Emofilt: the Simulation of Emotional Speech by Prosody-Transformation

Felix Burkhardt

T-Systems International GmbH
TZ Engineering Networks, Products & Services
Germany
felix.burkhardt@t-systems.com

Abstract

Emofilt is a software program intended to simulate emotional arousal with speech synthesis based on the free-for-non-commercial-use MBROLA synthesis engine. It acts as a transformer between the phonetisation and the speech-generation component. Originally developed at the Technical University of Berlin it was recently revived as an open-source project written in Java (http://emofilt.sourceforge.net). Emofilt's language-dependent modules are controlled by external XML-files and it is as multilingual as MBROLA which currently supports 35 languages. It might be used for research, teaching or to implement applications that include the simulation of emotional speech.

1. Introduction

Emofilt is a software program intended to simulate emotional arousal with speech synthesis based on the free-for-non-commercial-use MBROLA synthesis engine ([1]). It acts as a transformer between the phonetisation and the speech-generation component. Originally developed at the Technical University of Berlin in 1998 it was recently revived as an open-source project and completely rewritten in the Java programming language. Some of the experiments done with Emofilt are described in [2].

Emofilt's language-dependent modules are controlled by external XML-files and it is as multilingual as MBROLA which currently supports 35 languages. The emotional simulation is achieved by a set of parameterized rules that describe manipulation of the following aspects of a speech signal:

- Pitch changes
- Duration changes
- Voice Quality (simulation of Jitter and support of multiple-voice-quality database)
- Articulation (substitution of centralized/decentralized vowels)

Emofilt consists of two main interfaces:

- Emofilt-Developer: a graphical editor for emotion-description XML-files with visual and acoustic feedback (see figure 1).
- Emofilt itself, taking the emotion-description files as input to act as a transformer in the MBROLA framework.

This paper is organized as follows. In the next section we give a short overview on text-to-speech synthesis and the current situation regarding emotional speech synthesis. Section 3 describes the interface format that Emofilt is operating on and the preprocessing steps it undertakes. In the following section the parameters are described that are currently modifiable. A final section describes the configuration possibilities that Emofilt offers. Mainly this alludes to the storage of modification sets as "emotions" and the voice-description file that encapsulates multilingual differences. We close with an outlook section, where ideas for further development are described.

2. Foundations

The simulation of emotional speech by means of speech synthesis started soon after the first mature speech synthesizers were developed (e.g. [3]) and is gaining rising attention with the more widespread use of speech synthesis in voice-portals and multimodal user interfaces. For an overview on the history of emotional speech synthesis the reader may be referred to [4].

A speech synthesizer consists generally of two main components (following the terminology in [5]).

- The so-called NLP (Natural Language Processing)-module does the conversion of orthographical text into a phoneme-alphabet and calculates a prosody description.
- In a second step the DSP (Digital Speech Processing)-module does the synthesis of speech signal from output of NLP-component.

Non-prosodic features are usually not part of the NLP-DSP interface, because they carry little semantic information. In the future the extension of this interface by emotion-related information on different layers is foreseeable. The detail of such a markup will depend on the technology of the DSP-component (see next paragraph). Formant synthesizers might accept phonation-related specification, because voice-quality is dependent on emotional expression. Unit-selection synthesis engines on the other hand might take a direct description of a desired emotional expression as input that leads them to select units from a database annotated with emotional markers.

The engines that synthesize the speech (DSP-component) are based mainly on three main technologies:

- **Non-uniform unit-selection** : best fitting chunks of speech from large databases get concatenated, thereby minimizing a double cost-function: best fit to neighbor unit and best fit to target prosody. Because signal manipulation is reduced as much as possible, the resulting speech sounds most natural (similar to the original speaker) as long as the utterance to synthesize is close to the original domain of the database. A transformation-program between NLP- and DSP module like Emofilt would be difficult to integrate, because the two components typically are heavily intertwined (both rely strongly on the same database).
- **Diphone-synthesis**: speech concatenated from diphone-units (two-phone combinations). The prosody-fitting is

