Generation of Emotions by a Morphing Technique in English, French and Spanish

Philippe Boula de Mareüil, Philippe Célérier, Jacques Toen*

Elan TTS & LIMSI-CNRS*
{mareuil,pcelerier,jtoen}@elan.fr; mareuil@limsi.fr

Abstract

Generating variants becomes a priority for text-to-speech (TTS) synthesis. In particular, additional mark-ups inserted within the text may be used to communicate emotions. Within the framework of a European project linked to the MPEG4 standard (INTERFACE),our purpose is the synthesis of six emotions (anger, disgust, fear, joy, surprise and sadness): this was performed by applying a morphing technique, from the sequence of phonemes and their corresponding prosodic characteristics, for a "neutral" style, generated by a multilingual TTS system. We dispose of corpora declined under these six emotions by professional actors in English, French and Spanish: some trends may be drawn, as the inversion of fundamental frequency slopes for disgust and the pruning of melodic movements for sadness. We tend to think that the perceptual identification of the different emotions will be facilitated, within the framework of MPEG4, by the addition of a visual component: a talking head.

1. Introduction

A recent ISCA workshop in Belfast testimonies of the fact that emotions in speech receive an increasing deal of interest [1]. Even if it has long been neglected by engineers, the idea that melodic and rhythmic structures convey emotions and attitudes is not novel though [2,3]. For this, we may find physiological reasons, a higher articulatory effort provoking an increase of the speech rate and of energy, for instance. This way, regret and sadness are often linked to a slow tempo, to a diminution of mean energy, and to a monotonous and low melody, with respect to admiration or surprise [4,5]. Interest is manifested by an increase of energy standard deviation; fear, joy and anger are characterised by an increase of energy, of speech rate, of average pitch and of pitch range [6] (see also table 1 for the English language). Incredulity towards what has just been established is characterised by a low start, and by emphasis on the penultimate syllable, followed by a rising last syllable; as for suspicious irony, it goes back on the interlocutor's statement at a high pitch level, with an abrupt fall of fundamental frequency over the last syllable [7] (for the French language). But these contours, on long sentences, seem to split into several parts, and additional modulations may appear.

It may be argued that voice quality is a more relevant feature than prosody to identify emotions such as boredom, joy and cold anger (e.g. [8] for the Spanish language). The recognition of shame, disgust and amusement relies even more on visual clues, and it is wished that the identification of emotions would not be too language- or speaker-dependent. But this is beyond the scope of this paper (on this issue, see [9]).

mean value	anger	fear	joy	sadness
speech rate	faster	faster	faster	slower
height & pitch range	higher wider	higher wider	higher wider	lower narrower
intensity	louder	normal	louder	lower

Table 1: handled prosodic parameters, from Murray & Arrnott [10].

Within the framework of the European project INTERFACE, linked to the MPEG4 standard, our goal is the synthesis of the following six emotions: (a) anger, (d) disgust, (f) fear, (s) surprise and (t) sadness; A morphing technique is applied to automatically modify the speech — the sequence of phonemes and their corresponding prosodic characteristics — generated by a PSOLA-based diphone-concatenation multilingual text-to-speech (TTS) synthesis for a neutral type of speech. This way, we have access to this information, and a postprocessor shall have to modify the values of pitch, duration and energy. We can play on their mean values and possibly on their dispersion with respect to regression lines. Taking more advantage of linguistic knowledge, but still keeping microprosody, we can rely on stressed syllables, to play on melodic schemes, to distribute duration modifications, to reinforce energy. For this human-machine interaction project, we adopted an empirical, hybrid approach, based on statistics on the one hand, and on rules motivated by observations (listening tests and linguistic expertise) on the other hand. In English, French and Spanish, we studied the expression of emotions by actors — even though this non spontaneous type of discourse has been widely debated (see [1]). The material and the protocol are described in section 2. The main features of what we did for synthesising our six emotions by a morphing technique are then presented, while section 4 concludes with evaluation issues.

