COMPLEX EMOTIONS - THE SIMULTANEOUS SIMULATION OF EMOTION-RELATED STATES IN SYNTHESIZED SPEECH

Felix Burkhardt[1] *and Benjamin Weiss*[2]

[1]*Deutsche Telekom AG,* [2]*Technische Universität von Berlin*
felix.burkhardt@telekom.de

Abstract: We describe an approach to simulate first and secondary emotional expression in synthesized speech simultaneously by targeting different parameter categories. The approach is based on the open-source system "Emofilt" which utilizes the diphone-synthesizer "Mbrola". The evaluation of the approach by a perception experiment showed that the pure emotions were all recognized above chance. Whereas the results are promising, the ultimate aim to validly synthesize two emotions simultaneously was not fully reached. Apparently, some emotions dominate the perception (fear), and the salience or quality of synthesis does not seem to be equally distributed over the two feature bundles.

1 Introduction

Current state-of-the-art synthesizers support the simulation of specific speaking styles in one way or the other. A specific form of speaking style is emotional speech. Many articles in the literature can be found on strategies on how to simulate a single emotional expression described by a categorical designation or by single point in an emotion-dimensional space. The expression of only one emotional state in speech is a first step towards more naturalness. Nonetheless, it is an over-simplification to only model one emotional state at every given time. In the real world, there are many situations conceivable where at least two emotion-related states influence the speaking style. Especially when the term "emotion" gets broadened to "emotion-related state", i.e. includes mood, alertness or personality.

Psychologists have been very interested in the topic of mixed emotions, emphatically debating the degree to which conflicting emotions can be simultaneously experienced. One perspective suggests that the ability to experience conflicting emotions simultaneously is limited, as positive and negative emotions represent opposite dimensions on a bipolar scale. A second perspective argues the opposite, namely, that emotional valence is represented by two independent dimensions. Thus, not only can one simultaneously experience conflicting emotions, such a joint experience may be natural and frequently occurring [1].

Although the research on the simulation of affective speaking styles with speech synthesis has a long history [2, 3, 4], to our knowledge until now no one reported on the attempt to find a strategy to display more than one affective state at the same time.

We describe an approach to simulate more than one emotion utilizing the open source program "Emofilt" which in itself is based on the diphone synthesizer "Mbrola" [5] as well as a text-to-phoneme converter, for example the text-to-speech framework "Mary" [6]. The approach is based on the idea to mix configurations for several feature categories during the synthesis process. Feature categories are for example: articulation, phonation, pitch or duration parameters. We evaluated this approach with a perception experiment. In a systematic confusion, each of Darwin's four "basic emotions" (joy, sadness, fear and anger) was combined with all other emotions and used as an emotional model to synthesize four target phrases taken from

the Berlin emotional database EmoDB. The two German target phrases were generated with a male and female Mbrola voice (de6 and de7).

In a forced-choice evaluation experiment using the Speechalyzer Toolkit [7], 32 subjects categorized each stimulus with a primary and alternative label, applying either one of the four emotions or "neutral". The alternative labels were introduced as second opinion, no complex emotions were inferred. The "neutral" emotion was introduced as default in case of uncertainty. The pure emotions were all recognized above chance. Results for the complex emotions indicate that one parameter bundle is significantly eliciting the target emotions, whereas the second bundle reveals mixed results. In particular, the second rating was dominantly "neutral". Nevertheless, when analyzing the non-neutral ratings, the intended complex-emotions work especially well for combinations with fear.

This article is structured as follows. Firstly we describe the speech synthesizer in section 2. We then report on the way we approached the simultaneous simulation of two affective states in section 3. The next section 4 describes the perception experiment that was used to verify our approach. Lastly, section 5 discusses the results and insights that could be gained from the experiment. We conclude the paper with an overview and some ideas for improvements in section 6.

2 Emofilt

Emofilt [8] is a software program intended to simulate emotional arousal with speech synthesis based on the free-for-non-commercial-use MBROLA synthesis engine [5]. 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 revived in 2002 as an open-source project and completely rewritten in the Java programming language.

