SyntAct: A Synthesized Database of Basic Emotions

Felix Burkhardt[1], **Florian Eyben**[1], **Björn W. Schuller** [1,2,3]
[1]audEERING GmbH, [2]University of Augsburg, [3]Imperial College London
Germany, Germany, UK
{fburkhardt, fe, bs}@audeering.com

Abstract

Speech emotion recognition is in the focus of research since several decades and has many applications. One problem is sparse data for supervised learning. One way to tackle this problem is the synthesis of data with emotion-simulating speech synthesis approaches. We present a synthesized database of five basic emotions and neutral expression based on rule-based manipulation for a diphone synthesizer which we release to the public. The database has been validated in several machine learning experiments as a training set to detect emotional expression from natural speech data. The scripts to generate such a database have been made open source and could be used to aid speech emotion recognition for a low resourced language, as MBROLA supports 35 languages.

Keywords: emotional, database, synthetic, speech synthesis, simulation

1. Introduction

The recognition of affect in speech is quite an old research field that gains momentum with the spread of vocal interfaces as numerous applications appear. Examples are natural human machine interaction, gaming and the security domain. The overwhelming majority of approaches to machine learning is still based on supervised learning, which is even stronger for emotional arousal as there is no clear definition, compared to other speech features like speaker age or textual contents. Obviously, labeling emotional data manually is costly and methods to multiply existent data are needed.

With the dawn of modern deep learning based speech synthesizers, emotional expression is usually learned with the data, see Section 2 for some references.

About a decade ago, we have chosen a different approach by simulating emotional arousal with manually adapted prosody rules. In (Schuller and Burkhardt, 2010; Schuller et al., 2012), we already tried successfully to add the synthesized samples to the training of emotional recognition models.

We now re-synthesized five emotional categories plus neutral versions and published them to the research community. We call the database “SyntAct” because it displays basic emotions with always the same prosodic expression, like a bad actor would do.

This paper describes the process of generation of the samples in Section 3, the format of the database in Section 4, and an evaluation experiment in Section 5.

Contributions of this paper are:

- We introduce a new dataset of simulated emotional expression that is available to the public.
- The simulation is based on rules that can target specific emotions.
- We release the scripts that generate the database as well so researchers can extend the data as needed.
- We evaluate the database with respect to its usefulness to train a machine learning model to detect prototypical emotional categories in natural data.

2. Related Work

The idea to synthesize training data is not a new one. Based on an existing training database, with deep learning techniques like variational autoencoders (VAEs) (Baird et al., 2019; Deng et al., 2014; Deng et al., 2013) or generative adversarial networks (GANs) (Latif et al., 2020; Eskimez et al., 2020), the generation of new training data has been realised. In some of these approaches only non-audible acoustic parameters have been generated, in other ones, actual speech samples (though not necessarily with semantic content).

An alternative approach is to generate new training data from scratch with traditional speech synthesis approaches. In (Baird et al., 2018), we investigated the likability and human likeness of synthesized speech in general and found that many systems are acceptable.

The approaches to synthesize speech can be categorized like this:

- articulatory synthesis
- formant synthesis
- diphone synthesis
- non-uniform unit selection synthesis
- HMM based synthesis
- deep learning based synthesis

in the order of historic importance. Basically, there has been a trade-off between flexibility and naturalness for the algorithms. While formant synthesis for example is quite flexible with respect to signal manipulation, for the non-uniform unit-selection approach, all envisaged emotional expression must already be contained in the training database. The highest quality of speech synthesizes approaches (with respect to out-of-domain naturalness) are these days based on artificial neural networks (ANN) under the label deep learning (DL) (Zhou et al., 2020). Although these DL based systems deliver very natural speech, it is difficult to determine which emotional expression will be simulated, as they usually are generated by manipulation of the latent space inside the network.

