Expressive Speech Synthesis: Evaluation of a Voice Quality Centered Coder on the Different Acoustic Dimensions

Nicolas Audibert[1], Damien Vincent[2], Véronique Aubergé[1] & Olivier Rosec[2]

[1] Institut de la Communication Parlée
CNRS UMR 5009, Grenoble, France
[2] France Telecom, R&D Division
{audibert, auberge}@icp.inpg.fr, {damien.vincent, olivier.rosec}@francetelecom.com

Abstract

Expressive speech is intrinsically multi-dimensional. Each acoustic dimension has specific weights depending on the nature of the expressed affects. The quantity of expressive information carried by each dimension separately (using Praat algorithms), as well as the processing implied to carry it (global value vs. contour) has been perceptively measured for a set of natural mono-syllabic utterances (Audibert et al, 2005). It has been shown that no parameter alone is able to carry the whole emotion information, F0 contours or global values revealed to bring more information on positive expressions, voice quality and duration conveyed more information on negative expressions, and the intensity contours did not bring any significant information when used alone. These selected stimuli, expressing anxiety, disappointment, disgust, disquiet, joy, resignation and sadness were resynthesized with an LF-ARX algorithm, and evaluated in the same perceptive protocol extended to the three voice quality parameters (source, filter and residue). The comparison of results between natural, TD-PSOLA resynthesized and LF-ARX resynthesized stimuli (1) globally confirms the relative weights of each dimension (2) diagnoses local minor artifacts of resynthesis (3) validates the efficiency of the LF-ARX algorithm (4) measures the relative importance of each of the three LF-ARX parameters.

1. Introduction

In order to study the expressive speech, whether for theoretical purposes or for applications in synthesis or recognition, one first has to face the fundamental problem of the acoustic dimensions of the affective prosody. From the old and unsolved debate about the specificities of some acoustic dimensions possibly devoted to particular affects, it can be at least retained that all prosodic dimensions, i.e. F0, intensity, "voice quality" and duration, must be tracked to model all kinds of affects (moods, emotions, attitudes, …). A central open question is in particular to understand if the voice quality must be considered globally as a single dimension, or should be described as several dimensions. The glottal articulatory-to-acoustic modes (breathy, creaky voice, etc.) are quite well described [9], but even if some studies link one of these modes to some affect characteristics (e.g. breathy to intimacy [9] or care [6]), most studies globally describe the voice quality by a great number of acoustic parameters. In spite of the complexity of speech inversion, expressive speech inversion using the LF model [7] theoretically makes possible the characterization of the signal without using redundant and not significant acoustic parameters. Our aim is to evaluate how the LF model can encode some affective information in real spontaneous data, not globally, but specifically for each dimension (F0, intensity, duration, voice quality), by comparing for each dimension the LF coded stimulus and the original stimulus.

The methodology used is based on copy synthesis, i.e. an acoustic analysis of reference stimuli is performed prior to synthesizing new stimuli using the analyzed parameters as an input. Those parameters may be voluntarily altered, and either the whole set of parameters or only part of them may be used, according to the aim of the experiment. Eventually, generated stimuli can be perceptively evaluated to assess tested hypotheses.

Such a method was already used in experiments focusing on affective expressions in speech. For instance, Gobl et al. [8] evaluated the role played by voice quality in emotional expressions using stimuli synthesized from a reference glottal flow waveform modified to express different voice qualities. Stimuli used in this study were used as a basis to synthesize new stimuli with different voice qualities and prototypical expressive F0 contours applied separately and together, in order to evaluate the relative influences of these dimensions on the perception of synthesized expressions [13]. In another experiment [5], prosodic parameters (limited to F0 variations and duration) analyzed from emotional expressions in English of anger, happiness and sadness as well as neutral expressions were applied to diphones recorded with the same set of emotional expressions in concatenative synthesizers. Stimuli were built with prosodic parameters matched to diphones set, as well as mismatched ones. The authors concluded from identification scores of mismatched expressions that anger was mainly carried by diphones, sadness by prosodic parameters, when no clear pattern could be observed from expressions of happiness. Eventually, in studies such as [3], stimuli were synthesized from multiparametric measurements in order to evaluate the relevance of the extracted parameters for the perception of emotional expressions.