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 following 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
```

Emofilt's language-dependent modules are controlled by external XML-files and it is as multilingual as MBROLA which currently supports 35 languages.

Emofilt consists of three 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.
- A storyteller interface that can be used to mark phrases in a dialog with colors that correspond to emotional expression [9].

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). The valid phoneme-names are declared in the MBROLA-database for a specific voice and must be known by Emofilt.

In a first 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.

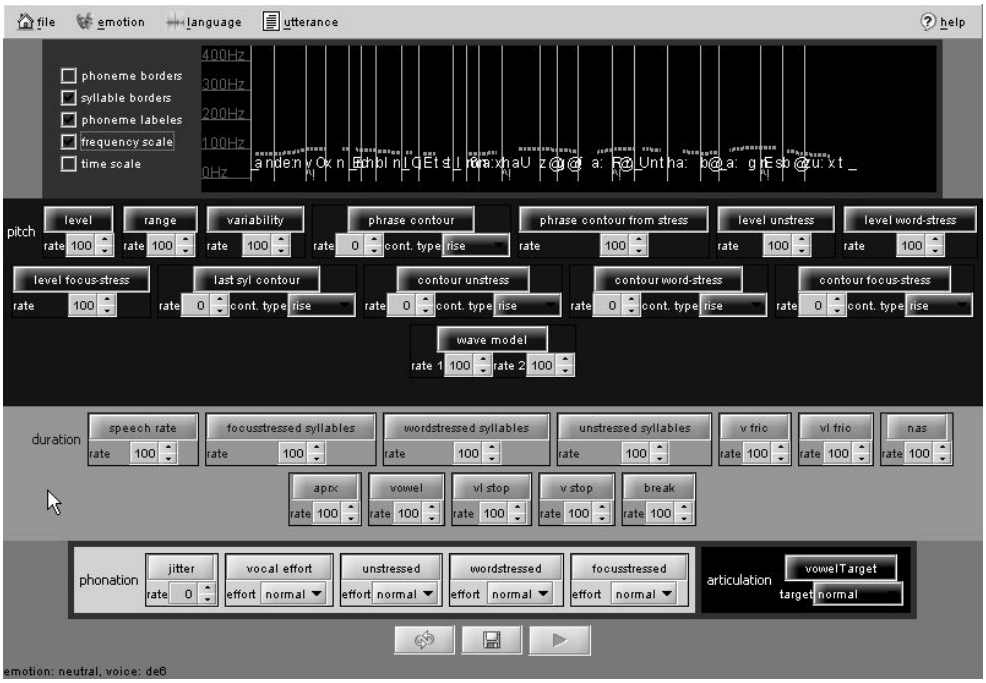

Figure 1 – Emofilt Developer Graphical User Interface

The emotional simulation is achieved by a set of parameterized rules that describe manipulation of the following aspects of a speech signal:

- Pitch changes, for example: “Model a rising contour for the whole utterance by ordering each syllable pitch contour in a rising manner”.
- Duration changes, for example: “Shorten each voiceless fricative by 20%”.
- Voice Quality, for example the simulation of jitter by alternating F0 values and support of a multiple-voice-quality database.
- Articulation precision changes by a substitution of centralized and decentralized vowels.

The rules were motivated by descriptions of emotional speech found in the literature [3]. As we naturally can not foresee all modifications that a future researcher might want to apply, we extended Emofilt by an extensible plugin-mechanism that enables users to integrate customized modifications more easily.

3 Data Generation

As stated in section 2, Emofilt's modification rules are categorized into four modification categories: pitch, duration, voice-quality and articulation.

Now the first naive idea on how to simulate two different states at the same time would perhaps be to simply fuse the modification parameters for each desired expression by using the average value. For example if anger leads to an increase of stressed syllables by 20% and sadness leads to a decrease of 20%, use 0% modification because it's the average value. But, as can be seen directly from the example, this may easily lead to an equalization between the two expressions and thus neither expression would be detectable.

So instead we used the distinction between prosodic features (i.e. pitch and duration) to express the more "foreground" emotion and the other feature categories, namely voice-quality and articulation, to express the secondary emotional state. This distinction lacks a basis in psychological models, but was motivated purely by pragmatic motivation.

The following example displays the configuration for happy as a primary and sadness as a secondary emotion.