3. Preparing the Speech Samples

The synthesised database consists of samples that were generated with the rule-based emotion simulation software "Emofilt" (Burkhardt, 2005). It utilises the diphone speech synthesiser MBROLA (Dutoit et al., 1996) to generate a wave form from a phonetic description (SAMPA symbols with duration values and fundamental frequency contours) and the MARY text-to-speech (TTS) system (Schröder and Trouvain, 2003) to generate the neutral phonetic description from text.
Emofilt acts as a filter in between to 'emotionalise' the neutral phonetic description; the rules are based on a literature research described in (Burkhardt, 2000). All six available German MBROLA voices – *de2*, *de4*, and *de7* as female, and *de1*, *de3*, and *de6* as male – were used.

3.1. Text Material

With respect to emotional speech, usually three kinds of texts are distinguished:

- Emotional texts, where the emotion is also expressed in the words. This happens often in real world conversations.
- Mundane texts, where strong emotions seem inappropriate.
- Emotional arbitrary texts , that might indicate an emotional event but it's not clear which emotion.

For the experiment at hand, the intention is to soften the problem of limited prosodic variability with a large number of different texts. Therefore we utilized a German news corpus from the University of Leipzig[1] (Goldhahn et al., 2012). No special preprocessing was applied.
The following lists three sentences (in the original language: German).

```
1) Um die in jeder Hinsicht zufrieden
  zustellen , tueftelt er einen Weg
  aus , sinnlose Buerokratie wie Laden−
  schlussgesetz und Nachtbackverbot
  auszutricksen .
2) Um die gesamte Insel zu erkunden ,
  empfiehlt es sich , ein Auto zu mieten .
3) Sowohl die Landesregierung als
   auch die Kreise tragen Schuld .
```

3.2. Rule-Based Emotion Simulation

Emofilt differentiates four different kinds of acoustic features:

- Intonation: Here, we model intonation contours that can be specified for the whole phrase or for specific kinds of stressed syllables, with a special treatment of the last one.
- Duration: General duration as well as duration on syllable (differentiated for stress type) and phoneme level can be specified.
- Voice quality: Although voice quality is inherently fixed for diphone synthesis, some databases have voice quality variants (Schröder and Grice, 2003). In addition, a simulation of jitter is achieved by shifting alternating F0 values.
- Articulation: Because with diphone synthesis a manipulation of format tracks is not directly possible, we achieve a simulation of articulatory effort by substituting tense with lax vowels in stressed syllables and vice versa, following an idea by (Cahn, 2000).

The exact values that we decided upon to generate the samples per emotion are detailed in the following subsections.
The scripts to generate such a database have been made open source and could be used to aid speech emotion recognition for a low resourced language, as MBROLA supports 35 languages[2]. It must be noted though, that many MBROLA languages miss implementations of a natural language processing (NLP) component which means that "emotionally neutral" input samples would have to be specified in the native phonetic MBROLA format. Also, for most languages, only two or even only one voice has been made publicly available.

3.2.1. Simulation of Sadness

We applied the following configuration to simulate sadness:

```
<pitch>
 <variability rate=”80” />
 <f0Range rate=”80” />
 <contourFocusstress rate=”30”
   type=”straight” />
 <lastSylContour rate=”10”
   type=”rise” />
 </pitch>
<phonation>
 <jitter rate=”10” />
 <vocalEffort effort=”soft” />
</phonation>
<duration>
 <speechRate rate=”140” />
</duration>
```

[1] https://wortschatz.uni-leipzig.de; we used the list "deu_news_1995_10K-sentences.txt".

[2] https://github.com/felixbur/syntAct

```
<articulation>
 <vowelTarget
    target=”undershoot” />
</articulation>
```

This means that the variability of F0 in general has been reduced to 80 %, the F0 contour of the stressed syllables is now straight and the last syllable rises by 10 %. The F0 range has also been reduced by 20 %. With respect to phoneme duration, the speech rate (syllable per second) has been made slower by 40 %. The voice quality has been set to soft vocal effort, meaning that the respective samples from voices “de6” and “de7” were used.

