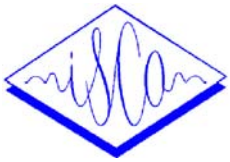

5th ISCA Speech Synthesis Workshop
Pittsburgh, PA, USA
June 14-16, 2004

# TOWARDS EMOTIONAL SPEECH SYNTHESIS: A RULE BASED APPROACH

*Enrico Zovato, Alberto Pacchiotti, Silvia Quazza, Stefano Sandri*

Loquendo S.p.A., Vocal Technology and Services, Turin - Italy
enrico.zovato@loquendo.com

**ABSTRACT**

This note describes a framework used to simulate three basic emotional styles by means of prosodic transplantation techniques applied to the output of a corpus based speech synthesis system. The target pitch profiles together with duration and energy constraints have been obtained applying simple rules inferred from the analysis of a small corpus, recorded in three emotional styles. Results of perceptual tests show that styles are well recognized even if the acoustical quality, in some cases, degrades.

## 1. INTRODUCTION

Despite the quality and intelligibility reached by corpus based text-to-speech synthesis systems, still much work has to be done at prosodic level. What could be improved is the capability to replicate different styles related to emotional attitudes. When the goal is working in a hypothetical continuous space, a solution could be the application of prosodic-driven rules to manipulate waveforms. This work consists of the prosodic analysis of an Italian emotional speech database and the following extrapolation of some correlates between emotions and acoustic parameters. In order to synthesize speech with emotions, an attempt was also made to extract rules in terms of variations with respect to a prosodically neutral situation. The emotional speech corpus contains the audio recordings of an Italian female professional speaker, and it is part of a larger corpus used to develop one of Loquendo TTS voices [1]. The emotion driven rules were thus used to manipulate the neutral output of this TTS voice. In the next two sections the speech database and its acoustic analysis are described, while section 4 reports the adopted speech synthesis solutions, and results of a perceptual test are reported in section 5.

## 2. EMOTIONAL SPEECH DATABASE

The emotional speech database consists of 25 sentences recorded in four different styles (angry, happy, sad and neutral). Sentences were about 10 words long and well representative of the Italian phonetic alphabet, while the semantic content was rather neutral and therefore could not provoke any particular emotional attitude. Talent had to simulate each style, and a director was always present during the recording sessions to control her pronunciation and her prosody and to avoid emphatic performances. Signals were recorded in an acoustically treated room, a high quality directional microphone was used, and waveforms were digitally acquired at 44.1 kHz sampling rate (16 bit). For the following analysis, these signals were down-sampled to 16kHz.

A perceptual test has been proposed to 10 volunteers in order to evaluate the corpus, and verify if it could be successfully used to extract style dependent rules. They interactively had to listen to the signals and had to give a preference among the three plus one choices. Four samples for each style were selected from the entire database. The results of this evaluation test are reported in the confusion matrix of Table 1. Recognition rates are quite high even if some appreciable confusion values occur in the pairs neutral-sad and happy-angry.

| | neutral | angry | happy | sad |
|---|---|---|---|---|
| neutral | 77.5% | 0% | 0% | 22.5% |
| angry | 7.5% | 87.5% | 0% | 5% |
| happy | 2.5% | 10% | 87.5% | 0% |
| sad | 0% | 0% | 0% | 100% |

*Table 1:Emotional speech database recognition rates.*

## 3. PROSODIC FEATURES EXTRACTION

For this task, syllable was chosen as the reference acoustic unit and for each syllable of the database, prosodic parameters such as minimum F0, maximum F0, F0 mean, F0 range and RMS energy were thus calculated.

Segmentation into syllabic units was achieved automatically. Firstly, a rule based phonetic transcriber (the same used by our TTS system) was used to convert graphemes to phonemes. Then an HMM based phonetic aligner segmented the signals and aligned the phonetic labels to waveform instants. Expert phoneticians manually and accurately annotated a set of training signals produced by the same speaker that was recorded to collect the emotional database. The acoustic features used to train these models were: 8 MFC coefficients, 1 spectral variation parameter and 1 energy coefficient, together with their first and second order derivatives [2].

In order to get the syllabic boundaries, the phonemes of each sentence were weighted according to the sonority scale. This particular scale assigns growing values to phonemes, beginning from unvoiced plosives and ending to open vowels. Syllable boundaries were selected in correspondence of local minima in the resulting weighted sequence [3].

