ANALYZING THE SPEECH EXPRESSIVENESS USING PROSODIC DYNAMIC CONTROL

Valentin Eugen Ghișa*, Lavinia Nicoleta Ghișa**

* Department of Automatics and Information Technology, Faculty of Electrical Engineering & Computer Science, "Transylvania" University of Brasov, Brasov, Romania, valentin.ghisa@unitbv.ro

** Department of Regional Studios, Romanian Radio Broadcasting Co., Tg. Mures, Romania, nicoghisa@yahoo.com

At the level of verbal communication, the prosodic support and emotional space is modelled as a nonlinear system described through some parameters extracted from the spectral model of vocal wave, respectively the outline of the fundamental frequency, the time and energy of sonorous segments, the duration of non-acoustic segments and breaks, the voice timbre etc. Through the discretised addressing of the spectral model, the aim is to optimise the prosodic characteristics extracted from local variations of the fundamental frequency, by a method of dynamic control.

Keywords: *dynamic control, prosody, message, communication*

1. INTRODUCTION

On their way for a superior positioning more and more entities with interests in the area of socio-economics applied in policy, looking to optimize the communication in various ways, updating classical techniques of persuasion or those of management media and experimenting with new ones in order to maintain and increase public support, sometimes discovering innovative methods to communicate more effectively. In the present, the pursuit of ratings and profit from the *fourth power in the state*, attract new challenges in terms of the use of the mass communication forms that have as target the citizen, media and socio-political vectors being polarized excessive on the direct and fast influence expository of the target audience. As a result, we observe an unprecedented discrepancy between the policies on the provision of necessary information as a citizen to effectively participate in the democratic life and those more oriented towards *infoshow*, characterizing a more commercial media.

Voice, music, verbal information, sound effects, silence and prosody are all fundamental elements that are operating in public communication. Despite the fact that verbal information is of utmost relevancy, prosody is what defines the message and guides the communication partner with its speed rhythm, intonation and its variation in accent and emphasis. Prosody, as part of linguistics, examines certain aspects of spoken language that reflects some features of the verbal message related to the emotional or intentional state of the speaker. Also, the prosodic component of a verbal message signals the presence of some aesthetic or philosophical categories, the speaker's concentration within the communication on certain ideas or structures with a differentiation or persuasion role. In this context, restrictions can be given by grammatical rules, semantics of the edited text or the capabilities and vocal peculiarities of the speaker.

A mathematical model that increases the efficiency of a complex verbal information transfer, in the presence of restrictions, is based on several conditions and criteria that a point in the allowable solutions space has to satisfy in order to qualify as the best solution. In essence we are dealing with a mathematical programming problem, which consists in

determining the values of the variables vector $x \in R^n$, which has achieved the minimum target of a function *f(x)*, under conditions which must satisfy an array of restrictions (equalities, inequalities or simple edges). Following the procedure in which the conditions are resolved for a point x* to qualifying as the solution of the problem, different algorithms for this method of approach are obtained. To develop an optimisation model of direct speech, there are a number of algorithms with a high degree of complexity. These algorithms are based on high performance computing techniques, which take into account the geometry admissibility domain, the presence of nonlinearities and even of non-convex functions, for the objective function as well for restrictions or lack of feasibility or boundlessness of the problem.

There are three different approaches to obtain optimal conditions. The first one is based on separation and support of convex sets theorems (theorems of the Hausdorff type), the second is based on penalty functions and the third and final approach is based on the classical theory of Lagrange multipliers. These optimal conditions, assuming differentiability, are the Karush-Kuhn-Tucker conditions. In addition to their intrinsic value to characterise the optimal solutions of a convex programming problem, they define the theoretical foundations for the development and analysis of mathematical programming algorithms [1].

At the vocal signal level, the prosodic and emotional descriptions are analyzed with the help of some nonlinear models, based on some features extracted from the vocal wave, respectively by the contour of the fundamental frequency, period and energy of sound segments, the period of non-sonorous segments and breaks, voice timbre etc. In the synthesis of expressive speech, the emotional level is a very important element. Unlike other classical analyses that were focused on meshing emotional states, this study focuses on finding ways to improve the discourse from an emotional point of view, regardless of the classification of its intensity: strong, weak or medium.