3.2.2. Simulation of Happiness

We decided on the following configuration to simulate happiness:

```
<pitch>
 <f0Mean rate=”120” />
 <f0Range rate=”130” />
 <contourFocusstress rate=”40”
    type=”rise” />
 <levelFocusstress rate=”110” />
</pitch>
<duration>
 <durVLFric rate=”150” />
 <speechRate rate=”70” />
 <durVowel rate=”130” />
</duration>
<phonation>
<vocalEffort effort=”loud” />
</phonation>
```

We enlarge the F0 range by 30 % and raise the whole contour by 20%. The stressed syllables are raised by additional 10 %. The stressed syllables gets an upward pitch direction by 10 %. The speech rate gets faster by 30 % in general, but voiceless fricatives and vowels get an extra speed accelerator by 50 and 30 % respectively. The vocal effort gets stronger.

3.2.3. Simulation of Anger

We applied the following configuration to simulate anger:

```
<pitch>
 <f0Range rate=”140” />
 <levelFocusStress rate=”130” />
 <variability rate=”130” />
 <contourFocusstress rate=”10”
     type=”fall” />
 <levelFocusstress rate=”130” />
</pitch>
<duration>
<durVowel rate=”70” />
 <speechRate rate=”70” />
 <durationFocusstressedSyls
    rate=”130” />
 </duration>
<phonation>
 <vocalEffort effort=”loud” />
 <jitter rate=”2” />
</phonation>
<articulation>
 <vowelTarget target=”overshoot” />
</articulation>
```

To simulate anger the F0 range is compressed by 20 %, the contour of the stressed syllables gets a downwards direction and they are raised by 30 %. The speech is made faster by 30 % for all non-stressed syllables whereas the stressed syllables are made longer by 30 %. In addition we apply jitter simulation and the “loud” phonation type.

3.2.4. Simulation of Fear

Additionally, we simulated two emotional states that were discussed in (Burkhardt, 2000), though we did not test them against real databases within the work reported here.

```
<pitch>
 <phraseContour rate=”10”
     type=”rise” />
 <contourFocusstress rate=”10”
   type=”straight” />
 <lastSylContour rate=”10”
     type=”rise” />
 <f0Mean rate=”200” />
</pitch>
<duration>
 <speechRate rate=”70” />
 <durationFocusstressedSyls
     rate=”80”/>
 <durPause rate=”200” />
</duration>
<phonation>
 <jitter rate=”5” />
 <vocalEffort effort=”loud” />
</phonation>
<articulation>
 <vowelTarget target=”undershoot” />
</articulation>
```

Fear is characterized by a rising phrase pitch contour, straight stressed syllables and an additional rise at the end. The speech rate is faster, especially for the stressed syllables and the duration of pauses longer. The articulation vowel target is undershot, meaning that stressed vowels get replaced by unstressed ones.

3.2.5. Simulation of Boredom

The second additional emotion we simulate is boredom.

```
<pitch>
 <f0Mean rate=”120” />
 <phraseContour rate=”40”
     type=”fall” />
 </pitch>
```

```
<duration>
 <speechRate rate=”120” />
 <durationFocusstressedSyls
    rate=”120” />
</duration>
<phonation>
 <vocalEffort effort=”soft” />
</phonation>
```

Boredom has primarily a falling pitch contour, a slower speech rate, especially with the stressed syllables, and a soft vocal effort.

4. Description of the Database

The database is downloadable[3] as a zip file. The format of the data is in the audformat style being described in the next section.

4.1. The audformat Package

audformat[4] defines an open format for storing media data, such as audio or video, together with corresponding annotations. The format was designed to be universal enough to be applicable to as many use cases as possible, yet simple enough to be understood easily by a human and parsed efficiently by a machine.

A database in *audformat* consists of a header, which stores information about the database (source, author, language, etc.), the type of media (sampling rate, bit depth, etc.), the raters (age, mother tongue, etc.), the schemes (numerical, categories, text, etc.), and the splits (train, test, etc.). It also keeps reference of all tables that belong to the database, which hold the actual annotations and are stored in separate files in text and/or binary format.

