Formal expressive indiscernibility underlying a prosodic deformation model

Ioana Suciu[1], Ioannis Kanellos[2] and Thierry Moudenc[1]
[1]TECH/SSTP/VMI, France Telecom R&D, Lannion, France
[2]Computer Science Department, ENST Bretagne, Brest, France

https://doi.org/10.36505/ExLing-2006/01/0051/000051

Abstract

We are here concerned by the setting up of a model and a formalism for expressive speech synthesis under the paradigm of a corpus-based approach. Our objective is to apply prosodic expressive forms, acquired from natural human-reading recordings, on a new textual matter. We outline a general model for speech expressiveness. Then we deal with some formal aspects of expressive representation. We point out the core transformational aspects and the indiscernibility criteria allowing comparisons be-tween forms. We finish by some interpretational issues of such an approach.

Introduction

In speech synthesis one cannot for long avoid the prosodic deformation model requirement, especially when expressiveness becomes a central research theme for natural speech (Keller 2002). The arguments are important and massive (epistemological, theoretical, applicative...) and increase their intelligibility once one opts for a corpus-based synthesis approach. Indeed, acquired corpora, even extended, cannot give high quality acceptability results for all expressive forms one encounters in everyday life linguistic uses. Thus, the only issue is the generative-transformational one: on the basis of a kernel of forms one has to generate a relevant range of the expressive universe one typically encounters in a practice. This last defines a research program in which the model architecture and the choice of the formal representation condition an important part of both the expected results and their validation protocols. The model gives the framework in which the conception of the speech expressivity is thoroughly thought out, while the formality gives the scope of its transformational potentiality likely to be implemented. Such remarks give evidence to the structure we subsequently follow for our talk.

A model for speech synthesis expressivity

Without any exigency for exhaustiveness, we outline here a general model for the expressive synthesised speech, as shown in (Kanellos & *al.* 2006). It is set up through five consecutive steps:

ExLing 2006: Proceedings of 1st Tutorial and Research Workshop on Experimental Linguistics, 28-30 August 2006, Athens, Greece

Firstly, one has to start by positioning a discursive form in a multidimensional space defined by three global characteristics, extrinsic to the textual matter: the textual genre *(tg)*, the discursive situation *(ds)* and the reader's profile *(rp)*. Implicit for a human speaker and essential for the speech production, perception and interpretation, they are required as entries for an expressive speech handling.

Once situated, the text to synthesize is analysed through complementary points of view giving rise to lexical, morpho-syntactical, typographical, semantic, punctuation etc. treatments.

Then, the levels of the textual analysis are chosen. For complexity efficiency and adequacy to current techniques, three of them seem sufficient: the syllable *(syl)*, the syntagm *(syn)* and the phrase group *(phg)* level. The expressive transposition may concern the deformation of a unit of any of these levels.

The next step deals with prosodic representation that concerns precisely the units of these three levels. It allows the description of the melodic *(F)*, temporal *(T)* and intensity *(I)* movements (local and global), to be associated with them.

Finally, an expressive discursive form is defined as a choice strategy among the values of these three prosodic parameters.

Formal representation of expressiveness

Let us now see a possible formal description for the expressive phenomena.

One naturally starts by representing the space vector of the extrinsic characteristics: S =df <tg, ds, rp>. Therefore, any expressive form type E has to be envisaged as a situated complex vector in such a space: Esit =df <E, S> =df <E, <tg, ds, rp>> (cf. step 1 above).

For a correct synthesis realisation of a textual unit U, one generally needs to know its phonological level of analysis l (i.e. syl, syn or phg; cf. step 3), its compositional structure C, as well as some linguistic aspects D describing it. We represent these by the linguistic unit description vector: U =df <l, C, D>. C informs about the number n of the units of l-1 level composing U (number of syllables for a syntagm, number of syntagms for a phrase group) and the (ordered) list of their identifiers: C =df <n, id1...n>; for reliability reasons, these identifiers are supposed to be unique. Finally, D gives information on the syntagm nature (noun, verb etc.), the focus localisation(s), some punctuation or typographical specifications etc. (cf. step 2).

Local or global, each expressive unit carries pieces of information corresponding to the prosodic movements in speech (*cf.* step 4). We represent them by a three-dimensional prosodic vector: $P =_{df} <F, T, I>$. *F, T, I* are typed data; their type is decided by the *l* level (for instance T_{syl}, T_{syn} and T_{phg}).

We define an expressive form type as a triplet: $E =_{df} <id, U, P>$, where, *id* is its unique identifier. Thus, $E =_{df} <id, <l, C, D>, <F, T, I>>$ (1) and clearly: $E_{sit} =_{df} <<id, <l, C, D>, <F, T, I>>, <tg, ds, rp>>$ (2).