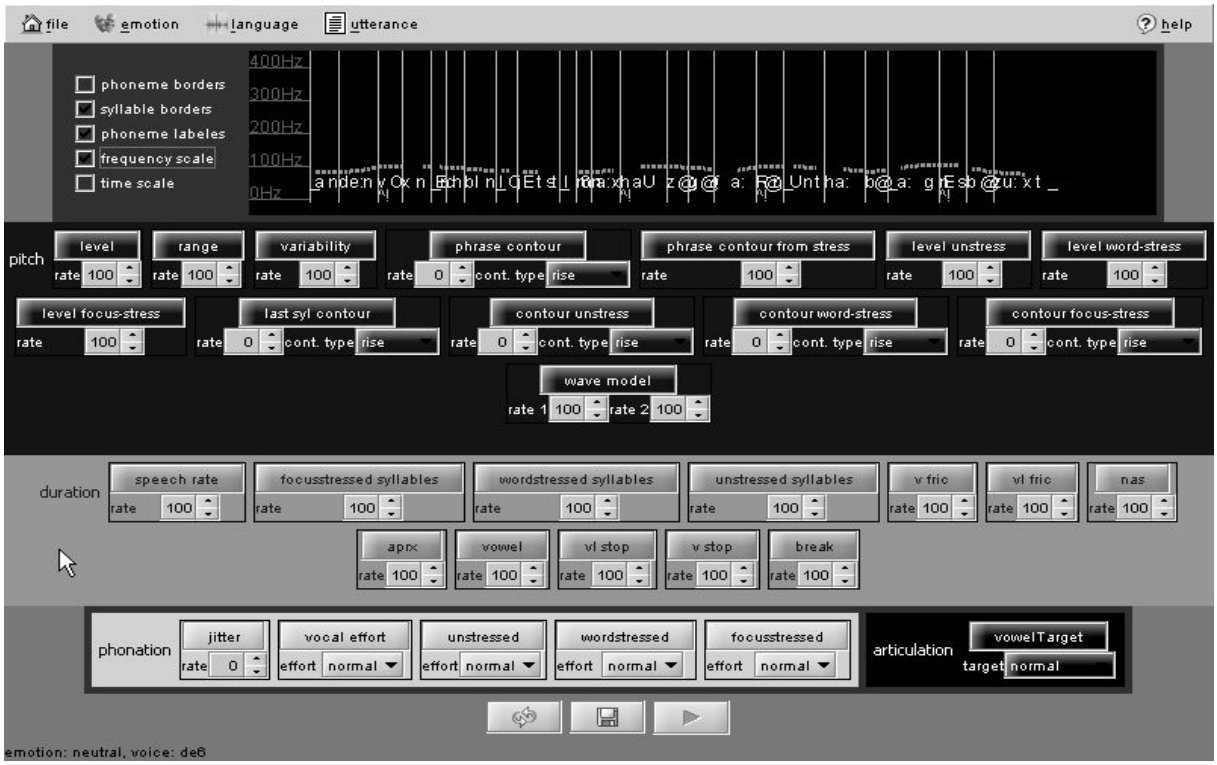

Figure 1: Emofilt Developer Graphical User Interface

done by signal-manipulation algorithms that depend on the coding of the diphone-units. This approach provides for a relatively small footprint but does not sound very natural, because the signal manipulation leads to artifacts. The DSP engine used with Emofilt is based on diphone-synthesis.

- **Formant-synthesis**: speech synthesized by physical models (formants are resonance frequencies in the vocal-tract). Very flexible and smallest footprint, but sounds very unnatural because the models are not yet good enough. Nonetheless this approach is most promising when it comes to simulation of emotional expression as emotional models can directly be integrated in the calculation of the synthesis parameters. As shown in [6] speech indistinguishable from natural data can be synthesized when the parameters are carefully modeled. Although formant synthesis has attracted little interest by the research community in the last years (perhaps best illustrated by the fact that to our knowledge no wide-band formant-synthesizer has been developed), we believe therefore that this approach will have a renaissance.

One of the most pressing tasks today for speech synthesis is to find solutions for the trade-off between the naturalness of data-concatenation engines and the flexibility of signal generation synthesis. This problem is evident in the case of emotional speech synthesis, where voice quality features become very important and it is often not practicable to extend the database by emotion-related variants. Several approaches tackle these task:

- In [7], we describe an emotional speech synthesis engine based on formant-generation technology. It's based on a hybrid formant/unit-selection approach in the sense that the formant tracks are derived from a natural database and the noise part of the sounds are computed by rules. This approach is gaining attention (e.g. [8]). The main problems arise from the difficulty to derive the formant tracks from natural data without manual correction.
- An interesting approach is described in [9]: doing unit-selection on extremely large databases covering all sorts of speaking situations. This approach provides also for the generation of the large number of extra-linguistic sounds that appear in natural language.
- In [10] the authors propose "speaking style modeling" with HMM-based speech synthesis trained on emotional data.
- [11] describes the preparation of multiple-voice quality databases for MBROLA. Two German databases (male and female) were recorded, each providing for soft, normal and load voice-quality. Emofilt provides for an interface to this voices.

The approach described here does not handle the above mentioned problem but does as far as possible without manipulation of non-prosodic features. Nonetheless it might be useful in research, teaching and applications that utilize emotional speech simulation.