```
<emotion name="happySad">
        <phonation>
                <jitter rate="10" />
                <vocalEffort effort="soft" />
        </phonation>
        <articulation>
                <vowelTarget target="undershoot" />
        </articulation>
        <pitch>
                <waveModel rate1="150" rate2="100" />
        </pitch>
        <duration>
                <durVLFric rate="140" />
        </duration>
</emotion>
```

As modifications to display happiness, the pitch-contour gets assigned the so-called "wave model" (which means a fluent up-and-down contour between stressed syllables, see [8] for details) and the duration of the voiceless fricatives gets lengthened by 40%. At the same time, the phonation and articulation parameters get altered according to the emotion model defined for sadness, i.e. jitter is added, the vocal effort is set to "soft" and the articulation target values are set to "undershoot".

To generate test samples for evaluation in a systematic confusion, each of Darwin's four "basic emotions" (joy, sadness, fear and anger) was combined with all other emotions and used as primary as well as secondary emotional state. As a reference we added neutral versions, but did not combine neutral with the emotional states. This resulted in 17 samples (4 emotions by 4 + neutral). The target phrases were taken from the Berlin emotional database EmoDB [10]. We used two short and two longer ones.

Table 1 – Confusion matrix for the single basic emotions only. Primary rating in %. Highest values bold.

Prim. Rating / Emotion	Anger	Fear	Joy	Neutral	Sadness	F1
Anger	**.496**	.156	.117	.211	.020	.536
Fear	.223	**.367**	.180	.133	.098	.411
Joy	.066	.180	**.383**	.320	.051	.435
Neutral	.043	.039	.082	**.582**	.254	.488
Sadness	.023	.043	.000	.141	**.793**	.716

Table 2 – Confusion matrix for the emotions synthesized with prosody. Primary rating in %. Highest values bold.

Prim. Rating / Emotion Set 1	Anger	Fear	Joy	Sadness	F1
Anger	**.375**	.337	.239	.049	.3866
Fear	.173	**.518**	.202	.108	.4977
Joy	.248	.206	**.427**	.119	.4340
Sadness	.184	.085	.031	**.700**	.6976

All target phrases were synthesized with a male and female Mbrola German voice (de6 and de7). The resulting number of samples was thus 134 (17*4*2).

4 Perception Experiment

In a forced-choice listening experiment, 32 listeners (15 males, 15 females, 20–39 years old, M=27.26, SD=3.75) assigned all stimuli to one of the four emotions or "neutral". A second rating was asked for as "alternative" categorization. The "neutral" emotion was introduced as default in case of uncertainty. The evaluation was done with the Speechalyzer Toolkit[7]. For playback of the stimuli in randomized order, AKG K-601 headphones were used. One single session took about 40 minutes.

A validation of the full emotions (256 ratings per category) confirmed the synthesis quality for basic emotions, as all five synthesized categories are labeled on average with 52,4% as intended (see table 1).

The intended complex emotions were categorized with a primary label 3072 times. Excluding all full single emotions, and thus also all primary ratings for "neutral", resulted in 2244 answers. The complex emotions as intended with set 1 (prosody) are recognized most frequently. However, anger is equally often confused with fear (table 2).

A similar confusion matrix for the second intended emotion (voice quality, articulation) however, shows no identification by the listeners except for anger (table 3).

The alternative ratings are dominantly "neutral", indicating difficulties to assign two separate emotions to the stimuli (tables 4 and 5). The remaining data without any "neutral" responses, i.e. actually assigned to the four emotions in question, account only for 38% of the

Table 3 – Confusion matrix for the emotions synthesized with voice quality and articulation. Primary rating in %. Highest values bold.

Prim. Rating / Emotion Set 2	Anger	Fear	Joy	Sadness	F1
Anger	**.343**	.222	.215	.220	.346
Fear	**.336**	.176	.238	.250	.163
Joy	.168	**.325**	.199	.308	.214
Sadness	.130	**.483**	.247	.140	.141

Table 4 – Confusion matrix for the emotions synthesized with prosody. Secondary rating in %. Highest values bold.

Sec. Rating Emotion Set 1	Anger	Fear	Joy	Neutral	Sadness	F1
Anger	.123	.196	.066	**.511**	.104	.176
Fear	.136	.202	.097	**.392**	.173	.277
Joy	.090	.194	.100	**.498**	.117	.177
Sadness	.116	.211	.035	**.525**	.112	.115

Table 5 – Confusion matrix for the emotions synthesized with voice quality and articulation. Secondary rating in %. Highest values bold.

Sec. Rating Emotion Set 2	Anger	Fear	Joy	Neutral	Sadness	F1
Anger	.151	.201	.082	**.435**	.131	.206
Fear	.104	.234	.067	**.475**	.120	.283
Joy	.108	.189	.072	**.521**	.110	.136
Sadness	.106	.178	.081	**.481**	.153	.172

3072 responses. Still, there are systematic results visible (table 6): Within the limits of those actually rating a secondary emotion, combinations of anger and fear as well as fear and sadness are dominantly classified irregardless of the assignment of emotions to the features. Joy combined with fear is most often correctly rated for joy synthesized with prosodic information. In sum, fear was the best performing emotion to be combined with others. Interestingly, all confusions had one emotion in common, whereas another was dominantly replaced with fear.

5 Discussion

