Expressive text-to-speech approaches

Productive homologies and irreducible discrepancies between natural and singing voice synthesis modelling.

Ioannis Kanellos[1], Ioana Suciu[1&2], Thierry Moudenc[2]
[1] Computer Science Department, ENST Bretagne, Brest, France, ioannis.kanellos@enst-bretagne.fr
[2] DIH/IPS/VMI, France Telecom R&D, Lannion, France, {ioana.suciu, thierry.moudenc}@orange-ftgroup.com

***Abstract* — The core concern of this paper is the modelling and the tractability of expressiveness in natural voice synthesis. In the first part we quickly discuss the imponderable gap between natural and singing voice synthesis approaches. In the second part we outline a four level model and a corpus-based methodology in modelling expressive forms—an essential step towards expressive voice synthesis. We then try to contrast them with recurrent concerns in singing voice synthesis. We finally undertake a first reflection about a possible transposition of the approach to singing voice. We conclude with some program considerations in Research and Development for the singing voice synthesis, inspired from natural voice synthesis techniques.**

***Keywords* — Natural voice synthesis, expressiveness, corpus based voice synthesis.**

I. INTRODUCTION

There is undoubtedly a "significant" difference between natural and singing voice synthesis from the point of view of the text to speech (TTS) techniques they involve. In the first case, we search treats making the synthetic voice as near as possible to natural conditions of elocution; we designate this kind of proximity by the term of *expressiveness*; elocution is generally considered in the framework of various known discourse genres—or, at least, composite or idiosyncratic genres, recomposed by the linguistic competence of the receptor. In the second case, the voice is considered as a part of a general music composition system where expressiveness is thoroughly evaluated in terms of musical categories. In the first case, the principal objective is simulation. Without loss of generality, we can slightly perceive a discreet persistence of the Turing's imitation test: starting from a text, the machine has to be so natural that the hearer could not distinguish between a natural human elocution and its synthesized version. In the second, natural speech is merely secondary, sometimes even accessory or irrelevant, insofar as it is completely redefined into suitable musical genres. What is expressive in natural speech is not necessarily in music, and conversely. The musical construction acts as a global constraining system, external to linguistic competence and performance, that submits the voice expressiveness to irreducible handlings, inherent to musical production and reception schemata.

Of course, it could not be differently. Linguistic and musical systems are semiotic systems that entertain certainly tight relationships, but are built up on quite independent communication categories hard to unify. On the other hand, the voice synthesis concerned in both does not respond to the same practices. TTS systems are driven by production requirements where the conception and the development of technological bricks able to accomplish real-life services are pre-eminent; they are commonly motivated by industrial exigencies and configured by economical rationalities, in which they bear evidence and acquire pertinence. Singing voice synthesis, even if sometimes becomes the satellite of industrial objectives, still respects creation principles and practices and inherits its legitimacy and intelligibility only in artistic contexts, generally—but not exclusively—referring to live performance ecologies.

However, such a difference is not slight: it even has determinant implications for the approaches we choose when we deal with expressive voice synthesis, insofar as both the research and the development concerns are generally motivated by different application objectives. It might be thought that, liberated from the simulation exigency, the singing voice is more easy to treat as far as every sound, in particular, a synthesized voice—even not quite faithful—may be and function as musical feature. It might also be thought that the expressive discourse voice, as restricted and usually simulated from already existing forms, is straightforwardly reproducible. Currently, many voice synthesis systems are able to produce quite intelligible synthetic discourse. But more than intelligibility, the question is about expressiveness (see, for instance, [2] and below): discourse is not only *logos* but also *pathos* (and *ethos*). As a matter of fact, both questions are equivalent in complexity; they simply are in a different way difficult. An expressive natural voice superposes semiotic levels that are globally unknown and in multifaceted antagonisms. For instance: vocal, *epivocal* (sounds added to the vocal data and produced by the speaker) and *paravocal* (sounds coming from the real-life context) versus verbal information, incidence of rhythm in reception, morphology and/or syntax and/or semantics mutual determinations, *sociolectal* and *idiolectal* regulations, text genre and figures of style interdependencies, discursive situation and practice complementarities etc. On the other hand, a singing voice necessitates a wider range of parameters in order to respect genre principles; it does not suppose interaction scenarios between production and reception, it often has to find solutions to real-time constraints, it acquires sense and subtleties into musical traditions and genres that may be diachronically and cross-culturally extremely diversified—and yet not precisely understood—it calls composition and notation categories that are not precisely investigated etc. In brief: although both are human voices, singing and natural voice do not make reference to the same semiotic competences.

