# Prosody of Expressive Speech: Bringing to Light Some Discursive Situation Influences

Ioana Suciu[1], Ioannis Kanellos[2], Thierry Moudenc[1]

[1] TECH\SSTP\VMI, France Telecom R&D, 22300 Lannion, France
{Ioana.Suciu, Thierry.Moudenc} @Orange-ftgroup.com
[2] Computer Science Department, Ecole Nationale Supérieure des Télécommunications,
Technopôle de Brest-Iroise, CS 83818, 29238 Brest Cedex 3, France
Ioannis.Kanellos @Enst-bretagne.fr

**Abstract.** We propose in this paper a computational view of expressive speech synthesis in the general framework of human-computer interactions. We start by reviewing two essential aspects of a formal representation for the expressive phenomena in speech: the prosodic vector $P$ and the textual-extrinsic characteristics vector $S$. Then, we point out some influences of one of the $S$ dimensions, $ds$ (discursive situation), on the $P$ prosodic realisation of the synthesized speech. We describe these influences by the $O_{ds\text{-}P}$ algebraic applications over three phonological levels and illustrate them by examples. We discuss some prosodic convergence and divergence scenarios of the $O_{ds\text{-}P}$ application effect and the two other dimensions of $S$ vector: $O_{tg\text{-}P}$ and $O_{rp\text{-}P}$. We conclude by setting our work in an expressive-oriented computing background.

**Keywords:** speech synthesis, expressivity, prosody, discursive situation.

## 1 Introduction

Speech technologies have expanded interaction modalities between humans and computer-supported communicational artefacts (robots, PDAs, mobile phones…) and services. Improvements in speech synthesis have gradually redefined the computers' position with regard to humans and consequently, the humans' position with regard to computers. New paradigms of human-computer interaction have been established and new ergonomic exigencies have been identified on the subject of synthesized speech quality: once its intelligibility assured and thus distances in its reception reduced, the artefactual speech was required to be more "natural", *i.e.* more touching, more moving, more affective, more expressive. Computing expressive speech has therefore become a wide project aiming to improve the quality of human-computer interaction.

Nevertheless, modelling and implementing speech expressivity are far of being some trivial or obvious programs. This is certainly due to the difficulty of defining and measuring expressivity, but also to its misunderstanding and its rather frequent reduction to some emotional dimensions. Whereas multiple research investments (dealing most of the time with basic emotion typology) were settled down in the latter framework [1], [8], [12], we propose here an approach handling speech expressivity

in a more general framework. We prefer talking about "expressive" rather than "affective" speech modelling, as we suppose the second to be merely a particular case of the former. Thus, while speech affectivity seems to deal with emotion expression in speech and voice, speech expressivity outlines a wider range of significant speech phenomena. Conveying dimensions often ignored or deliberately neglected in synthesized speech production, these expressive phenomena are essential for every human reception and interpretation of speech. Within a general framework describing a formal representation of expressive speech [10], we handle some pragmatic characteristics [5] on the view of expressive speech synthesis. More specifically, following a computational perspective, we analyse and interpret some discursive situation (elocution) influence on the prosody of expressive speech.

## 2   A Computational Representation of Speech Expressivity Review

Based on a model proposed in [4], we have pointed out in previous work [10] a formal representation for the expressive phenomena in speech. Thus, the formal signature of an expressive form is defined as a multi-dimensional vector, such as:

$$E_{sit} = <id, U, P, S> . \qquad (1)$$

In (1), $id$ is the form identifier, while $U$, $P$ and $S$ are, respectively, the linguistic unit description vector, the prosodic representation vector and the textual-extrinsic characteristics vector; all of these correspond to the given form $id$. Interested reader may find in the same reference [10], more details about these algebraic descriptors. In this paper, we are particularly interested on the $P$ and $S$ dimensions of the expressive signature vector $E_{sit}$ that we precise below:

$$P = <F, T, I> . \qquad (2)$$

In (2), the left term is a three-dimension vector relating to the three main prosodic parameters [4]: thus, $F$ corresponds to the frequency (melody), $T$ to the temporal and $I$ to the intensity movements in speech. They explicit prosodic phenomena presented at three phonological levels [5]: syllable *(syl)*, syntagm *(syn)* and phrasal group *(phg)*.