In a previous study [2], monosyllabic stimuli carrying 8 emotional expressions were used as a basis to generate synthetic stimuli using the Praat software [3], by projecting analyzed parameters separately. The perceptive evaluation of the generated stimuli revealed that F0 contours carried more information on positive expressions, voice quality and duration carried more information on negative expressions, and intensity brought no information when used alone. However the synthesis method did not make possible the separate evaluation of the influence of voice quality vs. duration. The aim of the present study is to replicate a similar evaluation on stimuli generated from the same set of natural stimuli with an LF-ARX algorithm [12] able to process voice quality and duration independently, and to compare perceptive results with those previously obtained.

2. Experimental framework for copy synthesis

2.1. Speech model

Many speech production models hypothesize that a speech signal can be considered as the result of passing an excitation through a linear filter. In this source-filter approach, the source part refers to the so called glottal flow derivative (GFD) which corresponds to the signal produced at the glottis and integrating the effect of lip radiation approximated by a derivation. On the other hand, the filter models the vocal tract resonances. When voiced sounds are produced, the vocal fold vibration results in a quasi periodic GFD for which classical models exist. The model used in this paper is the LF model [7], which allows a parameterization of the shape of the GFD waveform with three parameters. Figure 1 depicts a waveform obtained from this model.

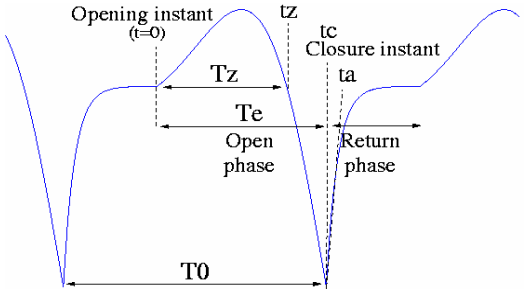

Figure 1: *One period of the glottal flow derivative.*

A stochastic component is also present to model noise-like effects (irregularity of the GFD, fricative noise, etc...). Thus, given the above assumptions, a voiced sound *s(n)* can be modeled by an ARX (Auto Regressive eXogenous) process defined as:

$$s(n) = -\sum_{k=1}^{p} a_k(n).s(n-k) + u(n) + e(n)\,,$$

where *u(n)* and *e(n)* respectively denote the deterministic (LF) and the stochastic parts of the GFD, and where $a_k(n)$ are the coefficients of the order-p filter characterizing the vocal tract.

Given this speech model, the analysis falls down to estimating the vocal tract filter, the fundamental frequency, the energy coefficient and the LF waveform parameters as well as a residual component. The joint estimation of this information is not straightforward as the optimization over the LF parameters is not a linear problem. However one can notice that when the LF parameters are known, the estimation of the filter and the residue can be achieved by least square methods. Based on this fact, an efficient method was proposed for solving this estimation problem by an exhaustive search over a space of quantized LF waveforms followed by local optimizations [12].

2.2. Implementation issues for copy synthesis

In this section the process of copy synthesis is presented. Let us consider a text message uttered by a speaker in two configurations: one considered neutral and the other one carrying an emotional content. As we are interested in identifying the relevant correlates for conveying emotion, copy synthesis will essentially consist here in replacing some of the parameters in the neutral utterance referred to as the source stimulus by their counterparts in the emotional target stimulus. For this purpose, two tasks are necessary: first an alignment procedure so as to map events between source and target stimuli and second a synthesis algorithm which enables the transformation of the desired correlates.

The alignment procedure is phonetically constrained and thus a prerequisite for our experiments is that both stimuli have the same phonetic content and that the phoneme segmentation is available. Then, the correspondence between source and target frames is done firstly by matching the phoneme boundaries and secondly, by relating the analysis instants within each phoneme by means of a linear interpolation mechanism. It is worth noting that problems can occur when the voicing information of source and target stimuli are different. However, after careful inspection of all speech signals, we did not find this kind of mismatch.

During the synthesis step, once the alignment between the source and target stimuli is done, the synthesis instants can be determined by an algorithm similar to the one used for TD-PSOLA based prosodic modifications [9]. Thus, this algorithm provides for each synthesis instant a pair of analysis frames respectively from the source and target stimuli. Given this mapping, the copy synthesis of any model parameter becomes straightforward.

