Voice analysis in FaceReader

Ruben Seggers
Machine Learning Engineer, Vicarious Perception Technologies BV

A white paper by Noldus Information Technology

WHAT IS VOICE ANALYSIS?

Voice analysis is a powerful tool for understanding human emotions, offering a new dimension to affective computing. So, how does it work?

Speech Emotion Recognition (SER) refers to the process of identifying and classifying emotional expressions based on vocal characteristics, using machine learning models trained on diverse datasets. By analyzing features such as pitch, loudness, intonation, and speech rate, SER can estimate the expressed emotional state of a speaker, independent of the actual words spoken.

SER has a wide range of research applications, including human-computer interaction, psychological research, and human factors studies. It enables the investigation of emotional responses in social interactions, cognitive load during task performance, and user experience in conversational systems.

By integrating voice analysis with existing affective computing technologies, researchers can gain a more comprehensive understanding of human emotions.

This multimodal approach allows for a more holistic assessment of affective states by combining facial and vocal expressions.

VOICE ANALYSIS IN FACEREADER

FaceReader, which already provides facial expression analysis, eye tracking, heart rate, and breathing rate estimation, now also supports voice-based emotion recognition.

This multimodal approach allows for a more holistic assessment of affective states by combining facial and vocal expressions. With this enhancement, FaceReader can better capture subtle emotional cues, making it an even more powerful tool for emotion research and applied behavioral studies.

This white paper discusses the methodology and expected performance of FaceReader's voice analysis techniques, including the assessment of potential bias and practical guidelines for optimal results.

HOW DOES IT WORK?

> Vocal features such as pitch, loudness, and speech rate can only be measured over time.

Unlike video, where each frame can be analyzed independently, audio-based emotion recognition requires a time window to extract patterns. A single moment of sound lacks information to determine emotional state, as vocal features such as pitch, loudness, and speech rate can only be measured over time.

To account for this, each video frame is analyzed alongside the preceding second of audio, meaning that voice analysis begins only after enough audio has been processed to detect a signal.

DETECTING VOICE ACTIVITY

Voice analysis in FaceReader employs Voice Activity Detection (VAD) based on a Gaussian Mixture Model (GMM) [1] to filter out audio that clearly falls outside the typical range of human speech. Although it does not explicitly classify segments as speech or non-speech, this approach helps reduce the impact of background noise and other irrelevant sounds.

While GMM-based VAD is generally reliable under typical conditions, it is not immune to misclassifications, particularly in environments with overlapping speech, high levels of ambient noise, or non-speech sounds within the speech frequency range. When a valid voice signal is detected, the following vocal measures will be calculated and reported by FaceReader.

EMOTIONS

Emotions are estimated based on a deep learning model that analyzes a broad set of vocal and audio features. These features are extracted from diverse datasets covering multiple languages and include both natural and acted speech to reflect a wide range of expressive variations. The model accounts for differences in gender, ethnicity, and speaking style, aiming to provide a robust estimation of the expressed emotional state.

The model focuses on prosodic features such as pitch, loudness, speech rate, and intonation, rather than the semantic content of speech. This approach is less dependent on the spoken language, as it emphasizes how something is said rather than the specific words used. However, because the model was trained primarily on English-language data, its performance is highest for English speakers.