Defined as overlapping structures, the *F, T* and *I* prosodic dimensions refer not directly to some numerical values (expressed, for example, in Hz, ms or dB), like in the pure phonetic approaches, but to their formal representations, as in the (surface) phonological ones. These representations are literal (inspired by INTSINT alphabet [2]) and numerical descriptions of *F, T* and *I* different level movements. Therefore, they translate either local (for instance, the lexical emphasis) or global (as the discursive situation influence on the discourse) expressive phenomena in speech.
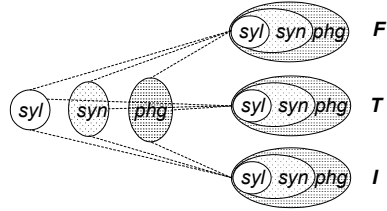
Fig. 1. The *P* vector dimensions are three prosodic overlapping structures: *F, T,* and *I.*

Long before it is explicitly identified, speech expressivity is situated, by locating the linguistic data it involves in a complex semiotic space. Such data furnish precious introductory clues for the receptor's interpretative strategies [5], [6]. They are formally represented by the textual-extrinsic characteristics vector *S,* as below:

$$S = <tg, \, ds \, ,rp>. \tag{3}$$

Here, *tg* specifies the textual genre (for example, "letter", "story", "horoscope", "receipt", "proverb"); *ds* makes reference to the discursive situation, informing about the speaker's permanent or temporary traits (such as "angry", "hysteric", "drunken", "ironical"…), about some rhetorical or argumentative intensions (as "trying to incite compassion", "searching to seduce"…) or even some situational-typed manner of speaking ("commentary on football", "airport-typed voice" and so on); and *rp* defines the reader's personal speaking profile (where such features as his gender, age, accent, personal way of speaking, social membership or level of language, are illustrated).
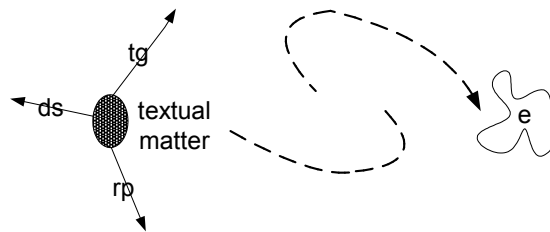


**Fig. 2.** *tg, ds* and *rp* extra-textual parameters situate each $E_{sit}$-type expressive form *e*

Certainly, other textual-extrinsic dimensions could be envisaged for expressive speech handling. But for computational reasons, we have reduced this semiotic space at only three dimensions.

# 3 Extra-textual influences on the prosodic parameters selection

## 3.1 First-order operators $O_{S-P}$

Essential for each linguistic oral expression, the three extra-textual characteristics mentioned above are implicit and comprehensive dimensions for a human speaker, but not for a machine. Their specification in the input of every speech systems and their consequent formal handling is therefore required for the expressive quality of the output speech. Perceptually, this quality is directly related to the prosodic realization of the artefactual speech. Therefore, in order to create a valid $E_{sit}$-type expressive form $e$, the choice of *tg, ds* and *rp* has an impact on the *F, T* and *I* determination (see the figure below):
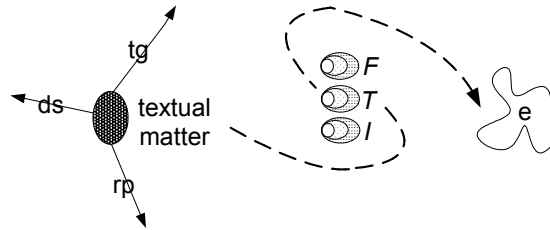


**Fig. 3.** The *S* parameters *tg, ds* and *rp* have influence on the *P* dimensions *F, T* and *I*

In computational terms, these influences are represented by an $O_{S-P}$ algebraic application over an expressive form $e$ (figure 4). We remind the reader [10] that the O-family applications are reversible unary transformations. They request as argument a unique $E_{sit}$-type vector $e_1$ and they return a different $E_{sit}$-type expressive vector, $e_2$.
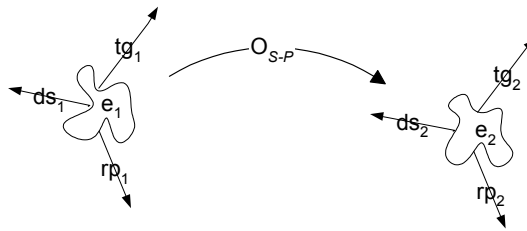