It seems already obvious that a model for natural voice synthesis cannot cover all features needed for the singing voice, and vice-versa. But partial recovering (in analysis or in methodology, in modelling or in implementation…) may be found adequate in special cases; they even may be epistemologically productive.

This is the dialectic idea of our paper. Let us see, at the very beginning, the way we work to model (in the spirit of [8]) and implement (in the spirit of [1]) an automatic expressive voice, which is nowadays a recurrent question in research and development. To begin, it is useful to remind that current systems are quite efficient in rendering an intelligible voice: the hearer understands the linguistic content without effort, but such a voice still lacks in relief (it is monotonous, flat, non sensible to context…). This last is clearer when the synthesis goes beyond the phrase level, and deals with real, long texts. The shift from phrase to text may be seen as evident, but it is not: it resumes alone the entire epistemological inquiring about the local/global regulations and the relevance of levels of analysis. No matter: in facing the expressiveness problem, we have to deal with texts and semantics of texts ([6]). This last means that we have to choose our levels of analysis to be coherent with the text nature. There is no synthesis system likely to do this for the moment.

II. Expressiveness in discourse

A. Model

In communication situations, expressiveness concerns certainly both production and reception; but, since it is validated only at the level of reception, it may be seen as a general interpretative problem ([4], [5] between others), inescapably related to a specific discursive practice. By "expressiveness" we owe to understand a class of complex semiotic extensions and refinements that the oral speech endows a piece of linguistic information by superposing to it various complementary contents—contents that the written text inhibits (affective, cognitive, social, intentional, evaluative, communicative, aesthetic etc.).

More technically, for a given text, expressiveness needs to be qualified on, at least, 3 dimensions: (a) text genre, (b) discursive situation and (c) reader's profile. Indeed, a novel is not read in the way a poem is declaimed or a political program is announced, we do not dictate a recipe the way we give voice to a love letter etc. (a); the same textual material may be said under different locution objectives and/or conditions that may be permanent or incidental (determined, anxious, angry, imploring, indifferent, tired, ironic, humoristic, solemn, hysterical, religious etc.) (b); and, of course, two different speakers will necessarily introduce proper specificities in speaking (idiolectal) depending on the linguistic performance each of them typically demonstrates (c). Thus, in order to acquire the necessary expressive prerequisites, we have to furthermore determine the input (written) text, where such elements inevitably are missed. This constitutes the first level of any expressive policy. Of course, it is not sufficient.

Voice expressiveness concerns furthermore different aspects (or, better, points of view) of a text, such as lexical, morphological, syntactical, semantic, stylistic, thematic, tactic, typographical or concerning punctuation etc.). Expressiveness usually selects some of these points of view and elaborates its forms on textual items in respect with them. Such items come out from different levels of analysis of the text. This is a crucial—and traditionally unsolved—question: clearly, any collection of data (the text being such) may be developed under a boundless number of levels of analysis, bearing legitimacy and rationality (technical, applicative, epistemological etc.). In our case, for efficiency and application reasons essentially, we have chosen only three: micro-items (syllables or *syl*), meso-items (syntagms or *syn*) and macro-items (phrase groups or *phg*). Clearly, alternatives are by any means possible (from diphones (pico-items) to intertexts (mega-items)). In all cases, it is important to conceive the levels of analysis from the most global to the most local ones. In other words: to adopt a top-down design as far as it is the global that determines the local. For us, the prime level is the text level (roughly identified to *phg*).

Voice plasticity is actually envisaged as an emerging dynamic relief implanted on such items; for our purposes, it is rendered under three classical prosodic parameters: melody (F), tempo (duration) (T) and intensity (I). Each of them is threefold as far as it concerns necessarily items coming out from our three levels of analysis of the textual matter (*syl*, *syn* and *phg*). An expressive form is a structure obtained by convenient choices over the prosodic elements of this third level (see just below).