For each utterance, the F0 contour was calculated by means of an autocorrelation based algorithm. Signals were hamming windowed every 5 ms and energy and voicing thresholds were used to skip speechless or unvoiced signal intervals. Finally, a gross errors detection and removal

procedure was applied to get a reliable fundamental frequency contour.
Once the syllabic segmentation and labeling was completed, for each syllabic unit the RMS energy was calculated, and given the pitch contour, F0 mean and range values were stored. At utterance level, minimum and maximum F0 values together with peak RMS energy were taken into account. Data were aggregated and elaborated in order to trace the variations of emotional styles parameters with respect to the neutral case. The following table reports a summary of these values.

| | angry | happy | sad |
|---|---|---|---|
| utterance maximum F0 | -7.0% | 16.9% | -14.7% |
| utterance minimum F0 | -6.1% | -0.4% | 0.8% |
| utterance peak RMS energy | 3.7% | 0.0% | -2.4% |
| stressed syllable F0 range | -12.7% | 63.1% | -37.5% |
| unstressed syllable F0 range | 3.8% | 56.4% | -8.3% |
| stressed syllable F0 mean | -4.4% | 15.4% | -14.6% |
| unstressed syllable F0 mean | -4.7% | 12.0% | -13.4% |
| stressed syllable length | -3.3% | -4.7% | 5.3% |
| unstressed syllable length | 0.1% | -2.5% | 4.5% |
| stressed syllable RMS energy | 2.9% | 2.8% | -2.5% |
| unstressed syll. RMS energy | 3.7% | 3.3% | -2.4% |

*Table 2: Percentage variations of some prosodic parameters for three emotional styles.*

## 4. SPEECH SYNTHESIS

To synthesize emotions, we applied prosody transplantation techniques to the output of our TTS system. Units selection was made observing phonetic and prosodic rules, depending on the input text, and if no particular constraint was imposed the output speech prosody was rather neutral [4]. To simulate the three emotions, a supra-segmental and a segmental processing stage were then applied to the TTS output data. These data consist of concatenated waveforms, phonetic labeling, pitch marks and F0 contour. Syllables boundaries were also calculated with the same technique used during the analysis of the emotional database. The fundamental frequency curve was stylized with linear regression, on a slope variation function basis. In fact, break points of the stylized curve correspond to abscissas in the original curve whose derivatives exceed a fixed threshold. A pruning procedure was also applied to remove redundant points.

In the supra-segmental processing stage, the stylized pitch contour was modified in order to adjust the global F0 range according to the utterance maximum and minimum F0 variation coefficients. In this way, macro-prosodic patterns were emphasized for high activation styles and de-emphasized for the sad one. In the following stage, segmental information was used and a more precise tuning was made distinguishing stressed from unstressed syllables. F0 mean and F0 range values were set according to the variation coefficients except in case of pre-pause patterns that were modified more softly. Syllables target durations were recalculated on the basis of the rules parameters. Different scaling factors were then applied to speech pauses setting them longer in the sad style and shorter in the other styles. Energy target values were also calculated according to the model parameters.

A time domain PSOLA-like technique was used to modify waveforms according to the new pitch and acoustic unit lengths values, while a gain function was used to set the correct energy values. In some cases, in synthesizing emotional speech samples, we had to tune the scaling coefficients to avoid disagreeable distortions. In particular, high activation styles have shown some critical aspects for this high-pitched female voice.

## 5. EVALUATION TEST

A perceptual test was carried out to evaluate the effectiveness of this paradigm, used to simulate the three basic emotions with our TTS system. A passage taken from a tale by a famous Italian writer was used to synthesize three emotional samples plus the neutral one. Also in this case, not to influence the listeners, the text had no emotional content. Eight volunteers were asked to listen to the four samples in random order and to evaluate how much sad, angry, happy or neutral each stimuli sounded. They were allowed to listen to the sentences more than once if necessary. Rating range was from 0 to 5 and results are shown in Table 3.

| | neutral | angry | happy | Sad |
|---|---|---|---|---|
| neutral (TTS) | 4.1 | 0.0 | 0.1 | 0.5 |
| angry (TTS) | 0.9 | 2.6 | 0.0 | 0.5 |
| happy (TTS) | 0.1 | 0.5 | 2.0 | 0.0 |
| sad (TTS) | 0.0 | 0.0 | 0.0 | 4.3 |

*Table 3:Evaluation rates of the synthesized emotional samples (rates range from 0 to 5).*

## 6. CONCLUSIONS

The described framework used to render emotions in a Text-To-Speech synthesis system has proven computationally efficient since most of the necessary information is stored in the TTS database. It is clear that this simulation has some limitations too, since it is not possible to stretch the waveform to any extent without producing perceivable distortions. Actually, more features should be investigated to improve this prototype, in particular spectral features should have an important role in describing articulatory alterations that characterize many emotional styles. On the other hand, the good recognition rates of the synthesized speech encourage future improvements of this work.

## 7. REFERENCES


[1] Quazza S., Donetti L., Moisa L, Salza P.L.,, "ACTOR: a Multilingual Unit-Selection Speech Synthesis System", *4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire Scotland, Sep. 2001.*

[2] Brugnara F., Falavigna D., and Omologo M., "Automatic Segmentation and Labeling of Speech based on Hidden Markov Models", *Speech Communication, Vol. 12, no. 4, pp. 357-370, August 1993.*

[3] Cutugno F., D'Anna L., Petrillo M., and Zovato E., "APA: towards an Automatic Tool for Prosodic Analysis". *Speech Prosody 2002, Aix-en-Provence, pp. 231-234.*

[4] Balestri M., Pacchiotti A., Quazza S., Salza P., and Sandri S., "Choose the Best to Modify the Least: a New Generation Concatenative Synthesis System", *Proceedings of EUROSPEECH '99, Budapest, Vol. 5, pp. 2291-2294.*