**Fig. 4.** An expressive form $e_1$ is transformed into $e_2$ by an $O_{S-P}$ application.

In view of (1), we can state here three first-order O-family applications: the $O_{S-P}$ operators, the $O_{U-P}$ operators and, more exceptionally, the $O_{P-P}$ operators (which are inherent for every $O_{S-P}$ or $O_{U-P}$ transformation). We have to mention that any $O_{S-P}$–

type transformation maintains unchanged the $U$ dimension of both the $e_1$ and the $e_2$ vector. Similarly, any $O_{U-P}$–type transformation keeps invariable the $S$ dimension of $e_1$ and $e_2$. Nevertheless, there are situations when $O_{P-P}$–type transformations may abolish the prosodic choice made by one or more of $U$ or $S$ dimensions (*i.e.* by the $O_{S-P}$ or $O_{U-P}$ operators). In this case, the $O_{P-P}$ application may provoke changes in the specified dimension(s) and consequently, in the global expressive signature (1). Hence, some inverse $O_{P-S}$ or $O_{P-U}$-type transformations may be induced.

### 3.2 Second-order operators $O_{ds-P}$

Even though it doesn't seem immediately obvious, the first-order $O_{S-P}$ decomposition into the second order operators $O_{ds-P}$, $O_{tg-P}$ and $O_{rp-P}$, has different consequences concerning their transformational impact on the $P$ dimensions. Their application effect varies depending on whether they have a convergent or a divergent result in terms of prosodic parameters selection.
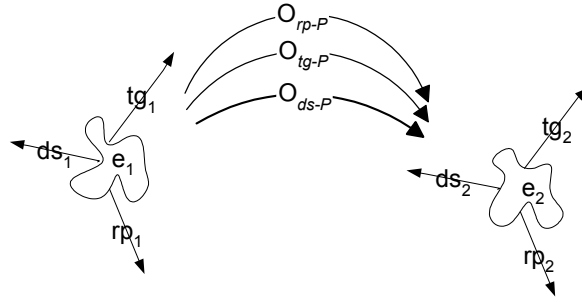


**Fig. 5.** The $O_{ds-P}$, $O_{tg-P}$ and $O_{rp-P}$ composed application on the $P$ dimension.

There are circumstances when $O_{ds-P}$, $O_{tg-P}$ and $O_{rp-P}$ application results are coherent and they produce a simple and non-ambiguous effect on the prosodic parameters. This relates to an integrative cooperative situation [9], where the three extra-textual characteristics determine a convergent choice on the $P$ dimensions. Hence, the *tg, ds* and *rp* semiotic dimensions are complementary effects on expressivity construction.

But there are cases when one or more of $S$ dimensions are prosodically in conflict with the others. This is typically the situation where divergent effects on the $F, T$ or $I$ parameters selection may be induced. Therefore, completely different expressive realisations of synthesized speech are possible. The prosodic outcome is either a global compromise between $O_{ds-P}$, $O_{tg-P}$ and $O_{rp-P}$ prosodic effect or a choice strategy result depending, for example, on an already-established preferential order over these operators (for instance, first $O_{ds-P}$, then $O_{tg-P}$ and finally, $O_{rp-P}$). This order relation translates somehow the privileged dimensions among the *tg, ds* and *rp* on the other(s) disadvantage. We point out in the following section some of the *ds* expressive impacts on the $P$ realisation of the synthesized speech, in the favour of *tg* and *rp* ones.

