

Listen Closely: Measuring Vocal Tone in Corporate Disclosures

Jonas Ewertz
Ruhr University Bochum
jonas.ewertz@rub.de

Charlotte Knickrehm
Goethe University Frankfurt
knickrehm@econ.uni-frankfurt.de

Martin Nienhaus
Ruhr University Bochum
martin.nienhaus@rub.de

Doron Reichmann
Goethe University Frankfurt
d.reichmann@econ.uni-frankfurt.de

Compliance with the Journal of Accounting Research data policy

August 2025

Item 1. Data Handling and Analyses

FinVoc2Vec was developed by Jonas Ewertz, Charlotte Knickrehm, and Doron Reichmann. Data handling and analyses were conducted by Jonas Ewertz, Charlotte Knickrehm, Martin Nienhaus, and Doron Reichmann.

Item 2. Generation of the Data

See the article for a detailed description of the procedures used to generate the data. Conference call audio data was manually collected from January 2022 to March 2022 via the Refinitiv Corporate Event Calendar (now LSEG). The data was then matched with Compustat, Execucomp, CRSP, and I/B/E/S.

Item 3. Proprietary Nature of Data

The data for this study were obtained from commercial data providers, including Refinitiv (now LSEG), Compustat, CRSP, and I/B/E/S. While the data are not proprietary to this research, they are subject to licensing agreements and intellectual property rights held by the providers. As such, we are unable to share the raw data directly. However, details about the data sources and how they were used in the analysis are provided in the article.

Item 4. Steps Necessary to Collect and Process Data

Refer to Section 2 of the article for a detailed description of the steps necessary to collect and process the data.

Item 5. Data Manipulations

All data processing and analyses were conducted using Python.

Item 6. Code Used to Conduct Primary Analyses

The paper is accompanied by two comprehensive online code repositories. The first repository contains the source code of an openly available package solution, ‘ccalign,’ developed in this paper. The repository contains the code used to convert the raw conference call transcripts and audio data to timestamped transcripts, indicating start and end times of each sentence spoken during the conference call. The second repository releases FinVoc2Vec, the model used to derive our main measures of vocal tone. The repository further describes the process of loading the model and applying the model to audio files.

The repositories are available here:

ccalign: <https://github.com/j-ewertz/ccalign>

FinVoc2Vec: <https://huggingface.co/waiv/FinVoc2Vec>

To reproduce tables, we provide the following code and data at <https://www.chicagobooth.edu/jar-online-supplements>:

- Identifiers.parquet contains firm and conference call identifiers. “tic” is the firm ticker retrieved from Refinitiv. “callid” is a conference call identifier from the Refinitiv Corporate Event Calendar. “gvkey” is a Compustat firm identifier.
- create_sample.py is Python code that reproduces the sample used to conduct analyses in Tables 4 to 9 of the paper.
- Tab2_ModelEval.py is Python code that reproduces the analyses for Table 2 of the paper.
- Tab3_InnerWorking.py is Python code that reproduces the analyses for Table 3 of the paper.
- Tab4_9_MarketOutcomes.py is Python code that reproduces the analyses for Tables 4 to 9 of the paper.
- reg_functions.py is Python code that contains supporting functions for statistical analyses and table production.

Item 7. Log File of Code Execution

The audio processing described in the online repositories was carried out on external servers starting in 2023 using the Google Cloud Platform, and required several months to complete. The file code_execution_log.log contains a comprehensive log of the code execution used to reproduce the tables in the paper.

Item 8. Maintenance of Data and Programs

The authors assure that the data and programs will be maintained for at least six years, consistent with National Science Foundation guidelines.