Prosodic analysis of a multi-style corpus in the perspective of emotional speech synthesis

Enrico Zovato, Stefano Sandri, Silvia Quazza, Leonardo Badino

Speech Technology Division
Loquendo S.p.A., Turin – Italy
{enrico.zovato,stefano.sandri,silvia.quazza,leonardo.badino}@loquendo.com

Abstract

This paper describes the collection and analysis of a multi-style emotional speech corpus, accomplished to study the variations of some acoustical parameters. Specifically, three emotional styles were considered: happiness, sadness and anger. Speech data in a neutral style were also collected, and prosodic differences of each style with respect to this neutral baseline were quantified. According to the analysis results, experiments were also carried out to synthesize emotional styles by modifying prosodically neutral signals both extracted from the original corpus and produced by our Text To Speech synthesis system. Perceptual tests were made to evaluate the effectiveness of the adopted method.

1. Introduction

Nowadays speech synthesis systems have reached a high degree of intelligibility and satisfactory acoustical quality. The goal of next generation speech synthesizers is to express in a natural way the variability typical of human speech or, in other words, to reproduce in a reliable way different speaking styles and particularly the emotional ones.

Many authors agree to classify emotions in a continuous multi-dimensional space whose dimensions are activation, valence and power. Many studies have also pointed out the correlations between emotional dimensions and acoustical parameters such as fundamental frequency, intensity and speech rate [1]. In order to reproduce emotional styles in synthetic speech systems, full control over these parameters is thus necessary. This means that prosody and voice quality has to be manipulated in real time, depending on the speech style to be simulated. Formant or rule-based synthesis systems allow a high degree of control over spectral and prosodic parameters even if the produced speech lacks in naturalness. In diphone concatenation synthesis systems, good control over prosodic parameters is possible but it is rather difficult to control voice quality. On the other hand, corpus based TTS systems could be designed in order to switch among a finite number of corpora, one for each style, depending on the situation, and preserving in this case naturalness and acoustical quality [2,3,4]. In this work, a different approach is experimented to simulate emotional styles in a corpus based speech synthesis system. More precisely, it is based on the use of morphing techniques to manipulate some parameters of prosodically neutral signals like those generated by a variable length unit selection TTS when no prosodic constraints are imposed.

Following this approach, an Italian emotional speech database was recorded and analyzed to verify the correlations and to quantify, for the emotional styles, the prosodic parameters variations with respect to a neutral situation. A set of variation coefficients was produced for each speaker of the database and by means of signal processing techniques some experiments were done to simulate these styles by modifying the speech waveforms produced by LoquendoTTS system [5]. Finally, perceptual tests were made to evaluate the efficacy and quality of this prototype.

Next section describes the composition and acquisition of the emotional speech corpus. Section 3 describes the analysis steps and their results; in section 4, the methodology used to synthesize speech samples in different styles is presented while its evaluation is reported in section 5, discussion and conclusions will follow.

2. Multi-style emotional speech database

Three professional Italian speakers, one female and two males, were recorded for this task. They were about 30 years old and familiar with this kind of recordings since they are the same speakers used for the collection of LoquendoTTS corpora [6]. They had to read 25 sentences in three emotional styles that were: happiness, anger and sadness. Besides, they recorded the same 25 sentences in a neutral way and this last set of signals was used as the baseline for successive elaborations. Sentences were about 10 words long, generally composed of two parts separated by a comma.

Scripts were designed in order to be phonetically balanced. As a consequence, most of the sentences were non-sense and had no semantic emotional content. In this way none of these texts could influence the talents provoking any emotional attitude. On the other hand, emphatic performances were avoided because we wanted these samples to be as natural and reliable as possible. Recordings were accomplished in a sound treated room using high quality microphone and digital acquisition equipment. Signals were originally sampled at 44.1 kHz (16 bit), and were subsequently downsampled to 16 kHz for analysis purposes.

A selection of 48 out of 300 samples was used to make some evaluations. A listening test was proposed to 10 volunteers. They had to listen to this subset containing sixteen utterances of different styles for each speaker. Utterances were proposed in random order and listeners interactively had to choose the closest style among the four possibilities: sad, happy, angry or neutral. As shown in table 1, in most of the cases emotional styles were correctly identified. Sometimes, however the neutral style was confused with the sad one and the happy style with the angry one. Nevertheless the good recognition percentages proved the reliability of this database as a basis for our study.

	rec. as neutral	rec. As angry	rec. as happy	rec. as sad
neutral	**90.0 %**	0.0 %	0.8 %	9.2 %
angry	3.3 %	**90.8 %**	0.0 %	5.9 %
happy	5.8 %	9.2 %	**83.3 %**	1.7 %
sad	5.8 %	0.0 %	0.0 %	**94.2 %**

Table 1:Emotional speech database recognition rates.

