	$0.9_{0.3}$ ***T**	$0.9_{0.3}$ ***T**	$0.8_{0.2}$ ***T**

Statistics for sentence duration (in s,) are given in the following style: $[\mathrm{Mean}]_{(\mathrm{standard-deviation})}$; $^* = p < 0.05$: alternative hypothesis is true when comparing data between child groups, i.e., T, A, P, and S; REF: reference sentences; TD (T): typically developing; AD (A): autism disorder; PDD (P): pervasive developmental disorders not-otherwise specified; SLI (S): specific language impairment.

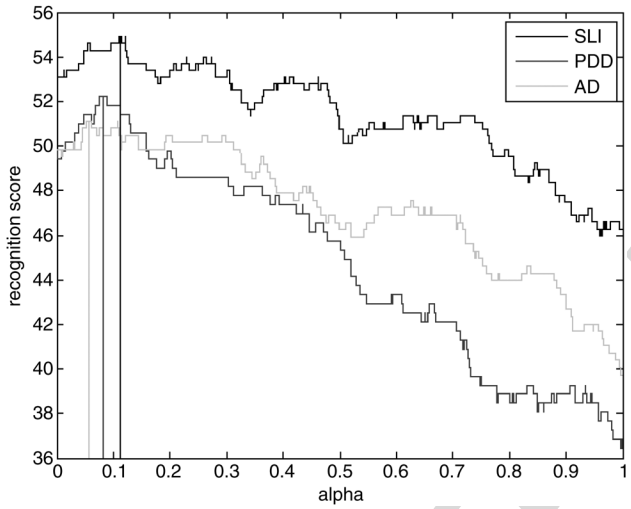


Fig. 6. Fusion recognition scores as function of weight alpha attributed to both static ($\alpha = 1$) and dynamic classifier ($\alpha = 0$).

TABLE XI
Q STATISTICS BETWEEN STATIC AND DYNAMIC CLASSIFIERS

Measure	TD	AD	PDD-NOS	SLI
$Q_{stat,dyn}$	0.4166	0.6539	0.4521	0.5542

to be more efficient than the static one for comparing the LIC's intonation features with those of TD children.

The Q statistics between the classifiers were higher for LIC than TD children so that even after recognizing that dynamic processing was most suitable for LIC, both the static and dynamic intonation recognition methods had less dissimilarity than for TD children (Table XI).

LIC recognition scores were close to those of TD children and similar between LIC groups for the "*Descending*" intonation while all other intonations were significantly different ($p < 0.05$) between TD children and LIC (Table XII). However, the system had very high recognition rates for the "*Rising*" intonation for SLI and TD children whereas it performed significantly worse for both AD and PDD-NOS ($p < 0.05$). Although some differences were found between LIC groups for this intonation, the LIC global mean scores only showed dissimilarity with TD.

TABLE XII
FUSION INTONATION RECOGNITION PERFORMANCES

Intonation	TD	AD	PDD-NOS	SLI
Descending	64	64	70	63
Falling	55	$35^{*\mathbf{T}}$	$45^{*\mathbf{T}}$	$39^{*\mathbf{T}}$
Floating	72	$48^{*\mathbf{T}}$	$40^{*\mathbf{T}}$	$31^{*\mathbf{T}}$
Rising	**95**	$57^{*\mathbf{T,S}}$	$48^{*\mathbf{T,S}}$	$\mathbf{81}^{*\mathbf{T,A,P}}$
All	70	$56^{*\mathbf{T}}$	$53^{*\mathbf{T}}$	$58^{*\mathbf{T}}$

Performances are given as percentage of recognition; $^* = p < 0.05$: alternative hypothesis is true when comparing data from child groups, i.e., T, A, P, and S; TD (T): typically developing; AD (A): autism disorder; PDD (P): pervasive developmental disorders not-otherwise specified; SLI (S): specific language impairment.

TABLE XIII
CONFUSION MATRIX OF THE INTONATION RECOGNITION FOR AUTISTIC DIAGNOSED CHILDREN

Intonation	Descending	Falling	Floating	Rising
Descending	61	14	20	0
Falling	39	33	20	2
Floating	16	9	23	0
Rising	5	23	2	40

Tested intonations are given in rows while recognized ones (I^*) are given in columns. Diagonal values from top-left to bottom-right thus correspond to sentences that were correctly recognized by the system while all others are miscategorized.

