EMOTIONAL FESTIVAL-MBROLA TTS SYNTHESIS

Fabio Tesser, Piero Cosi**, Carlo Drioli**, Graziano Tisato***

*Istituto Trentino di Cultura – Centro per le Ricerca Scientifica e Tecnologica, Trento, Italy
**Istituto di Scienze e Tecnologie della Cognizione, C.N.R. Padova, Italy
tesser@itc.it - [cosi, drioli, tisato]@pd.istc.cnr.it

Abstract

The topic of this work is an extension of our previous research on the development of a general data-driven procedure for creating a neutral "*narrative-style*" prosodic module for the Italian FESTIVAL Text-To-Speech (TTS) synthesizer, and it is focused on investigating and implementing new strategies for building a new emotional FESTIVAL TTS. The new emotional prosodic modules, similarly to the neutral case, are still based on the "Classification And Regression Tree" (CART) theory. The extension to the emotional speech synthesis is obtained using a differential approach: the emotional prosodic modules learn the differences between the neutral (without emotions) and the emotional prosodic data. Moreover, due to the fact that Voice Quality (VQ) is known to play an important role in emotive speech, a rule-based FESTIVAL-MBROLA VQ-modification module, for control of temporal and spectral characteristics of the synthesis, has also been implemented. Even if emotional synthesis still remains an attractive open issue, our preliminary evaluation results underline the effectiveness of the proposed solution.

1. Introduction

Speech synthesis is a by now stable technology that enables computers to talk. While existing synthesis techniques produce speech that is intelligible, few people would claim that listening to computer speech is natural or expressive. Therefore, in the last years, research in speech synthesis has been strongly focused on producing speech that sounds more natural or human-like. Meanwhile, the emotions and their role in human-to-human, human-to-machine (and vice versa) communication has become an interesting research topic.

Recently, new expressive/emotive human-machine interfaces are being studied that try to simulate the human behavior while reproducing man-machine dialogues, and various attempts to incorporate the expression of emotions into synthetic speech have been made [1][2].

The goal of this work is to investigate and implement strategies allowing a synthesizer to produce emotional speech.

This goal is relevant to both Text to Speech (TTS) and Concept to Speech (CTS) synthesis, and there are many possible applications scenarios ranging from human-machine interaction in general to speaking interfaces for impaired users, electronic games, virtual agents, and story-telling scenarios.

2. Prosodic Data-Driven Module

The task of a prosodic module in a TTS synthesizer is to compute the values of a set of prosodic variables, starting from the linguistic information contained in the text that has to be synthesized. In up to date TTS technologies, synthesis control has been mainly focusing on phoneme duration and pitch, which are the two main parameters conveying the prosodic information. Results on the learning of different speaking styles by modeling the main prosodic parameters through CARTs (*Classification And Regression Trees*) [3] have been reported in [4] and [5] for Italian.

More recently, the speech synthesis community is showing an increasing interest in the control of a broader class of voice characteristics. As an example, voice quality is known to play an important role in emotive speech, and some recent studies have addressed the exploitation of source models within the framework of articulatory synthesis to control the characteristics of voice phonation, or the use of signal processing to control the voice quality characteristics of pre-recorded voice [6][7].

This investigation relies on the FESTIVAL text-to-speech synthesis framework developed at CSTR [8], and on the MBROLA synthesis engine [9]. CART and VQ techniques were embedded into this framework, and adapted to the emotional case which is the central topic of this work.

3. Emotions

3.1. Emotional database

In order to collect the necessary amount of emotional speech data to train the TTS prosodic models, a professional actor was asked to produce vocal expressions of emotion as based on emotion labels and/or typical scenarios [10]. The Emotional-CARINI (E-Carini) database recorded for this study contains the recording of a novel ("Il Colombre" by Dino Buzzati) read and acted by a professional Italian actor, in different elicited emotions. According to Ekman's theory [11] six basic emotions, plus a neutral one, have been taken into consideration: anger, disgust, fear, happiness, sadness, and surprise. The duration of the database is about 15 minutes for each emotion.