2.3. Generation of synthetic stimuli

The 10 stimuli used as a basis for the copy synthesis were the same monosyllabic stimuli carrying emotional expressions as those used in the previous resynthesis study carried out at ICP [2], extracted from the E-Wiz / Sound Teacher corpus [1] and perceptively validated [11], phonemes boundaries being manually labeled. The stimulus expressing satisfaction was discarded from this set as the aforementioned framework failed to give a sufficiently good quality for generated stimuli, this problem being under investigation. Thus, this set was restrained to 7 stimuli expressing anxiety, disappointment, disgust, disquiet, joy, resignation and sadness on the French monosyllabic color names [ʁuʒ] and [sabl], as well as neutral expressions on each of these words.

From the analysis of different stimuli, the 6 following sets of parameters could be set independently to the value either of the expressive stimulus or of the corresponding neutral expression: F0, intensity, phonemic duration, source, residue and vocal tract filter. All 64 combinations of the 2 possible values of these 6 sets of parameters were systematically generated from each of the 7 expressions. However, only 7 synthesis conditions were selected for the perceptive evaluation: (i) a control condition obtained by applying all the parameters extracted from the expressive stimulus and labeled *'full resynthesis'* (ii) a *'VQ & duration'* condition obtained by applying the source, residue and vocal tract filter of the expressive stimulus, as well as the phonemic durations, F0 and intensity being extracted from the neutral expressions (iii) a *'VQ'* (voice quality) condition with the source, residue and vocal tract filter of the expressive stimulus (iv) a *'source & residue'* condition with only the source and residue of the expressive stimulus (v) a *'source'* condition (vi) a *'duration'* condition and (vii) an *'F0 & intensity'* condition.

Consequently the selected subset contains 49 stimuli. Moreover 2 additional stimuli were selected in the control condition, generated as a copy synthesis of the neutral expression stimuli, for a total of 51 stimuli. The control, *'VQ & duration'* and *'F0 & intensity'* conditions were selected to enable direct comparison with the previous results [2].

3. Perceptive evaluation

The 51 generated stimuli were perceptively evaluated by 25 judges (7 male, 18 female, aged 25.7 in average) at ICP, in a soundproof room with high quality headphones, with 3 presentations of each stimulus. Stimuli were presented to each judge in a different random order, the same stimulus being not presented twice consecutively. Stimuli presentation and judges' answers recording were performed using an automated interface: judges had to select either an expression within the 7 proposed (anxiety, disappointment, disgust, disquiet, joy, resignation and sadness) or the neutral expression. Moreover they were asked to rate the perceived emotional intensity on a 1-10 scale by moving a slider.

4. Results

The high Cronbach's alpha value (α=.92) indicates that answers given by different judges are quite coherent with each other. Results were then distributed into confusion matrices and analyzed separately for each of the 7 synthesis conditions. Most of the analysis presented hereunder concerns identification scores (without taking rated emotional intensities into account). As a matter of fact perceived emotional intensities do not bring much additional information, as they are significantly correlated to identification scores (r^2=.889). However consequences of the inter-judge effect on the expression of disgust could be observed as in [2], since it was attributed the highest emotional intensities both in control and *'duration'* conditions though other expressions were better identified. A Chi-square test on transformed data shows that the distribution of answers for different conditions are independent (p=.001).

As the confusions between different labels were similar to those observed in [2] in control condition, and in order to enable comparisons with previous results, the same clustering was applied: anxiety and disquiet were grouped together, as well as resignation, disappointment and sadness, while joy, disgust and neutral remained separate categories. Chi-square tests for clustered matrices indicate that distributions of answers differ significantly from chance distribution for all conditions (p=.001). Since most of the relevant information appears on the matrix diagonals (i.e. identification scores) after clustering, clustered data were converted to right or false answers and normalized to make possible further statistical evaluation. Transformed data were used as the input of an ANOVA of repeated measures with synthesis condition and expression as fixed factors. This analysis of variance shows a significant effect (p=.01) of condition and expression, as well as of the condition*expression interaction. Analyses of variance of repeated measures were computed for each synthesis condition, revealing a significant effect (p=.01) of the expression for all conditions except *'full resynthesis'*, showing that all expressions were identified with comparable scores in this control condition. Inter-expression contrasts were also systematically tested for each condition, as well as contrasts between different conditions for a given expression. Contrasts are detailed in the discussion of results, the level of significance being p=.01 when not specifically stated.

