Harvesting spontaneous speech data from digital reservoirs to study prosody

Aviad Albert[1], Constantijn Kaland[1], T. Mark Ellison[1], Francesco Cangemi[2], Bodo Winter[3], and Martine Grice[1]

[1]University of Cologne, [2]Tokyo University of Foreign Studies, [3]University of Birmingham

This paper presents a proof-of-concept workflow that collects and analyzes speech corpora from public online databases. The complete workflow (code and data) is open-source, available in a dedicated OSF repository [1], which is an integral part of this submission. The following description provides a general overview of the online content. We demonstrate our workflow by collecting different renditions of the lexical string "can I ask you a question?" from the TV News Archive [2] (a public online database of U.S. television news broadcasts, searchable by captions; see also [3]). We analyze the varying intonation patterns of this fixed lexical string to closely observe intonation patterns in the wild. To achieve that in a practical way, we use the *ProPer* toolbox [4] to derive acoustic metrics that reflect the pitch trajectory over syllabic intervals. Finally, we submit the aggregated ProPer values to a cluster analysis system [5] that splits the data into groups of similar items. By sharing this workflow, we hope to contribute to a developing knowledge base devoted to methodologies advancing the ecologically valid study of phonology using speech corpora (see [6] for a brief discussion on the topic).

Data harvesting. We start by saving as a local html file a fully populated TV News Archive search results page obtained using a standard web browser. We then run the first of many R [7] scripts in this workflow. The *Data harvesting* component in the OSF repository includes the html file we used, the R file that we ran on it, *DownloadingData*, and the resulting table, *Raw_harvest_df*, that features 3,935 rows/results. The table was designed to be very rich in information, including multiple links to each video, the entire caption text, timing of the target phrase, and more. The *Raw harvest* table was edited to screen out the results that were deemed as "invalid" given a set of criteria (see *Criteria for valid results* in the same directory). We conducted this validation process until we reached 250 valid cases to start our analysis. The resulting annotated table, *Edited_harvest*, was used by the two ensuing R files to download the 1-minute video clips (*Download Videos*) and automatically trim the size and save a corresponding audio file (*Trim Clips & Make Audio*). The results of this data harvest (253 video clips and corresponding audio files), are provided in a separate directory in the *Data harvesting* component – *Initial set of valid results*. For the ensuing acoustic analysis, we annotated the syllabic intervals of the target phrase using Praat [8] TextGrids. The list of annotators' comments, as well as the 190 audio files and corresponding TextGrid files that were chosen for analysis, are provided in a separate directory – *Selected items for analysis*.

Analyses. We used the ProPer toolbox for our acoustic analysis. ProPer uses measurements of periodic energy and F0 to provide visualization and quantification of prosodic information. The entire ProPer workflow for this paper is provided in the *Analyses* component as an ordered set of Praat and R scripts. We use ProPer to produce rich and informative F0 visualizations (*Periograms*) and to derive values for the two ProPer metrics that quantify aspects of the F0 contour – *Synchrony* and Δ*F0* (see Fig. 1). After further technical exclusions, the ProPer metrics were computed for 165 items, which were then submitted to cluster analysis. Synchrony and ΔF0 values were submitted to a multivariate complete linkage hierarchical cluster analysis [9]. This process resulted in further exclusions, ending up with 125 items. We used a modified version of the *Contour clustering* tool [5], where instead of analyzing data points along the F0 contour, the values were analyzed as time-series data with length 6 (one data point per syllabic interval for each measure). The separately generated and scaled distance matrices of each measure were summed, resulting in a single final matrix (giving equal weight to the two measures). Several rounds of clustering were carried out ranging from 2 to 15 clusters. An assessment of the within- and between-cluster variance was done for each round and showed that with 8 clusters, both types of variance reached an optimal divergence, i.e. high between-cluster variance for a low within-cluster variance, without having to assume more clusters. The results of the round with 8 clusters are shown in Fig. 1D. As we hope to have shown here, combining cluster analysis with prosodic metrics can bootstrap further perceptual exploration of prosodic categories (see [10] for a study of nuclear tunes).

