Going Retro: Astonishingly Simple Yet Effective Rule-based Prosody Modelling for Speech Synthesis Simulating Emotion Dimensions

Felix Burkhardt[1], *Uwe Reichel*[1], *Florian Eyben*[1], *Björn Schuller*[1,2,3]

[1]*audEERING GmbH, Germany,*
[2]*Chair EIHW, University of Augsburg, Germany,*
[3]*GLAM, Imperial College London, UK*
fburkhardt@audeering.com

Abstract: We introduce two rule-based models to modify the prosody of speech synthesis in order to modulate the emotion to be expressed. The prosody modulation is based on speech synthesis markup language (SSML) and can be used with any commercial speech synthesizer. The models as well as the optimization result are evaluated against human emotion annotations. Results indicate that with a very simple method both dimensions arousal (.76 UAR) and valence (.43 UAR) can be simulated.

1 Introduction

Affect-modulated speech synthesis of a text can be achieved amongst others by modifying the prosody of the utterance accordingly [1]. In this work, emotions will be represented in terms of the dimensional approach of Schlosberg [2], who identified the three emotion dimensions valence, arousal, and dominance. *Valence* is referred to as *pleasure* in the following. For this paper, we neglect the dominance dimension for the benefit to focus on the main topic: the control of emotional expression in speech synthesis with a very limited set of prosodic rules.

1.1 Acoustic correlates of emotions

For each of these dimensions, several acoustic correlates have been found. These findings are summarized in [3], [4], and [5] (for further details please see the references therein).

High as opposed to low arousal is characterized by higher speech rate, higher intensity mean and variability, higher fundamental frequency (F_0) mean and variability, higher spectral balance indicating increased vocal effort, and a higher first formant due to an increased mouth opening.

Positive as opposed to negative pleasure is amongst others characterized by higher speech rate and by lower intensity mean and variability. In addition, pleasure is positively correlated with the second formant due to more lip spreading caused by smiling [6]. The relation between pleasure and pitch is more complicated as found by [7, 1, 8]: Higher F_0 characterizes both elation joy (positive) and fear (negative), while comfort (positive) and boredom (negative) are both reflected by lower F_0 [1].

In general, many results from the literature, for example [9, 10, 11], indicate that it is difficult to predict and simulate the valence dimension by acoustic cues alone, as opposed to linguistic ones, an assumption that is confirmed also in this investigation.

1.2 Emotions in speech synthesis

There are many articles that deal with the simulation of emotional speech and even many that review them, for example [12, 13, 14]; we refer to these for a deeper discussion. Historically, first algorithms to simulate emotional expression were based on prosody rules and categorical emotions. Later, and in line with the new statistical techniques to synthesize speech, data based approaches were used and emotional dimensions as well as speaking styles targeted. Triantafyllopoulos et al. review deep learning based approaches in [15]. Marc Schröder was the first on to target emotional dimensions with prosody rules in his dissertation [11] and his work is one of the foundations of this paper. An approach to simulate emotion dimensions with learned features was presented by Hamada [10] by mapping acoustic features to the valence-arousal space. Later, Stanton et al. [16] showed how to target the latent space within a Tacotron architecture to generate expressive speaking styles.

1.3 Emotional simulation with SSML

Within the scope of the European H2020 EASIER project [17], we faced the problem to enable a, not-yet emotional but available in many languages, commercial speech synthesizer to simulate emotional expression. The obvious way to do this, in a way that is agnostic to a specific synthesizer, is to utilize the W3C's Speech Synthesis Markup Language (SSML) [18] which is, at least in parts, interpreted by almost all speech synthesis engines that are available. This has also been done for categorical emotions by Shaikh et al. [19].

SSML amongst others allows for specifying prosodic modifications of an utterance along the prosodic dimensions pitch, energy, and duration. Thus, speech synthesis can be affect-modulated by mapping emotion dimensions to prosodic parameters based on the findings above, and by passing on these parameters to the Text to speech (TTS) engine via SSML. We tested this for the commercial Google Speech API[1] and the open source MARY TTS engine [20], where the support is only partial.

