Automatic Prosodic Disorders Analysis for Impaired Communication Children

Fabien Ringeval - Mohamed Chetouani
Institute of Intelligent Systems and Robotic, Pierre et Marie Curie University - Paris 6
3 rue Galilée, 94200 Ivry sur Seine, France
33-1-44276216
fabien.ringeval@isir.fr , mohamed.chetouani@upmc.fr

David Sztahó - Klara Vicsi
Laboratory of Speech Acoustics, Budapest University of Technology and Economics, TMIT
Stoczek u. 2, 1111 Budapest, Hungary
36-1-4631940
sztaho@tmit.bme.hu , vicsi@tmit.bme.hu

ABSTRACT

This paper is devoted to the study of a pseudo-phonetic approach to characterize prosodic disorders of children with impaired communication skills. To this purpose, we have designed with the help of the clinicians' staff a database containing autistic children. Another database with non disordered speech is used as a control one. Concerning the characterization of the prosodic disorders, we extract the features from phonemic units such as vowels. These segments are provided by a pseudo-phonetic speech segmentation phase combined with a vowel detector. Since the pseudo-phonetic segments convey a lot of prosodic features, such as duration and rhythm, many differentiations can be made between children from the two studied databases. As a conclusion, correlations between prosodic particularities got in this study and those described in the literature are given.

1. INTRODUCTION

In spoken conversation, the speech is produced in a segmental timing by the use of the phonemes, while prosody is supra-segmental. Prosody helps listeners to locate phrase boundaries and word emphasis, but also to identify the pragmatic structure of a given utterance: (e.g. interrogative vs. declarative). It also conveys paralinguistic information such as affect, personality, culture and ethics, which are the most important components of the emotions [1]. The ability to perceive and express emotions, through the prosodic expressions of the face and the voice, has an essential role in the development of the intersubjectivity, and is developed during the early stages of the children's life. Consequently, many children who have speech disorders may have limited social interactions, contributing to social isolation.

As a part of the communication impairment, children may have also prosodic disorders: they may sound different from their peers, adding an additional barrier to both social interactions and integration. Since prosodic disorders are seen as contributing to problems in communication and may lead to social isolation, some researchers have attracted their attention on atypical prosody in individuals with speech disorders [2,3]. They believe that prosodic awareness underpins language skills, and deficit may continue to affect both language development and social interaction.

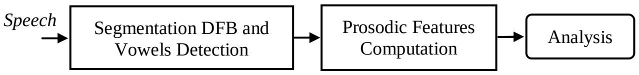
Speech
Segmentation DFB and Vowels Detection
Prosodic Features Computation
Analysis

Figure 1: Pseudo-phonetic approach for prosodic disorders characterization.

In this paper, we exploit a pseudo-phonetic approach to study prosodic disorders for impaired communication analysis, such as autism (figure 1). Prosodic features are usually extracted from the voiced segments. Since vocalic nucleus has been proved to be the most perceptive speech unit [4], the key idea of our approach is to extract the features from vowel segments. Moreover, some recent works have shown the relevance of a feature extraction at the phonetic level for emotion recognition [5,6].

In our approach, vowels are identified by a segmentation of stationary segments (Divergence Forward Backward algorithm – DFB [7]) combined with a vowel detector [8]. This process is language independent and does not aim at the exact identification of phonemes as it can be done by a phonetic alignment. As a result, the obtained segments are termed pseudo-phonetic ones.

Two databases including both spontaneous and read speech were studied. We used the database of the USIT Project which contains speech data from four autistic children (see [9] for details on this corpus). Another database with non disordered speech was studied too (table 1). This database was designed in an elementary school where children were asked to tell a story. Since children were also discussing to each others, the records contain many spontaneous speech. Obtained data were transcribed in turn speakers by their authors. From these data we collected the speech of the children by rejecting those with unsatisfactory quality.

Table 1: Characteristics of the studied databases

Database	Disorder	Number of Children	Speech Quality	Duration
"USIT" Project	Autism	4	Clean	9 ' 17
Elementary School	None	many (~10)	Clean	9 ' 11

While the "USIT' database contains speech disorders: autism is a pervasive developmental disorder (PDD) characterized by the association of communication and socialization impairments, and by repetitive stereotyped behaviours. Communication impairment includes absence or delay in language, lack of non verbal communication, and others specific features such as pronoun reversal, stereotyped and repetitive language, and prosodic abnormalities. In any case, language is not functional, and is not used appropriately to communicate.