The entire model of natural voice expressiveness may thus be represented by the Figure 1. For an extended discussion about the program motivations of our modelling approach, see, for instance [3].

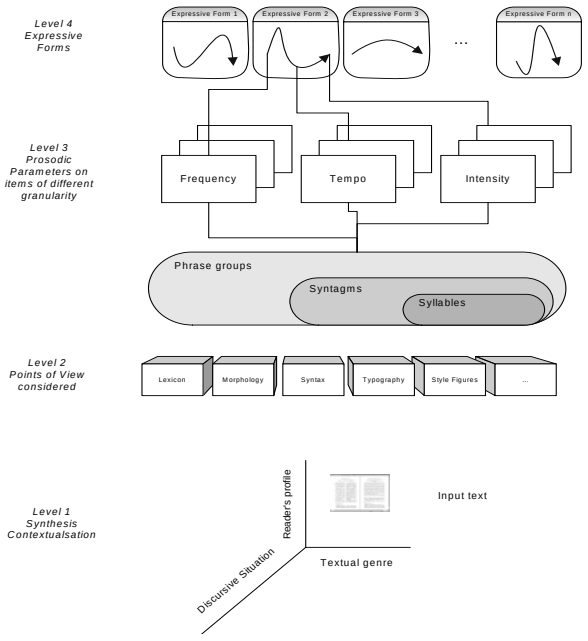

Fig. 1: A Model for expressive natural voice

B. Methodology

It is a direct extension of the *corpus-based* approach in voice synthesis, developed by Orange/France-Telecom (see [1] for an on-line demonstrator). We briefly remind that this approach consists in the constitution of extended sound corpora of read text, from which, after a first series of signal treatments and the use of rather sophisticated and robust algorithms, as well as decision making techniques,

the system extracts and combines the best phonetic contextual candidates in order to synthesize new input texts. For voice expressiveness, the objective is not different, but the corpus is also filled (not with diphones but rather) with expressive forms. Intuitively, an expressive form may be understood as a vocal pattern, ready-to-apply on a linguistic item of a certain level. In a certain sense, it functions as a mould of the human verbal performance. From a cognitive point of view, it is nowadays considered as some prelinguistic basis in language acquisition. In all cases, it constitutes an unavoidable verbal dimension. We skip here the technical part of the formalization of such expressive forms (see, for instance, [7] for more details); it will be sufficient for the moment to understand them as the result of a selection process which operates on the third level of the outlined model. In other words, an expressive form selects (i) some points of view, (ii) some item levels and, finally, (iii) some values for the prosodic structure *<F, T, I>* of the linguistic input, and builds something similar to a schema. An expressive form concerns mainly the macro-linguistic level (that generally exceeds the phrase range).

C. Corpus constitution

We have chosen two textual genres: fairy tails and horoscopes (20 texts from each) and worked with professional actors asking them to read each text in different manners—attested or not in real life practices. Improvisation was admitted, insofar as intelligibility was guaranteed, and many times we asked a reading applying expressive patterns coming out from quite different discursive practices.

The whole approach may be schematized by the following figure:

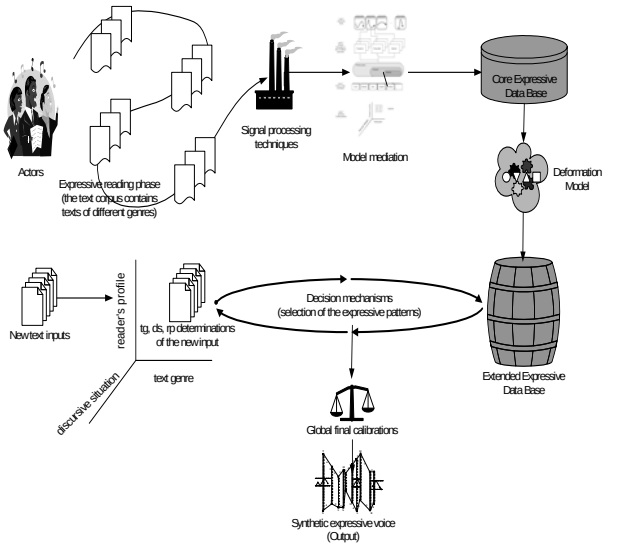

Fig. 2: Corpus based expressive voice synthesis: Global approach.

