voice: A Comprehensive R Package for Audio Analysis

05 May 2025

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, tag, 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 tools like `wrassp` (Winkelmann et al. 2024) and `tuneR` (Ligges et al. 2023), `voice` offers a unified solution for audio data analysis.

Statement of Need

Audio data analysis is complex due to variability in file formats and the lack of integrated tools. While packages like `seewave` (Sueur, Aubin, and Simonis 2008) provide foundational capabilities, they often require specialized knowledge. The `voice` package addresses these challenges by combining existing functionalities with novel features such as *Formant Removals*, which enhance predictive accuracy for tasks like sex classification.

Its user-friendly design 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, `voice` 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, Allaire, and Tang 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 and 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, Spanbauer, and McCulloch 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 for typical audio lengths. 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

The authors gratefully acknowledge Renfei Mao for their technical support and guidance in implementing the gm library (Mao 2025).

References

Ardila, Rosana, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. "Common Voice: A Massively-Multilingual Speech Corpus." *arXiv Preprint arXiv:1912.06670*. https://arxiv.org/abs/1912.06670.

Bredin, Hervé, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2019. "Pyannote.audio: Neural Building Blocks for Speaker Diarization." https://arxiv.org/abs/1911.01255v1.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf.

Cao, Houwei, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset." *IEEE Transactions on Affective Computing* 5 (4): 377–90. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4313618/.

Cramer, Jan Salomon. 2002. "The Origins of Logistic Regression." http://hdl.handle.net/10419/86100.

Ligges, Uwe, Sebastian Krey, Olaf Mersmann, and Sarah Schnackenberg. 2023. *tuneR: Analysis of Music and Speech*. https://CRAN.R-project.org/package=tuneR.

Livingstone, Steven R., and Frank A. Russo. 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, Renfei. 2025. *Gm: Create Music with Ease*. https://github.com/flujoo/gm.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sparapani, Rodney, Charles Spanbauer, and Robert McCulloch. 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://www.jstatsoft.org/article/view/v097i01.

Sueur, J., T. Aubin, and C. Simonis. 2008. "Seewave: A Free Modular Tool for

Sound Analysis and Synthesis." *Bioacoustics* 18: 213–26. https://doi.org/10.1080/09524622.2008.9753600.

Ushey, Kevin, JJ Allaire, and Yuan Tang. 2023. *Reticulate: Interface to 'Python'*. https://CRAN.R-project.org/package=reticulate.

Vapnik, Vladimir. 2000. *The Nature of Statistical Learning Theory*. 2nd ed. Springer Science & Business Media. https://link.springer.com/book/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, Nikolaos, Rigas Kotsakis, Aikaterini Liatsou, Charalampos A Dimoulas, and George Kalliris. 2018. "Speech Emotion Recognition for Performance Interaction." *Journal of the Audio Engineering Society* 66 (6): 457–67. https://www.researchgate.net/profile/Nikolaos-Vryzas/publication/326005164_Speech_Emotion_Recognition_for_Performance_Interaction/links/5b97e33b299bf14ad4ce3ee5/Speech-Emotion-Recognition-for-Performance-Interaction.pdf.

Vryzas, Nikolaos, Maria Matsiola, Rigas Kotsakis, Charalampos Dimoulas, and George Kalliris. 2018. "Subjective Evaluation of a Speech Emotion Recognition Interaction Framework." In *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, 1–7. https://dl.acm.org/doi/pdf/10.1145/3243274.3243294.

Winkelmann, Raphael, Lasse Bombien, Michel Scheffers, and Markus Jochim. 2024. *Wrassp: Interface to the 'ASSP' Library*. https://CRAN.R-project.org/package=wrassp.

Zabala, Filipe J. 2025. *Voice: Voice Analysis, Speaker Recognition and Mood Inference via Music Theory*. https://github.com/filipezabala/voice.