



EmoDB 2.0: A Database of Emotional Speech in a World that is not Black or White but Grey

Felix Burkhardt^{1,5}, Oliver Schrüfer¹, Uwe Reichel¹, Hagen Wierstorf¹, Anna Derington¹, Florian Eyben^{1,2}, Björn Schuller^{1,3,4,6}

¹audEERING GmbH, Germany, ²Agile Robots, Germany, ³TU-Munich, Germany, ⁴Imperial College London, England, ⁵TU-Berlin, Germany, ⁶Munich Data Science Institute, Munich, Germany
{fburkhardt | oschruefer | ureichel | hwierstorf | aderington | fe | bs}@audeering.com

Abstract

In 2004, the Berlin Database of Emotional Speech (Berlin EmoDB) was made publicly available, and since then, it has been utilized in numerous studies on emotional speech, with over 3,000 citations. We now extend this database with original material that was previously absent from the distributed versions within the research community. This includes ambiguous samples that were not agreed upon by at least 80 % of the raters, the addition of glottograms for all samples, and a new rater label indicating the perceived naturalness of the samples. The paper provides detailed descriptions and reports on preliminary studies conducted using this extended data. For instance, incorporating glottogram information has been shown to improve the average recall of an SVM classifier by 8.1 % UAR.

Index Terms: speech, emotion, database, glottograms, uncertainty

1. Introduction

This paper focuses on the extension of EmoDB, a German database of emotionally acted speech recorded in 1998 in the anechoic chamber at the Technical University of Berlin [1]. As one of the earliest publicly available databases without a formal license agreement, it has been utilised in over 3,000 studies worldwide, according to Google Scholar. Interestingly, despite its original purpose, the database has been predominantly employed in machine learning research. The funding project, *Phonetic Reductions and Elaborations in Emotional Speech*, carried out by the TU Berlin Institute of Communication Science and funded by the German Research Foundation (DFG), was originally intended to manually examine the acoustic correlates of emotional speech.

The database follows a categorical model of emotions, distinguishing six basic emotions—anger, joy, sadness, fear, disgust, and boredom—alongside neutral speech. Ten professional native German actors (5 female and 5 male) performed these emotions by imagining relevant situations, producing 10 utterances (5 short and 5 longer sentences), which were semantically appropriate for everyday communication and interpretable in all the applied emotions. An additional linguistic criterion was the potential for phonetic reduction or elaboration. The recordings were made using a Sennheiser MKH 40 P 48 microphone and a Tascam DA PI portable DAT recorder in an anechoic chamber. Additionally, glottograms were captured with a portable laryngograph (Laryngograph LTD). Figure 1 shows an actor during the recording session, with the glottograph sensor visible as the belt around their neck.

If the voice actors produced more than one convincing portrayal of an emotion, these were included as additional versions. The recorded speech material, consisting of 817 sentences (7

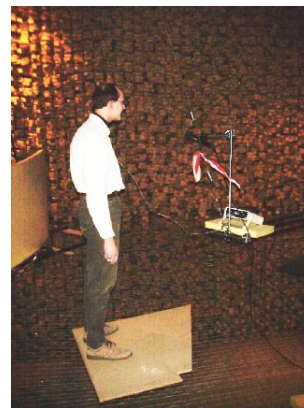


Figure 1: An actor in the anechoic chamber of the Acoustic institute of TU Berlin. One can see the laryngograph and its sensor around the actor's neck.

emotions * 10 actors * 10 sentences + additional versions), was evaluated in terms of rater agreement and naturalness through a forced-choice automated listening test by 20 judges (students of communication science, of mixed gender). In the first version of this database, which was distributed over the years via online platforms such as Kaggle or Zenodo, only those utterances where the emotion was recognised by at least 80% of the raters were included. We now present three additions:

- The release of the complete set of samples, including those with ambiguous emotional portrayals;
- The accompanying glottograms;
- Labels for naturalness in addition to rater agreement.

Unfortunately, the individual labels from each rater have been lost, so we only report mean values.