This paper is structured as follows: In section 2, we introduce two rule-based model variants in order to adapt the prosody of an utterance accordingly. In sections 3 and 4, we describe the perceptual evaluation procedure and their results, respectively. Section 5 concludes the paper with an outlook.

Contributions of this paper are as follows:

- We present two approaches to simulate emotional dimensions with SSML, which has to our knowledge not been done before.
- We simulate the valence dimension by a very simple pitch manipulation approach.

2 Rule-based affect modulation

In our study, emotion scores are mapped to speech prosody parameters in two rule-based algorithms based on the findings introduced in section 1.

2.1 Method syntact

As a naive baseline we simply implemented the prosody rules as being positively correlated with pitch and speech rate. To distinguish between arousal and valence, we simply assigned speech rate to arousal and pitch to valence, an approach that worked surprisingly well.

[1] https://cloud.google.com/text-to-speech

Of course, we can not be sure if the outcomes are specific to the Google synthesizer that was used to generate the samples.

2.2 Method Schroeder

To try out a more complex approach than Syntact, we implemented a very reduced version of the approach that Marc Schröder described in his dissertation [21]. To this end, we analyzed Schroeder's MARY TTS [20] sources [2]. According to the sources, we extracted the rules displayed in Listing 2.2, originally in Java, and implemented them in the Python language.

Listing 1 – Prosody rules according to the MARY emotion module

```
pitch => 0.3 * arousal + 0.1 *  valence - 0.1 * power
pitch-dynamics => -15 + 0.3 * arousal - 0.3 * power
range (in Semitones) => 4 + 0.04 * arousal
range-dynamics (min 100) => -40 + 1.2 * arousal + 0.4 * power
accent-prominence => 0.5 * arousal - 0.5 * valence
preferred-accent-shape => when valence < -20: falling ,
    when valence > 40: alternating , else rising
accent-slope => 1 * arousal - 0.5  * valence
rate => 0.5 * arousal + 0.2  * valence
number-of-pauses => 0.7 * $arousal pause
duration => -0.2 * arousal
vowel-duration => 0.3 * valence + 0.3 * power
nasal-duration => 0.3 * valence + 0.3 * power
liquid-duration => 0.3 * valence + 0.3 * power
plosive-duration => 0.5 * arousal - 0.3 * valence
fricative-duration => 0.5 * arousal - 0.3 * valence
volume => 50 + 0.33 * arousal
```

To implement this in SSML, we filtered the list for pitch and speech rate global values resulting in the two rules shown in Listing 2.2. These rules had already been tested in the scope of a project to generate an appropriate robot voice for children with the autistic spectrum [22].

The resulting values were then scaled as described in the next Section. Of course, this is only a very small subset of the rules defined by Marc Schröder and this might well be the main reason that this approach did not show to be very successful.

Listing 2 – Reduced prosody rules according to the MARY emotion module

```
pitch => 0.3 * arousal + 0.1 * valence
    - 0.1 * power
rate => 0.5 * arousal + 0.2 * valence
```

2.3 Mapping from dimensions to rules

We adapted the emotion to prosody mapping approach of [23]: Values for the emotion dimensions arousal, and pleasure are mapped to the *pitch*, *rate*, and *volume* attributes of the SSML element <prosody>. This mapping is carried out in the following way:

1. rescale the scores of emotion dimension $e \in \{\text{pleasure, arousal}\}$ to the range $[-1, 1]$

[2] https://github.com/marytts/marytts/blob/79e4edef3f478dcef0aad3609ba77090e91f0b6d/marytts-client/src/main/resources/marytts/tools/emospeak/emotion-to-mary.xsl

2. calculate each of the prosody parameters $y \in \{\text{pitch, rate, volume}\}$ by the following linear combination

$$y = \sum_{e \in \{\text{pleasure, arousal}\}} w_{e,y} \cdot e \tag{1}$$

3. rescale y to a range defined by still natural sounding minimum and maximum values of the respective prosody dimension.