Table 1 summarizes identification scores for each label and each synthesis condition after clustering in the present evaluation (labeled 'ARX'), together with identification scores obtained for stimuli synthesized with Praat [2] (labeled 'Praat') when comparison is possible. As joy and satisfaction were mutually confused in this study, the confusions of joy with satisfaction were taken into account for the calculation of the identification score of joy. Identification scores of natural stimuli for each cluster of expressions are also presented in this table (labeled 'natural'). These scores were derived from [11]. It should be noted however that in this study 14 different labels were proposed to judges, who were allowed to rate stimuli as a blend of several of the proposed labels. In order to convert these results to identification scores and make comparisons with present results possible, a stimulus was considered as identified when at least one of the appropriate labels (i.e. from the same cluster as the presented stimulus) was selected. Confusions of joy with satisfaction were considered in this identification scores, as well as confusions of joy with amusement, also largely confused with joy and satisfaction in this first perceptive evaluation. Identification of natural stimuli appears to be higher than identification of synthetic stimuli, except for the expression of disgust. However this difference could be explained as a consequence of the inter-judge effect observed in [11]. As the structure of collected data does not make possible a statistical evaluation of differences between results obtained in these 3 studies, comparisons across sets of results are only qualitative.

Considering identification scores of the present study, a first observation is that major trends observed in [2] are confirmed in these data. Indeed manipulated stimuli were generally not identified as well as the corresponding control stimuli. However a few exceptions were observed: for instance the identification score of the sadness, resignation and disappointment expressions, as well as of the anxiety and disquiet expressions, were not significantly different in

Table 1: *Identification scores compared with those obtained in [2] and [11], both after clustering.*

		natural	control	F0+int	VQ+dur.	duration	VQ	source+res	source
joy	Praat	80.9%	70.8%	42.5%	6.7%				
	ARX		77.3%	58.7%	30.7%	0%	10.7%	4%	4%
sadness, resign. disapp.	Praat	82%	59.6%	26.7%	55.8%				
	ARX		56%	27.1%	52.9%	56.9%	44.4%	44%	41.8%
anxiety disquiet	Praat	76.1%	55.6%	21.4%	47.2%				
	ARX		67.3%	40.7%	60%	46%	46%	35.3%	24%
disgust	Praat	55.7%	61.7%	3.3%	34.2%				
	ARX		70.7%	1.3%	42.7%	49.3%	8%	5.3%	1.3%
neutral	Praat	66.1%	31.7%						
	ARX		52.7%						

'VQ & duration' vs. *'full resynthesis'* condition. In control condition the identification scores show a pattern similar to the one previously observed [2], most of the expressions being slightly better identified, with a large improvement for the neutral expression (52.7% vs. 31.7%). In *'F0 & intensity'* condition most of the expressive information on the expression of joy is retained, though the identification score is significantly lower than in control condition (58.7% vs. 77.3%): joy is significantly better identified than all other expressions in this synthesis condition. As stated above most of the affective information on the expressions of anxiety and disquiet, as well as of sadness, resignation and disappointment was retained in the *'VQ & duration'* condition. However for the expression of anxiety and disquiet this tendency is stronger than previously (60% vs. 47.2%), suggesting that voice quality of these expressions was better retained using the LF-ARX algorithm than with Praat. In this synthesis condition the part of the retained information on the expression of disgust (identified at 42.7% vs. 70.7% in control condition) is comparable with what was previously observed. This quite low identification, whereas F0 and intensity carry few information, confirms the observation that disgust is more sensitive to manipulations than other expressions [2].

On the other hand observations could also be done from resynthesis conditions not evaluated before. In particular the relative influence of voice quality vs. duration could be evaluated. In the *'duration'* condition the expressions of sadness, resignation and disappointment were as well identified as in control condition (no significant difference). However these expressions were identified significantly fewer times in the *'VQ'* condition but still well over chance level. For anxiety and disquiet, duration and VQ appear to carry the same amount of affective information (both identified at 46%). Eventually, phonemic duration carries most of the information on the expression of disgust (no significant difference when compared to the *'VQ & duration'* condition).

Moreover the relative weights of the different sets of parameters used for the modeling of voice quality by the LF-ARX algorithm could be evaluated by comparing scores obtained in *'VQ'*, *'source & residue'* and *'source'* conditions. For the expressions of sadness, resignation and disappointment, as well as joy, the source parameters appear to carry the whole information on voice quality. Indeed these expressions were not significantly better identified when residue and filter information was used vs. source only. For the expression of disgust this difference is hardly significant (p=.05), the score in the *'VQ'* condition remaining below chance level. On the other hand the expressions of anxiety and disquiet were significantly better recognized in *'source & residue'* vs. *'source'* condition, and significantly better in *'VQ'* vs. *'source & residue'* condition.