A number of studies have shown that there is a strong correlation between emotions and stress. For example, the emphasis is the most prominent element found in a speech or "semantic concentration", which is the central factor that shows the attitude of the speaker. Contextual information derived from the linguistic particularities is also very important for the emotional type of expression. To determine an optimum point for the Prosodic type expression, it will operate on local pitch spectral pattern through a method of dynamic control, using Lagrange multipliers method and variational calculus.

2. PROSODIC ANALYSIS TECHNIQUES OF THE VOICE SIGNAL

Interpersonal communication is almost exclusively based on affective and emotional components. The detection of emotions can be achieved by analyzing facial expressions, gestures, physiological characteristics, but mainly, from the speech process. For the recognition of the nature of emotions and emotional level from verbal communication act, it is not very important as to know what is communicated but, mainly, **how** communication is realized. Amplitude, intensity, intonation, rhythm are intrinsic features of the voice from which identification and extraction of emotion can be achieved. It's about the so-called prosodic component of speaking. For example, a low vocal intensity could denote low motivation, sign of the existence of sadness or disgust feelings. On the other hand, a high intensity or an alert speech pace could mean the existence of nervousness or fear. Current automated systems of speech analysis can operate with about 17 vocal features, the accuracy hovering around the level of 80%.

In general, the studies on prosody focus on the evolution of intensity and fundamental frequency curves, but also on temporal features - rhythm, marked by breaks and the duration of syllables. With sensitive differences from a model to another, syllables can be acute, grave, infra-grave or over-acute. Outside prosodic transcription, each

prosogram, namely the visual representation of sound sequences, allows sound intensity and the speech frequency to be highlighted. The processing of prosodic aspects of verbal communication follows, in essence, the identification of some patterns and of some rules that describe the in-time evolution of the elements that present these characteristics existent in a voice signal.

For example, in studying the evolution of speech characteristics and of its synthesis, scaling algorithms depending on time or on sound level are intensively used. The aim of voice modification by scaling depending on time is the recovery of speech speed without the modification of its original content. The alteration in time can be made uniform by changing the speed for some factors/features, according to the prosody or sound characteristics, for different parts of the speech. A function to scale depending on time, named time-warping function, assigns time series in the original signal for entering in correspondence with the time series of the new signal. The unequally scaling depending on time can increase intelligibility of the voice signal under analysis. The same technique is used in the synthesizing concatenation of speech, where voice properties segments subject to concatenation are changed in accordance to the language restrictions. A good technique for voice signal processing, from a prosodic point of view, is NSM (*Non Linear Springing Method*) [2]. This method, which operates in the domain of time, is based on the prosody prediction (prosodic matrix) and on signal processing and generates, as a result, a synthesized speech signal, which already contains the targeted prosody. In order to maintain the sound length, some periods must be repeated or omitted. In this way, there exists a control of sound frequency and duration. NSM algorithm contains descriptors of the fundamental period (pitch markers), a prosodic matrix and a re-sampling method with variable frequency. As initialization, the program computes the number of periods for every vocal waveform:

$$N_{nr.per} = \int_{t_i}^{t_f} f_{0\,prd}(t)dt \qquad (1)$$

where $f_{0\,prd}(t)$ represents the predicted intonation curve on which it operates the prosodic predictor, and t_i and t_f represent the initial moment, respectively, final moment of evolution for the selected sound. The output signal is determined by knowing the sound duration, the sampling frequency and the number of samples. Then, it follows the concatenation operation of needed bi-phonemes and the calculus of each signal periods. The manipulation of the main period's number is performed in concordance with the pitch markers. Obtaining the concatenated signal s_L from a number of L-samples, re-sampling is subsequent necessary in N samples for s_N.

