Speech Synthesizing Simultaneous Emotion-Related States

Felix Burkhardt[1(✉)] and Benjamin Weiss[2]

[1] audEERING GmbH, Friedrichstraße 68, 10117 Berlin, Germany
fxburk@posteo.de
[2] Technische Universität Berlin, Ernst-Reuter-Platz 7, 10587 Berlin, Germany
benjamin.weiss@tu-berlin.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.

Keywords: Speech synthesis · Emotion simulation · Mixed emotions

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. Since decades 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, see [1] or [12] for some more recent examples. 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 or blended 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

A. Karpov et al. (Eds.): SPECOM 2018, LNAI 11096, pp. 76–85, 2018.
https://doi.org/10.1007/978-3-319-99579-3_9

[2,18]. For the case of facial expressions, mixed emotions have been successfully acted by providing situational descriptions and prototypical pictures [8], and even models to blend basic emotions exist [13].

The research on the simulation of affective speaking styles with speech synthesis has a long history [3,14,15] and started with the simulation of one single speaking style or emotional expression. Mixing two speaking styles has later also been studied, for example [17] interpolated the HMM models of two different emotional speaking styles to generate a mixed expression. They did not report on the success of the method with respect to an expression that is perceived by listeners as a mixture between two emotions.

In a similar fashion, [11] learned parameter clusters for HMM speech synthesis to model speaker identity and emotional expression. This method was used to model expression even for speakers whose model was not trained on emotional data by using prosodic models trained on speakers that included expressive samples, while the spectral features are meant to encode the speaker identity. So the foremost aim of this research was to transplant expressive speaking styles from one source speaker to another.

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 not using interpolation between speaker expression models.

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" [9] as well as a text-to-phoneme converter, for example the text-to-speech framework "Mary" [16]. The approach is based on the idea of mixing 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 Darwins 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).

This article is structured as follows. Firstly we describe the speech synthesizer in Sect. 2. We then report on the way we approached the simultaneous simulation of two affective states in Sect. 3. Section 4 describes the perception experiment that was used to verify our approach. Lastly, Sect. 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 Sect. 6.

2 Emofilt

Emofilt [4] is a software program intended to simulate emotional arousal with speech synthesis based on the free-for-non-commercial-use MBROLA synthesis engine [9]. It acts as a transformer between the phonetisation and the speech-generation component. Originally developed at the Technical University of Berlin

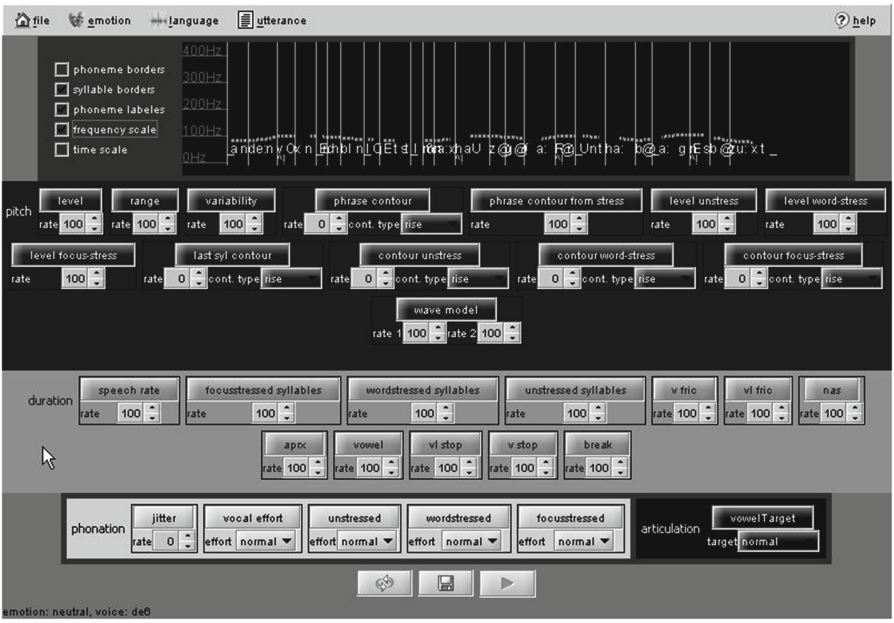

Fig. 1. Emofilt Developer Graphical User Interface.

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 milliseconds). 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 Fig. 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 [6].

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.

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

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 [4] 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 Darwins 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 [5]. We used two short and two longer ones.

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 (16 males, 16 females, 20–39 years old, mean = 27.26, standard deviation = 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 min.

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).

Table 1. Confusion matrix for the single basic emotions only. Primary rating in % divided by 100. Highest values bold.

Prim. Rat. / Emotion	Anger	Fear	Joy	Neutr.	Sadn.	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

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 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 pairs of non-neutral ratings, the intended complex emotions including fear work especially

well. 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.

Table 2. Confusion matrix for the emotions synthesized with prosody. Primary rating in % divided by 100. 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

Table 3. Confusion matrix for the emotions synthesized with voice quality and articulation. Primary rating in % divided by 100. 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

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.