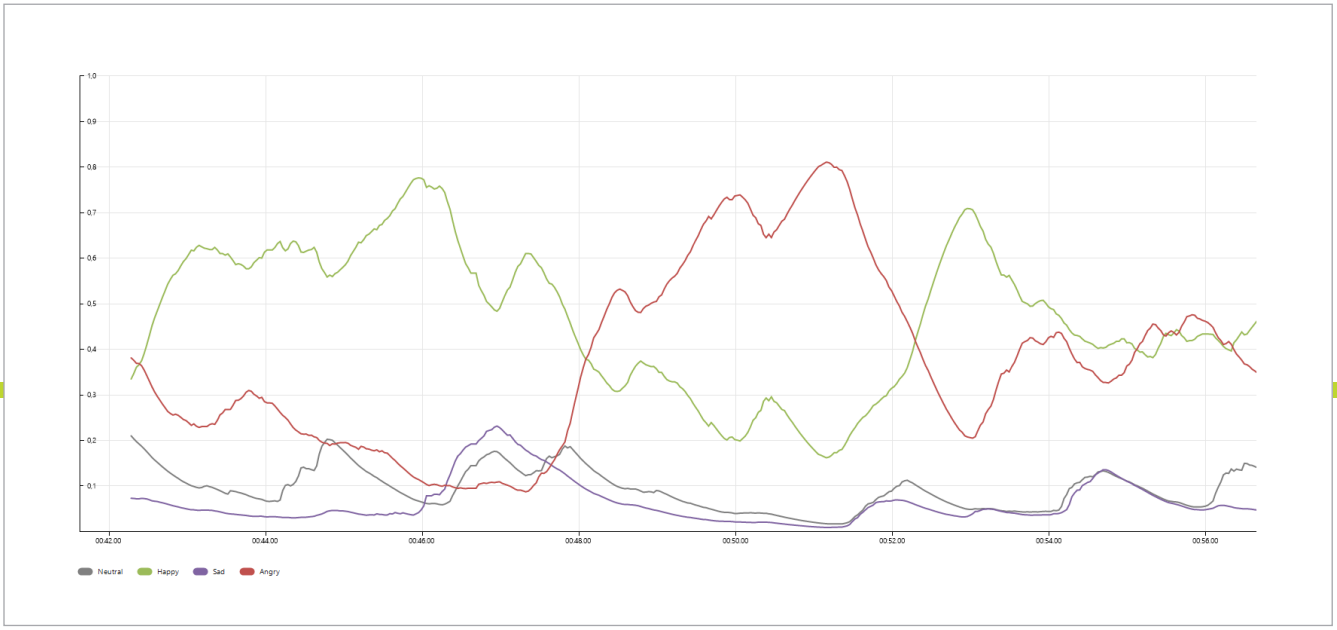

Figure 1. *Voice-based emotion predictions over time (neutral, happy, sad, angry), shown in a FaceReader line chart.*

VALENCE

Valence represents the emotional positivity or negativity of speech and is derived from speech emotion recognition.

Positive emotions, such as happiness, contribute to a higher valence score. Negative emotions, including anger and sadness, result in a lower valence score.

Valence is normalized to a scale from –1 to 1, where –1 indicates strong negative affect, 0 is neutral, and 1 reflects strong positive affect.

> Higher loudness and faster speech are typically associated with excitement or stress, while lower loudness and slower speech indicate calmness or disengagement.

AROUSAL

Arousal is estimated using both loudness and speech rate, which are also included as individual vocal measures in FaceReader (see definitions below).

Higher loudness and faster speech are typically associated with excitement or stress, while lower loudness and slower speech indicate calmness or disengagement.

Arousal is normalized to a 0 to 1 scale, with 0.5 being average arousal. When interpreted alongside valence, arousal helps differentiate between emotions that share a similar emotional direction but vary in intensity. For example, sadness (low arousal) versus anger (high arousal), both of which have negative valence.

LOUDNESS

Loudness is a custom volume-independent measure of sound intensity. Variations in microphone sensitivity and speaker positioning can significantly affect the absolute volume of a recording.