2. PROSODIC FEATURES EXTRACTION

Since both vocalic onset and offset are much more related to articulatory phenomena than prosodic ones, we extracted prosodic features from the pseudo-phonetic segments (vowels segments).

2.1 Prosodic Features Computation

Figure 1 describes the used approach for the characterization of the prosody: prosodic features are computed from automatically detected vowel segments. The two main components of the prosody (pitch and energy) are characterized by a set of 28 statistic's measures. Some of them are basic ones (e.g. maximum, quartile and standard-deviation), and others are more complex: relative positions of the maximum and minimum, jitter - shimmer, etc…; derivates Δ and ΔΔ were also computed for both pitch and energy.

Whereas many descriptors have been proposed to characterize the two main components of the prosody, a few can be found for the duration. Since many different concepts exist for rhythm, with sometimes specific units, such as phonemes, syllables, words and sentences, its characterization appears as a difficult task. Indeed, rhythm can be defined by variations from perceptual phenomena related to both pitch and energy. Moreover, pauses or silences between speech units are also considered as rhythmic events.

Despite of this apparent complexity, rhythm has been successfully modelled in a dialect characterization task of the Britain English [10]. The Pairwise Variability Indices (PVI) [11] was used in this study. The PVI quantifies intra or inter duration variability $|d_k - d_{k+1}|$ from N successive vocalized intervals (equation 1). In order to avoid a bias due to speech rate, a normalisation by the mean duration $(d_k + d_{k+1})/2$ is proposed. Since our approach uses vowel segments provided by both DFB pseudo-phonetic speech segmentation and vowel detection phases, we could have employed the PVI.

$$nPVI_{duration}\left(k\right) = 2*\frac{\left|d_k - d_{k+1}\right|}{d_k + d_{k+1}} \quad \textbf{(1)}$$

In order to study dynamic from a supra-temporal point of view of both pitch and energy features, and additionally to the proposed PVI method, we suggest including statistics from both pitch and energy in the computation of the PVI measures (equation 2 and 3).

$$\underset{k=1:N-1,\ s=1:28}{nPVI_{pitch}\left(k,s\right)} = 2*\frac{\left|stat(f0_k,s)-stat(f0_{k+1},s)\right|}{d_k + d_{k+1}} \quad \textbf{(2)}$$

$$\underset{k=1:N-1,\ s=1:28}{nPVI_{energy}\left(k,s\right)} = 2*\frac{\left|stat(en_k,s)-stat(en_{k+1},s)\right|}{d_k + d_{k+1}} \quad \textbf{(3)}$$

While the original PVI is characterized by a mean of the N-1 PVI values, we propose to extend this statistic to more complex ones for describing PVI in a finer way. We thus characterized the PVI measures (equation 1, 2 and 3) with the same set of statistics used for both pitch and energy characterization.

Table 2 presents the studied prosodic features. Rhythmic features are computed for each speech file differing from short sentences (less than 2 seconds), to long ones (maximum duration of 30 sec.). While both pitch and energy features are computed for each vowel segments. This explains why the number of samples between both prosodic features sub-groups differs so much.

Table 2: Characteristics of the groups of prosodic features

Group of Prosodic Features		**Size of the Feature Matrix**
Pitch		84 measures x 2461 samples
Energy		84 measures x 2461 samples
Rhythm	**Duration**	58 measures x 508 samples
	PVI Duration	112 measures x 508 samples
	PVI Pitch	4704 measures x 508 samples
	PVI Energy	4704 measures x 508 samples

2.2 Features Selection

A feature selection phase was employed to provide automatically a priori relevant features from the many computed ones (more than 9e4). Since each feature selection algorithm has both its own advantage and inconvenient, the prosodic features were ranked by two different algorithms.

The former is termed Fisher Discriminant Ratio and is based on statistics computing which assume a Gaussian modality from the data [12]. While the second used algorithm is RELIEF-F [13]. It is based on the computation of both *a priori* and *a posteriori* entropy through a basic classifier. The k-nearest-neighbours algorithm is used to this purpose. RELIEF-F feature selection algorithm is well known for correctly estimating feature's quality in classification problems. But on the other hand, it does not take into account correlations between features, and can thus not detect redundant ones.