## 4 Privileged *ds* influences over *P* parameters

"It is now the time to talk about the elocution (here, we translate the elocution in more general terms by discursive situation *ds*). As disposing of one's discourse matter is not sufficient, the adoption of an appropriate manner of speaking is mandatory. It is there an important condition to give to discourse a good appearance." [3] In the same reference, we can notice the connection between the elocution (*ds*) and the three mentioned above dimensions of *P:* "This action lies in the voice, in the manner of speaking, which will be sometimes loud and piercing, sometimes soft and gentle and sometimes moderate or temperate. We must examine the way we serve it in order to express each state of spirit, the way we use intonations in order to render it high-pitched, deep or moderate by turns and the way we use some rhythms for each discursive circumstance, because there are three things to consider: the magnitude, the harmony and the rhythm". Hence, the magnitude relates to voice force *(I)*, the harmony to intonation degree *(F)* and the rhythm to articulated sounds duration *(T)*.

In order to illustrate some discursive situation influences on the prosodic realisation of speech, we have used the same working protocol settled down in [4]. According to this protocol, expressive signatures are extracted from recorded texts (here, our corpora is constituted of horoscopes and short stories) and represented in a (1)-type form. Once these forms acquired some $O_{ds-P}$, $O_{tg-P}$ and $O_{rp-P}$ influences scenarios are established for each of them, depending on the selected phonological levels (see section 2). In computational terms, these influences are translated either by some very-precise prosodic rules or by schematic global constraints on the *F, T* and *I* parameters selection. We are interested here on the $O_{ds-P}$ algebraic transformation by considering some prosodic convergence and divergence scenarios at its junction with the $O_{tg-P}$ and $O_{rp-P}$ application effects. We point it up by some examples.

Let us consider a so-called "typical" discursive situation (with no particular imposed effect on the choice of *P* parameters). This is classically the case when *tg* and *rp* dimensions of *S* vector are free to manifest their influence on the prosodic selection. Thus, while the textual genre imposes its global constraints, the reader's profile reveals some local or global particular determinations on the speech prosody. We can exemplify it by different prosodic realisations corresponding to two different textual genres of our corpora: a short story and a horoscope. Therefore, a macroscopic analysis reveals a global slow speech rate, a progressive rhythm (depending especially on the narrative structure) and a wide range of pitch variation in short stories oral expression. In opposite, horoscopes are described by a global fast speech rate, an almost steady rhythm with regular accentuations and a more normalized melodic range. Similarly, we can mention TV News, another particular textual genre *tg* with strong prosodic restrictions, characterized by very regular intonation patterns [7]. We can notice that in TV or radio news, the discursive situation influence is almost extinguished (only the "typical" *ds* is allowed) and the reader's profile is roughly standardized (most of the time, teached in some famous journalism school). Formally, these examples convey the prosodic $O_{ds-P}$ fading away in the favour of $O_{tg-P}$ and, on occasion, $O_{rp-P}$ application effects. Nevertheless, prosodic convergence of all textual-extrinsic influences is present, even if, in part, some *S* dimensions are vanished.

Furthermore, we can easily envisage some generic violation scenarios where the *ds* discursive situation influence on the speech prosody, partially or totally revoke the

textual genre effect (or even the reader's profile). This is typically the case when elocution choice abolishes the listener's expectations in terms of local or global prosody movements for a specific *tg* (which in a linguistic-like vocabulary correspond to some *tg* prosodic isotopies presumptions). We let the reader imagine a drunken, hysterical or a military-typed reading of a short story, as well as a raging, ironic or reproachful-tone performance of a horoscope. In the best set of circumstances, these situations will be considered as parodist ones (this case supposes prior detection of the referential *tg*). Otherwise, they will produce a weird or an incongruity effect among the audience. The same impression may arise in the case of radio news presented, for example, with an exaggerated joy, an overstated exuberance or even a scornful tone. This kind of parodies operates precisely in terms of violation of prosody expectations which the listener has culturally and socially acquired, adopted and activated because they made sense for him. Formally, this situation translates expressive divergence mentioned in the previous section. Here, $O_{tg-P}$ and $O_{rp-P}$ application effect is vanishing in the favour of $O_{ds-P}$ prosodic impact, which seems to be largely privileged.

Obviously, various expressive elocution scenarios situated between these two extremes can be imagined. Therefore, there are cases when the choice of the discursive situation may reinforce the already-selected prosodic parameters. Perceived neither as "typical" or parodist, these performances can be easily accepted as having a non-exaggerate *ds* for a specified *tg*. To illustrate, we let the reader imagine a seductive recital of a horoscope, a nasty reading of a story or a completely indifferent news presentation. Formally, this scenario corresponds to a convergent prosodic effect of second-order operators, with $O_{ds-P}$ underlining the $O_{tg-P}$ and $O_{rp-P}$ application.


