What the voice reveals

audEERING®

Abstract

This white paper highlights the background of the audEERING **devAIce® v3.14.0-v4.1.1** product (as of June 8th, 2026), focused on the prediction of speaker expression and age. We discuss AI model generation and especially evaluation and present typical use cases of the technology.

Expression in the voice is an essential building block of human communication, it's being used to express urgency, distress or joy even before humans developed language, as already pointed out by Charles Darwin [1]. These expressions don't only appear involuntarily, but even more so in a social context to express meaning. A simple example is irony, which often is expressed by a divergence between meaning and expression of an utterance. As such it is a valuable build block for systems that deal with human verbal communication.

Contents

Contributors

Felix Burkhardt, Dagmar Damazyn, Anna Derington, Florian Eyben, Mehrzad Mashal, Soroosh Mashal, Uwe Reichel, Milenko Saponja, Rahul Swaminathan and Hagen Wierstorf contributed to this report.

1 Overview

This white paper describes the audEERING speaker prosody, expression, age and gender recognition model. We give an overview of these features and how the model is trained, what exactly is predicted and how we evaluate the model in terms of accuracy, robustness, fairness and efficiency, concluded by a section on possible use cases.

Generally, speaker characteristics recognition is a sub field of AI based on **machine learning**, and technically most often labeled as **pattern recognition**. The basic idea is to model the data of interest (in our case speech samples) by some features (for example, the melody of the voice). In case of **supervised learning** (which is typically used as a start), you collect so-called **labels** for the data samples and then use some statistical methodology to *teach* a model to predict labels for previously unseen data. This is the training process. With state of the art machine learners, the features that get extracted are determined by machine learning as well, and the statistical methodology is some artificial neural net approach based on **deep learning**.

2 Vocal analysis features

devAIce® technology encompasses many aspects of vocal analytics:

- Prosody: vocal biomarkers
- Expression dimensions: arousal, dominance, valence
- Expression categories: angry, happy, neutral, sad
- Speaker attributes: self-reported gender and age
- Automatic Speech Recognition (ASR): transcribed text

2.1 Prosody

Prosodic analysis is the classic approach to analyzing voices, and it is typically used for clinical and linguistic research. These features are being used as **vocal biomarkers**, meaning they are affected by health conditions and can thus be used to assess medical conditions.

Fundamental Frequency (F0) Pitch refers to the perceived frequency of a sound, specifically how high or low a sound is perceived to be. It refers to the number of vibrations of the vocal folds or

cycles of the fundamental frequency that occur in one second. Average F0, for example, can be used to differentiate between male and female voices (85-155 Hz vs 165-255 Hz, respectively). A high maximum F0 with a normal average F0 could indicate spontaneous vocal outbursts with high pitch, such as expressions of surprise or disgust.

Loudness Loudness is a measure of how loud a human perceives a sound. It is related to the amplitude of the sound wave, but accounts for physiological properties of our hearing, which are known to experts as psychoacoustics. A sound wave with twice the amplitude is not perceived as twice as loud by us; rather, it is perceived as being approximately 1.4 times louder.

Speaking Rate Speaking rate refers to the rate with which one speaks and is measured as the number of syllables per second and its variation within one speech segment. When the variation of speaking rate is low, it indicates that the voice has a steady pace, while a high variation indicates changing pace, e.g. slowing down on important utterance parts and speeding up on less important parts. A high variation could also indicate the presence of filled pauses and hesitations, due to thinking aloud or higher cognitive load.

Intonation Intonation is the measure of the rise and fall of the voice (e.g. a measure of how much pitch is being varied within one speech segment). It is useful for measuring how monotonously (low intonation value) or lively (high intonation value) a person is speaking.

2.2 Expression

Expression can be measured in a dimensional way (values on a continuous scale between -1 and 1) and categorical way (class labels). We provide both solutions in our dimensional and categorical expression modules. Measuring expression in a continuous way allows for nuanced comparisons, whereas categories have the benefit that speech can be clearly classified without requiring additional thresholds or clustering to form groups. The dimensional and categorical modules are designed to be consistent with one another, meaning that both can be used simultaneously without leading to contradicting results.

2.2.1 Dimensional Values

Arousal measures how passive or active a speaker sounds. High urgency, screaming, or shouting coincides with high arousal, whereas a calmness and low urgency in the voice are signs of low arousal. Specifically, panic or exaltation are high arousal expressions, whereas satisfaction and melancholy are low arousal expressions [2, 3].

Dominance measures the level of control or potency of a speaker. A voice that sounds angry, proud, or interested has a high dominance value, and a voice that sounds fearful, sad, or like the speaker feels moved has a low dominance value [2].

Valence measures how negative or positive a speaker sounds. Expressions of anger or sadness are examples of low valence, and happiness or joy are examples of high valence [4].

2.2.2 Categorical Values

The categorical emotion module predicts the categories *anger*, *happiness*, *neutral*, and *sadness*, so-called 'primary emotions', which, according to Darwin [1], are being recognized in a culture-universal manner, and are represented in the majority of labeled databases. Many theoretical emotion models, e.g. [5], include these categories as basic emotions. Associated with each of the four categories, our module also outputs a value between 0 and 1 that indicates the level that the category is expressed with.