3.2. Speech and Emotion in TTS Synthesis Environment

The speech characterisation of a certain emotion must be defined by the measure of its associated acoustic correlates, which directly derive from the physiologic constraints. For the prosodic correlates of emotional speech, such as f0, intensity and timing, it is easy to adopt the data-driven CART-based statistical modeling. A specific individual learning phase can be conceived for each emotional mood, such as those extracted from the E-Carini corpus, in order to obtain the prosodic modules for each emotion we want to be able to synthesize. Concerning speech timbre correlates of emotions, in general today's TTS technologies do not provide

a way to control such parameters. Various approaches are being explored to convert the timbre of a neutral voice into that of an emotional one, including source modeling in articulatory synthesis, and voice conversion algorithms [7] based on post-processing modules. Some experiments for Italian have been done on a limited emotive corpus of vowel-consonant-vowel sequences [12].

3.3. Emotional Prosodic Data-Driven Modeling: a Differential Approach

A wide number of studies on speech and emotions investigate the differences of emotional states with respect to a "neutral" state [10], and the transformation of a neutral utterance (real or synthetic) into an emotional one has been attempted with various techniques. Here a CART data driven approach will be used, to design an emotional prosodic module that learns the differences between the "neutral" prosody and the emotional one. For each prosodic parameter x (F0, duration, intensity), the parameter difference is given by $\Delta x = x_E - x_N$, where x_E is the emotional value for parameter x, as given by the acoustic analysis of the emotional database, and x_N is the "neutral" one, as predicted by a prosodic module trained on a "neutral" database. In the synthesis stage, the emotional data will be obtained using the simple superimposing model $x_E = x_N + \Delta x$. To be able to separate the macro prosody factors from the micro-segmental prosody ones, and to reduce data sparseness, various solutions were adopted, including the use of z-scores [13], normalization with respect to value ranges, and the use of parametric models for intonation curves [12],[14].

The training of differential CARTs was preferred over the training of emotion-specific CARTs, because this approach allowed us to use smaller databases for the different emotions (15 minutes each in our case, whereas the "neutral" one had a duration of about 50 minutes). Moreover, with this approach it is straightforward to implement smooth transitions from neutral to emotional speech for each emotion.

Table 1: Phoneme duration means (μ) and standard deviations (σ) for the E-Carini database (Δ subscript indicates differences between the neutral and emotional durations).

Emotion	μ (s)	σ (s)	μ_Δ (s)	σ_Δ (s)
Neutral	0.094	0.045	-	-
Anger	0.077	0.034	-0.017	-0.010
Disgust	0.103	0.055	0.009	0.011
Fear	0.078	0.036	-0.016	-0.009
Joy	0.076	0.032	-0.018	-0.013
Sadness	0.104	0.052	0.010	0.007
Surprise	0.076	0.033	-0.018	-0.012

3.3.1 Duration E-model

The macro prosodic differences for duration are represented on Table 1, where the averages statistics of the duration of phones in the different emotions are shown. The full procedure used to build an emotional duration module is the following:

- a CART is trained on the neutral z-score duration z_{Npred}
- the emotional z-score duration data z_{Ereal} are calculated using the means and standard deviations of the emotion.
- the linguistic and structural features used for the regression are extracted from the E-Carini database (these features are the same of the neutral duration case [14])
- a CART is trained on the difference $\Delta z = z_{Ereal} - z_{Npred}$
- in the synthesis stage z_{Npred} and the predicted Δz are summed and denormalized using the means and standard deviations of the given phoneme i in the given emotion, obtaining the predicted emotional duration $d_{Ei} = \mu_{Ei} + (z_{Npred} + \Delta z_{pred})\, \sigma_{Ei}$.

Using this approach, the macro prosodic part is represented by the table of the means and standard deviations for each phoneme and emotion, and the segmental prosodic part is represented by the differential Δz CART.