The voice signal s_N will have the wavelength and intonation curve determined by the prosodic descriptor. For the fitting of the intonation curve to the predicted one, the re-sampling has to be done in accordance to the control curve $f_{0\,ccd}(t)$:

$$f_{0ccd}(t) = c(f_{0\,prd}(t), f_{0L}(t)) \qquad (2)$$

where $f_{0\,L}(t)$ is the intonation curve of s_L. The function *c* correlates both intonation curves of (2) and will calculate a third one, which is named re-sampling curve. An advantage of this method is the continuity between fundamental periods, without an angular point occurrence. Also, the obtaining of the intonation curve can be done with great accuracy. As disadvantages of the NSM method there can be mentioned the need of a high volume of data entry, the fundamental period markers, the prosody descriptor markers and especially, the fundamental period's deformation fact, that sometimes leads to voice signal distortion [3].

In recent years, new stochastic methods for modeling and transformation on short time intervals both, for spectral parameters and for prosodic features have been explored. In regard to the application of the spectral

conversion technique GMM (Gaussian Mixture Model), it was considered that this one alone is not sufficient to describe the emotional state. In this case it is better to use the hybrid model of conversion pitches GMM-CART, where the last one is a model based on the classification and regression tree. In the same direction of analysis, it is important the conversion unified model that uses a temporal pre-set Bi-HMM (Hidden Markov Model). This is used for converting the spectrums and decision trees in order to transform the syllable formants segments depending on the context and for each type of emotion. The system of analysis and processing of emotional expression is comprised of three main modules. The stage of spectral conversion generates, as output, a signal that contains the prosodic source and converted spectrum. The second stage is the modification of the duration of phonemes using CART relative trees. Finally, within the third module, the converted durations are used in generating the pitch contour for the whole speech, using a syllables sequence from HMM [4].

The emotional degrees are very important elements, tracked in synthesizing the speech expressiveness (*expressive speech synthesis*), indifferent to the use of classifications depending on terms like “positive” , ”negative" , ”poor" ,“neutral” or “stronger". Given that these aim at the framing of utterance prosody in typical emotional categories such as: fear, sadness, anger or joy.

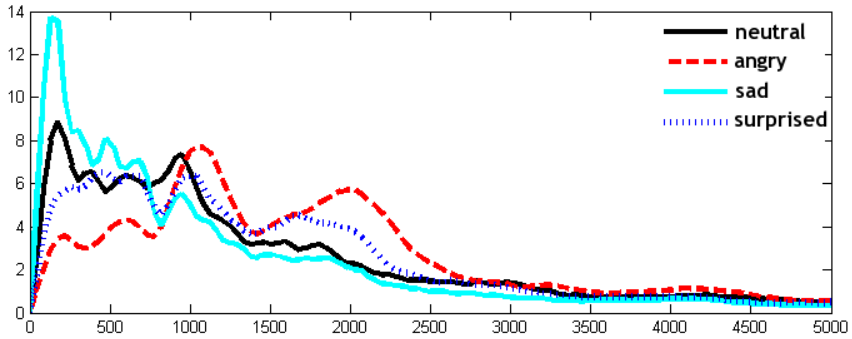

Fig. 1 - LATS analysis of phoneme /ae/ in the corpus of entrainment for different types emotional of speech

In most applications of analysis of sentiments / emotions, it is required to study diverse opinions, from several subjects engaged in the communication act, each opinion attesting a different level of subjectivity [5]. In this domain, one of the most interesting and up-to-date methods of message characteristics transformation, in the sense of bijection dependencies from the utterance characteristics and the text ones, is the LSI method (Latent Semantic Indexing), which transforms the space text in a new system of axes which is a linear combination among the original features of the words. The main disadvantage of the LSI method is the fact that it is an unsupervised technique that cannot realize the basic generated distributions. The FS methods family - of characteristics selection - also includes the techniques based on HMM and LDA models (Latent Dirichlet Allocation).