Table 4. Confusion matrix for the emotions synthesized with prosody. Secondary rating in % divided by 100. Highest values bold.

Sec. Rat. Emo. Set 1	Anger	Fear	Joy	Neutr.	Sadn.	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 % divided by 100. Highest values bold.

Sec. Rat. / Emo. Set 2	Anger	Fear	Joy	Neutr.	Sadn.	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

Table 6. Confusion matrix for the complex emotions separated for prosodic and non-prosodic feature order. Primary and Secondary ratings pooled (in % divided by 100). 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*

From a methodological point of view, hiding the true aim while assessing two emotions per stimulus seemed to be difficult. However, asking for only one emotion and analyzing the frequencies of replies would require comparable perceptual salience of each emotion involved. Fortunately, judging from conversations with the participants and the high amount of neutral second ratings, the cover story of asking for a first and an alternative impression worked.

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.

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 [4], could be enhanced when using a different synthesis technique. Data-based synthesis (like diphone synthesis or non-uniform unit-selection synthesis) is not well suited for manipulations of the articulation precision or voice quality. In this respect the simulation rules that were based on prosodic manipulation (set 1) were of course more effective.

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.

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

The approach did result in success with emotions that are neighbors with respect to the emotional dimensional space that's spanned by the PAD dimensions pleasure, arousal and dominance. For example the combination of sadness and anger as well as fear and sadness share two of the three dimensions and were recognized by the majority of the judges.

For future work it would be a possibility to try combinations of emotions that can be envisaged by the listeners more easy than systematic variation, for example by embedding the test sentences into situations that are appropriate for the targeted emotion mix.

It would also be an interesting research to investigate the acoustic manifestation of mixed emotions by analysis of natural data, for example the Vera am Mittag corpus [10]. As this corpus consists of real-life emotional expression happening in a TV-show, mixed emotions are very likely to occur. A set of clear representations would have to be identified by a new label process and then analysed for acoustic properties. The outcomes could then be synthesized to validate the findings in a more controlled environment.

References

1. Barra-Chicote, R., Yamagishi, J., King, S., Monero, J.M., Macias-Guarasa, J.: Analysis of statistical parametric and unit-selection speech synthesis systems applied to emotional speech. Speech Commun. **52**(5), 394–404 (2010)
2. Berrios, R., Totterdell, P., Kellett, S.: Eliciting mixed emotions: a meta-analysis comparing models, types, and measures. Front. Psychol. **6**, 428 (2015)
3. Burkhardt, F.: Simulation emotionaler Sprechweise mit Sprachsynthesesystemen. Shaker (2000)
4. Burkhardt, F.: Emofilt: the simulation of emotional speech by prosody transformation. In: Proceedings of Interspeech. Lisbon (2005)
5. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of German emotional speech. In: Proceedings of Interspeech. Lisbon (2005)
6. Burkhardt, F.: An affective spoken story teller. In: Proceedings of Interspeech. Florence (2011)
7. Burkhardt, F.: Fast labeling and transcription with the speechalyzer toolkit. In: Proceedings of LREC (Language Resources Evaluation Conference), Istanbul (2012)
8. Du, S., Tao, Y., Martinez, A.: Compound facial expressions of emotion. Proc. Natl. Acad. Sci. **111**(15), E1454–62 (2014)
9. Dutoit, T., Pagel, V., Pierret, N., Bataille, F., Van der Vreken, O.: The MBROLA project: towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In: Proceedings of ICSLP 1996, Philadelphia, vol. 3, pp. 1393–1396 (1996)
10. Grimm, M., Kroschel, K., Narayanan, S.: The Vera am Mittag German audio-visual emotional speech database. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Hannover (2008)
11. Latorre, J., et al.: Speech factorization for HMM-TTS based on cluster adaptive training. In: Proceedings of Interspeech. Portland (2012)
12. Lee, Y., Rabiee, A., Lee, S.: Emotional end-to-end neural speech synthesizer. CoRR (2017)
13. Martin, J.C., Niewiadomski, R., Devillers, L., Buisine, S., Pelachaud, C.: Multimodal complex emotions: gesture expressivity and blended facial expressions. Int. J. Humanoid Rob. **3**, 269–292 (2006)
14. Murray, I.R., Arnott, J.L.: Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. JASA **93**(2), 1097–1107 (1993)
15. Schröder, M.: Emotional speech synthesis - a review. In: Proceedings of Eurospeech 2001, Aalborg, pp. 561–564 (2001)
16. Schröder, M., Trouvain, J.: The German text-to-speech synthesis system mary: a tool for research, development and teaching. Int. J. Speech Technol. **6**, 365–377 (2003)
17. Tachibana, M., Yamagishi, J., Masuko, T., Kobayashi, T.: Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. IEICE Trans. Inf. Syst. **88**(11), 2484–2491 (2005)
18. Williams, P., Aaker, J.: Can mixed emotions peacefully coexist? J. Consum. Res. **28**(4), 636–649 (2002)