3.3.2 Intonation E-model

A VQ-PaIntE (Vector Quantized Parametric Representation of Intonation Events) model was adopted to effectively represent intonation curves [15]. The macro prosodic components of the intonation are the F0 mean and range values (Table 2).

Table 2: Pitch boundaries and means for the different emotions in the E-Carini corpus.

Emotion	LB(Hz)	μ(Hz)	UB(Hz)	R(Hz)	LB_Δ(Hz)	μ_Δ(Hz)	UB_Δ(Hz)	R_Δ(Hz)
Neutral	62	105	213	169	-	-	-	-
Anger	66	122	258	192	4	17	45	23
Disgust	53	81	238	185	-9	-24	25	16
Fear	66	114	223	157	4	9	10	-12
Joy	63	129	308	245	1	24	95	76
Sadness	53	89	208	155	-9	-16	-5	-14
Surprise	66	136	250	184	4	31	37	15

A pitch range normalization was performed in order to get rid of the influence of different pitch-range levels in the different emotions. This normalization was done using the LB and UB values of the Table 2. If we call $\mathbf{PN}_{Ereal}$ the real normalized PaIntE parameter vector in the emotional case and $\mathbf{PN}_{Npred}$ the predicted normalized PaIntE vector in the neutral case, the difference is given by $\Delta_{\mathbf{PN}} = \mathbf{PN}_{Ereal} - \mathbf{PN}_{Npred}$. The whole procedure for the design of the emotional intonation module is given by the following steps:

- a CART is trained to model the neutral PaIntE intonation;
- the emotional and the neutral PaIntE intonation data are normalized using the LB and UB values of the emotion, resulting in the vectors $\mathbf{PN}_{Ereal}$ and $\mathbf{PN}_{Npred}$.
- the vectors $\Delta_{\mathbf{PN}}$ are calculated for each emotion;
- a Principal Component Analysis is performed after a z-score normalization for each component of vectors $\Delta_{\mathbf{PN}z}$;
- the linguistic and structural features used for the regression has been extracted from the E-Carini database
- a CART is trained on the first PCA component, approximated by $\Delta_{\mathbf{PN}z} \cong \mu_{\Delta\mathbf{PN}z} + \alpha\, e_1$, where $\mu_{\Delta\mathbf{PN}z}$ is the mean of $\Delta_{\mathbf{PN}z}$, α is the predicted CART value and e1 is the first eigenvector;
- the predicted $\Delta_{\mathbf{PN}z}$ is denormalized using the inverse of the z-score transform obtaining $\Delta_{\mathbf{PN}}$. $\mathbf{PN}_{Npred}$ and $\Delta_{\mathbf{PN}}$ are then summed and denormalized using the LB and UB of the given emotion: $\mathbf{P}_{Ereal}$ = denormalize ($\mathbf{PN}_{Npred} + \Delta_{\mathbf{PN}}$).

3.3.3 Intensity E-model

The modeling of segmental intensity by CARTs has been the subject of a previous study [12]. However, since with the current diphone synthesizers the segmental intensity seems to have a small perceptual relevance if compared to the other prosodic cues, here we only took into account the overall mean intensity for the different emotions. These were stored in a table and used during the synthesis, as discussed in more details in [14].

3.3.4 Voice Quality E-model

In order to realize a VQ control in diphone concatenative synthesis, the FESTIVAL-MBROLA speech synthesizer was extended to allow for control of a set of low-level acoustic parameters that can be combined to produce some desired voice quality effects [16]. Some of the additional controls are: spectral tilt ("SpTilt"), implemented with a reshaping function in the frequency-domain that enhances or attenuates the low- and mid- frequency regions; F0 flutter ("F0Flut"), i.e. random low frequency fluctuations of the pitch, obtained by adding band-pass filtered (4-10 Hz) random noise to F0 patterns; spectral warping ("SpWarp") for the shift of upper formants, obtained by warping the frequency axis of the spectrum. The choice of the signal transformations was based on observations made on our emotional speech recordings, in which changes of spectral envelope, formants, and pitch and amplitude modulations, were relevant non-prosodic cues.