2.2.3 Relation of Expression Dimensions to Categories

Expression dimensions and categories can be visualized in a 3D space. Fig. 1[6] shows approximate areas where in this space the 4 expression categories angry, happy, neutral, and sad occur. The exact placement of expression categories in the dimensional expression space is subjective and the results of different studies vary [7, 8, 9, 10]. For example, anger can be expressed as furious, heated anger with high arousal, or as cold contempt with lower arousal. Happiness can range from a calmer, low arousal expression of contentment to a high arousal expression of joy. Therefore, our dimensional model maps to the categorical expression values broadly as shown in Figure 2. Measuring expression in terms of these categories allows for a clear cut separation and potentially a simpler analysis. On the other hand, the dimensional model values allow a more nuanced and precise capturing of expression. If desired, one can also divide the basic 4 categories further based on their dimensional value. For example, if the expression of happiness should be more specific, the categories could be split into positive arousal (excitement, joy) and negative arousal (contentment, relaxation).

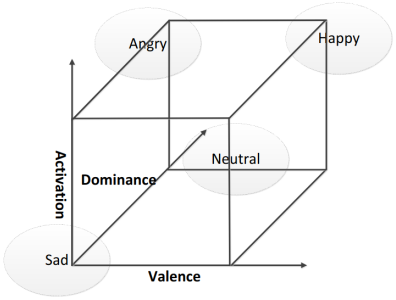

Figure 1. Correlation between the arousal (=activation), dominance, valence dimensions and the 4 categories angry (high arousal and dominance, low valence), happy (high valence), neutral, sad (low arousal, dominance and valence) [6].

2.3 Speaker Attributes

devAIce® outputs a gender and age attribute based on voice features and audEERING's deep AI, which has seen over 2 million speech segments. The gender attribute is not related to the gender identity of the speaker. It is purely an estimate of the perceived biological gender attribute of the voice, i.e. if the voice sounds like a male voice or female voice. For each speaker, the **devAIce**® model will give three output metrics (male, female and child) with a confidence level from 0-1. The

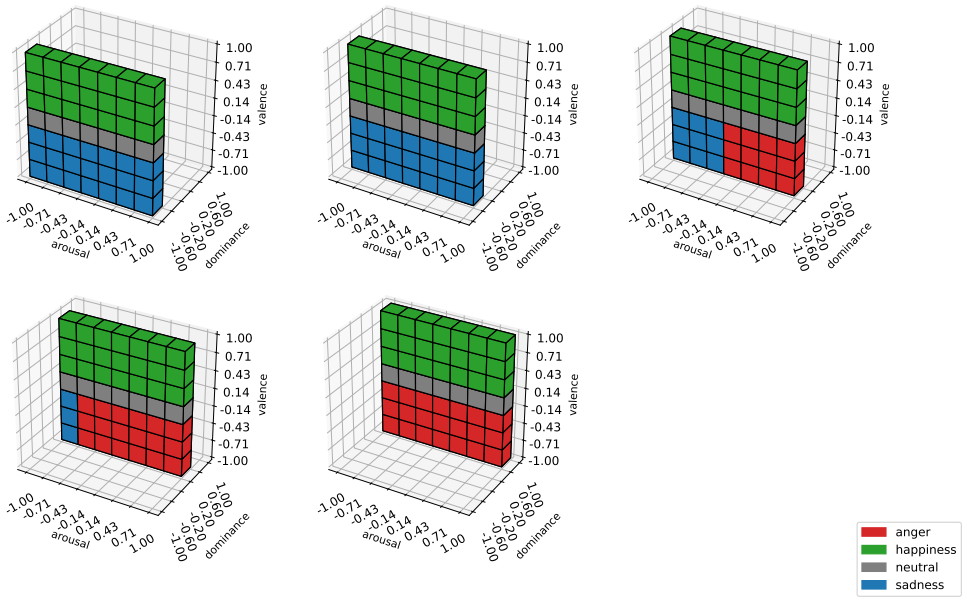

Figure 2. The **devAIce**® expression module's mapping of arousal, dominance, valence dimensional space to the 4 categories angry, happy, neutral, sad. Note that the "sad" category encapsulates both a low arousal variant (e.g. sounding depressed) and a high arousal variant (e.g. sounding distressed). The "angry" category contains both a low arousal variant (cold anger, contempt), as well as a higher arousal variant (hot anger, fury).

voice of young children differs significantly from the voice of both male and female adults, which is why we use these three output metrics. The age feature estimates the predicted age of the speaker in years.

2.4 Automatic Speech Recognition

The ASR (Automatic Speech Recognition) module transcribes spoken language into written text. Its foundation lies in the high-performance whisper.cpp API [11], which provides state-of-the-art transcription quality of OpenAI's Whisper models [12] with efficient processing. As can be seen in Table 1, the accuracy in terms of word error rate (WER) improves with the size and resource requirements of the model. Additionally, the English-only models perform better than their multilingual counterpart. For more detailed information on the available models' accuracies, plus documentation on more recent variants such as large-v3-turbo, please refer to OpenAI's documentation.[1]

Word-level timestamps are generated alongside the transcribed text. These timestamps allow a precise alignment of the text to the segment-level outputs of all **devAIce**® modules, ensuring a smooth comparison between text and prosodic features.