3. Database analysis

The analysis of this corpus consisted in calculating different parameters relative to each utterance and to each acoustical unit. Syllable was chosen as the reference acoustic unit as it can be considered the most natural speech unit. Fundamental frequency, RMS energy and syllables durations were the main features calculated.

Segmentation into syllabic units was automatically achieved, exploiting a phonetic transcriber, a phoneme to syllable converter and a phonetic aligner. The Italian grapheme to phoneme transcriber used in this framework is based on rules that scan graphemes left to right and, depending on the phonetic context, the word position and morpheme boundaries produce the output phonetic string. Lexicon files are used to manage transcription exceptions. At this stage, lexical stress is also assigned by means of a word classifier trained on large lexicons whose words are clustered according to their stress position. The phonetic string and the relative signal were the input to an HMM-based phonetic aligner [7]. Hidden Markov models were trained with annotated signals produced by the same speakers that recorded the emotional database. In fact, these speakers were formerly recorded to produce larger corpora for the development of some TTS Italian voices and at this early stage the training material was selected and manually annotated.

Once the phonetic alignment was completed, segmentation into syllables was obtained by exploiting the sonority scale. In this scale phonemes are ordered according to their voicing properties and consequently they are associated to numerical coefficients. Voiceless phonemes like unvoiced plosives have the lowest values while open vowels have the highest values. In this way numeric sequences are associated to phonetic strings and local minima of these sequences are likely syllabic boundaries that surround a vocalic nuclei [8]. To get more accurate data, expert phoneticians checked the phonetic transcriptions and syllables segmentations. In some cases extra-linguistic phenomena or other artifacts, such as glottal stops, speech pauses, or hesitations occurred. All these phenomena had to be manually annotated since they were not present in the automatically generated phonetic string.

For each utterance, the fundamental frequency contour was calculated on the basis of a frame analysis with 5 ms overlap interval. For each frame, the positions of most prominent autocorrelation peaks were selected and, exploiting dynamic programming techniques, the optimum pitch path was calculated. A residual error detection and correction procedure was also used to get a more reliable F0 contour. Finally, for each acoustic unit, the root mean square energy was calculated. Once the phonetic labeling, the segmentation into syllabic units and pitch and energy contours calculation were completed, for each speaker and for each emotional style some data, both at utterance and syllable level, could be extracted. Regarding utterances, average values of maximum F0, minimum F0, mean F0, mean and peak energy, were extracted. As concerns syllables, for each speaker and for each style, mean values over the whole corpus of the following parameters were calculated: maximum F0, minimum F0, mean F0, F0 range, RMS energy, syllable duration and the duration of its voiced part. Syllables were also classified into categories depending on their position inside the sentence and whether they were stressed or not. In this context, we considered as stressed those syllables containing a phoneme with lexical stress. The categories were: stressed, unstressed, first stressed (the first stressed syllable of the sentence or after a speech pause), last stressed (the last stressed syllable of the sentence or before a speech pause). Speech pauses were also taken into account and their mean duration values were calculated for each style.

For the extracted parameters of each emotional style, variation coefficients with respect to the baseline were calculated. In this way speaker dependent numeric rules were available in view of speech synthesis. Many variations are common in sign to all the considered speakers even if absolute values may significantly differ. In next subsections analysis results are discussed and in the following tables mean variation values of some prosodic parameters are reported.

3.1. Fundamental Frequency

In the happy style, F0 contour significantly increases, both mean and range values grow at syllable level. In a symmetrical way the same parameters decrease in the sad style. In the angry style, mean F0 is almost unchanged with respect to the neutral case, while range values slightly decrease. Actually, different behaviors were found among speakers. The samples of one of the two male speakers were characterized by increasing values of F0 mean and F0 range with respect to the neutral style, while the other two speakers were characterized by a small variation of the F0 mean values and slight decreasing of its range. This is due to the fact that speakers expressed anger in different ways, resulting in hot anger in some cases and cold anger in others.

For high activation styles, utterance maximum F0 values grow more than minimum values, while they both decrease almost the same quantity in the sad style.

		angry/ neutral	happy/ neutral	sad/ neutral
utterance	min F0	1.04	1.09	0.83
utterance	max F0	1.08	1.20	0.82
utterance	mean F0	1.01	1.19	0.83
first stressed. syllable		0.96	1.11	0.77
stressed syllable		1.04	1.25	0.82
last stressed syllable		1.02	1.31	0.90
unstressed syllable		1.01	1.16	0.81
first stressed syllable	F0 range	0.90	1.39	0.57
stressed syllable		0.93	1.61	0.57
last stressed syllable		0.77	1.58	0.50
unstr. syllable		1.00	1.50	0.74

Table 2: Mean variations of F0 parameters for three emotional styles with respect to the neutral one.

