# voice: A Comprehensive R Package for Audio Analysis

Filipe Jaeger Zabala ©[1*] and Giovanni Abrahão Salum ©[1,2*]

1 Graduate Program of Psychiatry and Behavioral Sciences, UFRGS, Brazil ROR **2** Child Mind Institute, New York, NY 10022, USA ROR ***** These authors contributed equally.

Summary

The `voice` package (Zabala, 2025) for R (R Core Team, 2024) is a free, open-source toolkit designed to streamline audio analysis by integrating music theory and advanced computational techniques. It enables researchers to extract, summarize, and analyze voice data efficiently, supporting applications such as speech recognition, speaker identification, and mood inference. The package simplifies workflows through three core functions: `extract_features`, `tag`, and `diarize`. By bridging gaps in existing R tools, `voice` offers a unified solution for audio data analysis.

Statement of Need

Tools like `reticulate` (Ushey et al., 2023) and `rpy2` (Gautier, 2025) enable interoperability between R and Python, allowing users to leverage external libraries for audio processing. While R provides foundational packages like tuneR and tuneR (Ligges et al., 2023), seewave (Sueur et al., 2008) and wrassp (Winkelmann et al., 2024), Python's ecosystem (e.g., `Librosa` (McFee et al., 2015), pyannote-audio (Bredin et al., 2019)) is more extensive, as evidenced by the Awesome Python Audio and Music collection by Andrei Matveyeu. However, R's state-of-the-art time-series infrastructure (CRAN Task View: Time Series Analysis), offers unique advantages for analyzing audio signals as temporal data. Our implementation combines these strengths, providing an R-centric workflow with optional Python integration for specialized tasks.

`voice` was designed with a user-friendly approach that makes it accessible to researchers in linguistics, psychology, and bioacoustics, where audio data remains underutilized. There are currently work fronts in these areas making use of `voice` functionalities. By simplifying the extraction and analysis of audio features, the package lowers the barrier to entry for researchers and expands the potential for audio data in scientific studies.

Features

Core Functions

1. **`extract_features`**:
   Extracts standardized audio features from files (e.g., *F0*, *Formant Dispersion*, *Gain*, *MFCC*), leveraging `wrassp` and `tuneR` while introducing new metrics to capture vocal tract characteristics.

2. **`tag`**:
   Attaches summarized audio features to datasets, supporting anonymization and privacy-aware analysis via a *6-number summary* (mean, median, standard deviation, coefficient of variation, interquartile range and median absolute deviation).

3. **`diarize`**:
   Identifies speaker segments using Python's `pyannote-audio` (Bredin et al., 2019), generating RTTM files for transcription and analysis.

Novel Contributions

- **Formant Removals**:
  Isolates fundamental frequency (F0) from formants, improving feature interpretability for classification tasks.

- **Integration of R and Python**:
  Uses `reticulate` (Ushey et al., 2023) to combine R's statistical power with Python's tools.

Example Applications

Predicting Sex from Voice

The package was tested on open datasets (AESDD (Vryzas, Kotsakis, et al., 2018; Vryzas, Matsiola, et al., 2018), CREMA-D (Cao et al., 2014), Mozilla Common Voice (Ardila et al., 2019), RAVDESS (Livingstone & Russo, 2018) and VoxForge (VoxForge, 2023)) to predict sex from voice features. Results showed high accuracy across multiple model classes (Binary Logistic (Cramer, 2002), SVM (Vapnik, 2000), Random Forest (Breiman, 2001), and BART (Sparapani et al., 2021)), with formant removals ranking among the top predictive features.

Speaker Diarization

The `diarize` function has been used successfully, and as a didactic example was applied to a LibriVox recording of *The Adventures of Sherlock Holmes* by Conan Doyle, successfully segmenting the audio into speaker turns. This demonstrates the package's utility for applications in transcription and audio analysis.

Performance

The `voice` package efficiently processes audio files, with `extract_features` allowing parallelization and generating feature-rich data frames in seconds. The `diarize` function, while computationally intensive for long recordings, provides accurate segmentation and integrates seamlessly with R workflows.

Availability

The `voice` package is available on CRAN (https://CRAN.R-project.org/package=voice) and GitHub (https://github.com/filipezabala/voice). Documentation, including vignettes and examples, is provided to facilitate adoption.

Acknowledgments

References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv Preprint arXiv:1912.06670*. https://doi.org/10.48550/arXiv.1912.06670

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., & Gill, M.-P. (2019). *Pyannote.audio: Neural building blocks for speaker diarization*. https://doi.org/10.48550/arXiv.1911.01255

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, *5*(4), 377–390. https://doi.org/10.1109/TAFFC.2014.2336244

Cramer, J. S. (2002). *The Origins of Logistic Regression*. https://doi.org/10.2139/ssrn.360300

Gautier, L. (2025). *rpy2* (Version 3.6.1). https://pypi.org/project/rpy2/

Ligges, U., Krey, S., Mersmann, O., & Schnackenberg, S. (2023). *tuneR: Analysis of music and speech*. https://doi.org/10.32614/CRAN.package.tuneR

Livingstone, S. R., & 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. https://doi.org/10.1371/journal.pone.0196391

Mao, R. (2025). *Gm: Create music with ease*. https://doi.org/10.32614/cran.package.gm

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). Librosa: Audio and music signal analysis in python. *Proceedings of the 14th Python in Science Conference*, *8*. https://doi.org/10.25080/Majora-7b98e3ed-003

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://doi.org/10.1080/10618600.2000.10474900

Sparapani, R., Spanbauer, C., & McCulloch, R. (2021). Nonparametric machine learning and efficient computation with bayesian additive regression trees: The BART r package. *Journal of Statistical Software*, *97*(1), 1–66. https://doi.org/10.18637/jss.v097.i01

Sueur, J., Aubin, T., & Simonis, C. (2008). Seewave: A free modular tool for sound analysis and synthesis. *Bioacoustics*, *18*, 213–226. https://doi.org/10.1080/09524622.2008.9753600

Ushey, K., Allaire, J., & Tang, Y. (2023). *Reticulate: Interface to 'python'*. https://doi.org/10.32614/CRAN.package.reticulate

Vapnik, V. (2000). *The Nature of Statistical Learning Theory* (2nd ed.). Springer Science & Business Media. https://doi.org/10.1007/978-1-4757-3264-1

VoxForge. (2023). *VoxForge: An open speech dataset set up to collect transcribed speech*. http://www.voxforge.org/

Vryzas, N., Kotsakis, R., Liatsou, A., Dimoulas, C. A., & Kalliris, G. (2018). Speech emotion recognition for performance interaction. *Journal of the Audio Engineering Society*, *66*(6), 457–467. https://doi.org/10.17743/jaes.2018.0036

Vryzas, N., Matsiola, M., Kotsakis, R., Dimoulas, C., & Kalliris, G. (2018). Subjective evaluation of a speech emotion recognition interaction framework. In *Proceedings of the audio mostly 2018 on sound in immersion and emotion* (pp. 1–7). https://doi.org/10.1145/3243274.3243294

Winkelmann, R., Bombien, L., Scheffers, M., & Jochim, M. (2024). *Wrassp: Interface to the 'ASSP' library*. https://doi.org/10.32614/CRAN.package.wrassp

Zabala, F. J. (2025). *Voice: Voice analysis, speaker recognition and mood inference via music theory*. https://cran.r-project.org/package=voice