[1]For detailed information on the models shown in Table 1, refer to [12]. For multilingual accuracies of the large-v3 and the large-v3-turbo models, refer to https://github.com/openai/whisper/discussions/2363

Table 1. English WER of model variants reported by OpenAI [12, Appendix D1.2] on the LibriSpeech test-clean benchmark [13].

Model	WER
Whisper tiny.en	5.4
Whisper tiny	6.7
Whisper base.en	4.1
Whisper base	4.9
Whisper small.en	3.2
Whisper small	3.3
Whisper medium.en	3.0
Whisper medium	2.7
Whisper large	2.8
Whisper large-v2	2.5

3 Model Training

We constantly adopt newest technologies in the deep learning field to build the best and most robust models for our customers. Currently the most stable and best performing architecture is based on the so-called transformer models [14], which also are the basis for large language models, for example ChatGPT. audEERING was the first company to outline its use for speech prediction to the public [15]. As mentioned, these models are trained with labeled data, The data itself is either commercial databases, open public data, scraped from public sources, recorded by audEERING or even synthesized. With respect to **data labeling** several approaches are being used.

- A pool of **trained human annotators** are labeling selected data to be used for evaluation or initial models.
- Data annotation platforms are used to add human labels from **crowdsourcing** in an international context. **Multicultural and multilingual** influences are an important factor when it comes to speaker expression prediction.
- So-called **soft labels** are predicted by strong but too large models to teach smaller models. They might also use additional information like facial expression, social situation or physical measures.
- Additional training data can be **synthesized** by speech synthesizers.

3.0.1 Training Data for Expression

Our training data for expression is generally based on subjective perception of speech. For our own databases, we employ a team of professionally trained annotators, and the average score of their labels is used to train and evaluate the model. Figure 3 shows the correlation on valence of a single, randomly selected human rating and the average of all human ratings on the database

MSP-Podcast [16], test set 1. A single human rating can vary quite significantly from the average of all ratings, highlighting that annotating dimensional expression is subjective and difficult.

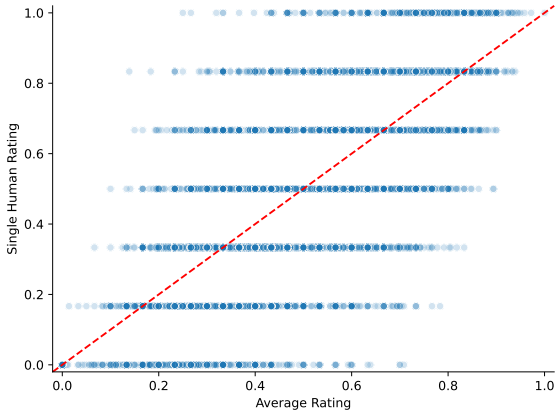

Figure 3. For each sample, we compare a randomly selected, single human rating to the average across all human ratings on that sample. The samples and ratings are taken from the valence dimension of the MSP-Podcast test set 1.

For both the training and evaluation, we use various multilingual public datasets, as well as some internal datasets. While a large portion of the training data is English, our expression model has also been trained on samples in the languages Bengali, Chinese. French, German, Italian, Kannada, and Portuguese. Further, our training data is balanced in terms of self-reported gender with 36% female, 42% male, and 22% unknown gender data.

3.0.2 Training Data for Age and Gender Estimation

The same goes with the data that is being used to train audEERINGs age and gender model, a heterogeneous mix of human populations is being used. Some of our approaches, as well as a (inferior to the commercial) public model for research, are being discussed in [17].

4 Model Evaluation

4.1 Expression Testing Framework

We validate our expression models with an extensive set of tests, which we present in [18]. Our evaluation process goes beyond the standard benchmark metrics and covers various aspects of correctness, fairness, and robustness.

Table 2. Correctness in terms of CCC for arousal (**A**), dominance (**D**), and valence (**V**) on MSP-Podcast test set 1.

A	**D**	**V**
0.75	0.65	0.64

Table 3. Correctness in terms of UAR on various languages.

Language	UAR
Bengali	0.89
German	0.70
English	0.70
Spanish	0.61
French	0.89
Italian	0.84
Kannada	0.62
Chinese	0.74

4.1.1 Correctness

For the classification of emotional expression, we mainly use the metric unweighted average recall (UAR) to measure correctness, and for the prediction of arousal, dominance, and valence, concordance correlation coefficient (CCC) is the most important metric. The definitions of these metrics are given in 8.1 and 8.2. Tab. 2 lists the correctness in terms of CCC for the dimensional expression on the MSP-Podcast [16] database (using samples from version 1.7 and test set 1).

Tab. 3 shows the correctness in terms of UAR for categorical expression, averaged across our many databases and grouped by language. Note that the databases' samples may have different types of speech (acting vs. natural conversation), different environments and context, which can make the task more or less difficult. Nonetheless, the results can be taken as a good indicator that our expression model works well across many cultures and languages, even though the largest portion of the training data is English.