III. WHAT ABOUT THE SINGING VOICE SYNTHESIS?

We can at this point tackle briefly the question of singing voice—certainly of greater interest for this conference. Can all this (or a part of this) approach be applied to music? Clearly, the approach we outlined may be directly used in the case of pieces or part of pieces or even in the case of musical genres where the voice is not sung; it may also be useful to undertake experiments on poetry whose metrics uses some musical categories (noticeably that of duration) but still remain closer to the natural speech (ancient Greek poetry, for instance). But our question is of general scope: can such an approach give interesting issues for singing voice synthesis?

The answer is yes and no—better: no for a start and, perhaps yes subsequently! There is an assortment of ways to precise this voluntarily ambivalence, in the light of our introduction anyway.

No—globally no—when creativity is concerned, and it is rather useless to argue at long extend about it, insofar as the application objective is not the same.

The proposed model could perhaps appear of a certain help when applied, for study purposes, to a defined musical tradition over already existing recordings (a musical genre, a given performance condition, a singer). It is easy to give and to multiply examples. But it is pretentious—if already arrogant—to undertake a direct transposition of this model for singing voice, making abstraction of centuries of refined and stabilized musicological knowledge and practice, impossible to encapsulate to our service-oriented modest model. The issue, that possibly makes still sense here, would be to revisit this model and the methodology that sustains it, under specific application targets of the singing voice synthesis similar to TTS contexts.

IV. MODEL TRANSPOSITION FOR THE SINGING VOICE

The singing voice is not an indivisible whole, of course. Its automatic reproduction by a Score To Song (STS) system is still an interpretative problem since its validation depends on various music reception norms. Generally, evaluation criteria lie on established reception traditions founded on proper musical genres. A genre—a musical genre in particular—is not simply a class. It is something that situates reception, enabling appropriate interpretation strategies during audition. Its profound function is to guarantee opposition judgments over cross-level and usually complex musical forms. It is a principal and inescapable "ecological" parameter even for the singing voice; clearly, for an automatic synthesis, it has to be given by an external means to the system, insofar as, like in the case of a text, a score is quite underdetermined compared to musical actualization. Thus, in the case of a STS system, the text genre exigency will naturally and necessarily be transformed to this of a musical genre. We can generalize for the rest: the discursive situation will be something like a "performance condition", and the reader's profile something like a "singer's profile".

In the case of music, the point of view level (second level in the representation of Fig. 1) has to take into account the specificities of the musical notation. Freed from linguistic constraints, the singing voice necessitates idiosyncratic segmentations since it is driven by the score and obeys to global composition intentions. This means that, in the case of the singing voice, the item levels of the linguistic information may be of any sort, going beyond usual linguistic categories of analysis—even if they seem thoroughly convenient.

The rest of the model may remain invariant (as long as the characterization and the formalization of the expressive singing forms are concerned).

The reader may legitimately wonder if there some gain after such a modelling investment. In other words, the question is: is there some additive value comparing such a construction with traditional scores? From an analytic point of view, *i.e.* when music is seen as the cumulative compositional effect of elementary sounds that the score is supposed to indicate (with eventual complementary interpretation prescriptions), the answer is unfortunately and definitely no. But a musical piece, as an actualization of a particular semiotic system, is also subject to hermeneutical principles as well, noticeably, the one concerning the determination of local structures by global ones. Said differently, a musical piece is merely an emergent form produced by negotiations between local and global sound structures realized at different levels—not only a sound concatenation; there, complex rhythm vectors seem to ensure weaving and integration of these local structures in a whole. For the intelligibility and the formal exploration of such an idea, the model we propose can still suggest interesting issues for music.

As in the case of natural voice synthesis, the main obstacle in singing voice synthesis is the dictatorship of a bottom-up compositionality, which renders indiscernible the reception effect of more global structures. For more reliable results, it would certainly be necessary to have at disposal means likely to characterize better musical cross-level expressiveness. Such a model is in fact generic, and has perhaps the additional—limited but real—quality to furnish formal material able to describe some of these global (and complex) structures and—what is a recurrent application demand—to drag them partially into devices supporting calculus.

