**Using fPCA and GAMMs to investigate categoriality and variability in intonation**

Stella Gryllia, Katherine Marcoux, Amalia Arvaniti
Radboud University, Netherlands

We used functional Principal Component Analysis (fPCA; Gubian et al. 2015) and Generalized Additive Mixed Models (GAMMs; Sóskuthy 2017) to investigate the contrast between H* and L+H* in Greek and examine effects of speaking task and individual variation on the realization of the two accents. H* and L+H*, when followed by L-L% edge tones, are used to mark broad and narrow focus respectively. A previous fPCA analysis has shown that the accents' curves differ in height (PC1) and shape (PC2) (Lohfink et al. 2019). However, those findings were based on controlled scripted data, raising the following questions: 1) Is the difference between H* and L+H* equally robust in unscripted as in scripted speech? 2) Are the accent differences stable across speakers and speaking tasks, especially between scripted and unscripted speech?

We used a sample of 1160 accents ($N_{H*}$ = 748; $N_{L+H*}$ = 412) elicited from 8 native Greek speakers recorded in quiet conditions (due to COVID-19 restrictions). Scripted data consisted of reading short dialogues (Q&A pairs with broad or narrow focus), a news item, and a fable. Unscripted data consisted in recounting the two texts and telling stories using "Story Dice", an app with which users throw virtual dice and devise stories featuring the items on the dice.

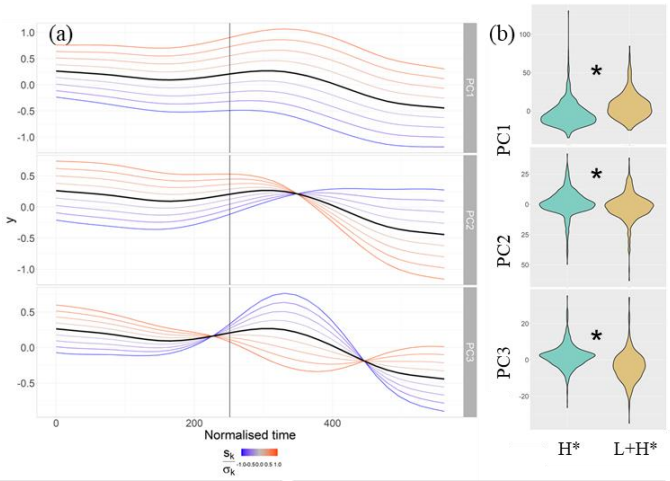
FPCA was used to uncover the principal ways in which the accents differ from each other. The resulting PCs were analysed using LMEMs (fixed factors: accent; task type [scripted, unscripted]). Task did not affect PC scores, but accent did: L+H* showed significantly higher scaling (PC1), an F0 dip (PC2), and a pitch rise ( PC3); see Fig. 1. The results were confirmed by downsampling (to keep equal numbers of H* and L+H* datapoints) and rerunning fPCA.

In addition, F0 landmarked registered curves were fitted in GAMMS using the *mgcv* (Wood 2017) and *itsadug* (van Rij et al. 2020) R packages (R Core Team 2020), with AccentTask (factor variable encoding accent and task) as a fixed intercept and as smoothed curves for AccentTask; speakers were included as a factor smooth. GAMMs confirmed the accent differences uncovered by fPCA, but also showed a significant task effect on L+H*: while H*s were virtually identical in scripted and unscripted data, L+H*s from scripted data exhibited a much more exaggerated shape (Fig. 2). We further split the data by accent to investigate speaker differences and ran GAMMs with TaskSpeaker (variable encoding speaker and task) as a fixed intercept and smoothed curves; word was included as a random intercept. This analysis showed that differences based on pooled results do not apply consistently across speakers (see Fig. 3 for illustrations of these speaker-specific differences).

The fPCA results replicated the findings of Lohfink et al. (2019), including the finding that H* is less prone to variation than L+H* (not shown here). As GAMMs indicate, the greater variability of L+H* could be (at least in part) due to participants exaggerating the features of this accent in scripted speech to clearly differentiate it from H*, especially when the two contrast, as they did here in successive dialogues (Q&A pairs). Such differentiation is less necessary in natural speech, since context and postfocal deaccentuation help listeners interpret the pragmatics of the accents. In addition, GAMMs quantified the extent of individual variation in the realization of the accents and the effect of speaking task on them.