In addition to evaluating with these standard metrics on multiple datasets, we also go beyond traditional evaluation of correctness. For example, we measure the consistency of the dimensional predictions, i. e., whether the dimensional model predictions are in the expected output range for certain expression categories. Fig 4 demonstrates the alignment of dimensional predictions with categories for the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) database.

4.1.2 Robustness

We evaluate the robustness of our models by testing various types of recording conditions and environments, including:

- Added background noise (human noises including coughing, sneezing, background chatter, environmental noise, music, artificial noise)
- Low quality phone simulation
- Different types of reverb by applying room impulse responses

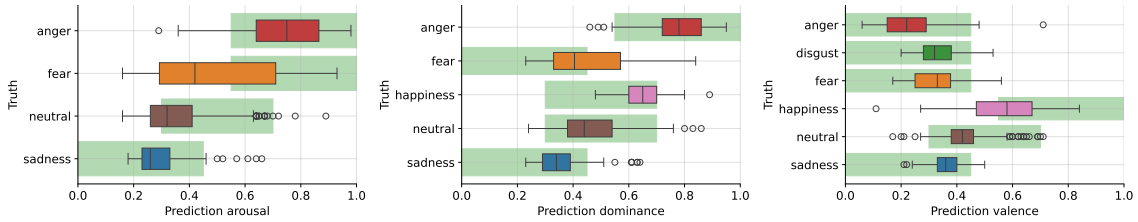

Figure 4. Model prediction for arousal (left), dominance (center), valence (right) on the CREMA-D test set, split by annotated categorical emotion. The green area marks the region in which a dimensional prediction falls into the expected, consistent range.

- Test data recorded simultaneously on different devices

Then, we measure how much the model output differs under the original condition compared to the condition with perturbation. For dimensional model outputs, we count the model output as unchanged, if the difference is less than 5%. Table 4 shows the average percentage of unchanged predictions under various robustness tests. Among the noises included in the Background Noise test,

Table 4. Average percentage of unchanged predictions for arousal (**A**), dominance (**D**), valence (**V**), and for emotional categories (**C**) for different robustness tests.

Test	A	D	V	C
Background Noise	.92	.94	.88	.88
Low Quality Phone	.86	.91	.76	.83
Recording Condition	.94	.92	.93	.90
Simulated Rec. Condition	.85	.86	.81	.86

coughing and sneezing noises have the biggest effect on predictions, especially for the valence dimension, where these noises can be confused with laughter. By applying specialized augmentation techniques for our model, approximately 70% of the predictions remain unchanged by this noise.

4.1.3 Fairness

Vocal expression may vary among people of different age, sex, culture, or other attributes. As such, we make it a priority to evaluate our emotional expression for different groups and evaluate:

- Fairness Accents
- Fairness Languages
- Fairness Pitch
- Fairness Sex

It is important to not only look at the raw difference in accuracy between two groups, but to also consider, for example, if there is a tendency to over-predict a certain class or value range for one group compared to the other groups (Principle of Equalized Odds [19]). For fairness on self-reported gender, our tests indicate that the correctness of each group compared to the entire test

set only differs up to 0.06. Further, considering the recall and precision (in the range of 0-1) for certain categories and dimensional ranges, we observe a difference of up to 0.1.

As mentioned in 4.1.1 the nature of the dataset can greatly vary in terms of context and environment, and the comparison in performance between languages can be skewed when samples from different languages originate from different databases. Therefore, we additionally employ a test using the dataset of Mozilla Common Voice [20], containing 2000 randomly selected samples from the languages German, English, Spanish, French, Italian, and Chinese. Although there are no expression labels for this dataset, each language should have a similar distribution, with mostly neutral samples. The mean value of our dimensional expression model on this data shifts only up to 4% between languages, indicating that the distributions are similar.

4.1.4 Efficiency

Since the publishing of our paper, we have continued to expand our product tests to also evaluate efficiency, for example with respect to memory or latency.

4.2 Summary

Speaker characteristics recognition is a sub field of AI based on machine learning. audEERING's devAIce® product predicts vocal biomarkers, expression and other speaker attributes like age or gender, based on models that are trained on large databases for culturally and population-wise diverse sources. These models get tested for accuracy, fairness, robustness and efficiency before being handed over to production.

5 Multimodality

The data that is being used to train and use expression predicting models, sometimes comes as video data. This data offers three modalities

- **Speech** (acoustics)
- **Text** (linguistics)
- **Images/Video** (facial expression)

For each of these modalities own models can be built[2]. With respect to expression dimensions (see Section 2.2.1), it has been shown that the *arousal* is more expressed in speech and the *valence* often clearer in the linguistic and facial modalities [15, 21]. The good news is, that linguistic features

[2]audEERING currently is focused on speech and text, video processing is only being used internally.

are also transcoded in the transformer models, the technology that is predominantly being used for acoustic expression prediction. Because the specific modalities might not be present during an incoming stream of data (the speaker might be silent, uttering extra-linguistic noises or the face might not be visible) synchronizations becomes an important issue. In addition, some special situations might apply:

- Several models using different modalities might **agree** on the prediction, which would make the prediction more certain.
- They might **disagree**, which could be a voluntary communication act, for example in the case of irony or sarcasm.