V. Conclusion: A R&D direction for the singing voice synthesis

We are convinced that the STS approaches may still find some interest in the TTS corpus-based approach. In fact, the model we proposed is nothing—or very little—without the methodology that makes it intelligible and whose it offers a synthetic view. Today's approaches to singing voice synthesis are uniformly founded on compositional approaches of natural voice synthesis (concatenation of diphones; many references all over the web; emblematically, MaxMBROLA [10] or VOCALOID [11]), where the singing voice is envisaged as a special case of the natural voice. Nowadays, we have at disposal systems able to generate an interesting variety of synthesized voices with convincing musical quality. But generally, all such systems "are challenged with respect to naturalness, range, the ability to synthesize both male and female voices, as well as the ability to capture the identity of the singer" ([12]).

It is rather straight that there is still much room for corpus-based expressive approaches of the singing voice. If the singing voice synthesis techniques are those we develop for natural voice synthesis, there seems to be no argument against the usefulness in importing the flow procedure sketched in Figure 2 for new STS system development. It could even take the form of two projects.

The first, conservative, would concern the direct implication of corpus-based extension of the usual concatenation of diphones. The second, more ambitious perhaps, would deal with a double corpus: one for the diphones in specified singing contexts, and a second one incorporating singing expressive forms, in the sense of our analysis. Explicitly, this last means to set up a data base of singing high-level expressive patterns of a targeted musical genre (obtained by professional singers and/or extracted from already existing recordings), to formalize them and to finally apply them on new scores through an adequate deformation model. Such a data base stands, in a sense, as the calculation counterpart of elementary valid reception scenarios. The objective of setting up patches implementing such expressive singing forms for STS applications may perhaps seem long and painful, but not less realistic (see, for instance [9], where the authors argue also about knowledge constitution possibilities of a patch; the model we present may par excellence assume such a use); in any case, they become more efficient and faster as the expressive base grows up.

References

[1] Baratinoo (TTS on line system of Orange/France-Telecom): http://tts.elibel.tm.fr/tts

[2] I. Kanellos, I. Suciu, Th. Moudenc, « Émotions et genres de locution. La reconstitution du pathos en synthèse vocale, » in M. Rinn (ed.) *Le Pathos en action*, Presses Universitaires de Rennes, 2007.

[3] E. Keller (ed), *Improvements in Speech Synthesis. COST 258: The Naturalness of Synthetic Speech*, John Wiley & Sons Ltd., 2002.

[4] F. Rastier, *Sémantique Interprétative*, Seuil, Paris, 1987. (For a quick introduction in English, *cf*. for instance: www.uqar.qc.ca/signo/rastier/a_semantique.asp).

[5] F. Rastier, *Meaning and textuality*, Toronto Buffalo: University of Toronto Press, 1997.

[6] I. Suciu, I. Kanellos, Th. Moudenc, "What about the text? Modelling global expressiveness in speech synthesis," *ICTTA Proceedings*, 2006, Damascus, Syria, pp. 177-178 (full version in the DVD of the Proceedings).

[7] I. Suciu, I. Kanellos, Th. Moudenc, "Formal expressive indiscernibility underlying a prosodic deformation model," *ISCA Proceedings*, pp. 229-232, 2006, Athens, Greece.

[8] V.-H. Zaldivar-Carrillo, V.-H., *Contributions à la formalisation de la notion de contexte. Le concept de « théorie » dans la représentation des connaissances*, Ph.D, University of Montpellier 2, France, 1995.

[9] A. Bonardi and J. Barthélemy, « Le patch comme document numérique : support de création et de constitution de connaissances pour les arts de la performance, » *CIDE'10 Proceedings*, September 2007, Nancy, France.

[10] MaxMBROLA (as well as MBROLA) projects: http://tcts.fpms.ac.be/synthesis/maxmbrola/

[11] VOCALOID, www.vocaloid.com

[12] M.E. Lee, M.J.T Smith, "Digital singing voice synthesis using a new alternating reflection model," *ISCAS Proceedings*, 2002, Vol. 2, pp. 863–866.