3. Preprocessing: Syllabication and application of stress

The input format for Emofilt is MBROLA's PHO-format. Each phoneme is represented by one line, consisting of the phoneme's name and its duration (in ms). Optionally follow-

ing is a set of F_0 description tuples consisting of a F_0-value (in Hertz) and a time value denoting a percentage of the duration. Here is an example of such a file:

```
_ 50
v 35 0 95  42 95   84 99
O 55 18 99 27 103 36 107 45 111
x 50
@ 30 0 178 16 175 80 160
```

The valid phoneme-names are declared in the MBROLA-database for a specific voice and must be known by Emofilt (see section 5.2). The first thing that happens if an utterance gets loaded is that the phonemes get assigned a phoneme class and a sonority-value. The so-called *sonority-hierarchy* (see e.g. [12]) divides phoneme-classes by their strength of resonance. The following types are used by Emofilt (ordered by sonority):

- vowels (long and short)
- approximants
- nasals
- fricatives (voiced or unvoiced)
- stops (voiced or unvoiced)

The phonemes get assigned to syllables following the sonority sequencing principle, that says that in each syllable there is a segment which is the syllable peak; any segment sequences preceding and following this segment have progressively decreasing sonority.

In the next step each syllable gets assigned a stress-type. Emofilt differentiates three stress-types:

- unstressed
- word-stressed
- (phrase) focus-stressed

As the analysis of stress involves an elaborate syntactic and semantic analysis and this information is not part of the MBROLA PHO-format, Emofilt assigns only focus-stress to the syllables that carry local pitch maxima. However, for research scenarios it is possible to annotate the PHO-files manually with syllable and stress markers.

As a further step, a constant frame rate (per default 10 milliseconds) is used to interpolate the F_0-values linearly. This is a preparation for modifications like adding jitter or changes to the pitch-contour.

4. Modification Parameters

In this section we describe the modification rules in detail. The rules were motivated by descriptions of emotional speech found in the literature. As we naturally can not foresee all modifications that a future user might want to apply, we plan to extend Emofilt by an extensible plugin-mechanism that enables users to integrate customized modifications more easily.

4.1. Pitch Modification

The following modifications are provided for pitch feature changes.

4.1.1. Pitch level

The overall level of the F_0 contour can be shifted by multiplying all values with a rate factor (rate=0 means no change). This means that high values undergo stronger changes than low values and was chosen to conform with the human logarithmic hearing.

4.1.2. Pitch range

The pitch range change was motivated by the peak-feature model mentioned in [13] and achieved by a shift of each F_0-value by a percentile of its distance to the mean F_0-value of the last syllable. If range=0, all values become the last syllable's mean value. The shifting corresponds to a contrast change in image processing.

Note that Emofilt currently assumes its input to consist of one "utterance" in the sense of a *short* part of speech that shall be uttered emotionally. This might lead to problems if several sentences are given as input, because utterance-global values like e.g. the "mean pitch value of last syllable" are currently computed only once for the whole of the input phoneme sequence.

4.1.3. Pitch variation

A pitch variation on the syllable-level is achieved by the application of the pitch range algorithm on each syllable separately. The reference value in this case is the syllable's mean pitch.

4.1.4. Pitch contour (phrase)

The pitch contour of the whole phrase can be designed as a rising, falling or straight contour. The contours are parameterized by a gradient (in semitones/sec). As a special variation for happy speech, the "wave model" can be used where the main-stressed syllables are raised and the syllables, that lie equally distanced in between, are lowered. It's parameterized by the maximum amount of raising and lowering and connected with a smoothing of the pitch contour, because all F_0-values get linearly interpolated.

4.1.5. Pitch contour (syllables)

A rising, falling or level contour can be assigned to each syllable-type. Additionally, the last syllable can be handled separately.

4.2. Duration Modification

The speech rate can be modified for the whole phrase, specific sound categories or syllable stress-types separately by changing the duration of the phonemes (given as a percentile). If the duration in consequence of a length reduction is shorter than the frame rate, the phoneme gets dropped. This may lead to a MBROLA error message if the resulting diphone combination of the left-over phonemes is not represented in the voice-database. This behavior should be changed therefore in a scenario were a valid PHO-file has to be guaranteed.

4.3. Articulation Modification

As a diphone synthesizer has a very limited set of phoneme-realizations and doesn't provide for a way to do manipulations with respect to the articulatory effort, as a work-around to change the **vowel precision** the substitution of centralized vowels with their decentralized counterparts and vice versa is possible. This operation was inspired by [3]. The substitution list for this operation must be specified in the language-description file (see section 5.2).

4.4. Phonation Modification

4.4.1. Jitter

In order to simulate Jitter (fast fluctuations of the F_0-contour) the F_0 values can be displaced by a percentile alternating down and up. The preprocessing (see section 3) guarantees that there is a F_0 value for each time frame.