TABLE XIV
CONFUSION MATRIX OF THE INTONATION RECOGNITION FOR PERVASIVE-DEVELOPMENTAL-DISORDER DIAGNOSED CHILDREN

Intonation	Descending	Falling	Floating	Rising
Descending	50	5	16	0
Falling	29	34	13	0
Floating	18	8	16	0
Rising	8	19	4	29

Tested intonations are given in rows while recognized ones (I^*) are given in columns. Diagonal values from top-left to bottom-right thus correspond to sentences that were correctly recognized by the system while all others are miscategorized.

The misjudgments made by the recognition system for LIC were approximately similar to those seen for TD children (Tables XIII–XV). For all LIC, the "*Floating*" intonation was similarly confused with "*Descending*" and "*Falling*" and was never confused with "*Rising*." However, the "*Rising*" intonation was rarely confused when two other intonations were tested. This intonation appeared to be very different from the other three but not for the TD group in which more errors were found when the "*Falling*" intonation was tested.

VII. DISCUSSION

This study investigated the feasibility of using an automatic recognition system to compare prosodic abilities of LIC (Tables II and III) to those of TD children in an intonation imitation task. A set of 26 sentences, including statements and

TABLE XV
CONFUSION MATRIX OF THE INTONATION RECOGNITION FOR SPECIFIC LANGUAGE IMPAIRMENT DIAGNOSED CHILDREN

Intonation	Descending	Falling	Floating	Rising
Descending	65	22	15	1
Falling	47	41	16	0
Floating	20	16	16	0
Rising	3	10	2	63

Tested intonations are given in rows while recognized ones (I^*) are given in columns. Diagonal values from top-left to bottom-right thus correspond to sentences that were correctly recognized by the system while all others are miscategorized.

questions (Table IV) over four intonation types (Fig. 4), was used for the intonation imitation task. We manually collected 2772 sentences from recordings of children. Two different approaches were then fused to characterize the intonation contours through prosodic LLD: static (statistical measures) and dynamic (HMM features). The system performed well for TD children excepted in the case of the "*Falling*" intonation, which had a recognition rate of only 55%. This low score may be due to the fact that too many ambiguous speech modalities were included in the "*Falling*" intonation group (e.g., question/order/counseling etc.). The static recognition approach provided a list of 27 features that almost represented dynamic descriptors, i.e., delta and delta-delta. This approach was contributed more than the dynamic approach (i.e., HMM) to the fusion.

Concerning LIC (AD, PDD-NOS, and SLI), the assessment of basic language skills [71] showed that 1) there was no significant difference among the groups' mean severity levels and 2) all three groups presented a similar delay when compared to TD children. In the intonation imitation task, the sentence duration of all LIC subjects was significantly longer than for TD children. The sentence lengthening phenomenon added about 30% for the first three intonations and more than 60% for the "*Rising*" intonation. Therefore, all LIC subjects presented difficulties in imitating intonation contours with respect to duration especially for the "*Rising*" intonation (short questions). This result correlates with the hypothesis that rising tones may be more difficult to produce than falling tones in children [16]. It also correlates with the results of some clinical studies for SLI [13], [19]–[21], AD [24]–[26], and PDD-NOS [27] children although some contradictory results were found for SLI [18].

The best approach to recognize LIC intonation was clearly based on a dynamic characterization of prosody, i.e., using HMM. On the contrary, the best fusion approach favored static characterization of prosody for TD children. Although scores of the LIC's intonation contours recognition were similar to those of TD children for the "*Descending*" sentences group, i.e., statements in this study, these scores have not yet been achieved in the same way. This difference showed that LIC reproduced statement sentences similar to TD children, but they all tended to use prosodic contour transitions rather than statistically specific features to convey the modality.