We publish this version in the open-source audio-data format audformat [2]¹. Given the extensive use of this database for machine learning, this version proposes a standard data split: 4 speakers (of mixed gender) for evaluation and 6 speakers for training purposes. The 281 previously excluded, ambiguous (not being recognized by sufficient number of raters) files are compiled in a separate table. To maintain backward compatibility, we did not alter the original tables. The raters' mean response to the second question, "How natural does the speaker sound?", scaled between 0 and 1, has been added as a second column alongside the emotional confidence value (which represents the percentage of rater agreement).

¹<https://github.com/audeering/emodb/tree/main/2.0.0>

Finally, all tables have been duplicated to include the glottograms (also referred to as laryngograms) that were recorded alongside the audio data. This paper is organised as follows: Section 2 provides examples of related work. We then describe experiments using the additional ambiguous samples to examine uncertainty in Section 3, followed by Section 4, which offers initial insights into the potential of glottograms for distinguishing emotional expressions. Section 5 explores the correlation between naturalness and emotional rater agreement. The paper concludes with Section 6, which presents a summary and outlook.

2. Related work

Since the inaugural release of EmoDB at Interspeech 2004 [1], numerous analogous speech databases have been developed across languages – sharing four principal characteristics:

- Inclusion of core emotion categories (typically *angry*, *happy*, *sad*, and *neutral*)
- Fixed textual content across emotional conditions
- Limited speaker cohorts (4–20 participants, with occasional larger samples)
- Acted rather than spontaneous emotional portrayals, though some incorporate film/TV excerpts

We present representative examples of such corpora. Most retain the three core emotions above, frequently extending to Ekman’s “Big Six” model [3] – originally devised for facial expressions. Notably, *fear* is often excluded, likely due to ethical considerations.

The IEMOCAP corpus [4] features emotionally acted dialogues captured via multimodal sensors, with 10 actors generating 12 hours of data. MELD [5] extends this paradigm to multi-party conversations by annotating excerpts from the television series *Friends*. RAVDESS [6] comprises 24 professional actors producing vocalisations at two intensity levels (normal, strong).

Several databases source material directly from media: CHEAVD [7], a Mandarin corpus with 238 speakers of diverse demographics, exemplifies this approach, though resultant textual and acoustic variability complicates analysis. The well-known CREMA-D [8] contains 91 speakers across US English dialects delivering fixed phrases for six basic emotions plus neutral.

Until recently, English dominated this field, but multilingual resources now proliferate:

- EMOVO (Italian): 6 actors simulating 7 emotions via 14 sentences [9]
- ASED (Amharic): 65 non-professionals portraying 5 emotions [10]
- CaFE (Canadian French): 12 actors performing 6 emotions + neutral [11]
- Polish Emotional Speech: 8 speakers delivering 5 sentences across 6 states [12]
- SUBESCO (Bangla): 22 speakers enacting Ekman’s “Big Six” with fixed text [13]
- KES is a corpus in Kannada: 13 speakers of varying age [14]
- DES (Danish): 4 professionals using varied phrases [15]
- BUEMODB (Turkish): 11 amateurs producing 11 sentences for 3 emotions + neutral [16]

Emergent meta-collections like EmoBox [17] now aggregate multiple databases (including EmoDB 1.0). A cross-

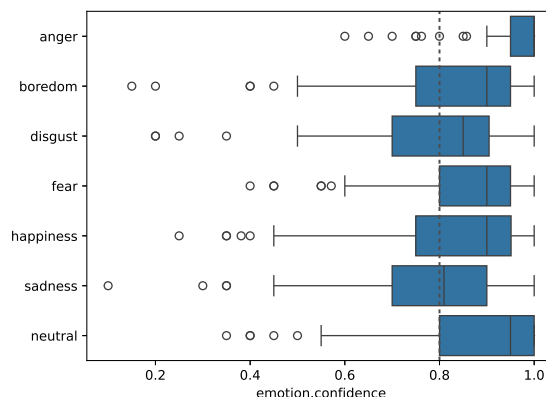


Figure 2: Confidence distribution per emotion, all samples left of the dotted line at 80% were previously discarded.

domain analyses of many of the above mentioned databases is presented in [18].