A corresponding Python implementation[5] provides tools to access the data, create statistics, merge annotations, and search / filter information.

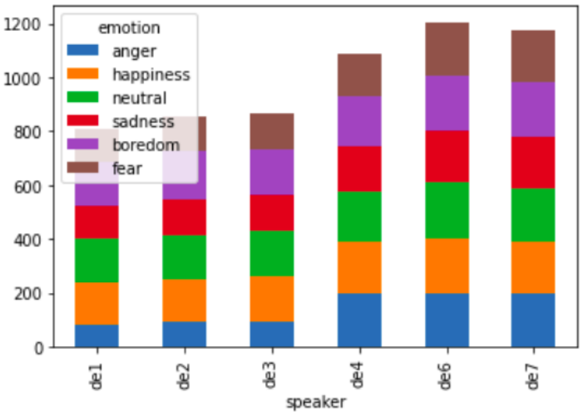
4.2. Specifics of the Database

We generated in total 6000 samples with different textual content, for all six German voices (de1, de2, de3, de4, de6, de7) and each emotion. There are two reasons why not all combinations could be synthesized:

- For some phrases and emotions, the modifications led to the total elimination of phonemes and the result did not adhere to the phonotactics of the voice.

- In some cases, the MARY software, being used to generate the “neutral” phoneme version, ignored the phonotactics of the MBROLA voices.

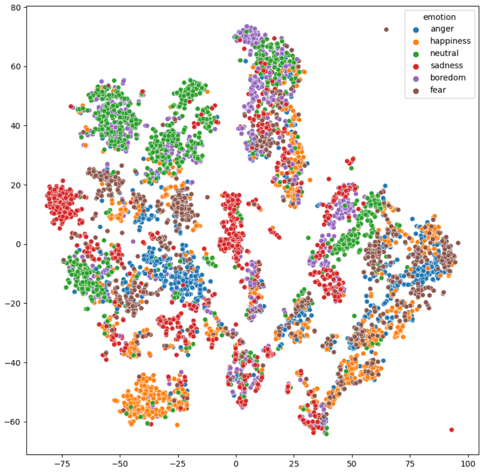
Figure 1: Frequencies of emotions per voice in the database



Both lead to an MBROLA error and no samples were generated. The resulting number of samples per emotion and synthetic voice in the database are shown in Figure 1. As can be seen especially the anger and fear emotions caused problems with the voices de1, de2 and de3, probably due to missing phoneme inventory caused by phoneme elisions.

Figure 2 shows a t-SNE plot (van der Maaten and Hinton, 2008) for the eGeMAPS (Eyben et al., 2015) features, colored by intended emotion. As can be seen by the colored clustered, the emotions can be separated based on acoustic features.

Figure 2: t-SNE plot for egemaps featues colored by emotion label



5. Evaluation

With respect to evaluation two approaches make sense:

- Evaluate the validity of the emotional expression by a human perception experiment.

[3] as zip file and from Zenodo `https://doi.org/10.5281/zenodo.6573016`

[4] `https://audeering.github.io/audformat/`

[5] `https://pypi.org/project/audformat/`

neutral	happiness	sadness	anger	mean
.6	.35	.5	.55	.5

Table 1: Results (total accuracy over all labels) per emotion in the perception experiment.

- Evaluate the usefulness with respect to machine learning as a training set.