To address this, the loudness measure is computed using a continuously updated, normalized audio signal. For each frame, the system analyzes the

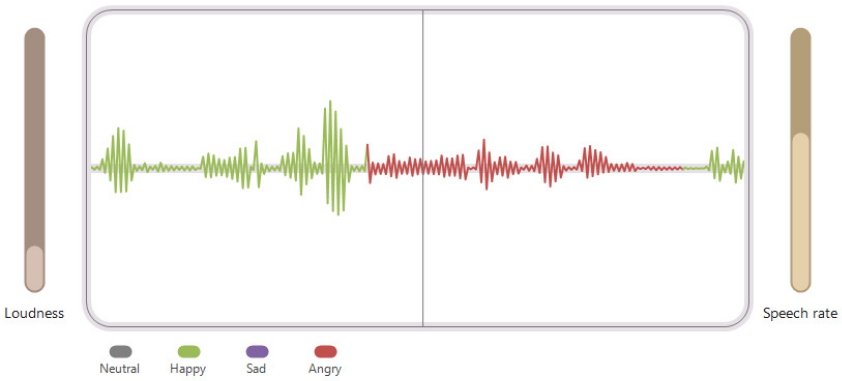

Figure 2. *Voice view in FaceReader, showing waveform segments labeled by emotion (neutral, happy, sad, angry), alongside loudness and speech rate indicators.*

past second of the audio signal. Within this one-second window, the audio is normalized to a maximum absolute amplitude of 1.0.

This ensures that the measurement reflects the speaker's vocal dynamics. So how energetically a person is speaking within their own dynamic range, rather than the recording conditions.

> The algorithm detects peaks in the audio signal, typically corresponding to syllables, by identifying bursts of vocal energy.

SPEECH RATE

The speech rate measure estimates how quickly a person is speaking. The algorithm detects peaks in the audio signal, typically corresponding to syllables, by identifying bursts of vocal energy.

These peaks are counted over the past one-second window of audio, using an upper bound of 160 syllables per minute, which is suitable for capturing various emotional states [2,3].

The count is normalized to a value between 0 and 1. A higher value indicates faster speech, while a lower value reflects slower or more deliberate speaking.

CALIBRATION

Voice data is not automatically calibrated during acquisition. However, in the Project Analysis window, you can apply baseline correction to your voice data.

This works by normalizing each voice feature based on its average across the selected samples (e.g., within a subject, session or stimulus condition, depending on the analysis context).

The result is a mean-centered dataset where inter-subject or inter-session comparisons are more meaningful. This approach is particularly useful in group analyses where absolute voice feature values are less informative than relative changes or patterns.

HOW WELL DOES FACEREADER ANALYZE VOICE?

Table 1 summarizes the performance of FaceReader's voice analysis on a diverse English-language test set. This test set was constructed using both publicly available and proprietary datasets. It was also explicitly excluded from training to ensure unbiased evaluation.

Table 1. *Overall performance metrics on the English-language test set.*

Accuracy	87.1	Percentage of predictions that were correct across all emotion classes.
Precision	87.2	Measures how many of the predicted emotions were correct (focuses on minimizing false positives).
Recall	87.1	Measures how many of the emotions were correctly identified (focuses on minimizing false negatives).
F1 score	87.1	Harmonic mean of precision and recall; a balanced measure of model accuracy.
UAR	87.3	Unweighted Average Recall: average recall per class, giving equal weight to each emotion regardless of how often it occurs.

The test set includes a broad spectrum of recordings, covering both acted and spontaneous expressions of emotion. The dataset features over 4,000 recordings from speakers of varying gender, ethnicity, and dialects to ensure robustness across demographic variation.

Emotion labels were either predefined (in the case of scripted, acted speech) or obtained through crowd-sourced or expert annotation (for both spontaneous and acted speech). We used inter-annotator agreement to assess and improve labeling reliability.

Table 2 breaks down performance by emotion category, highlighting consistently strong and balanced results across all four target emotions.