It has been argued, and shown in scientific experiments [19] that acted emotions are not really comparable to real emotions, as speakers do not really feel the emotion and therefore tend to overact the expression. Nevertheless, we argue that acted databases are still relevant, mainly for two reasons: First, there are use cases where you are interested in acted emotional expression, for example, when using emotional expression as a communicative signal. Secondly, these expressions are usually much more clear than *daily life* recordings and are well suited to compare machine learning approaches. They are further easy to collect under acoustically undisturbed condition, which renders them attractive for studies under added noise in controlled ways and speech synthesis training.

3. Ambiguous samples and uncertainty

At the beginning of the era of automatic speech-emotion recognition (SER) the attention was primarily laid on the general capability of recognising emotions. For this formulation of the task, it is easiest to select only those samples where the raters could agree on a common prediction so that there is a labelled ground truth.

Since the introduction of Transformer models [20], near-perfect unweighted average recall (UAR) recognition has been achieved on EmoDB resulting in no notable performance improvements on simple classification tasks over the last years [21]. Consequently, recent research has shifted its focus towards more complex, real-world applications, which present challenges such as realistic recording conditions and the inherent ambiguity of emotions [22]. Additionally, the importance of predicting the confidence of any given prediction has increased, as making no prediction is often preferable to making an incorrect one. Models with robust uncertainty estimation should be capable of filtering out samples where raters could not reach consensus on a ground truth.

Uncertainty is commonly divided into two types: aleatoric and epistemic uncertainties [23, 24]. The epistemic uncertainty is primarily introduced by the predictor, while the aleatoric uncertainty is a result of the variability of the underlying data itself, and thus cannot be reduced by improving the model. The newly published ambiguous samples aim to contribute to further

research on the aleatoric/data uncertainty of emotional speech data.

In the context of Speech Emotion Recognition (SER) this aleatoric (stochastic) uncertainty especially comes from the ambiguity of emotions. The same utterance might be perceived by different listeners in a different way, leading to different emotion predictions. In addition, different speakers might utter the same emotion in other ways. Since there is hence no such thing as a perfect ground truth, every emotion prediction system will reach its limitations should the emotions become too blurry and (acoustically) too close to one another. For reliable SER predictions, it is therefore essential to understand these limitations. The new ambiguous samples in the Berlin EmoDB are intended to provide a better insight into the limitations of credible emotion prediction. Here, a potential task for a predictor could be to detect the ambiguous samples and mark them as unsuitable for an emotion prediction.

Another interesting factor regarding aleatoric uncertainty is that the task to act an emotion so that it is reliably decoded, is also not equally difficult across all emotions. As can be seen in Figure 2, some emotions are much easier to identify than others. Although nearly all of the anger samples were identified correctly, almost half of the sadness samples were discarded in the first version due to insufficient agreement. Since this is an acted dataset, it is difficult to say whether it is harder to act those emotions or to identify or differentiate them from other emotions.

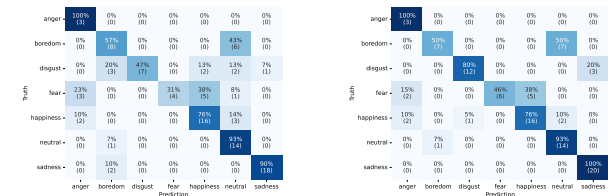
Other sources of aleatoric uncertainty could be a bad recording condition or language or cultural differences between the speakers and the raters. However, since the recording was done in an anechoic chamber and all speakers and raters are German native speakers, we can neglect those sources of uncertainty.

In order to compare ambiguous samples to ones the raters agreed on, we use the open source wav2vec2-large-robust model pruned to 12 layers [20] that has been fine-tuned for dimensional emotions (arousal, dominance, valence) on the MSP-Podcast data set [25]. Because in EmoDB emotion expression is encoded as categories, we cannot use the model directly, but use the penultimate layer as embeddings. We train a simple Support Vector Machine (SVM) classifier with RBF kernel and $C=1$ on the embeddings². Adhering to the suggested splits; we use 6 speakers for the training, leaving the remaining 4 speakers for evaluation (both sets gender distributed). We train two versions of this model: one trained on only (the original) non-ambiguous samples, and one trained on all samples. We then evaluate on both the non-ambiguous test samples and all test samples together. As shown in Tab. 1, as to be expected, the UAR is in general higher for non-ambiguous samples than for ambiguous samples. Further, the model that includes the ambiguous training samples is better than the model that does not on both test sets. Figure 3 shows the confusion matrix of both models on the ambiguous test samples. As can be seen, the prediction of disgust improves the most by including ambiguous samples in the training, but fear and sadness benefit as well. The only category that is slightly degraded is boredom which is in half of the cases confused with neutral. Whether these findings are merely specific to EmoDB or some underlying principle is shown, would be a topic for further investigation.