3. THE CURRENT METHODS IN PROSODIC MODELLING OF SPEECH

Currently, there are several effective methods of prosodic modeling of speech. These models are LMM (Linear Modification Model), GMM (Gaussian Mix Model) and CRT (Classification by Regression Tree). The LMM technique can perform a direct representation of frequency contours F0 and syllabic duration given by the sound

distribution of emotional speech (F0- peak level, F0-basic level, duration, intensity). A more detailed analysis shows that emotional speech is linked to the level of stress and to the information shown linguistically. Unlike LMM, the GMM and CRT methods try to map the prosody distribution of neutral speech and speaking with emotion. The studies show that out of all three methods, LMM offers the least desirable results [6].

The GMM method is more suitable for lower sets of training, while the CART method confers a better analysis of utterances in a more complex emotional context, if provided bigger training corpora and more contextual appropriated. There are also interesting implementations in the Emotional Speech Synthesis. For example, some researchers [8] added emotional control parameters to basic methods for sound analyzing, with remarkable results. Other researchers [9] did determinations on emotional speech, by operating a direct "black box" type, using an editor of acoustic and visual characteristics. Recently, good results have been obtained by using some prosodic corpora of large size. An interesting system [10] in expressive speech synthesis was conducted using a base of spoken texts recorded over a period of 5 years and that has led to impressive results. Other researchers [11] generated a motor TTS (text-to-speech) for expressive interpretation of texts. This could be programmed, by language annotation of speech synthesis, to use a variety of expressive styles arising from ten hours of interpretation of emotional "neutral" sentences. Finally, some researchers have used an array of emotionally keywords by using a TTS type system [12]. It was found that, in general, the emotional state of the utterance is mostly determined by factors related to the text made available for playback. One can say that the emotional state can be considered as generated by an emotional vector. Starting from this idea, unlike traditional methods, we can label text corpora according to four categories: strong, normal, weak and unpleasant, corresponding to the following status: happiness, sadness, fear and anger. The term “neutral speech” is used as source reference (for calibration) and it is not annotated according to the categories above.

In this manner, we can test prosodic conversion methods that aim to transform the prosody characteristics, duration or intensity of a certain utterances, to obtain an emotional type superior utterance. In this context you can use combined methods: LMM, GMM and CART. Because the LMM method does not provide very good outputs, a parallel application of the GMM and CART methods is often imposed. The GMM method achieved the mapping of the prosodic features distribution from a neutral state to various kinds of emotions, while CART model establishes some links between linguistic and prosodic characteristics. In contrast to the LMM method, the GMM and CART methods can not directly use the frequency F0 contours, therefore a special model must be introduced. It is based on the idea that the observed F0 values are not language units in itself. They constitute themselves surface achievements of functional language units like tone and accents. A recent analysis shows that LMM direct method offers the worst results of all three methods. The GMM method is suitable on corpora of short trainings, while CART proves more suitable for longer corpora [13]. Overall, the more profound analyses of emotions and stress in the speech show that they are closely linked to prosodic distribution.

4. THE METHOD OF PROSODIC CONVERSION

In the world there are some languages or dialects with emphasised tonal form, in which one syllable with a different tone can represent a different morpheme. There are four types of reference tone: high, increasing, low and decreasing. They manifest mainly in the contours of the F0 frequency. They are using two types of commands: with impulse shape, which gives an increased tone to the global intonation, and the accent command with step form, which confers great emphasis given by specific oscillations of each word. For example, the STEM-ML system, developed by Bell Labs, is a labelling system in which the

F0 contours are described by tags, including accent markers for local tone and level markers for global expression curves [14].

The problem is the difficulty of all these models to establish a relation between commands tags and different utterances. In the pitch identification model, the variations of its contour representation result not only from underlining the units pitch (syllables from the respective language), but from articulatory constraints. Local pitches are defined as the smallest units associated to functional linguistic units and these pitches can be static (high or low) or dynamic (with specification of motion, capturing their rapid increase or decrease). We consider that the local pitches are very weakly bound to the particularities of language but are closely related to the level of involvement in communication and the manner of control to the emotional component. The rules for implementing the local pitch model are based on possible articulatory constraints in F0 contour generation. We consider that the generation fundamental frequency of F0 contour is a process of continuous approximation of the local pitches along syllables. When the syllable limit is reached, the new approximation starts for the next syllable, with a new local pitch. We consider the time of a syllable as [0, D]. The LPM, local pitch model, is described by the following equations:

$$\begin{aligned} &\tau(t) = at + b \\ &y(t) = \beta \cdot e^{-\theta t} + at + b \\ &\theta \geq 0 \\ &0 \leq t \leq D \end{aligned} \qquad (3)$$

where $\tau(t)$ is the support of local pitch and $y(t)$ is the equation of F0 contour. The parameters a and b represent the slope and the intercept home for the local pitch. These two parameters describe the target intonation of the speaker, which can be totally different from the overall contour F0. The coefficient β is a parameter which measures the distance from a general outline of F0 and local pitch fixed at time $t=0$. The parameter θ describes how quickly the local pitch oscillates. As the value of θ is higher, as bigger the oscillation speed is. So, a model of local pitch for one syllable can be represented as a set of parameters (a, b, β, θ). Closely related with physiological limits of the human vocal tract, these parameters impose restrictions such as the maximum range of pitch or maximum speed of pitch change [15].

5. PROSODIC OPTIMISATION BY DYNAMIC CONTROL OF LOCAL PITCH MODEL (LPM)

At the speech level, the prosodic and emotional descriptions are shaped with the help of nonlinear systems, on the basis of some parameters extracted from the spectrum of voice-wave, respectively, the fundamental frequency outline, time and energy duration of non-acoustic segments and of breaks, voice timbre etc.

In general, optimal control problems associated to the dynamic evolutionary systems, as the process of free speech, consist of the intervention on their evolution with commands that lead to the satisfaction of performance indices. The choice of operating for spectral model of local pitch by dynamic control method is justified by the fact that the state described by this voice spectral model reflects very accurately the emotional state of the operator or, at least, the emotional charge of the issued informative content. The intention of operating on this spectral model is due to the idea that utterance prosody is reflecting, in turn, in pitch developments due to the correlation with the semantics of the communicated message. By reflecting accordingly to the local pitch pattern, we will have to do with formalization of the LPM model:

$$\dot{x} = f(t, \tau(t), u(t)), 0 \leq t \leq D \qquad (4)$$

where $\tau: [0, D] \rightarrow R^m$, $m \geq 1$, describes the state of the system at the time t € [0, D].

We define the function $u: [0, D] \rightarrow R^n$, $n \geq 1$, describing the decision of the process routing the dynamic speech/utterance system, which is taken at the time $t \in [0, D]$

and $\dot{\tau} = \frac{d\tau}{dt}$, for evolutionary process (4). The components $\tau_1, \tau_2, \ldots, \tau_m$ of the vector $\tau \in R^m$ are state variables and the components $u_1, u_2, \ldots, u_n$ of the vector $u \in R^n$ are the control variables which we consider directly associated with changes imposed by the volitional component of the speaker. The state variable also satisfies the initial condition:

$$\tau(0) = \tau^0 \tag{5}$$

of the initial entry point. The control variables are put under some restrictions described by functions (regulated) $g_i: R \to R$, $g_i(u(t)) \le 0$, t $\in$ [0, D], which set if a command $u(t)$ is or not admissible at time t $\in$ [0, D]. It can be noted that the conditions of regularity are satisfied:

(i) $\tau(t)$ is a continuous function and differentiable on parts

(ii) $u(t)$ is a continuous function on parts

(iii) function f is continuous in its arguments and differentiable in relation with t and τ

A problem of optimal control, or of speech control optimizing, associated to the previous process, consists of the admissible taken decision $u(t)$ such that a nonlinear state functional:

$$F(u) = \int_0^D f_0(t, \tau(t), u(t))dt \tag{6}$$

to be minimal (maximum) on states set $\tau(t)$, solutions of the system:

$$\begin{cases} \dot{\tau} = f(t, \tau(t), u(t)) \\ \tau(0) = \tau^0 \end{cases} \tag{7}$$