2. Material and protocol

In the following paragraphs, a model is described, designed on the basis of a corpus of 150 phonetically balanced sentences per language, read twice in different sessions by an actor and an actress — both professional. The text, the prosodic parameters per phoneme and a sound file sampled at 16 kHZ are available for the whole amount of data — recorded with more than two weeks of interval, in silent rooms with high quality microphones. Nevertheless, we only considered one of the male actors' recording: contrary to usual practice, we did not compute any average values across two repetitions. This was due to the voluntary absence of control over the pragmatic interpretation of utterances. Without such control, the speakers' interpretation of pragmatic information, and therefore the scaling of individual pitch targets, can vary greatly from one reading to another. We can already notice that our male actors exhibit quite different cross-language strategies, especially as far as fear is concerned: the English actor simulated it, playing on voice quality, somewhat whispering, whereas the French and the Spanish actors rather shouted and cried respectively — in addition to this, the English and Spanish languages can rely on stress more than French can. As for the Spanish actor, who has a rather low “normal” voice (see figure 1), he also dramatically modifies his timbre and mean pitch to communicate the various emotions.

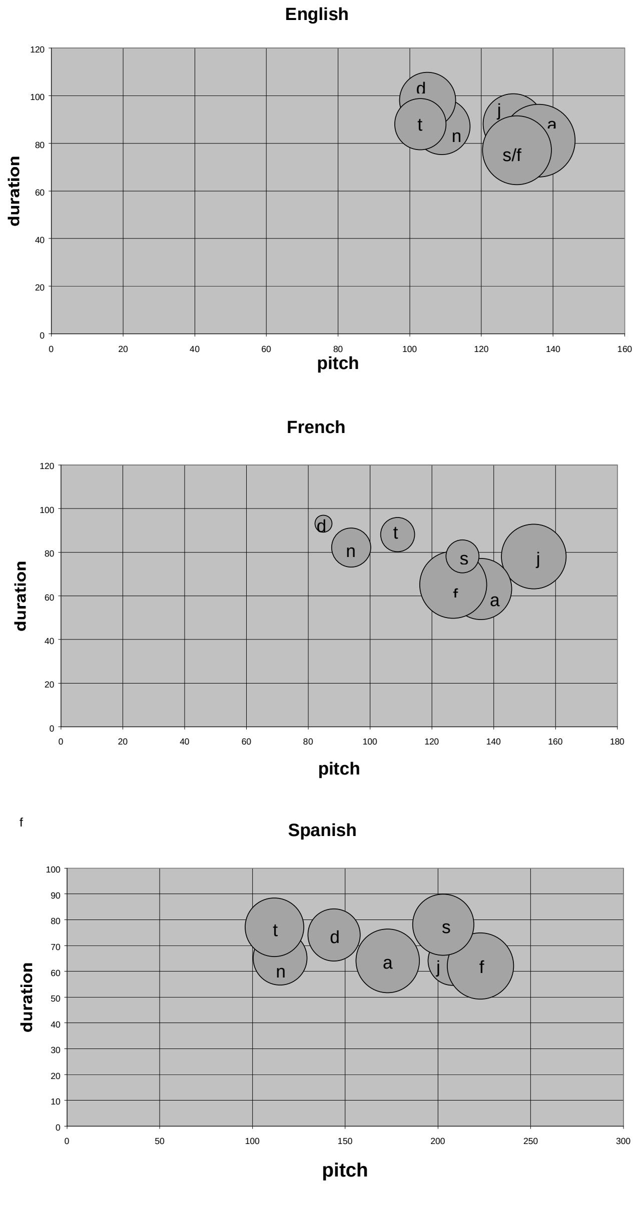


Figure 1: by emotion, average pitch of voiced parts (in Hz , average duration per phoneme (in ms) and average energy in English, French and Spanish.

In the representation of figure 1, with pitch on the *x*-axis and time on the *y*-axis (energy being denoted by the size of bubbles), we can see that two groups emerge in English: one for disgust, sadness and neutral — and the other one for the remaining (active) emotions. In French, we have three groups: joy and surprise are more easily distinguishable. The same happens if we consider the amplitudes of pitch variations (positive and negative). Indeed, melody is the most important parameter for perception [11]: the most monotonous emotions are the first three ones (disgust, sadness and neutral), even though the ranking is not as clear as that of the other languages — in French, for example, joy is more marked (see also figure 1).