5.1. Perception Experiment

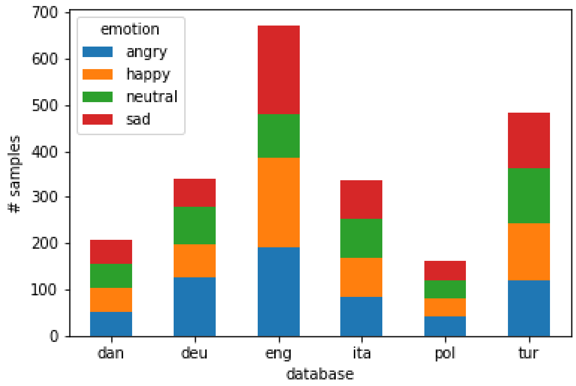
We conducted a perception experiment after we defined the modification rules as described in the literature (Burkhardt, 2000, chapter 5, pages 97-105) . It was a forced choice listening experiment with 20 participants who listened to the stimuli in random order. The results are displayed in Table 1. All emotions were recognized well above chance, the rather low value for happiness is mainly due to the fact that it was often confused with anger. This confusion is often seen in the literature (Yildirim et al., 2004) and caused by a similar level of arousal.

Interestingly, the mean accuracy for all emotions in this perception experiment for the rule-based simulation was highest compared to four others that used prosody copy from actors of the Berlin emotional database (Burkhardt et al., 2005).

5.2. Evaluation Databases

We investigate the usefulness of the data as a training set for machine classifiers by setting up a series of experiments with databases displaying acted basic emotions. Although these emotion expressions appear rarely in the real world, its detection still might be of practical value, as for example in Gaming scenarios or to teach children in the autism spectrum (Burkhardt et al., 2019).

Figure 3: Overview of databases with respect to basic emotion portrays



We look at the following six databases from different countries:

- ’emodb’ (Germany): The Berlin Emotional Speech Database (emodb)[6] (Burkhardt et al., 2005) is a well known studio recorded dataset.

- ’emovo’ (Italy): Italian Emotional Speech EMOVO[7] (Costantini et al., 2014) is a database consisting of the voices of six actors (three female, three male) who utter 14 Italian sentences simulating seven emotional states: anger, disgust, fear, joy, neutral, sadness, and surprise.

- ’ravdess’ (USA): The Ryerson Audio-Visual Database of Emotional Speech and Song (ravdess)[8] (Livingstone and Russo, 2018) contains recordings of 24 professional actors (12 female, 12 male), vocalising two English statements in a neutral North American accent. We excluded the songs.

- ’polish’ (Poland): The Database of Polish Emotional Speech (Powroźnik, 2017) consists of speech from eight actors (four female, four male). Each speaker utters five different sentences with six types of emotional state: anger, boredom, fear, joy, neutral, and sadness.

- ’des’ (Denmark): The Danish Emotional Speech (des) (Engberg et al., 1997) database comprises acted emotions of four professional actors – two males and two females – for five emotional states: anger, happiness, neutral, sadness, and surprise.

- ’busim’ (Turkey): For the Turkish Emotional Database (busim) (Kaya et al., 2014), eleven amateur actors (eight female, three male) provided eleven Turkish sentences with emotionally neutral content.

An overview of the databases is provided in Table 5.2. We tested these databases with a subset of the synthesized data being used solely as training. Therefore, all database emotion designations were mapped to the four target emotions of SyntAct, or removed if not part of the four target emotions. The resulting distributions per emotion category can be seen in Figure 3. For the four target emotions (angry, happy, neutral, and sad), out of the 1000 samples per speaker we selected randomly 30 samples per speaker and emotion, getting 720 samples with distinct texts.

We realize that emotional expression is culture and language specific (Neumann and Vu, 2018; Feraru et al., 2015; Burkhardt et al., 2006; Scherer et al., 1999) but as the expression of “basic, full-blown emotions” also is culturally universal and the acoustic feature set that we used for the evaluation does not model linguistics, we think it’s justified to use this German data to train