If the optimum criteria is given through a linear functional, then - through a variable shift in system status - the optimal criterion is written in a linear form:

$$\tau_0(t) = \int_0^t f_0(s, \tau(s), u(s)ds \tag{8}$$

Resulting in:

$$\tau_0(0) = 0 \text{ și } \dot{\tau}_0(t) = f_0(t, \tau(t), u(t)) \tag{9}$$

So the problem of optimal control has the form:

$$(\text{PCO}) \begin{cases} \dot{\tau} = f(t, \tau(t), u(t)) \\ \tau(0) = \tau^0 \\ u(t) = \beta \cdot e^{-\theta t} + at + b = \beta \cdot e^{-\theta t} + \tau(t) \\ J = \sum_{i=1}^{m} c_i \tau_i(D) = \langle c, \tau(D) \rangle \end{cases} \tag{10}$$

with $0 \le t \le D$; where $\tau \in R^{m+1}$ has the components $\tau_1, \tau_2, \ldots, \tau_m$, f $\in R^{m+1}$ has the components $f_0, \ldots, f_m$ and:

$$J = \tau_0(D) = (c, \tau(D)), \quad c = (1,0,...,0); \quad c \in R^{m+1} \tag{11}$$

The determination of the best solutions is done on a set of functions, namely on the set of the *problem Cauchy PCO* solutions, which are obtained by functional parameter $u(t)$ variation. We use the method of Lagrange multipliers, where multipliers $l_1, l_2, \ldots, l_m$ will be functions by $t \in [0, D]$. So for any $t \in [0, D]$ we build the Lagrange function:

$$L(\tau, u; l) = J[\tau(D)] + \sum_{i=1}^{m} l_i(t)[f_i(t, \tau(t), u(t) - \dot{\tau}_i \tag{12}$$

to which we attach the nonlinear functional:

$$L = J[\tau(D)] + \int_0^D \langle l(t), f(t, \tau(t), u(t) - \dot{\tau} \rangle dt \tag{13}$$

where $\langle \cdot, \cdot \rangle$ represents the scalar dot in R^m. By integrating by parts we deduce that:

$$\int_0^D \langle l(t), \dot{\tau} \rangle dt = \langle l(t), \tau(t) \rangle \Big|_0^D - \int_0^D \langle \dot{l}, \tau(t) \rangle dt \tag{14}$$

which - replaced in L - produces:

$$L = J[\tau(D)] - \langle l(t), \tau(t) \rangle \Big|_0^D + \int_0^D [H + \langle \dot{l}, \tau(t) \rangle] dt \tag{15}$$

We calculate the variation for $L(\tau)$ to a variation $\delta\tau$ for the state caused by some variation δu of the command, i.e. approximately:

$$\Delta L = L(\tau+\delta\tau) - L(\tau) = [J(\tau+\delta\tau) - J(\tau) - \langle l(t), \delta\tau\rangle]_{t=D} + \int_0^D [H(\tau+\delta\tau, u; l) - H(\tau, u; l) +$$

$$\langle \dot{p}, \delta\tau\rangle]dt + \|\delta\tau\| o(u, \delta\tau) = [\frac{\partial J}{\partial \tau} - l(t)\delta\tau]_{t=D} + \int_0^D [(\frac{\partial H}{\partial \tau} + \dot{p})\delta\tau + \frac{\partial H}{\partial u}\delta\tau]dt + \|\delta\tau\| o(u, \delta\tau) \quad (16)$$

Where $o(u,\delta\tau) \to 0$ *for* $\delta\tau \to 0$ and where:

$$\begin{cases} \dfrac{\partial H}{\partial \tau} = \left(\dfrac{\partial H}{\partial \tau_1}, ..., \dfrac{\partial H}{\partial \tau_m}\right) \\ \dfrac{\partial J}{\partial \tau} = \left(\dfrac{\partial J}{\partial \tau_1}, ..., \dfrac{\partial J}{\partial \tau_m}\right) \\ \dfrac{\partial H}{\partial u} = \left(\dfrac{\partial H}{\partial u_1}, ..., \dfrac{\partial H}{\partial u_n}\right) \end{cases} \quad (17)$$