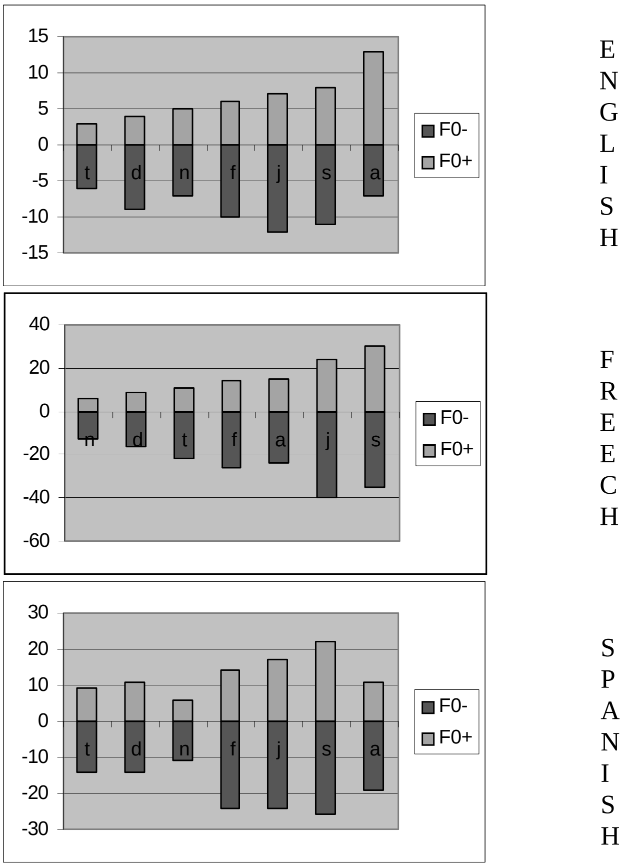
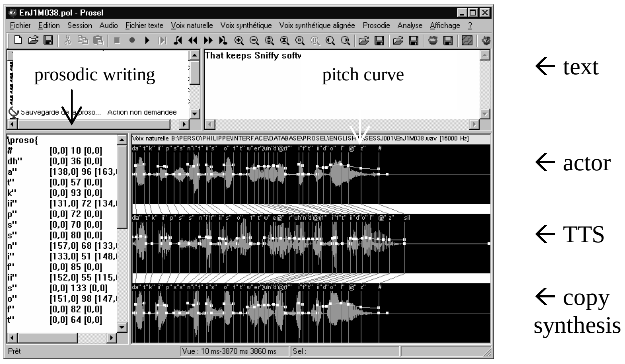


Figure 2: positive and negative pitch movements (in Hz) for neutral (n), anger (a), disgust (d), fear (f), joy (j), surprise (s) and sadness (t).

In the histograms above, the figures represent the cumulated sum of rising (resp. falling) movements, movements under a certain threshold being ignored. Whatever the calculation mode in French (by phoneme, by sentence, etc.), the results are actually the same in terms of ranking: small variations for disgust and sadness (passive emotions); medium variations for fear and anger; large variations for joy and surprise. Likewise in Spanish, movements range from small or medium for neutral, sadness, disgust and anger, to large for fear, joy and surprise. But in French, surprise goes down less than joy does, which may be interpreted if we look at ends of sentences: indeed, if by linear regression we interpolate F_0 curves by two segments, the final part of surprise even drops down less than neutral does.

To label the data phonetically with as few errors as possible, speech recognition and prosody transplantation tools were used (a window of the latter application, developed at Elan [12], is shown in figure 3): given an audio file and the text corresponding to what is pronounced, the system generates a file in “prosodic writing” and an audio file including the computed prosodic characteristics copied from the original [13]. In each language, about 25 sentences from 3 to 15 syllables were selected and examined in further detail. Declined under six emotions, they were sorted in such a way to listen to them emotion by emotion or sentence by sentence — about ten interrogative sentences per language, especially, were isolated. Re-synthesised, their F_0 curves were plotted: we were then able to draw and generalise regularities. The latter are presented in next section. However, they should be tempered by the following fact: to mimic emotions, actors modulate their voice qualities a lot, which is not allowed by PSOLA [13]. Emotions are all the more difficult to recognise through prosody transplantation — this was already the case with the original, which is well documented in the literature. This is one of the reasons why the percentages of pitch and speech rate increase/decrease applied below do not necessarily correspond to the ratios which would be given by figure 1 (the latter referring to male actors’ voices on the overall corpus): it is sometimes required to exagerate, to caricature, to make certain features more salient to facilitate the discrimination. Another reason is that too important a modification (e.g. a variation of +70% for pitch, to express joy) is badly supported by our coder, which leads to distortions.