Based on acoustical analyses of the E-Carini database and referring to other previous studies on the voice quality of the emotion [12], a mapping between emotions, voice quality, and low-level acoustic parameters was designed. Anger, characterized by a loud-harsh voice, was implemented by enhancing the high frequencies (by SpTilt) and by lowering the upper formants (by SpWarp). For fear voice quality (tremulous and breathy), F0Fluct was used. In Joy and surprise (loud and breathy), SpTilt was used to enhance the high frequencies. For disgust, a nasalized voice quality was simulated by enhancing the high frequencies and a rising of upper formants. For sadness, a darker voice was produced by attenuation of the higher frequencies, and by a substantial lowering of upper formants.

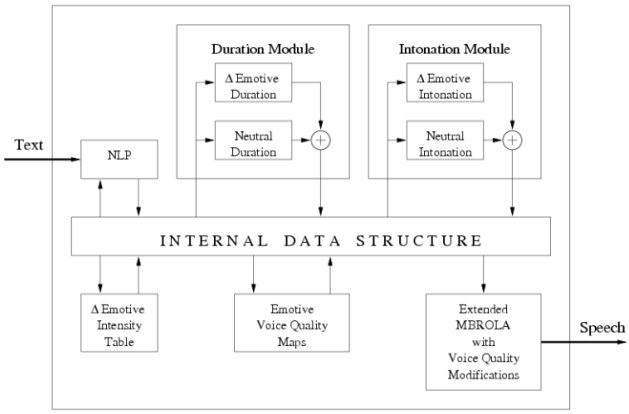

Figure 4: Overall Functional diagram of the E-TTS

4. Emotional FESTIVAL TTS

Figure 4 shows the overall architecture of the resulting E-TTS, in which the prosodic modules are embedded in a FESTIVAL-MBROLA framework For a given emotion and a given input text, the Natural Language Processing (**NLP**) module produces a phonetic-linguistic representation of the text. These data are used by the prosodic modules to predict the emotive prosody. Both the **Duration** and **Intonation Modules** use the differential approach: the internal data are used by both the neutral module and emotional differential one and subsequently summed to provide the emotional prosody patterns. The E-TTS also uses intensity values taken from the **ΔEmotive Intensity Table**, and low-level acoustic parameters of voice quality modification taken from the **Emotive Voice Quality Maps**.

5. Evaluation

Our prosody prediction models were assessed both with an objective and a subjective evaluation. Moreover, prosody prediction and voice quality modifications were assessed together and separately with a subjective evaluation.

5.1. Objective Evaluation

An objective evaluation of the prosodic modules was performed by splitting both the Carini and E-Carini database in a training set (90%) and a test set (10%), and measuring the differences between the synthetic prosody and the actual prosody in the test set. An indication of the performance can be given by the RMSE and Correlation ρ between the original prosodic signal and the predicted one, and by the absolute error |e| between the two prosody pattern. Table 5 shows the RMSE and the Correlation ρ between the original and the predicted values computed by the duration module for the different emotions. The mean and the variance of the absolute error |e| are also given. The values on the first three columns are expressed in z-score units, while the values on the last three columns are expressed in seconds.

Table 5: Duration prediction results for the different emotions.

Emotion	RMSE	$\mu_{\|e\|}$	$\sigma_{\|e\|}$	ρ	RMSE(s)	$\mu_{\|e\|}$(s)	$\sigma_{\|e\|}$(s)
Neutral	0.88	0.66	0.58	0.64	0.039	0.030	0.026
Anger	1.20	0.79	0.90	0.46	0.041	0.027	0.031
Disgust	0.99	0.65	0.74	0.50	0.054	0.035	0.041
Fear	1.03	0.70	0.75	0.56	0.037	0.025	0.027
Joy	0.88	0.62	0.63	0.61	0.028	0.020	0.020
Sadness	0.89	0.60	0.66	0.59	0.046	0.031	0.034
Surprise	0.93	0.65	0.66	0.62	0.030	0.022	0.022

Table 6: Intonation prediction results for the different emotions on the test set (the values on the first three columns are expressed in pitch normalized units).