All other tested intonations were significantly different between TD children and LIC ($p < 0.05$). LIC demonstrated more difficulties in the imitation of prosodic contours than TD children except for the "*Descending*" intonation, i.e., statements in this study. However, SLI and TD children had very high recognition rates for the "*Rising*" intonation whereas both AD and PDD-NOS performed significantly worse. This result is coherent with studies that showed PDD children have more difficulties at imitating questions than statements [24] as well as short and long prosodic items [25], [27]. As pragmatic prosody was strongly conveyed by the "*Rising*" intonation due to the short questions, it is not surprising that such intonation recognition differences were found between SLI and the PDDs. Indeed, both AD and PDD-NOS show pragmatic deficits in communication, whereas SLI only expose pure language impairments. Moreover, Snow hypothesized [16] that rising pitch requires more effort in physiological speech production than falling tones and that some assumptions could be made regarding the child's ability or intention to match the adult's speech. Because the "*Rising*" intonation included very short sentences (half the duration) compared with others, which involves low working memory load, SLI children were not disadvantaged compared to PDDs as was found in [13].

Whereas some significant differences were found in the LIC's groups with the "*Rising*" intonation, the global mean recognition scores did not show any dissimilarity between children. All LIC subjects showed similar difficulties in the administered intonation imitation task as compared to TD children, whereas differences between SLI and both AD and PDD-NOS only appeared on the "*Rising*" intonation; the latter is probably linked to deficits in the pragmatic prosody abilities of AD and PDD-NOS.

The automatic approach used in this study to assess LIC prosodic skills in an intonation imitation task confirms the clinical descriptions of the subjects' communication impairments. Consequently, it may be a useful tool to adapt prosody remediation protocols to improve both LIC's social communication and interaction abilities. The proposed technology could be thus integrated into a fully automated system that would be exploited by speech therapists. Data acquisition could be manually acquired by the clinician while reference data, i.e., provided by TD children, would have already been collected and made available to teach the prosodic models required by the classifiers. However, because intonation contours and the associated sentences proposed in this study are language dependent, they eventually must be adapted to intonation studies in other languages than French.

Future research with examine the *affective* prosody of LIC and TD children. Emotions were elicited during a story-telling task with an illustrated book that contains various emotional situations. Automatic systems will serve to characterize and compare the elicited emotional prosodic particulars of LIC and TD children. Investigations will focus on several questions: 1) can LIC understand depicted emotions and convey relevant prosodic features for emotional story-telling; 2) do TD children and LIC groups achieve similarly in the task; and 3) are there some types of prosodic features that are preferred to convey emotional prosody (e.g., rhythm, intonation, or voice quality)?

VIII. CONCLUSION

This study addressed the feasibility of designing a system that automatically assesses a child's *grammatical* prosodic skills,

i.e., intonation contours imitation. This task is traditionally administered by speech therapists, but we proposed the use of automatic methods to characterize the intonation. We have compared the performance of such a system on groups of children, i.e., TD and LIC (e.g., AD, PDD-NOS, and SLI).

The records on which this study was conducted include the information based on both perception and production of the intonation contour. The administered task was very simple because it was based on the imitation of sentences conveying different types of modality through the intonation contour. Consequently, the basic skills of the subjects in the perception and the reproduction of prosody were analyzed together. The results conveyed by this study have shown that the LIC have the ability to imitate the "*Descending*" intonation contours similar to TD. Both groups got close scores given by the automatic intonation recognition system. LIC did not yet achieve those scores as the TD children. Indeed, a dynamic modeling of prosody has led to superior performance on the intonation recognition of all LIC's groups, while a static modeling of prosody has provided a better contribution for TD children. Moreover, the sentence duration of all LIC subjects was significantly longer than the TD subjects (the sentence lengthening phenomenon was about 30% for first three intonations and more than 60% for the "*Rising*" intonation that conveys pragmatic). In addition, this intonation has not led to degradations in the performances of the SLI subjects unlike to PDDs as they are known to have pragmatic deficiencies in prosody.

The literature has shown that a separate analysis of the prosodic skills of LIC in the production and the perception of the intonation leads to contradictory results; [16]–[18] versus [13]–[15] and [19]–[21] for SLI children, and [3] versus [24]–[27] for the PDDs. Consequently, we used a simple technique to collect data for this study. The data collected during the imitation task include both perception and production of the intonation contours, and the results obtained by the automatic analysis of the data have permitted to obtain those descriptions that are associated with the clinical diagnosis of the LIC. As the system proposed in this study is based on the automatic processing of speech, its interest for the diagnosis of LIC through prosody is thus fully justified. Moreover, this system could be integrated into software, such as the SPECO [73], that would be exploited by speech therapists to use prosodic remediation protocols adapted to the subjects. It would thus serve to improve both the LIC's social communication and interaction abilities.