EnJ1M038.pol - Prosel
Fichier Edition Session Audio Fichier texte Voix naturelle Voix synthétique Voix synthétique alignée Prosodie Analyse Affichage ?
prosodic writing
Sauvegarde de la proso... Action non demandée
pitch curve
That keeps Sniffy softv
Voix naturelle B:\PERSO\PHILIPPE\INTERFACE\DATABASE\PROSEL\ENGLISH\SESSJ001\EnJ1M038.wav [16000 Hz]
Prêt
Vue : 10 ms-3870 ms 3860 ms
Sel :
← text
← actor
← TTS
← copy synthesis

```
\proso{
#       [0,0] 10 [0,0]
dh''    [0,0] 36 [0,0]
a''     [138,0] 96 [163,
t''     [0,0] 57 [0,0]
k''     [0,0] 93 [0,0]
ii''    [131,0] 72 [134,
p''     [0,0] 72 [0,0]
s''     [0,0] 70 [0,0]
s''     [0,0] 80 [0,0]
n''     [157,0] 68 [133,
i''     [133,0] 51 [148,
f''     [0,0] 85 [0,0]
ii''    [152,0] 55 [115,
s''     [0,0] 133 [0,0]
o''     [151,0] 98 [147,
f''     [0,0] 82 [0,0]
t''     [0,0] 64 [0,0]
```

Figure 3: example of prosody transplantation for the sentence that keeps Sniffy software under fifty dollars.

3. Synthesis of emotions

3.1. Anger

Speech rate, energy as well as F_0 mean values and slopes are increased over the sentence. In addition to this:

- in English, pauses are reduced, (part of) the stressed syllable is lengthened (by a simple linear model) and its energy enhanced;
- in French, the last vowel (as well as possible following consonants) are raised, if it is under a certain threshold;
- in Spanish, rises are shortened, to make them even more abrupt.

3.2. Disgust

Outcomes on disgust are rare: here are ours, for this emotion which is assumed to be difficult to model. In French and in Spanish, the average speech rate and pitch are lowered. Positive F_0 slopes are inverted, and the final frequency is brought to a value $F_{0\min}$ if it goes under this threshold (see figure 4d). Until the utterance-final vowel, negative slopes may also be inverted, this way transforming a descent into a rise, then limited to a value $F_{0\max}$ (e.g. 120 Hz for French and Spanish male voices). For instance, let two contiguous phonemes be defined respectively by their initial pitch (p_i and p'_i) and their final pitch (p_f and p'_f, with p_f=p'_i if $p_f \neq 0$ and $p'_i \neq 0$). If $p'_i > p'_f$ and if the slope inversion for p_f leads to a value smaller than p'_f, the descent over this second phoneme is replaced by a rise.

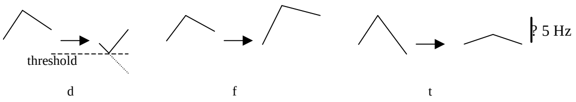


Figure 4: examples of modifications, with respect to the neutral original, of intonational movements for disgust (d), fear (f) and sadness (t).

In English, mean pitch and energy are lowered, stressed syllables are lengthened, as well as the labial plosives and approximants [p], [b], [m], [w], [ɹ].

3.3. Fear

In French, average pitch and speech rate are increased. Sentences start at a high pitch level (120 Hz on an average, for the French male voice). Rising movements are heightened (see figure 4f), and descents are attenuated.