5. Discussion

The score obtained for the expression of joy with LF-ARX in *'VQ & duration'* condition (30.7%) should a priori be directly compared to the score in *'VQ'* condition (10.7%), but 2 surprising results appear: this expression was not identified at all (0%) in *'duration'* condition, though the difference between *'VQ'* (10.7%) and *'VQ & duration'* (30.7%) is far above zero. On the other hand this expression in *'VQ & duration'* condition reached only a score of 6.7% using Praat *vs.* 30.7% using LF-ARX. This led us to look for a possible artifact, so we noticed that an unexpected very short, mean energy "closure" noise appears at the beginning of the stimulus generated with LF-ARX in *'VQ & duration'* condition. As the copy synthesis algorithm was originally designed for "clean" speech, where speech segments labeled as "silence" actually correspond to a true silence, generation of silence segments is carried out by copying the silence zone of either the source or target stimulus. Since this noise appears in the neutral expression, it was automatically copied to the generated stimulus. We assume that this noise was interpreted by judges as a laughter cue, making the identification score higher. This unexpected artifact points out a very interesting phenomenon, since it shows that the characterization of affects in speech cannot be reduced to quantifying (by qualifying) the signal information.

A more relevant comparison would thus be between *'VQ'* condition with LF-ARX (10.7%) and *'VQ & duration'* condition with Praat (6.7%), as we want to compare those 2 algorithms. Another artifact might have lowered the score obtained with Praat (6.7%): to generate this stimulus, the relative intensity contour of the neutral expression was applied and the signal was scaled to reach the target stimulus overall energy level. As this method does not control local intensity values, the generated expression shows lower intensity at the end when compared to the neutral expression, though their mean intensities are equal. We assume that this final low intensity was interpreted by judges as incongruent with a joyful expression. It can thus be expected that, when replicating this experiment with "clean" stimuli, expressions of joy generated with LF-ARX in *'VQ'* and *'VQ & duration'* conditions would get the same score, around 10%.

6. References

[1] Aubergé, V., Audibert, N. and Rilliard, A., 2004. E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab. *4th LREC*, Lisbon, 179-182.

[2] Audibert, N., Aubergé, V., and Rilliard, A., 2005. The prosodic dimensions of emotion in speech: the relative weights of parameters. *Interspeech 2005*, Lisbon, 525-528.

[3] Bänziger, T., Morel, M., and Scherer, K. R., 2003. Is there an emotion signature in intonational patterns? And can it be used in synthesis? *Eurospeech 2003*, Geneva, 1641-1644.

[4] Boersma, P., and Weenink, D. Praat: doing phonetics by computer. http://www.fon.hum.uva.nl/praat

[5] Bulut, M., Narayanan, S. S. and Syrdal, A. K., 2002. Expressive speech synthesis using a concatenative synthesizer. *7th ICSLP*, Denver, Colorado, 1265–1268.

[6] Campbell, N., and Mokhtari, P., 2003. Voice Quality: the 4th Prosodic Dimension. *15th ICPhS*, Barcelona, 2417-2420.

[7] Fant, G., Liljencrants, J., and Lin, Q., 1985. A four-parameter model of glottal flow. *STL-QPSR* (4), 1–13.

[8] Gobl, C., and Ní Chasaide, A., 2003. The role of the voice quality in communicating emotions, mood and attitude" *Speech Communication* (40), 189–212,

[9] Laver, J., 1980. The phonetic description of voice quality. Cambridge: Cambridge University Press.

[10] Moulines, E. and Laroche, J., 1995. Non-parametric techniques for pitch-scale and time-scale modifications of speech. *Speech Communication* (16), 175–205.

[11] Rilliard, A., Aubergé, V. and Audibert, N., 2004. Evaluating an Authentic Audio-Visual Expressive Speech Corpus. *4th LREC*, Lisbon, Portugal, 175-178.

[12] Vincent, D., Rosec, O., and Chonavel, T., 2005. Estimation of LF glottal source parameters based on ARX model. *Interspeech 2005*, Lisbon, Portugal, 333-336.

[13] Yanushevskaya, I., Gobl, C., and Ní Chasaide, A., 2005. Voice quality and f0 cues for affect expression: implications for synthesis. *Interspeech 2005*, Lisbon, Portugal, 1849-1852.