Models based on different modalities can be fused (combined) in two ways:

- **Late fusion** is a lightweight fusion method that agrees on a prediction based on several predictions, typically by averaging them or using the most often predicted category.
- **Early fusion** is much more complex and constructs one overarching model from several modality inputs. audEERING offers models with early fusion of text and audio.

A further possibility that arises with multimodal data is that it can be used to automatically add uni-modal annotations from models that are trained with more robust modalities. An example would be to train a model predicting valence for speech with data annotated by a video and text-based model.

6 Use Cases

In the following we discuss some fields of application. In all of these audEERING worked successfully with industry partners.

6.1 Case Studies

6.1.1 Liveliness

A call center operator asked audEERING to predict *liveliness* in call center agents' speech, and as can be seen in Figure 5, different steps of (manually annotated) liveliness, correlate well with acoustic biomarkers like mean pitch (melody of the voice). That means that you can use liveliness as a quality metric and predict it automatically with machine learning.

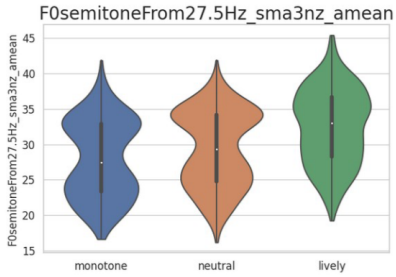

Figure 5. Distribution of mean pitch for three steps of liveliness.

6.1.2 Stress

Within the scope of an EU-project named WorkingAge[3], audEERING had the task to predict *stress* in elderly employees' speech, and as can be seen in Figure 6, different levels of (manually annotated) stress correlate well with predicted arousal level.

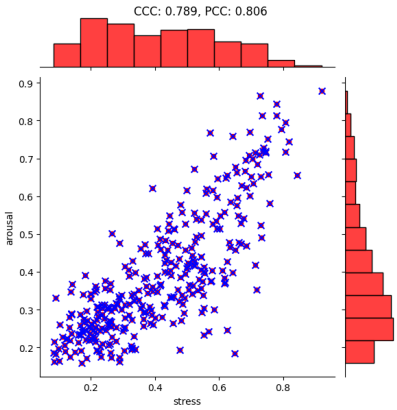

Figure 6. Correlation of predicted arousal with annotated stress with an internal database.

6.1.3 Customer Interest

For a larger German call-center operator, audEERING investigated the correlation of customer's *interest* in a product with predicted *friendliness* in the customer's as well as the agent's speech. In Figure 7 it can be seen that different steps of (manually annotated) interest correlate well with predicted friendliness (tone). Irrespective of the customer who are unsure, this can be used for an automated interest metric based on friendliness.

6.1.4 Intoxication

Alcohol-induced intoxication impairs muscle control and, consequently, alters speech patterns in measurable ways. Leveraging this connection, audEERING developed a voice-based intoxication

[3] https://www.workingage.eu/

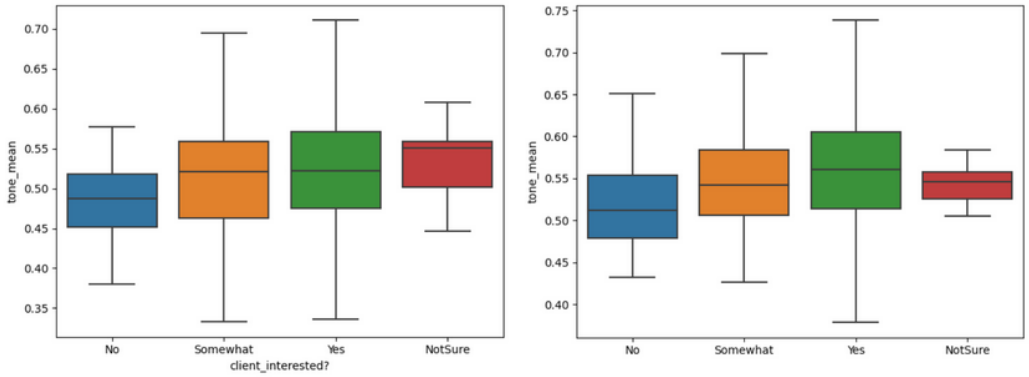

Figure 7. Distribution of predicted friendliness for customers (left) and agents (right) for four levels of customer interest in product.

prediction model with practical applications in safety-critical scenarios — such as alerting impaired drivers before they operate a vehicle, or serving as a non-invasive screening tool in workplace environments where sobriety compliance is required. Figure 8 shows a confusion matrix for the prediction of alcoholic intoxication, illustrating the model's classification performance across intoxicated and sober subjects. For this two-class classification task we achieve UAR values from 0.82 to .9 on our test sets.