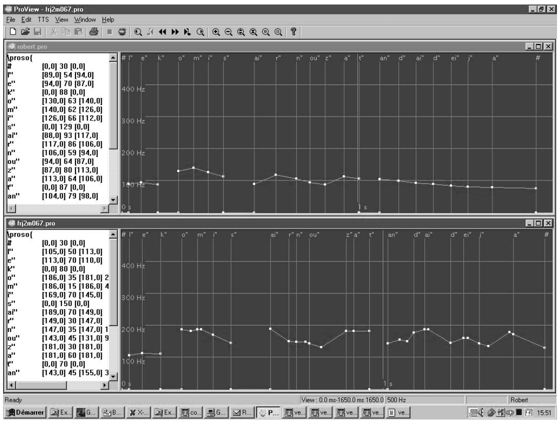
In English, F_0 mean values and slopes are increased. And the first phoneme of stressed syllables are doubled (except [l], [ɹ] and [w]). Another repetition with a 20 ms silence in between, and a 20 ms silence between words and after each stressed syllable are also applied. This way, it produces a kind of stammering, of tremor.

In Spanish, energy and F_0 mean values and slopes are increased (more than for anger). Unvoiced consonants (except [x] and [θ]) before stressed vowels and after pauses are doubled, and another repetition separated by a 20 ms silence is also applied. This was not done for French, due to the little markedness of stress in this language.

3.4. Joy

In French, the global pitch is increased — mere heuristic. Figure 5 illustrates the difficulty of morphing from neutral to joy.

In Spanish, the average pitch is increased, too; and pitch variations are replaced by straight lines between peaks (represented by stressed vowels) and valleys (represented by the median unstressed vowel between two peaks): the steepness of slopes is determined by the number of phonemes between two peaks. In English, pitch and energy are increased, especially on (part of) stressed syllables; rises are lengthened and increased.



```
\proso{
#      [0,0] 30 [0,0]
l''    [89,0] 54 [94,0]
e''    [94,0] 70 [87,0]
k''    [0,0] 88 [0,0]
o''    [130,0] 63 [140,0]
m''    [140,0] 62 [126,0]
i''    [126,0] 66 [112,0]
s''    [0,0] 129 [0,0]
ai''   [88,0] 93 [117,0]
r''    [117,0] 86 [106,0]
n''    [106,0] 59 [94,0]
ou''   [94,0] 64 [87,0]
z''    [87,0] 80 [113,0]
a''    [113,0] 64 [106,0]
t''    [0,0] 87 [0,0]
an''   [104,0] 79 [98,0]
```

```
\proso{
#      [0,0] 30 [0,0]
l''    [105,0] 50 [113,0]
e''    [113,0] 70 [110,0]
k''    [0,0] 80 [0,0]
o''    [186,0] 35 [181,0] 2
m''    [186,0] 15 [186,0] 4
i''    [169,0] 70 [145,0]
s''    [0,0] 150 [0,0]
ai''   [189,0] 70 [149,0]
r''    [149,0] 30 [147,0]
n''    [147,0] 35 [147,0] 1
ou''   [143,0] 45 [131,0] 9
z''    [181,0] 30 [181,0]
a''    [181,0] 60 [181,0]
t''    [0,0] 70 [0,0]
an''   [143,0] 45 [155,0] 3
```

Figure 5: F$_0$ *curves produced by the TTS synthesis (middle) and the actor portraying joy for the French sentence "le commissaire nous attendait déjà"* (the police officer was already waiting for us).

3.5. Surprise

In French and in Spanish, for statements, the modality is inverted (i.e. it becomes interrogative). For interrogative sentences (incredulous questions), the last vowel — as well as, possibly, following consonants — are heightened.

In English, pitch and energy mean values and variations are increased. Positive (resp. negative) F_0 slopes are multiplied by 2 (resp. –2) on sentence-final syllables. This also results in a final rise, and a pruning is applied.

3.6. Sadness

In the three languages, melodic movements are pruned to ±5 Hz (0.84 semitones) per phoneme, while keeping the shape of the movements (see figure 4t). To make the voice even more monotonous, with limited melodic movements, possible resettings after an unvoiced segment are avoided: this way, after a voiceless phoneme, the (voiced) phoneme takes the height at which the last voiced phoneme which precedes ends. In addition, a diminution of energy variations is applied.