4.4.2. Vocal effort

As stated in section 2, diphone synthesis engines usually don't allow the modification of voice-quality related acoustic features. For German although two voice-databases exist that were recorded in three voice-qualities: normal, soft and loud (see [11]). The change of voice quality can be applied to the whole phrase or specific syllable stress types only. It will only take effect if one of the two German voices (de6 and de7) are loaded.

5. Configuration Files

5.1. Emofilt's "Emotion" Concept

A set of parameter values can be stored in a XML-document and is called an "emotion". The emotion-file can be generated manually or with the Emofilt Developer (see figure 1) by storing the actual configuration. This file is read by Emofilt at startup in order to accomplish the modifications associated with an "emotion".

5.2. Multilinguality

As described in section 3, the valid phoneme-identifiers must be declared in a XML-file that gets also read by Emofilt at startup. It contains furthermore a list of possible substitutions of centralized resp. decentralized vowels. Each language (aka voice) to be processed by Emofilt requires, as it may contain its own phoneme set, an entry in the "language"-file.

5.3. Configuration

The program's fundamental configuration is stored in a central configuration-file. It consists of three sections:

- Paths: The paths to the emotion- and the language-file as well as to the MBROLA executable (for acoustic feedback in Emofilt-Developer).
- Colors and Fonts: The colors and fonts of Emofilt Developer.
- Labels: Labels and Help-messages for the Emofilt Developer. This is specially useful if the tool shall be used by non-English speaking persons.

6. Outlook

We think that Emofilt is specially useful to conduct multilingual comparison studies because of the large number of languages that MBROLA supports.

As a next step in the development it is planned to encapsulate modification-rules like e.g. "change pitch level" in a plug in-like framework with the aim to ease the integration of new modifications, especially for developers not familiar with the Emofilt source-code.

Other improvements will concern the possibility to calculate parameters for shades of an emotion, blending of two emotions or modeling the transition from one emotion to another.

7. Conclusions

Emofilt is an open-source tool to simulate emotional speech with the MBROLA text-to-speech synthesizer. It can be used to do experiments dealing with the effect of acoustic features on emotional perception or to implement applications that include emotional speech synthesis. It is particularly useful in studies investigating multilingual differences. In future Emofilt will support "manipulation plug-ins" that enables users to add functionality in an easy and straightforward manner.

8. Acknowledgments

This work was partially funded by the EU FP6-IST project HUMAINE (Human-Machine Interaction Network on Emotion).

9. References

[1] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van der Vreken, "The mbrola project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes," *Proc. ICSLP'96, Philadelphia*, vol. 3, pp. 1393–1396, 1996.

[2] F. Burkhardt, "Simulation emotionaler sprechweise mit sprachsyntheseverfahren," Ph.D. dissertation, TU-Berlin, 2001.

[3] J. E. Cahn, "The affect editor," *Journal of the American Voice I/O Society*, vol. 8, pp. 1–19, 1989.

[4] M. Schröder, "Emotional speech synthesis - a review," in *Proc. Eurospeech 2001, Aalborg*, 2001, pp. 561–564.

[5] T. Dutoit, *An Introduction to Text-To-Seech Synthesis*. Kluwe Academic Publishers, 1997.

[6] J. N. Holmes, "The influence of the glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Trans. Adio Electroacoust. Au 21*, pp. 298–305, 1973.

[7] F. Burkhardt and W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant-synthesis," in *Proc. of the ISCA Workshop on Speech and Emotion*, 2000, pp. 151–156.

[8] R. Carlson, T. Sigvardson, and A. Sjölander, "Data-driven formant synthesis," KTH, Stockholm, Sweden, Progress Report 44, 2002.

[9] N. Campbell, "Databases of expressive speech," in *Proc. Oriental COCOSDA Workshop 2003, Singapore*, 2003.

[10] M. Tachibana, J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Hmm-based speech synthesis with various speaking styles using model interpolation," in *Proc. Speech Prosody 2004, Nara, Japan*, 2004, pp. 413–416.

[11] M. Schröder and M. Grice, "Expressing vocal effort in concatenative synthesis," in *Proc. 15th International Conference of Phonetic Sciences, Barcelona, Spain*, 2003, pp. 2589–2592.

[12] E. O. Selkirk, "On the major class features and syllable theory," in *Language sound structure: Studies in phonology presented to Morris Halle by his teacher and students*, M. Aronoff and R. T. Oerhle, Eds. MIT Press. Cambridge, Mass., 1984, pp. 107–136.

[13] G. Bergmann, T. Goldbeck, and K. R. Scherer, "Emotionale eindruckswirkung von prosodischen sprachmerkmalen," *Zeitschrift für experimentelle und angewandte Psychologie*, vol. 35, pp. 167–200, 1988.