[6] https://www.tu.berlin/go22879/

[7] http://voice.fub.it/EMOVO

[8] https://doi.org/10.5281/zenodo.1188975

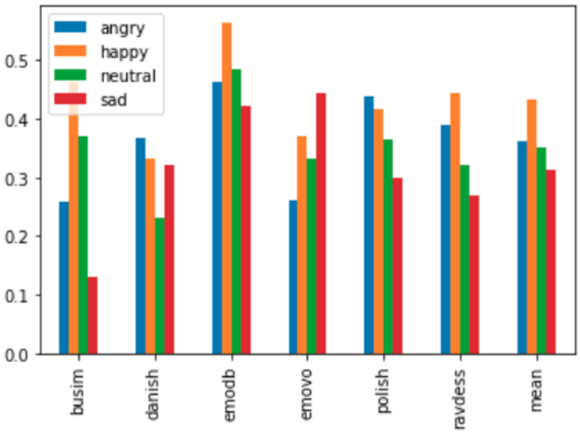
Name	Language	Year	#speakers	#emotions	#sentences	#samples
emodb	German	1999	10	7	10	484
emovo	Italian	2014	6	7	14	588
ravdess	N.-A. English	2018	24	8	2	1 440
Polish Emotional Speech	Polish	2014	8	6	5	240
des	Danish	1997	4	5	13	260
busim	Turkish	2014	11	4	11	484

Table 2: Overview of the emotional speech databases

busim	danish	emodb	emovo	polish	ravdess
.314	.317	.508	.360	.381	.366

Table 3: Results in UAR (unweighted average recall) per database when the synthesized data is being used as a training.

Figure 4: Overview of results per emotion for databases as F1 values



an emotion recognition classifier at least for European languages.

For the experiments we employed the Nkululeko framework[9] (Burkhardt et al., 2022) with an XGBoost classifier[10] (Chen and Guestrin, 2016) with the default meta parameters ($eta = 0.3$, $max\_depth = 6$, $subsample = 1$). This classifier is basically a very sophisticated algorithm based on classification trees and has been working quite well in many of our experiments (Burkhardt et al., 2021).

As acoustic features, we used the eGeMAPS set (Eyben et al., 2015), an expert set of 88 acoustic features for the openSMILE feature extractor (Eyben et al., 2010) that were optimised to work well to explain speaker characteristics and in particular emotions. These features are being used in numerous articles in the literature as baseline features (e. g., (Ringeval et al., 2018; Schuller et al., 2016)) as they work reasonably well with many tasks and are easy to handle for most classifiers based on their small number.

5.3. Results

In Figure 4 and Table 3, we present the result of our experiment. Note that in the figure, we use the F1 measure as a combination of recall and precision (because the results stand for one specific emotion), while in the table, we report unweighted average recall (UAR, which is the standard measure of the Interspeech Computational Paralinguistics Challenge).

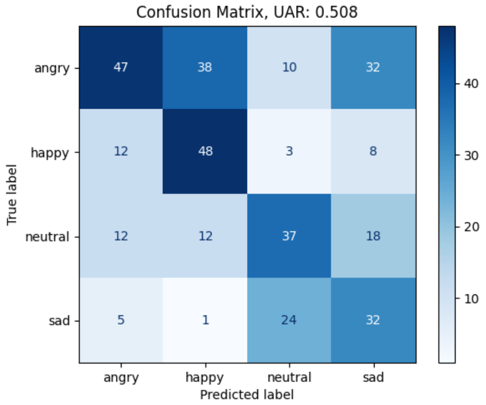
Following the numbers in the table, we can see that all the values are above the chance level (of .25 UAR for four emotions). Likewise, it seems that in general it is possible to detect emotional arousal with the database for several languages, but considering the results for the specific emotions in the figure, it is striking that the results depend extremely on which emotion is classified in which database. For example, while the simulation of sadness does work quite well fort the Italian database (emovo), this is not at all true for the Polish, the ravdess (American English), and especially the busim (Turkish) databases. On average, happiness simulation result in a much better model than sadness. The German database shows the best performance and it is probably not by chance that the database is also German. Although we did not use linguistic features, the expression of emotions is influenced by culture (Scherer et al., 1999; Burkhardt et al., 2006; Barrett et al., 2019)..

As an example, we present in Figure 5 the confusion matrix for the Emodb database as a test set. As can be seen, the classification mainly worked, especially well for happiness, which is in general the best working simulation based on Figure 4. “Angry” was often confused with happy, which is a quite typical confusion based on a similar level of arousal, but also with sadness, which we can’t explain really. “Neutral” was sometimes confused with sadness, and “sadness” consequently with “neutral”, probably based on the common low level of arousal.