Emotion	RMSE	$\mu_{\|e\|}$	$\sigma_{\|e\|}$	ρ	RMSE(Hz)	$\mu_{\|e\|}$(Hz)	$\sigma_{\|e\|}$(Hz)
Neutral	0.13	0.09	0.07	0.43	29	20	15
Anger	0.20	0.16	0.12	0.28	38	30	24
Disgust	0.15	0.11	0.10	0.22	28	21	19
Fear	0.25	0.18	0.17	0.16	39	28	26
Joy	0.22	0.18	0.13	0.23	54	43	32
Sadness	0.20	0.14	0.14	0.19	31	22	22
Surprise	0.27	0.21	0.17	0.22	49	39	30

Looking at RMSE and correlation columns the best performance is obtained for the neutral duration module, and the worst result is obtained for Anger. Surprise and Joy have an high correlation and their CARTs were the more complex in term of number of leafs. Sadness has a good performance too, and Surprise, Disgust and Fear have mid-low scores. Table 6 shows the results for the different emotions in the objective evaluation test for the intonation module. Also for intonation the best performance was obtained for Neutral. As for the emotions, the best RMSE performances were obtained for Disgust, and the worst results were obtained for Surprise. Further comments on these results can be found in [14].

5.2. Subjective Evaluation

The effectiveness of the prosodic modules and of the voice quality modifications was also assessed with perceptual

tests aimed at evaluating: a) the single contribution on the emotional expressiveness carried out separately by the emotional prosodic modules and the emotive voice quality modifications, and b) the synergistic contribution given by the union of these two correlates of the emotive speech. Four types of test sentences were generated:

(A) **neutral** prosody **without** emotive **VQ** modifications;
(B) **emotive** prosody **without** emotive **VQ** modifications;
(C) **neutral** prosody **with** emotive **VQ** modifications;
(D) **emotive** prosody **with** emotive **VQ** modifications.

For each emotion and for each of these four conditions, two utterances were produced by the new emotional TTS system for a total of 48 sentences, which were presented in a randomized order to 40 listeners, who judged, knowing the target emotion, the level of acceptability of the emotional synthesis, within a MOS scale (5=excellent, 4=good, 3=fair, 2=poor, 1=bad). Results are summarized in Fig. 5. Cases B, C, and D had always better results than those obtained for case A, signifying that emotive modules were quite successful. Case D shows always better MOS values and this is an indication that the created emotive prosodic modules quite improve the acceptability of the emotional TTS. Emotive VQ modifications alone were superior to the neutral case, except for Fear and Sadness. This can be the consequence of the fact that the contribution of prosody and voice quality might differ between different emotions [17], or might indicate that the chosen VQ acoustic modification should be modified for these emotions.

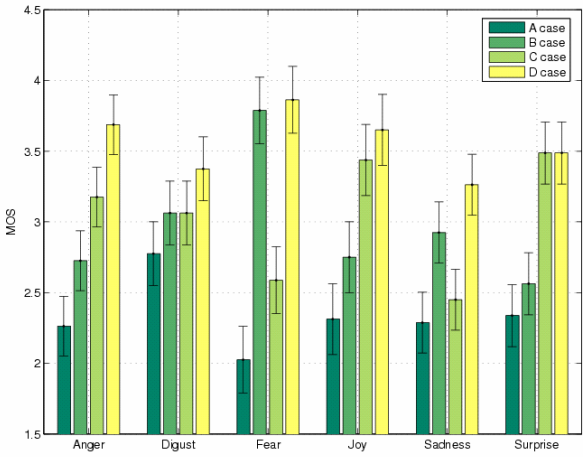

Figure 5*: Subjective Evaluation results (for A,B,C,D see text).*

6. Conclusions

It can be concluded that, emotional adequate speech can be obtained within a diphone-based approach, and that emotional prosodic modeling based on data-driven approaches has shown to produce appreciable results. Moreover Voice Quality processing has increased the quality of the perceived emotion on subjective test.