²We use the same approach as shown in the tutorial of the wav2vec2 model <https://github.com/audeering/w2v2-how-to>

Table 1: Results in terms of UAR for two SVM models using wav2vec2 embeddings, one trained on only non-ambiguous samples, and the other trained on all samples. We then evaluate on the non-ambiguous test samples, as well as all test samples.

training data	UAR non-ambig.	UAR all
non-ambig.	.89	.83
all	.95	.90



(a) Model trained only on non-ambiguous samples (b) Model trained on all samples

Figure 3: Confusion matrix for an SVM model trained on wav2vec2 embeddings, trained on only ambiguous samples (a) and trained on both ambiguous and nonambiguous samples (b). The models are evaluated on the ambiguous samples of the test speakers.

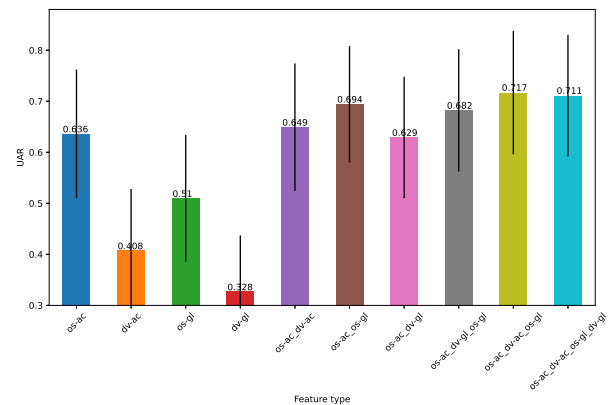


Figure 4: Results on the test partition obtained for all feature set combinations. os=eGeMAPS, dv=DisVoice, ac=Acoustic signal, gl=Glottogram.

4. Emotion recognition with glottograms

As mentioned in Section 1, one of the key additions in EmoDB 2.0 is the inclusion of glottograms alongside the acoustic waveforms. This raises the pertinent question, particularly in the context of machine learning, of how glottograms can contribute to SER. From each of the two speech signals—the acoustic waveform (**ac**) and the glottogram (**gl**)—we extracted features from two sets:

- **os**: openSMILE eGeMAPSv02 functionals and summary statistics [26]
- **dv**: DisVoice glottal source feature summary statistics [27] (originally referred to as *static features*)³

For DisVoice glottal feature extraction from glottograms,

³DisVoice features were extracted using the following code: <https://github.com/jcvasquezc/DisVoice>.

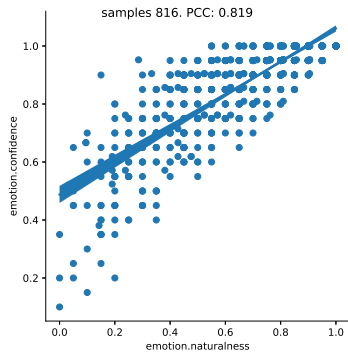


Figure 5: Correlation between majority rater agreement (named *emotion.confidence* in the distribution) and naturalness.

we first extracted the glottal closure instants (GCI) by local admittance peak detection within voiced regions. Voiced regions were identified by means of RAPT pitch extraction [28]. For GCI detection, the glottogram was smoothed by Savitzky Golay filtering, which well preserves the time locations of local maxima. Local peak detection was then guided by minimum distance, minimum height, and minimum valley constraints of (adjacent) peak candidates. Threshold values were adjusted dynamically for each voiced region.

Subsequently, for both **os** and **dv** feature extraction the glottograms were inverted in order to convert the admittance signal into a pseudo sound pressure signal. For the inverted glottograms we discarded all **os** features related to formants, namely formant frequencies, amplitudes, bandwidths, as well as formant-related spectral tilt summary statistics.

