Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Cutting down on manual pitch contour annotation using data modelling

Yuki Asano[12], *Michele Gubian*[3], *Dominik, Sacha*[1]

[1]University of Konstanz, Germany
[2]University of Tübingen, Germany
[3]University of Bristol, UK

yuki.asano@uni-tuebingen.de, mm14722@bristol.ac.uk, dominik.sacha@uni-konstanz.de

Abstract

When experimental studies on intonation are based on large data sets, manual annotation of F_0 contours using pre-defined categories such as a ToBI (Tones and Break Indices) system is tedious, costly and difficult to provide reliability. We present two data-driven modelling techniques that provide visual and quantitative maps of the F_0 contour data set. The maps can be used to determine which ToBI categories are present in the data and in what proportions. Importantly, parts of the map that are sufficiently homogeneous, i.e. they contain only one ToBI category, can be directly labelled without manual annotation, hence reducing overall annotation costs. The modelling techniques will be evaluated using a small data set where a complete manual ToBI annotation was carried out, hence providing a ground truth for the evaluation.

Index Terms: F_0, data-driven analysis methods, manual annotation, L2 production

1. Introduction

Production studies of intonation entail analysis of numerous F_0 contours extracted from speech material. The analysis is usually carried out manually and involves both listening to audio data and visual inspection of F_0 contours. The result of the analysis is the classification of each utterance into intonational categories based on an annotation system. A popular annotation system is ToBI [1], which is based on Autosegmental-Metrical Phonology [2]. ToBI, originally conceived for English, has been specialised to several languages, including German (G-ToBI, [3]) and Japanese (J-ToBI [4]), and is also commonly used for the annotation of L2 production data [5, 6, 7, 8, 9].

Like many manual procedures, the amount of human effort required to annotate using a ToBI system increases costs of both time and money. A further problem with ToBI is its inherent subjectivity. In fact, it is customary to employ more than one annotator in a study – hence increasing its cost – in order to monitor the agreement among annotators with annotator agreement reported using a Kappa coefficient. [10]. Several other potential difficulties arise when using ToBI annotation for the analysis of second language (L2) data since one selected ToBI system might not cover all deviant variabilities of L2 data that represent interlanguage characters [11]. A language-specific ToBI system does not account for the deviant L2 realisations with respect to first language (L1), so unexpected contours might be hard or even impossible to describe using either an L1 or L2 ToBI system. Moreover, the annotators' language backgrounds may influence the annotation.

This work presents, evaluates and compares two automatic data modelling and visualisation techniques that can be used as valuable support for annotating F_0 data. One is Functional Principal Component Analysis (FPCA), an application of Functional Data Analysis [12]. The other is an interactive visualisation of data clusters based on Self Organising Maps (SOM) [13, 14, 15].

While not proposing FPCA and SOM as a complete substitute for manual annotation, the current paper will show how they can be useful in substantially alleviating the two aforementioned costs of annotation. Both techniques take F_0 contours as input and produce a data model as output. The models are entirely data-driven and are not based on a language-specific theoretical model, but instead entirely on measured F_0 contours only. The models produced by either of the two techniques are composed of a numerical part (i.e. the model parameters, which can be used to navigate the data set as well as to produce statistics) and a visual part(, which links the parameters to the corresponding F_0 contour shapes). These models can be used by researchers as a guide to 1) determine which contour shapes are most common in the data, 2) detect outliers, 3) produce tentative ToBI labelling before the start of manual annotation, and finally to 4) take informed decisions regarding tokens that need to be manually annotated, and which may be safely left with their preliminary model-derived annotation. Emphasis will be given to this last point since it opens the door to a substantial reduction in the number of utterances that need to be manually annotated. To date, there have been a number of attempts to apply data-driven, (semi)automatic analysis methods to F_0 [16, 17, 18, 19, 20, 21]. Our analysis methods are at an advantage over the previous methods because of the following: First, FPCA and SOM do not require prior model training with larger data sets in order to develop the system (e.g. not like AuToBI [19]). Second, the methods are non-parametric and annotation system-independent making them easily generalizable and more flexible than many state-of-the-art alternatives. In fact, FPCA or SOM can be applied prior to and as guidance for the decision about which annotation system to use on a given data set. Finally, interactive SOM visualisations involve the expert's knowledge into an iterative analysis process (in contrast to methods that only "run once").