Also: $\tau(0) = \tau^0$ being given, we have that $\delta\tau|_{t=0} = 0$. For getting rid of the independent term of $\delta\tau$ from ΔL's expression, we choose one conveniently, namely:

$$\begin{cases} \dot{p} = -\dfrac{\partial H}{\partial \tau} \\ l(D) = \dfrac{\partial J}{\partial \tau}\Big|_{t=D} \end{cases} \quad (18)$$

In this case:

$$\Delta L = \int_0^D \left(\frac{\partial H}{\partial u}\delta\tau\right)dt + \|\delta\tau\| o(u, \delta\tau) \quad (19)$$

and because the variation δu is variable, we deduce as in the finite dimensional case that a necessary condition u^* to be an optimal command is as:

$$\left(\frac{\partial H}{\partial u_i}\right)_{u=u^*} = 0, \quad 1 \le i \le n, \quad t \in [0, D] \quad (20)$$

It is observed that once selected the command $u(t)$, the vector of Lagrange multipliers, $l(t)$, is unique determined from the Cauchy problem:

$$(PC) \begin{cases} \dot{p} = -\dfrac{\partial H}{\partial \tau} \\ l(D) = \dfrac{\partial J}{\partial \tau}\Big|_{t=D} \end{cases} \quad (21)$$

Also, it is observed that $\tau(t)$ and $l(t)$ are reciprocal dual, in the sense that:

$$\frac{\partial \tau}{\partial t} = \frac{\partial H}{\partial l} \text{ and } \frac{\partial l}{\partial t} = -\frac{\partial H}{\partial \tau} \quad (22)$$

Namely, they appear as solutions of a system in the form of classical Hamilton-Jacobi.

In the case in which the command of speech is subject of admissibility conditions $g_i(u(t)) \le 0,\ 1 \le i \le p,\ t \in [0, D]$, we can no longer consider some variations δu absolute, but we must choose such that $u^* + \delta u$ a supplementary command admissible for $\|\delta u\|$ to be sufficiently small. Then, the same as in the finite dimensional, we are led to a variational inequality:

$$\int_0^D [\frac{\partial H}{\partial u}(\tau, u; l)\delta u]dt \ge 0 \quad (23)$$

This must happen for any admissible variation δu. It shows that, for any admissible δu and any t € [0, D], we have:

$$H(\tau^*, u^* + \delta u; l^*) \ge H(\tau^*, u^*; l^*) \quad (24)$$

where τ^* and l^* are the solutions of the problem (PCO 10.1-10.4) and (PC) for $u=u^*$. So, for an admissible command of utterance $u(t) \in R^n$, for J to be minimum (maximum) in the case of dynamic communicative process (PCO), it is necessary to be a function $l(t) \in R^m$ solution of the system (PC) so that $u(t)$ to minimize (maximize) the Hamiltonian $H(\tau(t), u(t))$ for every $t \in [0, D]$.

6. CONCLUSIONS

A message which impacts you will be the one with a balanced structure, from the prosodic point of view, the asymmetric messages, from a linguistic but also phonetics perspective, risking not to achieve their goal, even partially.

By analyzing the characteristics of prosodic speech, with the help of the dynamic prosody conversion method, the expressivity evolution of utterance was followed from an emotional point of view, indifferent to the intensity level.

Emotional speech differs of neutral speech, not only in prosodic characteristics, but - in the same measure - in spectral characteristics. For the analysis of expressivity, some characteristics that describe the processes located in the vocal tract must be taken into account, with impact on the voice specificity.

The local pitch model parameters (a, b, β and λ), respectively neutral parameters and emotional parameters, describe, very well, the speech through a nonlinear mathematical system. The local pitch parameters can be extracted in a relatively simple way from the pitch contour of each syllable/phoneme and finally, the functions of mapping the parameters a, b, β and λ can be estimated using the regression model, CART. While GMM uses only acoustic characteristics, the CART model permits the integration of mapping linguistic characteristics.