The pure emotions were all recognized above chance. Results for the complex emotions indicate that the prosodic parameters significantly elicit the intended emotion, whereas the second bundle (voice-quality and articulation precision) reveals mixed results, even for the primary rating. In particular, the secondary rating was dominantly "neutral". Nevertheless, when analyzing the non-neutral ratings, the intended complex emotions work especially well for combinations with fear. Even the confusion pattern for the other targets show systematic effects in favor of fear, always retaining one of the intended emotions that is not dependent on the features bundle. Therefore, these results are most likely originated in the quality of the material and evaluation method at the current state of synthesizing complex emotions, and can not be taken to indicate invalidity of the concept of complex emotions.

Whereas the results are promising, the ultimate aim to validly synthesize two emotions simultaneously was not fully reached. Apparently, some emotions dominate the perception (fear), and the salience or quality of synthesis does not seem to be equally distributed over the two feature bundles.

From a methodological point of view, hiding the true aim of assessing two emotions per stimulus seems to be difficult. Just asking for one emotion expecting both emotions involved to surface, however, requires comparable perceptual salience of each emotion involved. This was found not to be the case judging from the identification rates. As alternative, openly asking for the mixture of emotions risks to induce effects of social desirability, which might still allow for testing the quality of synthesizing stereotypical emotion combinations, but not for testing validity of the complex emotions. Therefore, a more sophisticated evaluation paradigm applying social situations, in which complex emotions do occur, might be more meaningful.

Table 6 – Confusion matrix for the complex emotions separated for prosodic and non-prosodic feature order. Primary and Secondary ratings pooled (in %). Highest values bold, intended categories in italics.

Dual Ratings Complex Emotions	Anger:Fear	Anger:Joy	Anger:Sadness	Fear:Joy	Fear:Sadness	Joy:Sadness
Anger-Fear	***.461***	.113	.174	.148	.087	.017
Fear-Anger	***.424***	.094	.079	.180	.180	.043
Anger-Joy	**.418**	*.154*	.088	.143	.164	.033
Joy-Anger	**.308**	*.288*	.144	.115	.077	.067
Anger-Sadness	**.420**	.037	*.074*	.247	.198	.025
Sadness-Anger	.067	.053	*.400*	.000	**.413**	.067
Fear-Joy	.195	.076	.042	*.288*	**.373**	.025
Joy-Fear	.181	.108	.072	***.349***	.205	.084
Fear-Sadness	.227	.034	.034	.227	***.445***	.034
Sadness-Fear	.070	.020	.320	.020	***.480***	.090
Joy-Sadness	.108	.054	.068	.243	**.324**	*.203*
Sadness-Joy	.057	.014	.200	.000	**.629**	*.100*

6 Conclusions and Outlook

We described an approach to simulate first and secondary emotional expression in synthesized speech simultaneously. The approach is based on the combination of different parameter sets with the open-source system "Emofilt" which utilizes the diphone-synthesizer "Mbrola". An evaluation of the technique was done in a perception experiment which showed only partial results.

The ultimate aim to validly synthesize two emotions simultaneously was not fully reached, but, as the results are promising, the synthesis quality, especially for voice quality and articulation, needs to be optimized in order to establish comparable strength and naturalness of the emotions over both feature bundles. Especially the simulation of articulation precision, which is done by replacing centralized phonemes with decentralized ones and vice versa [8], could be enhanced when using a different synthesis technique.

As unrestricted text-to-speech synthesis is not of importance while this is still predominantly a research topic, one possibility would be to use articulatory synthesis where the parameter sets can be modeled more elaborately by rules than with data-based diphone synthesis.

After quality testing such optimizations, an improved evaluation methodology should be applied to study validity of complex emotions synthesized with "Emofilt".

References

[1] WILLIAMS, P. and J. AAKER: *Can mixed emotions peacefully coexist? Journal of Consumer Research*, 28(4), pp. 636–649, 2002.

[2] MURRAY, I. R. and J. L. ARNOTT: *Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. JASA*, 93(2), pp. 1097–1107, 1993.

[3] BURKHARDT, F.: *Simulation emotionaler Sprechweise mit Sprachsynthesesystemen*. Shaker, 2000.

[4] SCHRÖDER, M.: *Emotional speech synthesis - a review*. In *Proc. Eurospeech 2001, Aalborg*, pp. 561–564. 2001.

[5] DUTOIT, T., 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*, 3, pp. 1393–1396, 1996.

[6] SCHRÖDER, M. and J. TROUVAIN: *The German text-to-speech synthesis system MARY: A tool for research, development and teaching. International Journal of Speech Technology*, 6, pp. 365–377, 2003.

[7] BURKHARDT, F.: *Fast labeling and transcription with the speechalyzer toolkit. Proc. LREC (Language Resources Evaluation Conference), Istanbul*, 2012.

[8] BURKHARDT, F.: *Emofilt: The simulation of emotional speech by prosody transformation*. In *Proc. Interspeech 2005, Lisbon*. 2005.

[9] BURKHARDT, F.: *An affective spoken story teller*. In *Proceedings of Interspeech*. 2011.

[10] BURKHARDT, F., A. PAESCHKE, M. ROLFES, W. F. SENDLMEIER, and B. WEISS: *A database of German emotional speech*. In *Proceedings of Interspeech*. Lisbon, Portugal, 2005.