We then utilised all signal and feature set combinations in an emotion classification experiment, employing an SVM classifier with an RBF kernel and a C value of 10. The same training and test partition split, as outlined in Section 3, was applied. Figure 4 presents the UAR results for the test partition, with error bars representing the 95

Both the **dv** and **os** feature extractors were originally designed to operate on sound pressure signals, not inverted glottal admittance signals. This explains why in Figure 4, both **dv-gl** and **os-gl** exhibit lower UAR values compared to **dv-ac** and **os-ac**, respectively. In no case do features derived solely from glottogram signals outperform those derived from acoustic waveforms. However, in all cases, performance improves when glottogram-derived features are added, with a particularly notable improvement when eGeMAPS features based on glottograms are included. In the most extreme case, we observe an 8.1

Which emotional category benefits the most from adding glottogram information will be explored in future research, and additional data should certainly be included in such an investigation.

5. Labeler agreement and naturalness

As mentioned in Section 1, we incorporated the mean naturalness label into the database, raising questions such as: “Can naturalness be predicted automatically?”, “Is there a correlation between rater agreement and naturalness?”, and “Is naturalness biased by emotion category?”

To begin addressing these questions, we conducted exploratory experiments using the Nkululeko toolkit [29]. As il-

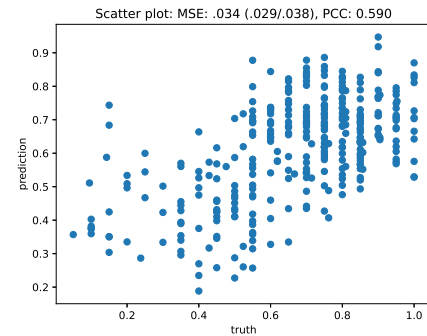


Figure 6: Prediction and truth of naturalness.

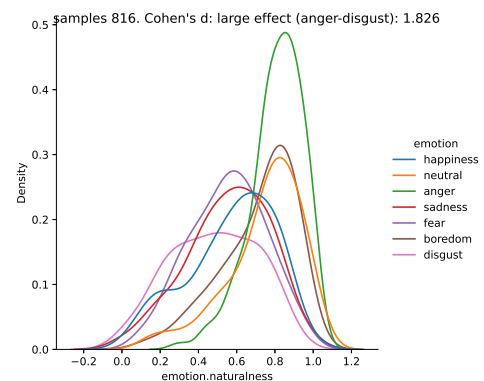


Figure 7: Density of naturalness distribution for emotional categories.

lustrated in Figure 5, we found a strong correlation between rater agreement and naturalness, with a Pearson correlation coefficient (Pearson’s correlation coefficient (PCC)) of .816.

We then attempted to predict naturalness based on acoustic features, using the openSMILE eGeMAPS feature set [26] and an SVM regressor (RBF kernel, C=10), employing the suggested training/evaluation splits. As shown in Figure 6, the correlation between the predicted and true naturalness values for this experiment yielded a PCC of .59.

Lastly, we plotted the distributions of mean naturalness across emotional categories, as shown in Figure 7. To assess the impact of emotion on naturalness, we computed pairwise Cohen’s D. The largest effect size was found between anger, which exhibited relatively high naturalness values, and disgust, which showed lower values (Cohen’s D = 1.826). All figures and analyses were generated using Nkululeko [29].

6. Summary and outlook

This paper introduced new contributions to a widely recognised database of acted emotional expressions and presents some preliminary findings based on this data. We acknowledge that further experiments, including those involving additional datasets, are required to establish statistically robust conclusions. These initial investigations serve primarily as baseline experiments for our research and potentially for the broader scientific community. One limitation of the current study is that several topics discussed would benefit from a more detailed analysis, which we intend to address in future work.

7. Acknowledgements

Part of this work received funding from the European SHIFT (Metamorphosis of cultural Heritage Into augmented hypermedia assets For enhanced accessibility and inclusion) project (Grant Agreement number: 101060660). The original data recording has been funded by Deutsche Forschungsgemeinschaft DFG (German Research Community), grant nr. SE 462/3-1.