[9] `https://github.com/felixbur/nkululeko/`

[10] Actually using the Python XGBoost package: `https://xgboost.readthedocs.io/`

Figure 5: Example confusion matrix for Emodb database (German) as a test set



6. Ethical Considerations and Broader Impact

With respect to ethical considerations, generally it must be stated that the processing of emotional states is of great severity (Batliner et al., 2022). It should be made transparent to users of such technology that the attribution of emotional states based on human signals by machines based on statistics and pattern recognition is simply a substitute technology, as the true emotional state can never be inferred by others. It is a large part of human-human communication and also human-machine communication benefits, but severe decisions, that affect users well-being, should definitely not be based on this.

Nonetheless, we do believe that the interpretation of the emotional channel is important for natural human-machine speech communication and hope, as stated above, that especially lower resourced languages might benefit from the idea to train emotion aware systems with simulated prosodic variation, which comes cheap and does not require much data.

7. Conclusion and Outlook

We described a database of prototypical emotional expression in German that has been synthesized with rule-based speaking style modifications and made accessible to the public. The application of this data to generate a training set for natural emotional expression has been investigated with six international databases.

With respect to the 35 languages the MBROLA supports, we plan to extend the database in the future. Also the number of emotion portrayals may be extended.

A very interesting approach would be the simulation of emotional dimensions like pleasure, arousal, and dominance because on the one hand, many natural databases have been annotated with these dimensions (Lotfian and Busso, 2017), and on the other hand, the dimensions might be mapped flexible to specific categories, like for example “interest”. At the time of writing, a first implementation of rule-based independent arousal and valence simulation has already been implemented and awaits evaluation experiments.

8. Acknowledgements

This research has been partly funded by the European EASIER (Intelligent Automatic Sign Language Translation) project (Grant Agreement number: 101016982).

9. Bibliographical References

Baird, A., Parada-Cabaleiro, E., Hantke, S., Burkhardt, F., Cummins, N., and Schuller, B. (2018). The perception and analysis of the likeability and human likeness of synthesized speech. pages 2863–2867, 09.

Baird, A., Amiriparian, S., and Schuller, B. (2019). Can deep generative audio be emotional? towards an approach for personalised emotional audio generation. pages 1–5, 09.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., and Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological science in the public interest*, 20(1):1–68.

Batliner, A., Neumann, M., Burkhardt, F., Baird, A., Meyer, S., Vu, N. T., and Schuller, B. (2022). Ethical awareness in paralinguistics: A taxonomy of application. *International Journal of Human-Computer Interaction*.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A database of german emotional speech. In *9th European Conference on Speech Communication and Technology*, volume 5, pages 1517–1520, 09.

Burkhardt, F., Audibert, N., Malatesta, L., Türk, O., Arslan, L., and Auberge, V. (2006). Emotional prosody - does culture make a difference? In *Proceedings of the International Conference on Speech Prosody*.

Burkhardt, F., Saponja, M., Sessner, J., and Weiss, B. (2019). How should pepper sound - preliminary investigations on robot vocalizations. Peter Birkholz, Simon Stone.

Burkhardt, F., Brückl, M., and Schuller, B. (2021). Age classification: Comparison of human vs machine performance in prompted and spontaneous speech. In Stefan Hillmann, et al., editors, *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2021*, pages 35–42. TUD-press, Dresden.

Burkhardt, F., Wagner, J., Wierstorf, H., Eyben, F., and Schuller, B. (2022). Nkululeko: A tool for rapid speaker characteristics detection. In *Proceedings of LREC 2022*.

Burkhardt, F. (2000). *Simulation emotionaler Sprechweise mit Sprachsynthesesystemen*. Shaker.

Burkhardt, F. (2005). Emofilt: The simulation of emotional speech by prosody-transformation. In *9th European Conference on Speech Communication and Technology*.