3.2. Syllables durations

Happy and angry styles are characterized by significant unit length shortenings, which means an increase in speech rate. In the sad style we found contrasting behaviors with one speaker who reduced the speech rate while the others left their speech rate almost unchanged. Regarding the durations of syllables voiced intervals we found that they generally tend to decrease in the sad and angry styles.

Speech pauses get shorter in the happy style and considerably longer in the sad one.

		angry/ neutral	happy/ neutral	sad/ neutral
first stressed syllable	mean duration	0.93	0.94	0.98
stressed syllable		0.90	0.92	0.96
last stressed syllable		0.93	0.91	1.00
unstressed syllable		0.94	0.96	1.00
pause		1.00	0.83	1.48

Table 3: Mean variations of duration parameters for three emotional styles with respect to the neutral one.

3.3. Intensity

RMS energy increases in the happy and angry styles while it decreases in the sad one. The high activation styles are characterized by bigger variance and sometimes, significant energy peaks occur in the final parts of the sentences too.

		angry/ neutral	happy/ neutral	sad/ neutral
Utterance	max energy	1.03	1.04	0.98
Utterance	mean energy	1.02	1.03	0.97
first stressed syllable		1.01	1.03	0.97
Stressed syllable		1.02	1.03	0.97
last stressed syllable		1.00	1.04	0.95
Unstressed syllable		1.02	1.03	0.96

Table 4: Mean variation of RMS energy parameters for three emotional styles with respect to the neutral one.

4. Synthesis

The analysis of the multi-style speech database was necessary to find experimental motivated rules to be applied when synthesizing the considered emotional styles. Our goal was finding simple rules to modify the prosodic parameters of concatenated speech waveforms, without degrading too much the final acoustic quality. To achieve this result, we imposed target values of F0, energy and acoustic unit durations, obtained by applying the appropriate variation coefficients to the original parameter values. Our starting point was a synthesized speech waveform having a neutral prosody. The optimum unit selection as obtained by our corpus based TTS system was not compromised, being the simulation of emotional styles achieved in a post-processing stage. The TTS output consisted of the waveform, together with phonetic labeling and segmentation information. The segmentation into syllables was also available as well as the partition of the input text into sentences.

As regards the fundamental frequency, target contour was calculated through different stages but the main idea was to maintain as much as possible the original morphology. Starting from the original pitch contour, a linear stylization was obtained. First, those abscissas in proximity of which significant slope variations occurred were classified as break points. Then a pruning procedure was applied to select those points whose pitch gap exceeded a fixed threshold in semitones. Each portion of the pitch contour between two target points was stylized with minimum squares linear regression. The final stylized pitch contour could be used in a computationally efficient way to modify F0 mean and slope values [9].

The target pitch contour was obtained by modifying the linearly stylized contour by means of variation rules applied both at utterance and syllable level. At the beginning the whole utterance F0 range was considered. Target maximum and minimum F0 values were calculated according to their variation coefficients, and all the pitch values of the sentence were recalculated according to the new utterance pitch range. Following this scheme pitch contour was compressed or expanded depending on the style to be synthesized. Macroprosody was thus modified by enhancing peaks and slopes in case of high activation styles while de-emphasizing them in the sad one.

Once macroprosody was set, acoustic units information was taken into account and for each syllable the pitch range was adjusted in the stylized F0 contour. The same analysis scheme was used to set these variation values; in fact syllables were classified depending on their position in the sentence and on their stress type. For each syllable a pitch range scaling factor was thus applied depending on the syllable classification. Syllables mean F0 values were set in order to be the same of the corresponding mean F0 syllable value as resulting from the previous pitch elaboration stage. Variation F0 range coefficients were also adapted not to exceed the utterance maximum F0 value and not to be lower than the utterance minimum F0 value.

Depending on the syllable classification, syllables duration values were recalculated by multiplying the original values for the appropriate variation coefficient. Finally, time and pitch target values together with the waveform were the input to the time/pitch scaling module. Speech rate and fundamental frequency modifications were obtained by means of a time domain pitch synchronous overlap and add procedure.

An adaptive gain function that assigned different gain values to each syllable was then applied to reproduce energy variations.

The coefficients used in the first experiments were exactly those resulting from the analysis of the emotional database. However we noticed that in some cases better perceptual results could be achieved by slightly changing these values. In general these changes regarded the absolute values of the variation coefficients but not their signs. In particular we found that significant improvements could be obtained by changing coefficients regarding the acoustic unit durations. This is probably due to the fact that the lack of voice quality modifications has to be compensated by increasing the variations of these prosodic parameters in order to get an adequately recognizable emotional style.