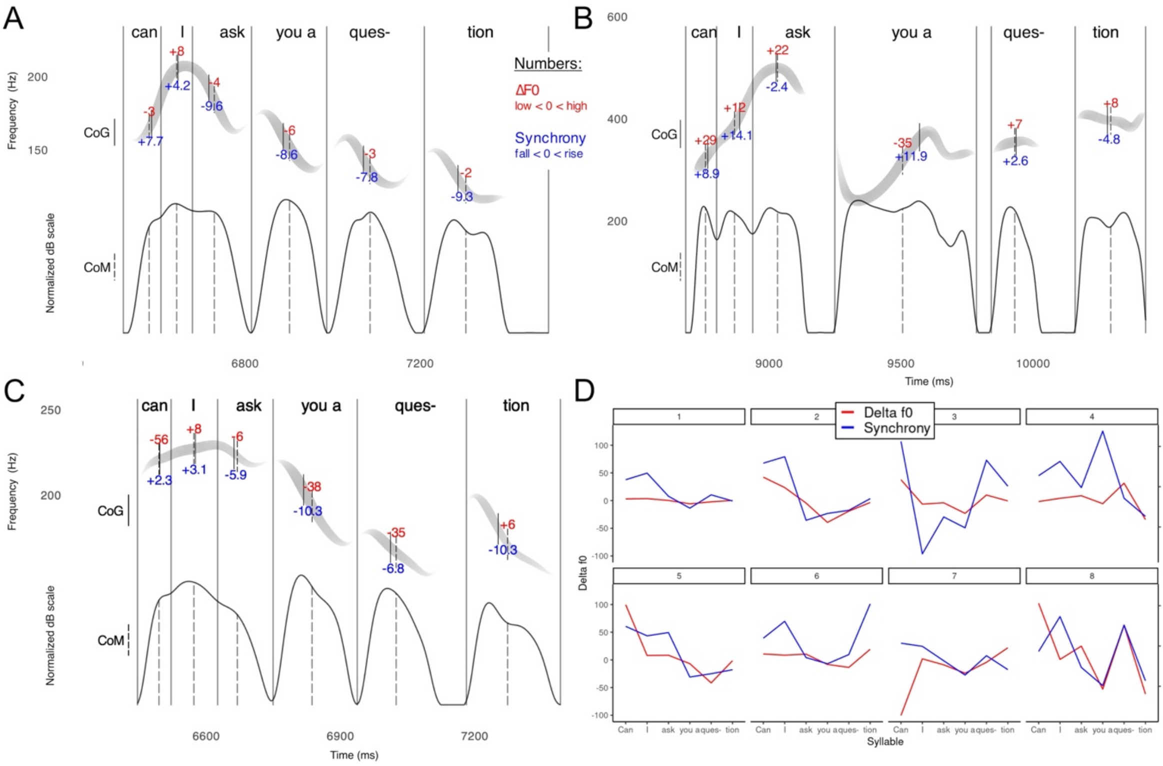

Fig. 1. Individual examples of Periograms (A–C) and a graph summary of the cluster analysis over aggregated results from 125 tokens (D). The Periograms show the F0 curve in grey in the upper half of the panel. The strength of the F0 signal is modulated by the periodic energy curve in the bottom half (controlling the thickness and darkness of the F0 curve above it). All plots demonstrate the ΔF0 metric in red and the Synchrony metric in blue. Each syllabic interval includes the Center of Mass of the periodic energy curve (CoM) and the Center of Gravity of the F0 curve (CoG), note 'you a' is treated as one interval. These landmarks are used to compute ΔF0 (change in F0 between CoMs across syllabic intervals) and Synchrony (distance between CoM and CoG within syllabic intervals). Periograms for representative examples from clusters 1, 4 and 7 are given in panels A, B and C respectively. The most prototypical items from each cluster are provided (audio, TextGrid and Periograms) in the OSF *Analyses* component under *Cluster prototypes*. The combination of the two ProPer metrics in the cluster analysis reveals patterns that are hard to detect from the shape of the F0 contours alone. For example, the overall shape of the contour in clusters 1 and 2 turned out to be quite similar (evident from the rather similar Synchrony pattern in blue) but they differ in that cluster 1 remained close to baseline with small F0 excursions compared to cluster 2, with more extreme movements across the F0 range (evident from the flat vs. dynamic ΔF0 patterns in red).

References

[1] Open Science Framework (OSF). Anonymized link to the OSF repository (this repository will be set to 'public' after review): https://tinyurl.com/yu8nnsrp

[2] The Internet Archive (archive.org). *The TV News Archive*. https://archive.org/details/tv

[3] Woodin, G., Winter, B., Perlman, M., Littlemore, J., & Matlock, T. (2020). “Tiny numbers” are actually tiny: Evidence from gestures in the TV News Archive. *PLOS ONE*, 15(11), e0242142. DOI: 10.1371/journal.pone.0242142

[4] Albert, A., Cangemi, F., Ellison, T. M., & Grice, M. (2023). *ProPer: PROsodic analysis with PERiodic energy* [computer software]. OSF. DOI: 10.17605/OSF.IO/28EA5. https://osf.io/28ea5/

[5] Kaland, C. (2023). Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours. *Journal of the International Phonetic Association*, 53(1), 159–188. DOI: 10.1017/S0025100321000049

[6] Cangemi, F., Grice, M., Jeon, H.-S., & Setter, J. (2023). Contrast or context, that is the question. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Scie*nces (pp. 1360–1364).

[7] R Core Team. (2023). *R: A Language and Environment for Statistical Computing* [computer softwarwe]. https://www.R-project.org/

[8] Boersma P., & Weenink D. (2024). *Praat: doing phonetics by computer* [computer softwarwe]. www.praat.org.

[9] Köhn, H.-F., & Hubert, L. J. (2015). Hierarchical Cluster Analysis. In *Wiley StatsRef: Statistics Reference Online* (pp. 1–13). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat02449.pub2

[10] Cole, J., Steffman, J., Shattuck-Hufnagel, S., & Tilsen, S. (2023). Hierarchical distinctions in the production and perception of nuclear tunes in American English. *Laboratory Phonology*, 14(1), Article 1. DOI: 10.16995/labphon.9437