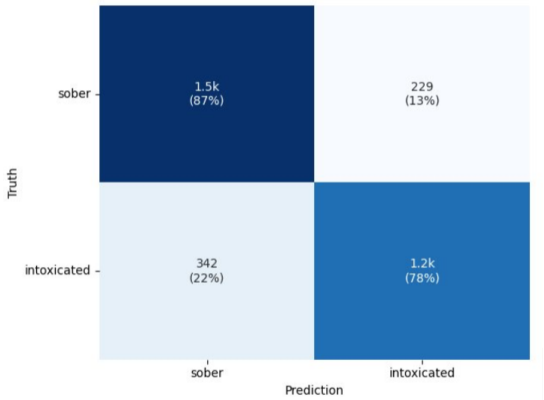

Figure 8. Confusion matrix for the prediction of intoxication (BAC $> .5$ ‰)

6.1.5 Respiratory Health

audEERING developed a respiratory health model that predicts an overall respiratory health score derived solely from voice samples. Such a model has broad practical applications — ranging from remote patient monitoring and early-stage screening tools to integration into consumer devices such as smartphones or wearables, enabling continuous, non-invasive health tracking. Figure 9 shows the score distribution across test subjects with and without respiratory symptoms, illustrating the model's discriminative capability.

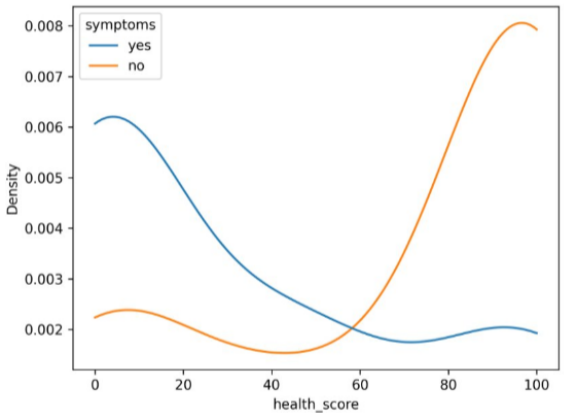

Figure 9. Respiratoiry health score distributions for test subjects with respiratory symptoms and without are separated by a large effect size level.

6.1.6 Dysphonia

In a project focused on voice disorders, we investigate the prediction of GRBAS [22], a five-parameter perceptual scale for clinically rating voice quality by evaluating Grade (overall severity), Roughness, Breathiness, Asthenia, and Strain. Traditionally, GRBAS ratings rely on the subjective judgment of trained clinicians, introducing variability across raters. An automated prediction model offers doctors an objective, reproducible metric to support diagnosis, track disease progression, and monitor treatment response over time — reducing inter-rater variability and enabling scalable assessments beyond the clinical setting. Figure 10 shows the correlations between ground truth values and model predictions for all GRBAS dimensions in our test set. We achieve a mean CCC of 0.813 over all dimensions ranging from 0.772 to 0.871

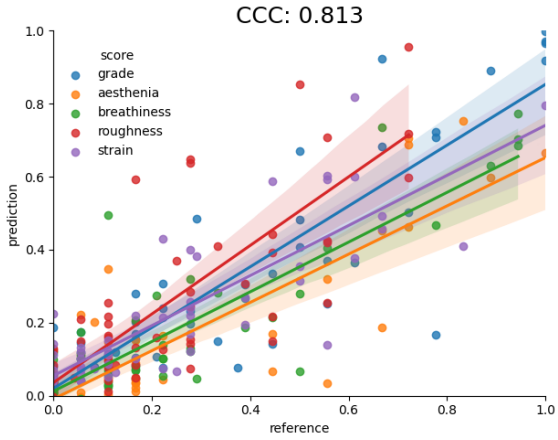

Figure 10. Correlation between ground truth values and model predictions for all GRBAS dimensions.

6.2 Market Research

With respect to market research, expression prediction can be used to infer end customers preferences in an unobtrusive way, simply by recording samples during the interaction of the customer with the product and measure speaker states such as interest or boredom based on expression prediction. The can be correlated for greater stability with results from other modalities such as facial expression or gaze detection, as for example described in the tracking study reported by Merkle [23].

6.3 Healthcare

Physical and mental states that effect speech production can be predicted by acoustic analysis. audEERING worked for example on COVID-19 prediction [24], multiple sclerosis [25], cognitive load [26] or cognitive decline [27]. Some results are mentioned in Section 6.1. Rather than targeting syndromes, audEERING's acoustic biomarker approach focuses on predicting individual symptoms — since a single feature such as hoarseness may stem from a wide range of underlying speaker states.

Parts of the analyses of the above studies are based on the openSMILE feature extractor [28, 29]. The **devAIce**® product [30] comprises several of these openSMILE speech feature sets that are of value for medical speech assessments. In addition, **devAIce**® provides prosodic features related to speech rate, intonation, and loudness.

6.4 Call Center

Call centers are all about communication and automatized expression detection can be used to track customer satisfaction automatically as well as a training tool for agents, as it is known that the tone of voice is more important than body language [31]. audEERING developed together with Jabra a product in this domain: EngageAI [32].

Research on over 1.5 million real-life calls has shown that using Engage AI improves the experience for both customers and agents. The increase in CSAT and call length reduction based on the analysis of the top 10 % of 700,000 Engage AI calls and measured when the agents perform at their best, for example, when their tone is in the top 10 % of their calls.