5. Evaluation Test

The scheme described in the previous section was used to synthesize some samples for a perceptual test. First, a set of neutral signals was collected and then prosodic rules were applied to synthesize the three emotional styles. The neutral signals that had to be manipulated were partially extracted from neutral samples of the original emotional corpus and partially generated by our TTS system whose prosody, as said before, is almost neutral. The test set consisted of 72 samples, styles and speakers being equally represented (18 samples for each style including the neutral one, 24 for each speaker). The sentences were proposed to listeners in random order and, in this case, they had to give a preference among five choices, the styles plus the "other" option. Results of the 72 samples evaluation test are reported in table 5 while table 6 contains the evaluation of the subset synthesized using the output of our TTS system as baseline.

	rec. as neutral	rec. as angry	rec. as happy	rec. as sad	rec. as other
neutral	**71.1%**	3.3%	10.0%	7.8%	7.8%
angry	17.8%	**52.2%**	16.7%	3.3%	10.0%
happy	5.6%	24.4%	**53.3%**	3.4%	13.3%
sad	10.0%	1.1%	1.1%	**82.2%**	5.6%

Table 5: Synthesized emotional speech samples recognition rates.

	rec. as neutral	rec. as angry	rec. as happy	rec. as sad	rec. as other
neutral	**51.1%**	6.7%	15.6%	11.1%	15.6%
angry	11.1%	**68.9%**	8.9%	2.2%	8.9%
happy	0.0%	40.0%	**46.7%**	2.2%	11.1%
sad	6.7%	2.2%	2.2%	**77.8%**	11.1%

Table 6: Synthesized emotional speech samples recognition rates (TTS baseline).

Sad style seems to be the best synthesized, since its recognition rates are quite high. More active styles are adequately recognized with rates beyond the chance level. Significant confusion values however occur, particularly in those samples whose baselines were generated by the TTS system. In these cases the happy style is often confused with the angry one.

6. Discussion and conclusions

This framework used to simulate emotional speech, has proven efficient from a computational point of view. Three basic emotions were simulated and recognition tests show that it is possible to partially characterize synthetic speech. On the other hand, not all the aspects of emotional speech production were considered. In fact, prosodic parameters such as fundamental frequency, energy and speech rate were modified while nothing was attempted to simulate the voice quality changing. This is mainly due to the fact that text to speech synthesis systems must be efficient from a computational point of view and accurately modifying spectral parameters requires a significant computational load. Moreover, in corpus based speech synthesis systems it is not always possible to stretch waveforms to any extent without producing annoying distortions. In particular, we have found that it is difficult to simulate very high active emotional styles starting from a neutral voice. On the contrary it is easier to simulate low activation styles like the sad one. During this work we have also experienced that each voice requires a precise tuning step. This means that there are no general rules to be easily applied to the baselines. It was also found that, for some speakers, the synthesis of particular styles was more convincing than others. This is probably due to the fact that the neutral style varies from speaker to speaker, depending on its own natural prosodic characteristics.

In corpus based synthesis systems, the use of different speech corpora for different emotional styles is a strategy to simulate speech variability while maintaining an acceptable acoustical quality [3]. This is however a discrete solution, in the sense that it is possible to simulate a limited number of emotional styles. Maybe a hybrid approach could be the good compromise to obtain both the variability typical of human speech and acceptable acoustical quality. Depending on the style to be synthesized, waveforms could be selected from the corpus whose characteristics are closest to the style to be simulated. Then, signal processing techniques could be applied to slightly modify prosodic parameters of concatenated waveforms in order to match the style target characteristics, thus obtaining a more reliable simulation.

7. References

[1] Schröder M., Cowie R., Douglas-Cowie E., Westedijk M & Gielen, S. "Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis", *Proceedings of EUROSPEECH 2001, pp. 87 - 90, Scandinavia, 2001.*

[2] Schröder, M. "Emotional Speech Synthesis: A Review", *Proceedings of EUROSPEECH 2001, pp. 561 – 564. Scandinavia, 2001.*

[3] Iida, A. Campbell, N. Higuchi, F. & Yasumura, M. (2003) "A corpus-based speech synthesis system with emotion", *Speech Communication*, n. *40, pp. 161-187.*

[4] Murray, I.R. and Arnott, J.L. "Synthesising emotions in speech: is it time to get excited?", *Proceedings of ICSLP 96, pp.1816-1819, 1996.*

[5] 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.*

[6] 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.*

[7] 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*.

[8] 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*.

[9] 'T Hart, J.; "F_0 stylization in speech: straight lines versus parabolas". *Journal of the Acoustical Society of America*, 90 (6), 3368-3370. 1990.