Before computing the feature selection algorithms, we group the prosodic features according to the three prosodic groups (table 2) and the two studied databases (table 1). Then we processed both fisher (equation 4) and RELIEF-F algorithms according to the two speech classes: disordered and non-disordered.

$$fisher\left(f\right) = \frac{\sigma\left(f, between\right)^2}{\sigma\left(f, whithin\right)^2} = 2*\frac{\left(\mu(f,2)-\mu(f,1)\right)^2}{\sigma\left(f,1\right)^2+\sigma\left(f,2\right)^2} \quad \textbf{(4)}$$

where μ(f,x) and σ(f,x) correspond to mean and standard-deviation values of a given feature f from class x.

Since the number of PVI measures is very high (see table 2), we ranked them in two different steps. Firstly, we kept from the three PVI subgroups the 100 best features, producing a global PVI of 300 features equally divided in the three main components of prosody. Then we processed a second time the features selection algorithms on the duration measures grouped with the global PVI.

3. PROSODIC DISORDERS ANALYSIS

Prosodic features presented in section 2.1 were extracted on the vowel segments identified from both disordered and clean speech databases (figure 1). Features from the three studied prosodic groups (table 2) were then ranked with two different algorithms (section 2.2). We obtain the final rank of the features by meaning those provided by the two selection algorithms. Table 3 presents the 5 mean best features according to the three prosodic features groups and to the two studied databases ("USIT" and "Elementary School" – "ELS"). Both mean and standard-deviation from these features are also given.

The best relevant prosodic features are issued from energy, while those from pitch appear as worst ones. Moreover, PVI rhythmic measures perform pretty well on energy, but not on pitch and duration (best duration and pitch features are ranked respectively 101 and 102). Nevertheless we must be careful with the results obtained by pitch. Indeed, children has much higher pitch than adults (ie. around 100Hz for adults against 300Hz for children), which involves pitch extraction errors (ie. octave jumps). Indeed, many harmonic candidates can be confused with the fundamental frequency value (pitch estimation reposed on the autocorrelation), occurring then jumps of the fundamental frequency, even if both pitch and energy values were filtered by a median to avoid micro-prosodic variations.

Interesting difference between statistic ratios (standard-deviation / mean) of the two databases can be yet noticed. Ratios from USIT are always lower than ELS for both pitch and energy, and upper for all of the rhythmic features. Concerning the PVI measures, both mean values and standard-deviation differ a lot according to the two databases: ratios from the mean values are upper than 2, and upper than 3 for the standard-deviation measures.

Data from the two speech databases are plotted in figures 2, 3, and 4 according to the two best features of each prosodic group: pitch, energy and rhythm[1] respectively. Data from Figure 2 show that the two speech classes ("disordered" and "non-disordered") are many mixed, which may be due to the pitch extraction errors as we said before, while these two classes are much more differentiated on both Figure 3 and Figure 4.

Table 3: Comparison of the 5 means best features of each prosodic group (mean and standard-deviation values) according to the two studied databases[2]

Prosodic Group	Best Feature		USIT		Elementary School	
	N	Name	Mean	Std	Mean	Std
Pitch	1	IQR (Δ)	4.61	4.78	3.42	3.73
	2	Jitter (ΔΔ)	8.14	17.71	13.95	37.36
	3	Kurtosis (Δ)	4.28	3.48	5.12	4.66
	4	Jitter	8.11	24.88	16.57	63.45
	5	Jitter (Δ)	6.53	14.42	10.84	28.33
Energy	1	Maximum	68.91	6.73	52.98	7.72
	2	3rd quartile	67.67	6.87	51.78	7.63
	3	Median	65.92	7.03	50.24	7.61
	4	Mean	65.42	7.00	50.02	7.60
	5	Onset value	64.77	7.58	49.13	7.88
Rhythm	1	$nPVI_{\Delta\Delta energy(\mathrm{minimum})}$ Standard-deviation	33.02	20.17	15.22	8.84
	2	$nPVI_{\Delta\Delta energy(\mathrm{1st\ decile})}$ Standard-deviation	32.93	20.22	15.15	8.89
	3	$nPVI_{\Delta\Delta energy(\mathrm{minimum})}$ Maximum	51.53	38.09	19.64	12.79
	4	$nPVI_{\Delta\Delta energy(\mathrm{minimum})}$ Last decile	51.51	38.09	19.64	12.79
	5	$nPVI_{\Delta\Delta energy(\mathrm{1st\ decile})}$ Maximum	51.40	38.14	19.57	12.83