From a methodological perspective these findings suggest that it is inadvisable to rely solely on scripted data to study a language's intonation system: scripted speech differs from unscripted speech when it comes to intonation, and the effects are not consistent across speakers. In turn, these findings have implications for standard practices, such as the collection of data that are highly controlled and the annotation of tonal categories in corpora based on category prototypes; rather, they argue in favour of less detail-oriented annotation protocols. Finally, our methods illustrate the complimentary ways in which fPCA and GAMMs address research questions in the study of intonation. While fPCA can help us understand essential differences among tonal categories, thereby helping with abstraction, GAMMs can capture within category differences, leading to a better understanding of variability and phonetic detail.

**Fig**. 1: *(a), first three PCs and their contribution to variability in the data( percentage in bottom left of each panel); (c), violin plots for PC scores by accent. Gray lines in (b) mark the accented vowel onset used for landmark registration*



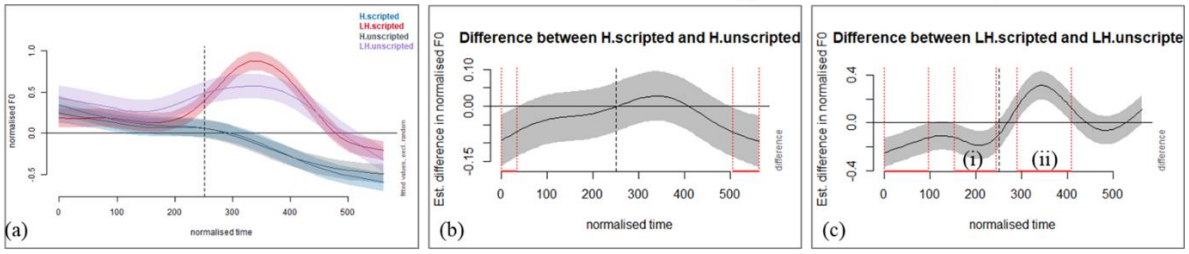




**Fig. 2**: *(a), Predicted F0 curves for scripted and unscripted H* and L+H*; (b) and (c), difference curves for H* (b) and L+H* (c); red bars indicate curve stretches where differences are statistically significant and show the direction of the difference relatively to the unscripted version of each accent; differences for H* are linguistically trivial, present only at the edges of the analysis window (which includes the entire word); for L+H*, significant differences pertain to the depth of the F0 dip (interval i) and the height of the pitch peak (interval ii).*

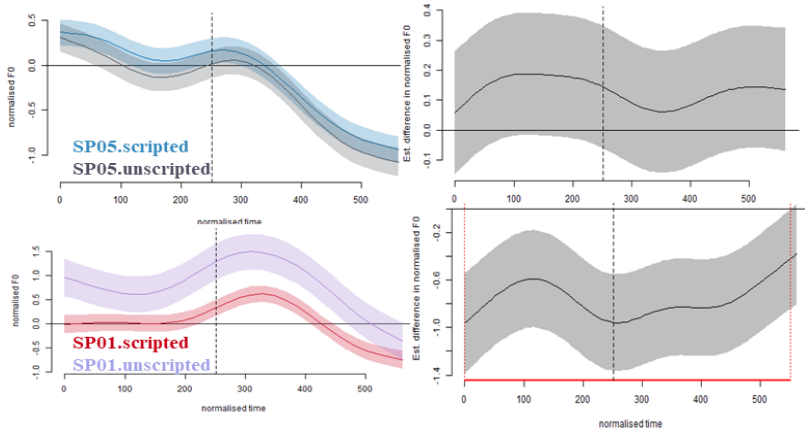



**Fig. 3:** *Predicted F0 curves (left) and difference curves (right) for the H*s of SP05 (top) and L+H*s of SP01 (bottom). SP05's H*s show no task-related differences, but both curves show a dip typical of L+H*; SP01 has typical L+H* curves, but their scaling is the reverse of that in the pooled data (Fig. 2).*

## References


Gubian, M., Torreira, F., Boves, L. 2015. Using functional data analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics* 49, 16–40.

Lohfink, G., Katsika, A., Arvaniti, A. 2019. Variability and category overlap in the realization of intonation. *Proceedings of ICPhS 2019*. httpss://assta.org/proceedings/ICPhS2019.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. [R v. 4.0.3]

Sóskuthy, M. 2017. *Generalised Additive Mixed Model for Dynamic Analysis in Linguistics: A Practical Introduction*. http://eprints.whiterose.ac.uk/113858/.

van Rij J., Wieling M., Baayen R., van Rijn H. 2020. "itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs." R package version 2.4.

Wood, S. N. 2019. mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. Computer software program, https://cran.rproject.org/ package=mgcv.