6.5 Gaming

To react on user emotional expression in a computer game is a key point to make games more entertaining and to enhance the *believability* of non-playable characters, an overview on approaches and methods is given in [33]. audEERING has a product especially designed as an interface for game designers [34].

In our recent study on the integration of Empathic AI in the video game industry, we conducted a comprehensive survey to gauge industry perceptions and potential applications of this emerging technology. The findings were overwhelmingly positive, with over 70% of respondents recognizing significant potential for Empathic AI to revolutionize gaming, particularly in enhancing player immersion and creating novel gameplay experiences (see Figure 11). The concept of emotion-aware NPCs and dynamic e-sports environments garnered strong support, indicating a promising future

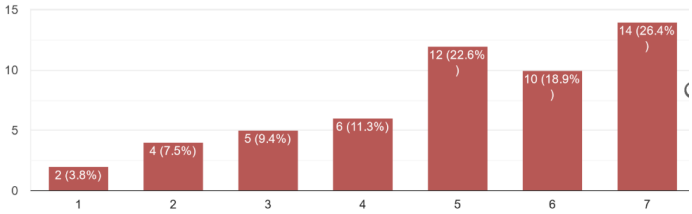

Figure 11. Likert-scale answers from 53 gaming industry participants to the question: 'I find NPCs that understand my emotions more appealing'.

for Empathic AI in driving innovation within the industry. This study underscores the readiness and enthusiasm among industry professionals to embrace Empathic AI as a transformative force in video gaming.

6.6 Conversational AI: Voicebots and Dialogsystems

As mentioned in the introduction, speaker expression is a fundamental building block of human communication and thus also has big potential within human-machine communication. Large neural audio models learn the emotional communication with the data, but so far the vast majority of dialog-systems are based on cascading architectures that integrate automatic speech recognition ASR, dialog-control, data backend and text-to-speech Text-to-speech (TTS) within a whole system. Emotional perception and synthesis, delivered by audEERING, are most helpful in making the interaction more natural and fruitful, when being regarded in the dialog managing [35].

7 Appendix

8 Metric Definitions

8.1 Unweighted Average Recall

UAR is computed as:

$$\text{UAR} = \frac{1}{K}\sum_{k=1}^{K}\frac{\text{true positive}_k}{\text{true positive}_k + \text{false negative}_k}, \tag{1}$$

which calculates the recall for each class $k \in 1, ..., K$ and computes the average over all scores. The benefit of using UAR over classification accuracy:

$$\text{accuracy} = \frac{\text{number of correct predictions}}{\text{number of total predictions}} \tag{2}$$

is that classes that occur less frequently in an evaluation set are given the same weight as classes that occur more often.

8.2 Concordance Correlation Coefficient

CCC is defined as

$$\rho_c = \frac{2\rho\sigma_{\text{prediction}}\sigma_{\text{truth}}}{\sigma^2_{\text{prediction}} + \sigma^2_{\text{truth}} + (\mu_{\text{prediction}} - \mu_{\text{truth}})^2}, \tag{3}$$

where ρ is the Pearson's correlation coefficient (PCC), μ the mean and σ^2 the variance [36]. Intuitively, the CCC measures how high the correlation between truth and prediction is, as well as how well the distribution is matched. For example, the random single human rating in Fig. 3 has a CCC of 0.66.

References

[1] C. Darwin and F. Darwin, *The expression of the emotions in man and animals*, 1872.

[2] C. Gillioz, J. R. Fontaine, C. Soriano, and K. R. Scherer, "Mapping emotion terms into affective space: Further evidence for a four-dimensional structure." *Swiss Journal of Psychology*, vol. 75, no. 3, p. 141, 2016. [Online]. Available: https://www.researchgate.net/profile/Christelle-Gillioz/publication/304184175_Mapping_Emotion_Terms_into_Affective_Space_Further_Evidence_for_a_Four-Dimensional_Structure/links/5770e03b08ae842225aad306/Mapping-Emotion-Terms-into-Affective-Space-Further-Evidence-for-a-Four-Dimensional-Structure.pdf

[3] H. Hoffmann, A. Scheck, T. Schuster, S. Walter, K. Limbrecht, H. C. Traue, and H. Kessler, "Mapping discrete emotions into the dimensional space: An empirical approach," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2012, pp. 3316–3320. [Online]. Available: https://www.researchgate.net/profile/Harald-Traue/publication/234063387_Mapping_discrete_emotions_into_the_dimensional_space_An_empirical_approach/links/545257fa0cf2cf516479c6e2/Mapping-discrete-emotions-into-the-dimensional-space-An-empirical-approach.pdf

[4] V. Sacharin, K. Schlegel, and K. R. Scherer, "Geneva emotion wheel rating study," *Center for Person, Kommunikation, Aalborg University, NCCR Affective Sciences. Aalborg University, Aalborg*, 2012.

[5] P. Ekman and D. Cordaro, "What is meant by calling emotions basic," *Emotion Review*, vol. 3, no. 4, pp. 364–370, 2011.

[6] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Speech emotion recognition with emotion-pair based framework considering emotion distribution information in dimensional emotion space," in *Interspeech 2017*, 2017, pp. 1238–1242.