REFERENCES

[1] S. Ananthakrishnan and S. Narayanan, "Unsupervised adaptation of categorical prosody models for prosody labeling and speech recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 17, no. 1, pp. 138–149, Jan. 2009.

[2] P. K. Kuhl, "Early language acquisition: Cracking the speech code," *Nature Rev. Neurosci.*, vol. 5, pp. 831–843, Nov. 2004.

[3] R. Paul, A. Augustyn, A. Klin, and F. R. Volkmar, "Perception and production of prosody by speakers with autism spectrum disorders," *J. Autism Develop. Disorders*, vol. 35, no. 2, pp. 205–220, Apr. 2005.

[4] P. Warren, "Parsing and prosody: An introduction," *Lang. Cognitive Process., Psychol. Press*, vol. 11, pp. 1–16, 1996.

[5] D. Van Lancker, D. Canter, and D. Terbeek, "Disambiguation of ditropic sentences: Acoustic and phonetic cues," *J. Speech Hear. Res.*, vol. 24, no. 3, pp. 330–335, Sep. 1981.

[6] E. Winner, *The Point of Words: Children's Understanding of Metaphor and Irony*. Cambridge, MA: Harvard Univ. Press, 1988.

[7] D. Bolinger, *Intonation and Its Uses: Melody in Grammar and Discourse*. Stanford, CA: Stanford Univ. Press, Aug. 1989.

[8] *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed. Washington, DC: American Psychiatric Assoc., 1994.

[9] I. Rapin and D. A. Allen, "Developmental language: Nosological consideration," in *Neuropsychology of Language, Reading and Spelling*, V. Kvik, Ed. New York: Academic Press, 1983.

[10] L. Wing and J. Gould, "Severe impairments of social interaction and associated abnormalities in children: Epidemiology and classification," *J. Autism Develop. Disorders*, vol. 9, no. 1, pp. 21–29, Mar. 1979.

[11] D. A. Allen and I. Rapin, "Autistic children are also dysphasic," in *Neurobiology of Infantile Autism*, H. Naruse and E. M. Ornitz, Eds. Amsterdam, The Netherlands: Excerpta Medica, 1992, pp. 157–168.

[12] J. McCann and S. Peppé, "Prosody in autism: A critical review," *Int. J. Lang. Commun. Disorders*, vol. 38, no. 4, pp. 325–350, May 2003.

[13] B. Wells and S. Peppé, "Intonation abilities of children with speech and language impairments," *J. Speech, Lang. Hear. Res.*, vol. 46, pp. 5–20, Feb. 2003.

[14] J. Morgan and K. Demuth, *Signal to Syntax: Bootstrapping From Speech to Grammar in Early Acquisition*. Mahwah, NJ: Erlbaum, 1996.

[15] S. Weinert, "Sprach- und Gedächtnisprobleme dysphasisch-sprachgestörter Kinder: Sind rhytmisch-prosodische Defizite eine Ursache?," in *[Language and Short-Term Memory Problems of Specifically Language Impaired Children: Are Rhythmic Prosodic Deficits a Cause?] Rhytmus Ein interdisziplinäres Handbuch*, K. Müller and G. Aschersleben, Eds. Bern, Switzerland: Huber, 2000, pp. 255–283.

[16] D. Snow, "Children's imitations of intonation contours: Are rising tones more difficult than falling tones?," *J. Speech, Lang. Hear. Res.*, vol. 41, pp. 576–587, Jun. 1998.

[17] D. Snow, "Prosodic markers of syntactic boundaries in the speech of 4-year-old children with normal and disordered language development," *J. Speech, Lang. Hear. Res.*, vol. 41, pp. 1158–1170, Oct. 1998.

[18] C. R. Marshall, S. Harcourt Brown, F. Ramus, and H. J. K Van der Lely, "The link between prosody and language skills in children with SLI and/or dyslexia," *Int. J. Lang. Commun. Disorders*, vol. 44, no. 4, pp. 466–488, Jul. 2009.

[19] P. Hargrove and C. P. Sheran, "The use of stress by language impaired children," *J. Commun. Disorders*, vol. 22, no. 5, pp. 361–373, Oct. 1989.