Table 2. *Classification performance per emotion category on the test set.*

Label	Precision	Recall	F1 score
Neutral	90.2	86.7	88.4
Happy	83.2	84.8	84.0
Sad	84.4	91.9	88.0
Angry	90.7	85.9	-88.2

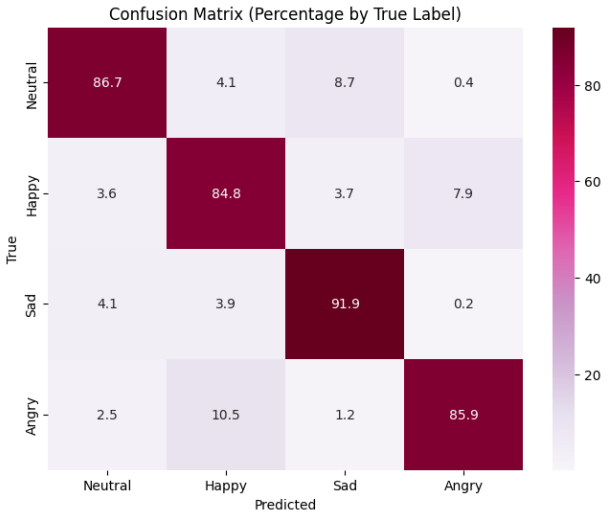

Figure 3. *Confusion matrix showing voice-based emotion classification accuracy (neutral, happy, sad, angry) based on true labels.*

MINIMIZING BIAS IN VOICE ANALYSIS

> Our datasets are carefully balanced to include a representative mix of male and female voices across a range of speaking styles and dialects.

To minimize gender bias in FaceReader's voice analysis, we take deliberate steps throughout data collection, training, and evaluation. Our datasets are carefully balanced to include a representative mix of male and female voices across a range of speaking styles and dialects.

We also evaluate model performance separately by gender to identify and address any disparities. Our current model shows a performance difference of 2.7% in accuracy and 3.9% in unweighted average recall (UAR), with slightly better results for female speakers.

The training data primarily consists of English-speaking adult voices; voices from children or elderly individuals are currently underrepresented. Performance will be lower for other languages, especially those that are linguistically distant from English.

We continuously monitor and refine our approach to support fairness and inclusivity, while improving generalization across diverse linguistic contexts.

PRACTICAL TIPS FOR VOICE ANALYSIS IN YOUR RESEARCH

The following recommendations can help you get high-quality input and reliable results when using FaceReader's voice analysisin your research.

SETUP

- *Use a high-quality microphone* in a quiet environment: voice analysis only works on speech and background noise can negatively affect results.
- *Ensure the recording level is appropriate*: neither too low (won't detect voice), nor too high (to prevent clipping/distortion).
- *Avoid overlapping speech*: voice analysis works best when one person speaks at a time.

RESEARCH

- *Try voice analysis on sample data* to make sure your setup is clean and the results are stable.
- *Results are relative,* so compare values across time or across speakers rather than relying on a single absolute number.
- *Use in combination with facial expressions* and the measurement of vital signs for more complete emotional insights.

Get the best for your research

Curious what more FaceReader can do? Read more on the website or contact us to discuss how FaceReader fits your research goals.

Learn more about:

- Applications in UX, consumer, and psychology research
- How to get objective data in no time
- Experiences from happy customers

LEARN MORE

CONTACT US

REFERENCES

1. Reynolds, D. (2009). Gaussian Mixture Models. In: Li, S.Z., Jain, A. (eds) *Encyclopedia of Biometrics*. Springer, Boston, MA.
2. Arnfield, S., Roach, P., Setter, J., Greasley, P., Horton, D. (1995) Emotional stress and speech tempo variation. Proc. *ESCA/NATO Workshop on Speech under Stress*, 13-15
3. Braun, Angelika & Oba, Reiko. (2007). *Speaking Tempo in Emotional Speech - a Cross-Cultural Study Using Dubbed Speech*.

INTERNATIONAL HEADQUARTERS
Noldus Information Technology bv
Wageningen, The Netherlands
Phone: +31-317-473300
E-mail: contact@noldus.com

NORTH AMERICAN HEADQUARTERS
Noldus Information Technology Inc.
Leesburg, VA, USA
Phone: +1-703-771-0440
Toll-free: 1-800-355-9541
E-mail: info@noldus.com

REPRESENTATION
We are represented by a worldwide network of distributors and regional offices. Visit our website for contact information.

NOLDUS.COM