For both variants, Schroeder, and Syntact, all weights $w_{e,y}$ of emotion dimension e for the calculation of the prosodic dimension y were set manually based on perceptual expert judgments how well the synthesized speech prosodically matches the intended emotions. [23] showed that emotion recognition performance can be increased by adding emotional speech samples synthesized this way to the training data.

For reproducibility, all code is open sourced in the Syntact gitlab repository[3].

3 Perceptual Evaluation

We conducted a perception experiment to validate the effectiveness of the approaches. We used the Google speech API as a speech synthesizer, with the standard male and female voices. As text material, we used two short sentences of the Berlin Emotional Database [24], which are meant to be emotionally undecided:

- ”*In sieben Stunden wird es soweit sein.*“ (*it will happen in seven hours.*)
- ”*Heute Abend könnte ich es ihm sagen.*“ (*i could tell him tonight.*)

The idea is that these sentences are neither too mundane nor already have a linguistic emotional connotation.

Fleiss' κ	**Arousal**	**Valence**
Schroeder	.467	.067
Syntact	.445	.121
All	.461	.112

Table 1 – Fleiss' kappa values for inter rater agreement for arousal and valence levels for methods Schroeder, Syntact, and all values.

These four combinations (two sexes times two sentences) were synthesised with both methods and with all combinations (9) of three valence and arousal levels: $.1, .5, .9$, with .5 being the neutral level. resulting in 72 samples $(2 \cdot 2 \cdot 2 \cdot 9)$.

The samples were annotated by 10 subjects employed by audEERING GmbH using the I-hear-U-play platform [25]. The labelers were 6 women and 4 men of mean age 34.87 years with 13.79 years standard deviation. After judging 10 test samples to get acquainted with the task, they answered for each sample the following two questions:

- ”*Please rate the arousal level on a scale of low, mid, and high.*“
- ”*Please rate the valence level on a scale of negative, neutral, and positive.*“

To measure the inter-rater agreement we used Fleiss' kappa, the results are depicted in Table 1. While the raters could agree on the arousal annotations, the values are only slightly agreed on for valence when simulated by the Syntact method but not for the Schroeder method.

[3] `https://github.com/felixbur/syntAct`

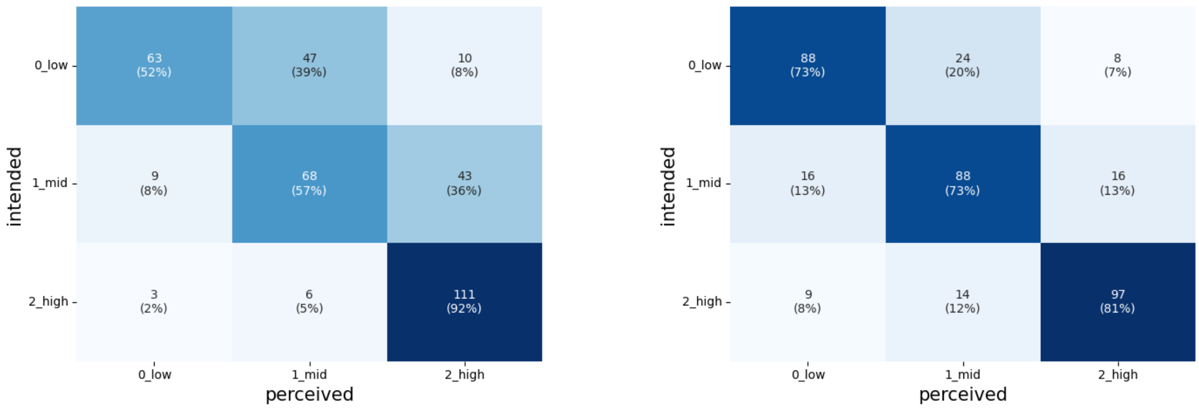

Figure 1 – Confusion matrices between intended and perceived arousal levels. Left: model Schroeder, Right: model Syntact

4 Results

The results of the perception experiment are shown in Table 2. The confusion matrix for the arousal dimension levels is shown in Figure 1 and for the valence dimension in Figure 2. The results were computed based on all listeners ratings, without a unified label.