[20] C. Samuelsson, C. Scocco, and U. Nettelbladt, "Towards assessment of prosodic abilities in Swedish children with language impairment," *Logopedics Phoniatrics Vocology*, vol. 28, no. 4, pp. 156–166, Oct. 2003.

[21] S. Van der Meulen and P. Janssen, "Prosodic abilities in children with Specific Language Impairment," *J. Commun. Disorders*, vol. 30, pp. 155–170, May–Jun. 1997.

[22] L. Kanner, "Autistic disturbances of affective contact," *Nervous Child*, vol. 2, pp. 217–250, 1943.

[23] R. Paul, L. Shriberg, J. Mc Sweeny, D. Ciccheti, A. Klin, and F. Volkmar, "Brief report: Relations between prosodic performance and communication and socialization ratings in high functioning speakers with autism spectrum disorders," *J. Autism Develop. Disorders*, vol. 35, no. 6, pp. 861–869, Dec. 2005.

[24] S. Fosnot and S. Jun, "Prosodic characteristics in children with stuttering or autism during reading and imitation," in *Proc. 14th Annu. Congr. Phonetic Sci.*, San Francisco, CA., Aug. 1–7, 1999, pp. 103–115.

[25] J. McCann, S. Peppé, F. Gibbon, A. O'Hare, and M. Rutherford, "Prosody and its relationship to language in school-aged children with high functioning autism," *Int. J. Lang. Commun. Disorders*, vol. 47, no. 6, pp. 682–702, Nov. 2007.

[26] M. T. Le Normand, S. Boushaba, and A. Lacheret-Dujour, "Prosodic disturbances in autistic children speaking French," in *Proc. Speech Prosody*, Campinas, Brazil, May 6–9, 2008, pp. 195–198.

[27] R. Paul, N. Bianchi, A. Agustyn, A. Klin, and F. Volkmar, "Production of syllable stress in speakers with autism spectrum disorders," *Research in Autism Spectrum Disorders*, vol. 2, pp. 110–124, Jan.–Mar. 2008.

[28] F. Volkmar, *Handbook of Autism and Pervasive Develop. Disorders*. Hoboken, NJ: Wiley, 2005.

[29] E. Fombonne, "Epidemiological surveys of autism and other pervasive developmental disorders: An update," *J. Autism Develop. Disorders*, vol. 33, no. 4, Aug. 2003.

[30] L. D. Schriberg, J. Kwiatkowski, and C. Rasmussen, *The Prosody-Voice Screening Profile*. Tuscon, AZ: Communication Skill Builders, 1990.

[31] D. Crystal, *Profiling Linguist. Disability*. London, U.K.: Edward Arnold, 1982.

[32] P. Martínez-Castilla and S. Peppé, "Developing a test of prosodic ability for speakers of Iberian-Spanish," *Speech Commun.*, vol. 50, no. 11–12, pp. 900–915, Mar. 2008.

[33] J. P. H. van Santen, E. T. Prud'hommeaux, and L. M. Black, "Automated assessment of prosody production," *Speech Commun.*, vol. 51, no. 11, pp. 1082–1097, Nov. 2009.

[34] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "PEAKS—A system for the automatic evaluation of voice and speech disorder," *Speech Commun.*, vol. 51, no. 5, pp. 425–437, May 2009.

[35] M. Black, J. Tepperman, A. Kazemzadeh, S. Lee, and S. Narayanan, "Automatic pronunciation verification of English letter-names for early literacy assessment of preliterate children," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 19–24, 2009, pp. 4861–4864.

[36] C. Min Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[37] G. P. M. Laan, "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and read speaking style," *Speech Commun.*, vol. 22, pp. 43–65, Mar. 1997.

[38] A. Potamianos and S. Narayanan, "A review of the acoustic and linguistic properties of children's speech," in *Proc. IEEE 9th Workshop Multimedia Signal Process.*, Chania, Greece, Oct. 23, 2007, pp. 22–25.

[39] R. D. Kent, "Hearing and believing: Some limits to theauditory-perceptual assessment of speech and voice disorders," *Amer. J. Speech-Lang. Pathol.*, vol. 5, no. 3, pp. 7–23, Aug. 1996.

[40] A. Tversky, "Intransitivity of preferences," *Psychol. Rev.*, vol. 76, pp. 31–48, Jan. 1969.