## 5   Conclusion and further work

Expressivity is an essential performance for every human oral communication, as well in production as in reception and interpretation. However, technologists have largely ignored it and created ergonomically non-satisfying services and often frustrating experiences for people in contact with communicational artefacts. Multiple research investments in speech synthesis were established in order to make this interaction more "intelligent", more "close to the human", more natural, more expressive.

The ergonomic requirements of expressive speech production made engagements in speech simulation by machine reformulate a well-known AI validation framework: the human-computer imitation challenge of Turing contest [11]. However, long before these imitation scenarios will be generally ensured, humans will be sensibly confronted with gradually improved technological services and products. Designed for limited-domain applications, these artefacts can be appropriate communicational tools for a specific practice. Thus, a research program aiming to assure some human-computer speech indiscernibility in restricted contexts does not seem a utopia.

In this paper, we have been particularly interested on short story telling and horoscopes uttering. Based on an already-defined general model for expressive speech computing, we have pointed out the importance of discursive situation influences on the expressive quality of synthesized speech. More precisely, we have considered here its privileged impact, in the favour of the textual genre or speaker's profile ones, on

the determination of local or global prosodic parameters. Certainly, a wider range of expressive speech phenomena depending on a more refined discursive situation typology might be envisaged for a more vigorous validation protocol. Some other directions can also be explored in future studies: the textual genre constraints for the linguistic unit influence over the expressive prosodic vector ($O_{U\text{-}P}$), the prosody general dependence on the reader's profile ($O_{rp\text{-}P}$) etc. Enlightening significant manifestations in speech, these influences proved to be important study clues for the improving of the expressive speech quality.

New theories and technologies will certainly advance fundamental understanding of speech expressivity and its role in human experience. There is no doubt that in a little while, we shall be able to talk about a large-scale expressive-oriented computing that could offer benefits in a broad range of applications.

# References

1. Cahn, J.: Generating Expression in Synthesized Speech. Master's thesis. MIT (1989)
2. Hirst, D., Di Cristo, A., Espesser, R.: Level of representation and levels of analysis for the description of intonation systems. In Horne, M. (ed.): Prosody: Theory and Experiment. Kluwer Academic Press (2000) 51-87
3. Les Belles Lettres (ed.): Rhétorique, Aristote (1968)
4. Kanellos, I., Suciu, I., Moudenc, Th.: What about the text? Modelling global expressiveness in speech synthesis. Proceedings of the ICTTA'05, Damascus, Syria (2005) 177-178
5. Kanellos, I., Suciu I., Moudenc, Th.: Émotions et genres de locution. La reconstitution du pathos en synthèse vocale. In Rinn, M. (ed.) Le Pathos en action, France (2006)
6. Rastier, F.: Art et science du texte. Presses Universitaires de France. Paris (2001)
7. Rodero, E.: Analysis of intonation of news presentation on television, In ExLing-2006. Athens (2006) 209-212
8. Scherer K. R., Johnstone T., Klasmeyer G.: Vocal Expression of Emotion. In Davidson R. J., Scherer K. R., Goldsmith H. H. (Eds): Handbook of Affective Sciences (2003) 433-456
9. Schmidt, K.: Cooperative Work: A Conceptual Framework. In Rasmussen, J., Brehmer, J. Leplat (eds.): Distributed Decision Making (Cognitive Models for Cooperative Work). Chichester: John Wiley and Sons (1991)
10. Suciu, I., Kanellos, I., Moudenc, Th.: Formal expressive indiscernibility underlying a prosodic deformation model. In ExLing-2006. Athens (2006) 229-232
11. Turing, A.: Computing Machinery and Intelligence. *Mind* Volume 59 (1950) 433-460
12. Zovato, E., Pacchiotti, A., Quazza, S., Sandri, S.: Towards emotional speech synthesis: a rule based approach. 5th Speech Synthesis Workshop. Pittsburgh (2004)