As can be seen, the simulation of arousal was successful with both approaches but to a clearly higher degree with the simpler Syntact method. For the Schroeder method, low arousal is often confused with the neutral versions and the neutrally meant samples with high arousal.

With respect to valence, we must admit that we were only partly successful with the Syntact method. As outlined above, this may largely be also due to the fact that it is mostly convey in linguistic information. This has recently again been shown also for deep representations of acoustics that succeed in better automatic valence recognition from speech due to their inherent encoding of linguistics [26]. Nonetheless, we think this is a valuable finding because this method basically hypothesizes that valence correlates positively with pitch which is an interesting approach based on its simplicity. The samples generated by the Schroeder method were labeled as neutral or high valence by the majority of the listeners, an outcome perhaps that indicates that the "normal" expression of the Google voices is rather friendly.

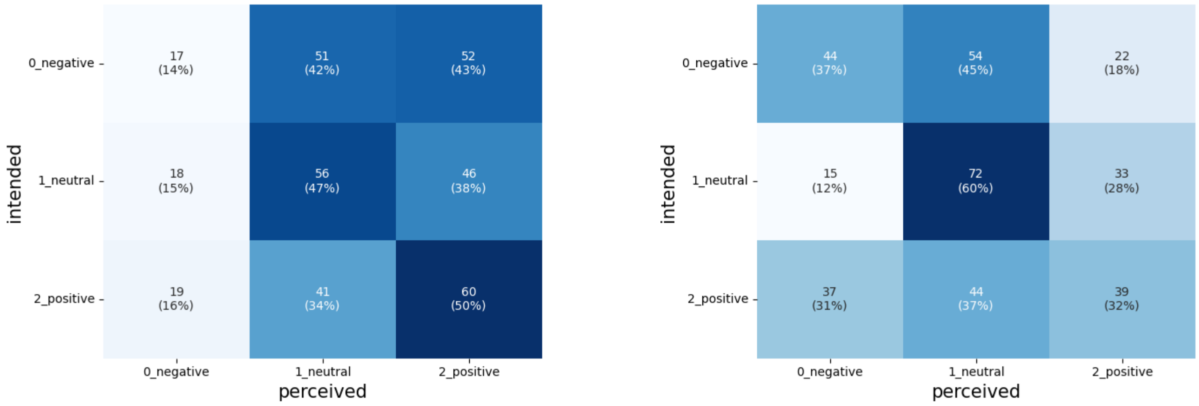

Figure 2 – Confusion matrices between intended and perceived valence levels. Left: model Schroeder, Right: model Syntact

As discussed in Section 1, it is quite difficult to simulate the valence dimension by acoustic cues alone and accordingly, we are satisfied to have reached even a partial success.

UAR	**Arousal**	**Valence**
Schroeder	.67	.37
Syntact	.76	.43

Table 2 – UAR values between intended and perceived arousal and valence levels for both methods Schroeder and Syntact.

5 Conclusion and Outlook

This paper investigated two methods to simulate emotional expression in speech synthesis by controlling prosody with SSML. The chosen method following a (strongly) reduced version of Marc Schröder's work did not outperform our rather naive baseline.

Of course, we cannot be sure if the outcomes are specific to the Google synthesizer that was used to generate the samples. Hence, a more general investigation that includes several speech engines will remain future work.

Also, it is much more promising to learn emotion-to-expression rules from data than to manually determine them based on isolated trials, amongst others because emotional expression is at least to a degree culture specific and the same rules can not be applied in all cultural and social contexts. This has to our knowledge not yet been done for SSML-based approaches and also remains future work.