[41] A. Pentland, "Social signal processing," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 108–111, Jul. 2007.

[42] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals," in *Proc. Interspeech ICSLP*, Antwerp, Belgium, Aug. 27–31, 2007, pp. 2253–2256.

[43] J. Nadel, "Imitation and imitation recognition: Functional use in preverbal infants and nonverbal children with autism," in *The Imitative Mind: Development, Evolution and Brain Bases*, A. N. Meltzoff and W. Prinz, Eds. Cambridge, MA: Cambridge Univ. Press, 2002, pp. 2–14.

[44] G. Szaszák, D. Sztahó, and K. Vicsi, "Automatic intonation classification for speech training systems," in *Proc. Interspeech*, Brighton, U.K., Sep. 6–10, 2009, pp. 1899–1902.

[45] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features and methods," *Speech Commun.*, vol. 48, no. 9, pp. 1162–1181, Sep. 2006.

[46] A. G. Adami, "Modeling prosodic differences for speaker recognition," *Speech Commun.*, vol. 49, no. 4, pp. 1162–1181, Apr. 2007.

[47] D. H. Milone and A. J. Rubio, "Prosodic and accentual information for automatic speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 321–333, Jul. 2003.

[48] A. Mahdhaoui, M. Chetouani, C. Zong, R. S. Cassel, C. Saint-Georges, M.-C. Laznik, S. Maestro, F. Apicella, F. Muratori, and D. Cohen, "Automatic motherese detection for face-to-face interaction analysis," *Multimodal Signals: Cognitive and Algorithmic Issues*, vol. LNAI 5398, pp. 248–255, Feb. 2009, Springer-Verlag.

[49] V.-M. Quang, L. Besacier, and E. Castelli, "Automatic question detection: Prosodic-lexical features and crosslingual experiments," in *Proc. Interspeech ICSLP*, Antwerp, Belgium, Aug. 27–31, 2007, pp. 2257–2260.

[50] S. Yildirim and S. Narayanan, "Automatic detection of disfluency boundaries in spontaneous speech of children using audio-visual information," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 1, pp. 2–12, Jan. 2009.

[51] H. Pon-Barry and S. Shieber, "The importance of sub-utterance prosody in predicting level of certainty," in *Proc. Human Lang. Tech. Conf.*, Poznan, Poland, May 31–Jun. 5 2009, pp. 105–108.

[52] D. Elenius and M. Blomberg, "Comparing speech recognition for adults and children," in *Proc. FONETIK*, Stockholm, Sweden, May 26–28, 2004, pp. 105–108.

[53] J.-F. Bonastre, C. Fredouille, A. Ghio, A. Giovanni, G. Pouchoulin, J. Révis, B. Teston, and P. Yu, "Complementary approaches for voice disorder assessment," in *Proc. Interspeech ICSLP*, Antwerp, Belgium, Aug. 27–31, 2007, pp. 1194–1197.

[54] M. Chetouani, A. Mahdhaoui, and F. Ringeval, "Time-scalefeature extractions for emotional speech characterization," *Cognitive Comp.*, vol. 1, no. 2, pp. 194–201, 2009, Springer.

[55] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, NJ: Wiley, 2004.

[56] E. Monte-Moreno, M. Chetouani, M. Faundez-Zanuy, and J. Sole-Casals, "Maximum likelihood linear programming data fusion for speaker recognition," *Speech Commun.*, vol. 51, no. 9, pp. 820–830, Sep. 2009.

[57] F. Ringeval and M. Chetouani, "A vowel based approach for acted emotion recognition," in *Proc. Interspeech*, Brisbane, Australia, Sep. 22–26, 2008, pp. 2763–2766.

[58] A. Mahdhaoui, F. Ringeval, and M. Chetounani, "Emotional speech characterization based on multi-features fusion for face-to-face communication," in *Proc. Int. Conf. SCS*, Jerba, Tunisia, Nov. 6–8, 2009.

[59] A. Mahdhaoui, M. Chetouani, and C. Zong, "Motherese detection based on segmental and supra-segmental features," in *Proc. Int. Conf. Pattern Recogn.*, Tampa, FL., Dec. 8–11, 2008.

[60] S. Ananthakrishnan and S. Narayanan, "Fine-grained pitch accent and boundary tones labeling with parametric f0 features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, Mar. 30–Apr. 4 2008, pp. 4545–4548.