Cahn, J. (2000). Generation of affect in synthesized speech. *J. Am. Voice I/O Soc.*, 8, 06.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Costantini, G., Iaderola, I., Paoloni, A., and Todisco, M. (2014). Emovo corpus: an italian emotional speech database. In *LREC*.

Deng, J., Zhang, Z., Marchi, E., and Schuller, B. W. (2013). Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *ACII*, pages 511–516. IEEE Computer Society.

Deng, J., Zhang, Z., Eyben, F., and Schuller, B. (2014). Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *Signal Processing Letters, IEEE*, 21:1068–1072, 09.

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

Engberg, I. S., Hansen, A. V., Andersen, O., and Dalsgaard, P. (1997). Design, recording and verification of a danish emotional speech database. In *EUROSPEECH*.

Eskimez, S., Dimitriadis, D., Gmyr, R., and Kumanati, K. (2020). Gan-based data generation for speech emotion recognition. In *Proc. Interspeech*, pages 3446–3450, 10.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). opensmile – the munich versatile and fast open-source audio feature extractor. *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*, pages 1459–1462, 01.

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., Andre, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., and Truong, K. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7:1–1, 01.

Feraru, S. M., Schuller, D., and Schuller, B. (2015). Cross-language acoustic emotion recognition: An overview and some tendencies. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 125–131.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12*.

Kaya, H., Salah, A., Gurgen, F., and Ekenel, H. (2014). Protocol and baseline for experiments on bogazici university turkish emotional speech corpus. In *2014 22nd Signal Processing and Communications Applications Conference, SIU 2014 - Proceedings*, 04.

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Epps, J., and Schuller, B. (2020). Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. *IEEE Transactions on Affective Computing*, PP:1–1, 04.

Livingstone, S. R. and Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):e0196391.

Lotfian, R. and Busso, C. (2017). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, PP:1–1, 08.

Neumann, M. and Vu, T. (2018). Cross-lingual and multilingual speech emotion recognition on english and french. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5769–5773, 02.

Powroźnik, P. (2017). Kohonen network as a classifier of polish emotional speech. *ITM Web of Conferences*, 15.

Ringeval, F., Cowie, R., Amiriparian, S., Michaud, A., Schuller, B., Kaya, H., Cummins, N., Çiftçi, E., Valstar, M., Schmitt, M., Lalanne, D., Güleç, H., Salah, A. A., and Pantic, M. (2018). AVEC 2018 Workshop and Challenge: Bipolar disorder and cross-cultural affect recognition. In *AVEC 2018 - Proceedings of the 2018 Audio/Visual Emotion Challenge and Workshop, co-located with MM 2018*.

Scherer, K. R., Banse, R., and Wallbott, H. G. (1999). Emotion Inferences from Vocal Expression Correlate across Languages and Cultures. *ICPhS 99*.

Schröder, M. and Grice, M. (2003). Expressing vocal effort in concatenative synthesis. *Proceedings of the 15th International Conference of Phonetic Sciences*, 09.

Schröder, M. and Trouvain, J. (2003). The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology*, 6:365–377.

Schuller, B. and Burkhardt, F. (2010). Learning with synthesized speech for automatic emotion recognition. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5150–5153.

Schuller, B., Zhang, Z., Weninger, F., and Burkhardt, F. (2012). Synthesized speech for model training in cross-corpus recognition of human emotion. *International Journal of Speech Technology*, 15, 09.

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J., Baird, A., Elkins, A., Zhang, Y., Coutinho, E., and Evanini, K. (2016). The interspeech 2016 computational paralinguistics challenge: Deception, sincerity and native language. In *Proceedings Interspeech 2016*, pages 2001–2005, 09.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Yildirim, S., Bulut, M., Lee, C., Kazemzadeh, A., Lee, S., Narayanan, S., and Busso, C. (2004). An acoustic study of emotions expressed in speech. 10.

Zhou, X., Ling, Z.-H., and King, S. (2020). The blizzard challenge 2020. pages 1–18, 10.