Thirdly, we restricted the investigation on two dimensions; valence and arousal. Future studies will take at least the dominance dimension into account, which is important to distinguish for example *anger* from *fear*.

It further appears interesting to measure how automatic speech emotion recogniser would recognise such rule- and SSML-based samples. In addition, one could evaluate if they could be used for model augmentation as was first suggested in [27].

On the opposing end – and likewise closing the circle between analysis and synthesis, one could implement a related rule- and SSML-based recognition of emotion from speech. Presumably, however, this would require some form of speaker normalisation grounded in neutral speech, hence, requiring an enrolment procedure.

6 Acknowledgements

This research has been partly funded by the European SHIFT (MetamorphoSis of cultural Heritage Into augmented hypermedia assets For enhanced accessibiliTy and inclusion) project (Grant Agreement number: 101060660).

References

[1] Scherer, K. R., R. Banse, H. G. Wallbott, and T. Goldbeck: *Vocal cues in emotion encoding and decoding. Motivation and emotion*, 15(2), pp. 123–148, 1991. doi:https://doi.org/10.1007/BF00995674.

[2] Schlosberg, H.: *Three dimensions of emotions. Psychological Review*, 61, pp. 81–88, 1954.

[3] Laukka, P., P. Juslin, and R. Bresin: *A dimensional approach to vocal expression of emotion. Cognition & Emotion*, 19(5), pp. 633–653, 2005. doi:https://dl.acm.org/doi/abs/10.1109/ICASSP.2017.7953135.

[4] SZAMEITAT, D. P., C. J. DARWIN, D. WILDGRUBER, K. ALTER, and A. J. SZAMEITAT: *Acoustic correlates of emotional dimensions in laughter: arousal, dominance, and valence.* *Cognition and emotion*, 25(4), pp. 599–611, 2011. doi:https://doi.org/10.1080/02699931.2010.508624.

[5] MÁDY, K., U. D. REICHEL, A. KOHÁRI, and A. SZALONTAI: *The role of accommodation in expressing emotions to newborn babies.* In *Proc. 1. International Conference on Tone and Intonation*. 2021.

[6] TARTTER, V. C.: *Happy talk: Perceptual and acoustic effects of smiling on speech.* *Perception & Psychophysics*, 27(1), pp. 24–27, 1980. doi:https://doi.org/10.3758/BF03199901.

[7] BANSE, R. and K. R. SCHERER: *Acoustic profiles in vocal emotion expression.* *Journal of personality and social psychology*, 70(3), p. 614, 1996. doi:https://psycnet.apa.org/doi/10.1037/0022-3514.70.3.614.

[8] JOHNSTONE, T. and K. R. SCHERER: *Vocal communication of emotion.* *Handbook of emotion*, 2, pp. 220–235, 2000.

[9] LOTFIAN, R. and C. BUSSO: *Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings.* *IEEE Transactions on Affective Computing*, PP, pp. 1–1, 2017. doi:10.1109/TAFFC.2017.2736999.

[10] HAMADA, Y., R. ELBAROUGY, and M. AKAGI: *A method for emotional speech synthesis based on the position of emotional state in valence-activation space.* In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pp. 1–7. 2014. doi:10.1109/APSIPA.2014.7041729.

[11] SCHRÖDER, M.: *Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions.* In E. ANDRÉ, L. DYBKJÆR, W. MINKER, and P. HEISTERKAMP (eds.), *Affective Dialogue Systems*, pp. 209–220. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[12] TITS, N., K. EL HADDAD, and T. DUTOIT: *Emotional speech datasets for english speech synthesis purpose: A review.* In *Proceedings of SAI Intelligent Systems Conference*, pp. 61–66. Springer, 2019.