[7] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, vol. 18, no. 12, pp. 1050–1057, 2007. [Online]. Available: https://doi.org/10.1111/j.1467-9280.2007.02024.x

[8] C. Gillioz, J. R. Fontaine, C. Soriano, and K. R. Scherer, "Mapping emotion terms into affective space," *Swiss Journal of Psychology*, 2016.

[9] H. Hoffmann, A. Scheck, T. Schuster, S. Walter, K. Limbrecht, H. C. Traue, and H. Kessler, "Mapping discrete emotions into the dimensional space: An empirical approach," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2012, pp. 3316–3320.

[10] G. K. Verma and U. S. Tiwary, "Affect representation and recognition in 3d continuous valence–arousal–dominance space," *Multimedia Tools and Applications*, vol. 76, pp. 2159–2183, 2017.

[11] "whisper.cpp," https://github.com/ggml-org/whisper.cpp, accessed: 2026-06-08.

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[13] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[15] 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*, pp. 1–13, 2023.

[16] 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, 2019.

[17] F. Burkhardt, J. Wagner, H. Wierstorf, F. Eyben, and B. W. Schuller, "Speech-based age and gender prediction with transformers," *Speech Communication - 15th ITG Conference*, 2023.

[18] A. Derington, H. Wierstorf, A. Özkil, F. Eyben, F. Burkhardt, and B. W. Schuller, "Testing speech emotion recognition machine learning models," 2023. [Online]. Available: https://arxiv.org/abs/2312.06270

[19] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, 2021.

[20] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[21] J. Russell, J.-A. Bachorowski, and J.-M. Fernandez-Dols, "Facial and vocal expressions of emotion," *Annual Review of Psychology*, vol. 54, pp. 329–349, 11 2003.

[22] M. Hirano, "Clinical examination of voice," *Disorders of Human Communication*, vol. 5, 1981.

[23] "Merkle study on emotion tracking," https://www.merkle.com/dach/en/topics-trends/insights/emotion-tracking, accessed: 2024-06-04.

[24] B. W. Schuller, D. M. Schuller, K. Qian, J. Liu, H. Zheng, and X. Li, "Covid-19 and computer audition: An overview on what speech & sound analysis could contribute in the sars-cov-2 corona crisis," *Frontiers in Digital Health*, vol. 3, 2021. [Online]. Available: https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2021.564906

[25] M. Gonzalez-Machorro, P. Hecker, U. D. Reichel, H. N. Hammer, R. Hoepner, L. Pedrotti, A. Zmutt, H. Sagha, J. van Beek, F. Eyben, D. M. Schuller, B. W. Schuller, and B. Arnrich, "Towards Supporting an Early Diagnosis of Multiple Sclerosis using Vocal Features," in *Proc. INTERSPEECH 2023*, 2023, pp. 1518–1522.

[26] P. Hecker, A. M. Kappattanavar, M. Schmitt, S. Moontaha, J. Wagner, F. Eyben, B. W. Schuller, and B. Arnrich, "Quantifying cognitive load from voice using transformer-based models and a cross-dataset evaluation," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, dec 2022, pp. 337–344.

[27] S. Kalabakov, M. Gonzalez-Machorro, F. Eyben, B. W. Schuller, and B. Arnrich, "A comparative analysis of federated learning for speech-based cognitive decline detection," in *Interspeech 2024*, 2024, pp. 2455–2459.

[28] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[29] 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.

[30] "audeering devaice," https://www.audeering.com/products/devaice/, accessed: 2024-06-04.

[31] M. W. Kraus, "Voice-only communication enhances empathic accuracy," *American Psychologist*, vol. 72.

[32] "Jabra engageai," https://www.jabra.com/software-and-services/jabra-engage-ai, accessed: 2024-06-04.

[33] G. Yannakakis and D. Melhárt, "Affective game computing: A survey," *Proceedings of the IEEE*, vol. PP, pp. 1–22, 10 2023.

[34] "devaice xr," https://www.audeering.com/de/products_alt/entertain-play/, accessed: 2024-06-04.

[35] S. Liu, C. Zheng, S. Sabour, Y. Li, Z. Yu, Y. Jiang, and M. Huang, "Towards emotional support dialog systems," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 3469–3483. [Online]. Available: https://aclanthology.org/2021.acl-long.269

[36] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, pp. 255–268, 1989.

Version History

Version	Date	Responsible	devAIce Version	Changes
V1.2	08.06.2026	Anna Derington, Felix Burkhardt Uwe Reichel	devAIce 3.14.0-current	Added details on expression categories, updated stress case study, added ASR section, updated use case section; Intoxication, Respiratory Health: Added performance details; Dysphonia: Updated results
V1.1	15.05.2025	Anna Derington	devAIce 3.14.0-current	Fixed contributors
V1	13.03.2025	Anna Derington	devAIce 3.14.0-current	Initial Version
V0.1	15.05.2025	Anna Derington	devAIce 3.9.1-3.13.0	Fixed contributors
V0	23.04.2025	Anna Derington, Felix Burkhardt	devAIce 3.9.1-3.13.0	Initial Version for devAIce 3.9.1-3.13.0