[61] A. Rosenberg and J. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level," in *Proc. Human Lang. Tech.: 2009 Annu. Conf. North Amer. Chapter Assoc. for Comput. Ling.*, Boulder, CO, May 31–Jun. 5 2009, pp. 81–84.

[62] Snack Sound Toolkit [Online]. Available: http://www.speech.kth.se/snack/

[63] R.-O. Duda, P.-E. Hart, and D.-G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2000.

[64] M. Robnik and I. Konenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learn. J.*, vol. 53, pp. 23–69, Oct.–Nov. 2003.

[65] L. Kuncheva and C. Whitaker, "Measure of diversity in classifier ensembles," *Mach. Learn.*, vol. 51, no. 2, pp. 181–207, May 2003.

[66] C. Lord, M. Rutter, and A. Le Couteur, "Autism diagnostic interview-revised: A revision version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders," *J. Autism Develop. Disorders*, vol. 24, no. 5, pp. 659–685, 1994.

[67] E. Schopler, R. Reichler, R. Devellis, and K. Daly, "Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS)," *J. Autism Develop. Disorders*, vol. 10, no. 1, pp. 91–103, 1980.

[68] R. Van der Gaag, J. Buitelaar, E. Van den Ban, M. Bezemer, L. Njio, and H. Van Engeland, "A controlled multivariate chart review of multiple complex developmental disorder," *J. Amer. Acad. Child Adolesc. Psychiatry*, vol. 34, pp. 1096–1106, 1995.

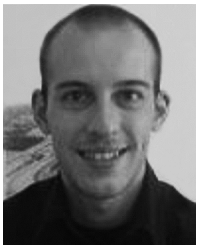
[69] J. Buitelaar and R. Van der Gaag, "Diagnostic rules for children with PDD-NOS and multiple complex developmental disorder," *J. Child Psychol. Psychiatry*, vol. 39, pp. 91–919, 1998.

[70] E. Rondeau, L. Klein, A. Masse, N. Bodeau, D. Cohen, and J. M. Guilé, "Is pervasive developmental disorder not otherwise specified less stable than autistic disorder?," *J. Autism Develop. Disorder*, 2010, to be published.

[71] A. Khomsi, *Evaluation du Langage Oral*. Paris, France: ECPA, 2001.

[72] K. Sjölander and J. Beskow, "WaveSurfer—An open source speech tool," in *Proc. 6th ICSLP*, Beijing, China, Oct. 2000, vol. 4, pp. 464–467 [Online]. Available: http://www.speech.kth.se/wavesurfer/

[73] K. Vicsi, A Multimedia Multilingual Teaching and Training System for Speech Handicapped Children Univ. of Technol. and Economics, Dept. of Telecommunications and Telematics, Final Annual Report, Speech Corrector, SPECO-977126 [Online]. Available: http://alpha.tmit.bme.hu/speech/speco/index.html, 09.1998–08.2001



Fabien Ringeval received the B.S. degree in electrics, electronic and informatics engineering from the National Technologic Institute (IUT) of Chartres, Chartres, France, in 2003, and the M.S. degree in speech and image signal processing from the University Pierre and Marie Curie (UPMC), Paris, France, in 2006.

He has been with the Institute of Intelligent Systems and Robotics, UPMC, since 2006. He is currently a Teaching and Research Assistant with this institute. His research interests concern automatic speech processing, i.e., the automatic characterization of both the verbal (e.g., intonation recognition) and the nonverbal communication (e.g., emotion recognition). He is a member of the French Association of Spoken Communication (AFCP), of the International Speech Communication Association (ISCA) and of the Workgroup on Information, Signal, Image and Vision (GDR-ISIS).

Julie Demouy received the degree of Speech and Language Therapist from the School of Medicine of Paris, University Pierre and Marie Curie (UPMC), Paris, France, in 2009.

She is currently with the University Department of Child and Adolescent Psychiatry at La Pitié Salpêtrière Hospital, Paris.

György Szaszák received the M.S. degree in electrical engineering from the Budapest University for Technology and Economics (BUTE), Budapest, Hungary, 2002 and the Ph.D. degree from Laboratory of Speech Acoustics, Department of Telecommunications and Media Informatics, BUTE in 2009. His Ph.D. dissertation addresses the exploitation of prosody in speech recognition systems with a focus on the agglutinating languages.