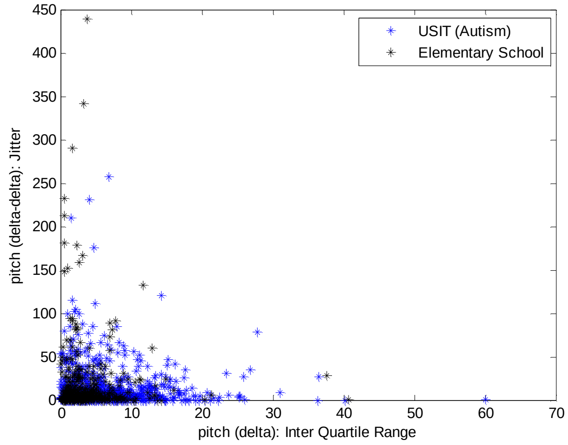

Figure 2: Two best prosodic features from "Pitch" group according to the two speech databases.

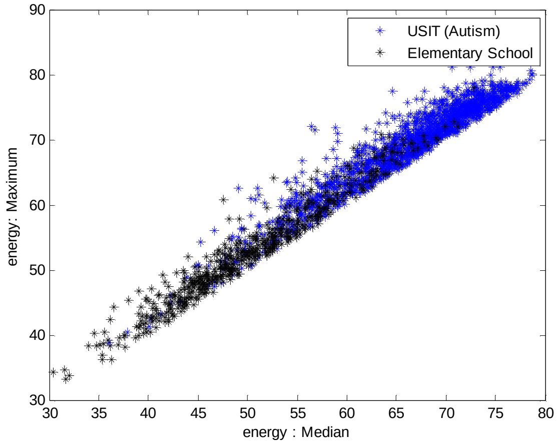

Figure 3: Two best prosodic features from "Energy" group according to the two speech databases.

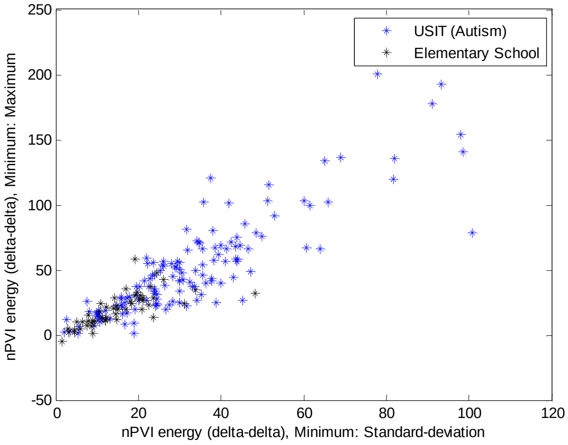

Figure 4: Two best prosodic features from "Rhythm" group according to the two speech databases.

[1] Redundant features were discarded (ie. minimum and 1st decile).

[2] IQR: Inter Quartile range; Onset value: value 10ms after the beginning of the feature vector. 10ms correspond to both pitch and energy extraction rates through the KTH snack toolbox.

4. CONCLUSION

Prosodic particularities from autistic children are both studied and compared to those produced by non-autistic children. A pseudo-phonetic approach was employed for the characterization of the prosody: prosodic features, such as pitch, energy and duration are extracted from automatically detected vowel segments (figure 1). Additionally to the well known PVI measures [11], which quantify time-variation of speech units, we introduced new rhythmic measures to describe the other components of prosody (pitch and energy). The prosodic features were ranked according to the two studied speech classes ("disordered" and "non-disordered") by two different selection algorithms (section 2.2). Best obtained features are then given in section 3. In agreement with descriptions found in literature on prosodic particularities of autistic children (high energy and low rhythmic variations) [3], we found many difference from our prosodic features, namely with those computed from energy. Indeed, these features appear as the most relevant to the prosodic disorders of the autistic children included in our database. But we must be careful since pitch estimation may be biased due to the specific voice of the children which include a lot of harmonics.

5. PERSPECTIVES

A computer assisted Multilingual Teaching and Training System was developed for Speech Handicapped Children, SPECO [14]. Several training blocks are in this audio-visual system, such as example teaching vowels, fricatives in words and sentences and teaching prosody too. Figure 5 presents intonation of a sentence in SPECO system. Such displays can and have been used to provide valuable pronunciation feedback to students. Experiments have shown that a visual F0 display of supra-segmental features combined with audio feedback is more effective than audio feedback alone [15,16], especially if the student's F0 contour is displayed along with a native model.