Indiscernibility and transformation scenarios

The expression (2) encapsulates the expressive transformational potentiality allowed by the model above (Zaldivar-Carrillo 1995). Indeed, any possible transformation modifies some (or all) of the components of the E_{sit} vector at different ranks of occurrence: *l, C, D* and *F, T, I* as well as *tg, ds, rp* at the first rank, *U* and *P* at the second, *E* and *S* at the third one. These transformations constitute an inescapable formal ingredient of two typical applicative orientations: the first corresponds to productive objectives and concerns the generation of expressive speech on the basis of existing forms; the second is interested in acquisition and learning purposes and deals mainly with the comparison of a new form with an existing one in the expressive data base.

Two types of formal operators are defined with respect to these applications: O (unary) and R (binary). The O-type operators generate new forms (of the same level *l*) following a given transformation scenario, such as compression, stretching, scaling, transgression, reversion, inversion, translation etc. These scenarios are different for different E_{sit} components (for instance, the stretching in time is not the same as the stretching in melody). An R-type operator searches to compare two expressive forms under some indiscernibility conditions, informing about their eventual complete or partial equivalence. Here again, compared forms must have the same level *l*. Partial equivalence renders the expressive indiscernibility over one or more of the *U*, *P* or *S* dimensions (thus $e_1 =_{F,T} e_2$ means that the two forms are indiscernible over both the melodic and temporal aspects; $e_1 =_C e_2$ that they are indiscernible over the *C* component etc.). These indiscernibility criteria are on the basis of any prosodic transformation procedure. They also found similarity and tolerance relationships (Ferret 1998) between expressive forms.

Clearly, one can easily envisage compositions of O-type operators. On the other hand, the R-type operators act as condition for the O transformations (they qualify the relationship between the original and the generated form by an O operator). Moreover, if e_1 and e_2 are expressive forms of the same level, it is possible to find at least one O-type transformational sequence $O_1, ..., O_k$ such that $O_k \circ \ldots \circ O_1 (e_1) = e_2$. In other words, once the level of analysis *l* is fixed, the universe $\mathcal{E}_l$ of the expressive forms becomes complete under the O-type transformations. The corollary of this is that starting from any form e_1, it is possible to generate an e_2 such that $R(e_1, e_2)$ satisfies a given indiscernibility criterion (*e.g.* $e_1 =_{F,T,I} e_2$). In both cases, it is possible to set up different criteria determining transformational costs.

Interpretational cues for indiscernible forms

We illustrate here some of the ideas above (*cf.* also (Suciu & *al.* 2006)).

The prosodic particularities of two speakers (Max *vs.* Tom) are studied in function of *F, T* and *I* variations, but supposing that the implied transformations preserve indiscernibility over C, D, tg and ds dimensions (=C, D, tg, ds). Detecting recurrences in these particularities may furthermore be a clue for a stylistic rp profile description and give rise to profile expressive universes.

Different reading manners (drunken vs. hysterical) are obtained by variations in T, F and I, while C, D, tg and rp remain indiscernible (=C, D, tg, rp).

An altered effect for the listener, going from strangeness to parody (a love letter read as a political discourse) may be produced by the same prosodic realisations for two different textual genres tg. Generally, it necessitates indiscernibility over U, P, ds and rp (=C, D, F, T, I, ds, rp).

On the other hand, for the same ds and tg values, the U and P indiscernibility scenario becomes a speaker imitation one. It is the case where the rp1 speaker appropriates the prosodic profile of the rp_2 (Max speaking as some famous president, for example).

Conclusions

As semantic extension of a written text, expressiveness formulates an indecidable problem insofar as it calls for extrinsic supplementary pragmatic information and concerns mainly the reception which strongly depends on the interpretation strategies of the listener. It seems however legitimate and even highly promising to extrapolate the corpus-based paradigm for expressive forms. In such a case, the set of transformations and the indiscernibility relationships give the essentials of the approach that can be implemented as an extension on a traditional corpus-based speech synthesis system.

References

Ferret, S., 1998. L'identité, Flammarion, Paris.

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

Kanellos, I., Suciu I., Moudenc, Th., 2006. Émotions et genres de locution. La reconstitution du pathos en synthèse vocale. In Rinn, M. (ed.) Le Pathos en action, Presses Universitaires de Rennes, France (to appear).

Suciu I., Kanellos, I., Moudenc, Th., 2006. What about the text? Modelling global expressiveness in speech synthesis. Proceedings of the ICTTA Conference, Damascus, Syria, 177-178 (extended version in the DVD of the Proceedings).

Zaldivar-Carrillo, V.-H., 1995. 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.