He has been with the Laboratory of Speech Acoustics, Department of Telecommunications and Media Informatics, BUTE, since 2002. His main research topics are related to speech recognition, prosody and databases, and both the verbal and the nonverbal communication.

Dr. Szaszák is a member of the International Speech Communication Association (ISCA).

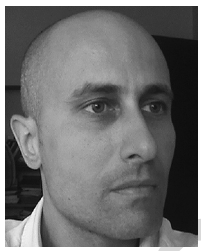
Mohamed Chetouani received the M.S. degree in robotics and intelligent systems from the University Pierre and Marie Curie (UPMC), Paris, France, 2001 and the Ph.D. degree in speech signal processing from UPMC in 2004.

In 2005, he was an invited Visiting Research Fellow at the Department of Computer Science and Mathematics, University of Stirling, Stirling, U.K. He was also an invited Researcher at the Signal Processing Group, Escola Universitaria Politecnica de Mataro, Barcelona, Spain. He is currently an Associate Professor in Signal Processing and Pattern Recognition at the UPMC. His research activities cover the areas of nonlinear speech processing, feature extraction, and pattern classification for speech, speaker, and language recognition.

Dr. Chetouani is a member of different scientific societies (e.g., ISCA, AFCP, ISIS). He has also served as chairman, reviewer, and member of scientific committees of several journals, conferences, and workshops.

Laurence Robel received the M.D. and Ph.D. degrees in both molecular neuropharmacology and developmental biology from the University Pierre and Marie Curie (UPMC), Paris, France.

She is currently coordinating the autism and learning disorders clinics for young children in the Department of Child and Adolescent Psychiatry, Hôpital Necker-Enfants Malades, Paris, France, as a Child Psychiatrist.



Jean Xavier received the Ph.D. degree in psychology from the University Paris Diderot, Paris, France, in 2008.

He is specialized in child and adolescent psychiatry and was certified in 2000. He is an M.D. in the Department of Child and Adolescent Psychiatry, Department of Child and Adolescent Psychiatry, Hôpital de la Pitié-Salpêtrière, Paris, France, and is head of an outpatient child unit dedicated to PDD including autism. He also works in the field of learning disabilities.

Dr. Xavier is a member of the French Society of Child and Adolescent Psychiatry.

David Cohen received the M.S. degree in neurosciences from the University Pierre and Marie Curie (UPMC), Paris, France, and the Ecole Normale Supérieure, Paris, in 1987, and the M.D. degree from the Hôpital Necker-Enfants Malades, Paris, France, in 1992.

He specialized in child and adolescent psychiatry and was certified in 1993. His first field of research was severe mood disorders in adolescent, topic of his Ph.D. degree in neurosciences (2002). He is Professor at the UPMC and head of the Department of Child and Adolescent Psychiatry, La Salpêtrière hospital, Paris. His group runs research programs in the field of autism and other pervasive developmental disorders, severe mood disorder in adolescents, and childhood onset schizophrenia and catatonia.

Dr. Cohen is a member of the International Association of Child and Adolescent Psychiatry and Allied Disciplines, the European College of Neuro-Psychopharmacology, the European Society of Child and Adolescent Psychiatry, and the International Society of Adolescent Psychiatry.

Monique Plaza received the Ph.D. degree in psychology from the University Paris Ouest Nanterre La Défence, Nanterre, France, in 1984.

She is a Researcher in the National Center for Scientific Research (CNRS), Paris, France. She develops research topics about intermodal processing during the life span, and in developmental, neurological, and psychiatric pathologies. In childhood, she studies specific (oral and written) language difficulties, PDD, and PDD-NOS. In adulthood, she works with patients suffering from Grade II gliomas (benign cerebral tumors), which the slow development allows the brain to compensate for the dysfunction generated by the tumor infiltration. Working in an interdisciplinary frame, she is specifically interested in brain models emphasizing plasticity and connectivity mechanisms and thus participates in studies using fMRI and cerebral stimulation during awake surgery. She develops psychological models emphasizing the interactions between cognitive functions and the interfacing between emotion and cognition. As a clinical researcher, she is interested in the practical applications of theoretical studies (diagnosis and remediation).