4. Conclusion

Text-To-Speech synthesis, which is the objective we are in keeping with, is nowadays readily intelligible, but lacks variability. It can be employed as a tool: the morphing technique proposed here enables us to modify "neutral" prosody in order to mimic six emotions. These ones, in a TTS application, may be indicated thanks to control mark-ups inserted within texts (markers such as `<sadness>`). And we think that their perceptual identification will be facilitated, within the multimodal framework of the INTERFACE project in which we participate, by the addition of a visual component: a talking head, of which a demonstration already exists at http://www-dsp.com.dist.unige.it//interface/. In this area, lip movements are a current research axis *per se*, to be explored in the field of emotions.

Added to the fact that an authoritative taxonomy of emotions does not exist [14] (they may be mixed to different degrees, in addition), listeners' judgement may be influenced by the meaning conveyed by sentences. We shall have to take care of this. And we may add possible answers such as "boredom", "indignation" or even "other" by way of distracters, if we envisage a forced-choice questionnaire, to compare the confusion matrices (between emotions) obtained with natural speech, prosody transplantation and synthesis — with respect to chance level. Results will eventually have to be nuanced with respect to the number of subjects: 50-60% of accurate answers are expected with this method, which needs to be confirmed. But our models are consistent with previous studies (see [11,9]), and the preliminary results of informal listening tests are encouraging.

5. Acknowledgement

This research has been carried out within the scope of the EC-funded INTERFACE project IST-1999-10036 ("Multimodal analysis/synthesis system for human interaction to virtual and augmented environments"). Grateful for this financial support, the authors are also thankful to our partners who participated in the corpus design, in the recruiting of speakers, in the annotation and in the formatting of the speech database.

6. References

[1] Cowie, R.; Douglas-Cowie, E.; Schröder, M., 2000, eds. *ISCA Workshop on Speech and Emotion: a Conceptual Framework for Research*. Belfast.

[2] Troubetzkoy, N.S., 1986. *Principes de phonologie.* Paris: Éditions Klincksieck.

[3] Bolinger, D., 1989. *Intonation and its uses, melody and grammar in discourse*. London: Edward Arnold.

[4] Carlson, R.; Granström, B.; Nord, L., 1992. Experiments with emotive speech — acted utterances and synthesized replicas. In *Proc. of ICSLP*. Banff, 671-674.

[5] Gérard, C. ; Rigaud, C., 1994. Patterns prosodiques et intentions des locuteurs : le rôle crucial des variables temporelles dans la parole. *Journal de Physique*, 4((4), 505-508.

[6] Whiteside, S.P., 1998. Simulated emotions: an acoustic study of voice and perturbation measures. In *Proc. of ICSLP*, Sydney, 699-692.

[7] Morlec, Y., 1997. Génération multiparamétrique de la prosodie du français par apprentissage automatique. PhD thesis, Grenoble.

[8] Montero, J.M.; Gutiérez-Arriola, J.; Colás, J.; Enríquez, E.; Pardo, J.M., 1999. Analysis and modelling of emotional speech in Spanish. In *Proc. of ICPhS*. San Francisco, 957-960.

[9] Schröder, M., 2001. Emotional Speech Synthesis: A Review. In *Proc. of Eurospeech*. Aalborg, 561-564.

[10] Murray, I.R.; Arnott, J.L., 1993. Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion. *Journal of the Acoustical Society of America*,93(2), 1097-1108.

[11] Scherer, K.R. ; Johnstone, T.; Sangsue, J., 1998. L'état émotionnel du locuteur: facteur négligé mais non négligeable pour la technologie de la parole. In *Proc. of JEP*. Martigny, 249-257.

[12] Boula de Mareüil, P. *et al.*, 2001. Elan Text-To-Speech : un système multilingue de synthèse de la parole à partir du texte. *Traitement automatique des langues,* 42(1), 223-252.

[13] Moulines, E.; Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication,* 9. 453-468), 1990.

[14] Mozziconacci, S., 1998. Speech Variability and Emotion: Production and Perception. Proefschrift, Eindhoven.