Choosing operation on the local pitch model by dynamic control method was justified by the fact that the state described by this spectral voice model reflects, very accurately, the emotional state of the operator or, at least, the emotional charge of the issued information.

Emotional support described and materialized by the use in communication of prosodic components reflected in the LPM model has been optimized by the application of the dynamic control method.

A prosodic optimization was performed through variational modeling of functional parameter u(t), which designates the decision function in the process of controlling the dynamic system of speech/utterance in verbal communication space.

Using the method of Lagrange multipliers and variational calculus, the determination of optimal solutions - in accordance to the evolution of the state parameters and utterance control parameters, introduced in the local pitch spectral model was successful.

REFERENCES

[1] Gobl, C., Chasaide, A.N., *The role of voice quality in communicating emotion, mood and attitude*, Speech Commun., vol. 40, 2003, pp. 189–212.

[2] Bodo, A. Zs., Buza, O.,Toderean, G., *"TTS Experiments: Romanian Prosody"*, Acta Tehnica Napocensis, Electronics and Telecommunications, Technical University of Cluj-Napoca, Cluj-Napoca, Romania, vol. 50, No. 1, 2009, p.31-36.

[3] Previdi, F., Lovera, M., *Identification of a class of nonlinear parametrically varying models*, Int. J. Adapt. Control Signal Process, vol. 17, 2003, pp. 33–50.

[4] Nar, V.V., Cheeran, A.N., Banerjee, S., *Verification of TD-PSOLA for Implementing Voice Modification*, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 3, Issue 3, May-Jun 2013, pp.461-465.

[5] Amir, N., *Classifying emotions in speech: A comparison of methods*, in Proc. Eurospeech. Holon, Israel, 2001, pp. 127–130.

[6] Falk, T.H., Chan, W.-Y., *A sequential feature selection algorithm for GMM-based speech quality estimation*, 13th European

Signal Processing Conference EUSIPCO, Antalya, Turkey, 4-8 sept. 2005.
[7] Sundberg, J., Nordenberg, M., *Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech*, The Journal of the Acoustical Society of America, 2006, 120(1), p.453-457.
[8] De Melo, C., Paiva, A., *Environment expression: Expressing emotions through cameras, lights and music*, In Proceedings of Affective Computing Intelligent Interaction (ACII'05), 2005, pp. 715–722.
[9] Campbell, N., *Perception of affect in speech—Toward an automatic processing of paralinguistic information in spoken conversation*, in Proc. ICSLP, Jeju, Korea, Oct. 2004, pp. 881–884.
[10] Tato, R., Santos, R., Kompe, R., Pardo, J.M., *Emotional space improves emotion recognition*, in Proc. ICSLP, Denver, CO, Sep. 2002, pp. 2029–2032.
[11] Schröder, M., Breuer, S., *XML representation languages as a way of interconnecting TTS modules*, in Proc. ICSLP, Jeju, Korea, 2004, pp. 1889–1892.
[12] Zimmerer, F., Reetz, H., Lahiri, A., *Place assimilation across words in running speech: Corpus analysis and perception*, The Journal of the Acoustical Society of America, Vol.125, No.4, april 2009, pp.2307-2322.
[13] Kang, Y., Shuang, Z., Tao, J., Zhang, W., Xu, B., *A hybrid GMM and codebook mapping method for spectral conversion*, in Proc. 1st Int. Conf. Affective Comput. Intell. Interaction, 2005, pp. 303–310.
[14] Kochanski, G.P., Shih, C., *STEM-ML: Language independent prosody description*, in Proc. ICSLP, Beijing, China, 2000, pp. 239–242.
[15] Xu, Y., Wang, Q.E., *Pitch targets and their realization: Evidence from mandarin chinese*, Speech Commun., vol. 33, 2001, pp. 319–337.
[16] Inanoglu, Z., Young, S., *A System for Transforming the Emotion in Speech: Combining Data-Driven Conversion Techniques for Prosody and Voice Quality"*, Interspeech conference, Antwerp, Belgium, aug.27-31, 2007, pp.2.