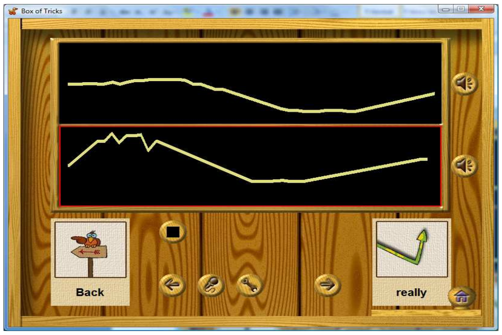

Figure 5: SPECO interface: audio-visual feedback

The feasibility of this type of visual feedback has been demonstrated by a number of simple prototypes [17,18] but nowhere is used automatic feedback to enhance the visual meaning. However this enhance could be important especially for impaired communication children. Consequently, vowel based features presented in this paper are planned to be included in SPECO [14], since they appear as relevant to prosodic disorders.

6. ACKNOWLEDGMENTS

Our thanks to the Valoria laboratory (Bretagne Sud – France) for allowing us to use their database termed "Elementary School" in this paper. The research was prepared in the framework of French and Hungarian Scientific Cooperation of BALATON project (EGIDE).

7. REFERENCES

[1] Cowie, R. 2005. Emotion-Oriented Computing: State of the Art and Key Challenges. Humaine Network of Excellence.

[2] Baauw, S., Ruigendijk, R. and Cuetos, F. 2003. The interpretation of contrastive stress in Spanish-speaking children. In van Kampen, J., Baauw, S. [Ed]: Proceedings of GALA, LOT Occasional Series, 103-114.

[3] McCann, J. and Peppé, S. 2003. Prosody in autism spectrum disorders: a critical review. International Journal of Language & Communication Disorders. 38, 4, 325-350.

[4] Pillot, C., Vaissière, J. 2006. Vocal effectiveness in speech and singing: acoustical, physiological and perspective aspects. Applications in Speech Therapy, Laryngol Otol Rhinol Journal, 127, 5, 293-298.

[5] Lee, C. and al. 2004. Emotion recognition based on phoneme classes. In Proceedings of ICSL.

[6] Ringeval, F. and Chetouani, M. 2008. A vowel based approach for acted emotion recognition. In Proceedings of Interspeech.

[7] André-Obrecht, R. 1988. A New Statistical Approach for Automatic Speech Segmentation. IEEE Transaction on ASSP, 36, 1, 29-40.

[8] Pellegrino, F. and André-Obrecht, R. 2000. Automatic Language Identification: an alternative approach to phonetic modelling. Signal Processing, 80, 1231-1244.

[9] Ringeval, F. and Chetouani, M. 2008. A pseudo-phonetic approach for prosodic disorders characterization of children with impaired communication skills. In A. Esposito and al. [Ed]: Cross-Modal Analysis of Verbal and Non-verbal Communication. Springer-Verlag Publishers.

[10] Grabe, E., Nolan, F. and Farrar, K. 1998. IViE – A comparative transcription system for intonational variation in English, in Proceedings of ICSLP.

[11] Low, E., Grabe, E. and Nolan, F. 2000. Quantitative characterization of speech rhythm: "syllable-timing" in Singapore English. Language and Speech, 43, 377-401.

[12] Fisher, R. M.A. 1936. The use of multiple measures in taxonomic problems. Annals of Eugenics, 7, 179-188.

[13] Robnik, M. and Konenko, I. 2003. Theoretical and empirical analysis of ReliefF and RReliefF, Machine Learning Journal, 53, 23-69.

[14] Vicsi, K. and al. 2001. A multimedia multilingual teaching and training system for speech handicapped children. Int. Journal of Speech Technology, 3, 289-300.

[15] de Bot, K. 1983. Visual feedback of intonation: effectiveness and induced practice behavior. In Language and Speech, 26, 4, 331-350.

[16] James, E. 1976. The acquisition of prosodic features of speech using a speech visualizer. In International Review of Applied Linguistics, 14, 227-243.

[17] Abberton, E., and Fourcin, A.J. 1990. The development of contrastiveness in profoundly deaf children's speech. In Clinical Linguistics and Phonetics, 4, 209-220.

[18] Anderson-Hsieh, J. 1996. A guide to computer-assisted aids for pronunciation improvement. In SPEAK OUT!, 19, 28-31.