8. References

- [1] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *9th European Conference on Speech Communication and Technology*, 2005.
- [2] H. Wierstorf, J. Wagner, F. Eyben, F. Burkhardt, and B. W. Schuller, "audb – sharing and versioning of audio and annotation data in python," *arXiv preprint arXiv:2303.00645*, 2023.
- [3] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17 2, pp. 124–9, 1971.
- [4] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.
- [5] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. ACL*, 2019, pp. 527–536.
- [6] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, p. e0196391, 2018.
- [7] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "CHEAVD: a chinese natural emotional audio-visual database," *J. Ambient Intell. Human Comput.*, vol. 8, pp. 913–924, 2017.
- [8] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affective Comput.*, vol. 5, no. 4, pp. 377–390, 2014.
- [9] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "Emovo corpus: an italian emotional speech database," in *LREC*, 2014.
- [10] E. A. Retta, E. Almekhlafi, R. Sutcliffe, M. Mhamed, H. Ali, and J. Feng, "A new amharic speech emotion dataset and classification benchmark," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 1, pp. 1–22, 2023.
- [11] P. Gournay, O. Lahaie, and R. Lefebvre, "A canadian french emotional speech dataset," in *Proceedings of the 9th ACM multimedia systems conference*, 2018, pp. 399–402.
- [12] P. Powroźnik, "Kohonen network as a classifier of polish emotional speech," *ITM Web of Conferences*, vol. 15, 2017.
- [13] S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "SUST bangla emotional speech corpus (SUBESCO): An audio-only emotional speech corpus for bangla," *PLOS One*, vol. 16, no. 4, p. e0250173, 2021.
- [14] A. Geethashree and D. Ravi, "Kannada emotional speech database: design, development and evaluation," in *Proceedings of International Conference on Cognition and Recognition: ICCR 2016*. Springer, 2017, pp. 135–143.
- [15] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a danish emotional speech database," in *EUROSPEECH*, 1997.
- [16] H. Kaya, A. Salah, F. Gurgen, and H. Ekenel, "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 2014.
- [17] Z. Ma, M. Chen, H. Zhang, Z. Zheng, W. Chen, X. Li, J. Ye, X. Chen, and T. Hain, "Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark," in *Proceedings Interspeech 2024*, 09 2024, pp. 1580–1584.
- [18] F. Burkhardt, A. Hacker, U. Reichel, H. Wierstorf, F. Eyben, and B. Schuller, "A comparative cross language view on acted databases portraying basic emotions utilising machine learning," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 1917–1924. [Online]. Available: <https://aclanthology.org/2022.lrec-1.204>
- [19] J. Wilting, E. Kraemer, and M. Swerts, "Real vs. acted emotional speech," in *Proceedings Interspeech 2006*, 2006, pp. paper 1093–Tue1A30.4.
- [20] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, 2023.
- [21] A. Triantafyllopoulos, A. Batliner, S. Rampp, M. Milling, and B. Schuller, "Interspeech 2009 emotion challenge revisited: Benchmarking 15 years of progress in speech emotion recognition," *arXiv preprint arXiv:2406.06401*, 2024.
- [22] O. Schrüfer, M. Milling, F. Burkhardt, F. Eyben, and B. Schuller, "Are you sure? analysing uncertainty quantification approaches for real-world speech emotion recognition," in *Proc. Interspeech 2024*, 2024, pp. 3210–3214.
- [23] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.
- [24] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher *et al.*, "A survey of uncertainty in deep neural networks," *Artificial Intelligence Review*, vol. 56, no. Suppl 1, pp. 1513–1589, 2023.
- [25] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October–December 2019.
- [26] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.
- [27] E. A. Belalcázar-Bolanos, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, T. Haderlein, and E. Nöth, "Glottal flow patterns analyses for parkinson's disease detection: Acoustic and nonlinear approaches," in *International Conference on Text, Speech, and Dialogue*. Springer, 2016, pp. 400–407.
- [28] D. Talkin and W. B. Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.
- [29] F. Burkhardt, J. Wagner, H. Wierstorf, F. Eyben, and B. Schuller, "Nkululeko: A tool for rapid speaker characteristics detection," in *2022 Language Resources and Evaluation Conference, LREC 2022*. European Language Resources Association (ELRA), 2022, pp. 1925–1932.