7. Acknowledgements

Part of this work has been sponsored by PF-STAR (Preparing Future multiSensorial inTerAction Research, European Project IST- 2001-37599, http://pfstar.itc.it). We wish to thank the MBROLA team for providing the source code of their synthesis engine.

8. References

[1] Schröder M., "Emotional speech synthesis: A review." *Proc. Eurospeech*, Aalborg, Denmark, 561-564, 2001.

[2] Murray I.R. and Arnott J.L., "Toward the simulation of emotion in synthetic speech: A review of literature on human vocal emotion", *Journal of Acoustical Society of America*, 93(2), 1097-1108, 1993

[3] Breiman L., Friedman J., Olshen R., and Stone C., *Classification and regression trees*. Wadsworth and Brooks, 1984.

[4] Cosi P., Avesani C., Tesser F., Gretter R., and Pianesi F., "On the use of Cart-Tree for prosodic predictions in the Italian Festival TTS", in Cosi P. Magno E. Zamboni A. editors, *Voce, Canto, Parlato- Studi in onore di Franco Ferrero*, UNIPRESS Padova, Italy, 73–81, 2002.

[5] Tesser F., Cosi P., Drioli C., and Tisato G., "Prosodic data driver modelling of a narrative style in Festival TTS", *CDROM Proc. of 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, U.S.A., 2004.

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

[7] Kain A. and Macon M.W., "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction", *Proc. of ICASSP*, Salt Lake City, UTAH, USA, Vol.2, 813 -816, 2001.

[8] Taylor P., Black A., and Caley R., "The Architecture of the Festival Speech Synthesis System", *Proc. 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves House, Blue Mountains, Australia, 147–151, 1998.

[9] Dutoit T. and Leich H., "MBR-PSOLA: Text-To- Speech Synthesis Based on an MBE re-Synthesis of the Segments Database", *Speech Commuication*, vol. 13, no. 3-4, 167–184, 1993.

[10] Scherer K. R., "Vocal affect expression: a review and a model for future research", *Psychological Bullettin*, 99: 143-165, 1986.

[11] Ekman P., "An argument for basic emotions", in *Basic Emotions*, Stein N.L. and Oatley K. (eds), Hove, UK, Lawrence Erlbaum, Edition, 1992.

[12] Drioli C., Tisato G., Cosi P., and Tesser F., "Emotions and voice quality: experiments with sinusoidal modelling", *Proc. of VOQUAL ESCA/workshop*, Geneva, Switzerland, 127-132, 2003.

[13] Campbell N. and Isard S., "Segment durations in a syllable frame", *Journal of Phonetics*, 37–47, 1991.

[14] Tesser F., "Emotional speech synthesis: from theory to applications", PhD Thesis, DIT- University of Trento, Trento, Italy, 2005.

[15] Cosi P., Tesser F., Gretter R., and Pianesi F., "A modified 'PaIntE Model' for Italian TTS", *CDROM Proc. of IEEE Workshop on Speech Synthesis*, Santa Monica, California, 2002.

[16] Drioli C., Tesser F., Tisato G., Cosi P., and Marchetto E., "Control of voice quality for emotional speech synthesis", *Proc. of 1st AISV (Italian Association of Voice Sciences) Congress*, Padova, Italy, 2004. In press.

[17] Montero J.M., Gutiérrez-Arriola J., Colás J., Enríquez E., Pardo J.M., "Analysis and modelling of emotional speech in Spanish", *Proc. of XIVth International Congress of Phonetic Sciences*, vol. II, pp 957-960, San Francisco August 1999