[13] BURKHARDT, F. and N. CAMPBELL: *Emotional speech synthesis 20.* In *The Oxford Handbook of Affective Computing*, pp. 286–290. Oxford University Press, 2014.

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

[15] TRIANTAFYLLOPOULOS, A., B. W. SCHULLER, G. İYMEN, M. SEZGIN, X. HE, Z. YANG, P. TZIRAKIS, S. LIU, S. MERTES, E. ANDRÉ, R. FU, and J. TAO: *An overview of affective speech synthesis and conversion in the deep learning era.* 2022. doi:10.48550/ARXIV.2210.03538. URL `https://arxiv.org/abs/2210.03538`.

[16] STANTON, D., Y. WANG, and R. SKERRY-RYAN: *Predicting expressive speaking style from text in end-to-end speech synthesis.* In *arxiv*, pp. 595–602. 2018. doi:10.1109/SLT.2018.8639682.

[17] MORGAN, H. E., O. CRASBORN, M. KOPF, M. SCHULDER, and T. HANKE: *Facilitating the spread of new sign language technologies across Europe*. In E. EFTHIMIOU, S.-E. FOTINEA, T. HANKE, J. A. HOCHGESANG, J. KRISTOFFERSEN, J. MESCH, and M. SCHULDER (eds.), *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pp. 144–147. European Language Resources Association (ELRA), Marseille, France, 2022. URL `https://www.sign-lang.uni-hamburg.de/lrec/pub/22026.pdf`.

[18] W3C: *Speech Synthesis Markup Language (SSML) Version 1.1*. Web page, 2010. `https://www.w3.org/TR/speech-synthesis11/`.

[19] SHAIKH, M. A. M., A. R. F. REBORDAO, K. HIROSE, and M. ISHIZUKA: *Emotional speech synthesis by sensing affective information from text*. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–6. 2009. doi:10.1109/ACII.2009.5349515.

[20] 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.

[21] SCHRÖDER, M.: *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis*. PhD thesis, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University, 2004.

[22] BURKHARDT, F., M. SAPONJA, J. SESSNER, and B. WEISS: *How should pepper sound - preliminary investigations on robot vocalizations*. In *Proc. of the Elektronische Sprachsignalverarbeitung (ESSV)*. Peter Birkholz, Simon Stone, 2019.

[23] BURKHARDT, F., F. EYBEN, and B. W. SCHULLER: *SyntAct: A synthesized database of basic emotions*. In *Proceedings of the 1st Workshop on Dataset Creation for Lower-Resourced Languages (DCLRL) at LREC2022*, pp. 1–9. Marseille, 2022.

[24] WAGNER, J., A. TRIANTAFYLLOPOULOS, H. WIERSTORF, M. SCHMITT, F. BURKHARDT, F. EYBEN, and B. W. SCHULLER: *Model for dimensional speech emotion recognition based on Wav2vec 2.0*. web downloaded, 2022. `https://zenodo.org/record/6221127`.

[25] HANTKE, S., T. APPEL, F. EYBEN, and B. SCHULLER: *ihearu-play: Introducing a game for crowdsourced data collection for affective computing*. In *Proc. 1st International Workshop on Automatic Sentiment Analysis in the Wild (WASA 2015)*, pp. 891–897. IEEE, 2015.

[26] TRIANTAFYLLOPOULOS, A., J. WAGNER, H. WIERSTORF, M. SCHMITT, U. REICHEL, F. EYBEN, F. BURKHARDT, and B. W. SCHULLER: *Probing speech emotion recognition transformers for linguistic knowledge*. In *Proceedings INTERSPEECH 2022, 23rd Annual Conference of the International Speech Communication Association*, pp. 1–5. ISCA, Incheon, South Korea, 2022.

[27] SCHULLER, B., Z. ZHANG, F. WENINGER, and F. BURKHARDT: *Synthesized Speech for Model Training in Cross-Corpus Recognition of Human Emotion*. *International Journal of Speech Technology*, 15(3), pp. 313–323, 2012.