The two data modelling techniques will be evaluated in a study in which semi-spontaneous production data were acquired from Japanese L1 speakers and German L2 learners. The evaluation takes advantage of the fact that a complete ToBI annotation has already been carried out on the data, so the results from FPCA and SOM can be compared against a ground truth.

2. Experiment

Participants: Fifteen speakers of Tokyo-Japanese (8 females) and 15 German learners of Japanese (6 females) participated in

 doi: 10.21437/SpeechProsody.2016-58

the experiment.

Materials: The target word used in the study was a very frequent Japanese word, *sumimasen* ([su.mi.ma.se.ɴ]), meaning *excuse me*. *Sumimasen* contains a lexically specified pitch fall associated with the penultimate mora in the word, [se]. Phonologically, this lexical pitch accent is described as H*+L in the J-ToBI system [4].

Procedure: The task was to produce the target word in a given context requiring participants to repeat the same word twice. Three attempts to produce the word were recorded from each speaker (3 attempts x 30 participants = 90 utterances in total) digitally onto a computer (44.1kHz, 16Bit). One utterance was discarded due to not-fulfilling task requirement.

F_0 computation and segmental boundary marking: F_0 contours were computed using the F_0 tracking algorithm in the Praat toolkit [22], with the default range of 70-350 Hz for males and 100-500 Hz for females. In order to minimise gender effects, F_0 values were expressed in normalised semitones. Then, segmental boundaries were marked applying standard segmentation criteria in Praat [23].

3. Data modelling

This section presents the two techniques used to support and facilitate ToBI annotation, namely FPCA and SOM.

3.1. Pre-processing

The same pre-processing procedure was applied to F_0 contours before loading them into either FPCA or SOM, namely *smoothing* and *landmark registration* [24]. Smoothing is an interpolation procedure that transforms contours sampled at discrete points into continuous and smooth functions. As a result, undesired detail from the sampled contours is removed. Small and rapid ripples that typically originate either from measurement errors or from microprosodic effects are removed. Landmark registration is a warping of the time axis that makes it possible to align corresponding events in time across contours. The operation allows us to interpret the variation of F_0 contour shapes across the data set in terms of the underlying segmental content, e.g. an observed shift in the onset of a falling gesture will not be confused with a difference in segmental durations. The procedure is conceptually equivalent to a segment-by-segment time normalisation, but the result is guaranteed to be smooth, i.e. without spurious discontinuities. In this case, all moraic boundaries of the word *sumimasen* were used as landmarks.

3.2. Functional Principal Component Analysis (FPCA)

FPCA [12] is an extension of conventional Principal Component Analysis [25, 26] that allows input in the form of continuous functions. It has already been proposed as statistical tool in intonation and phonetic studies [24, 27, 28]. FPCA provides a model of the (smoothed and landmark-registered) input curves in terms of a mean curve and a small number of Principal Component curves (PCs). Each PC curve represents a different deformation of the mean curve. Each input curve is associated with parameters called *PC scores*, each one determining the weight with which the corresponding deformation (PC) has to be applied in order to approximate that curve as closely as possible:

$$F_0(t) \approx \mu(t) + s_1 \cdot PC1(t) + s_2 \cdot PC2(t) + \cdots , \qquad (1)$$

$F_0(t)$ is the function of time representing a given F_0 contour, $\mu(t)$ is the overall mean curve, $PC1(t)$ and $PC2(t)$ are the first two PC curves, which are the same for all input curves and s_1 and s_2 are the specific PC scores that best approximate this particular $F_0(t)$. FPCA was applied to the 89 smoothed and landmark-registered F_0 contours. The first two PCs were retained, which explain 60% and 23% of the contour variance, respectively. This guarantees that the F_0 contours described by the FPCA model are a satisfactory approximation of the real ones. The distribution of PC scores s_1 and s_2 across the data set is shown by the 89 small empty circles in Figure 1 (ignore grid and legend for the moment).

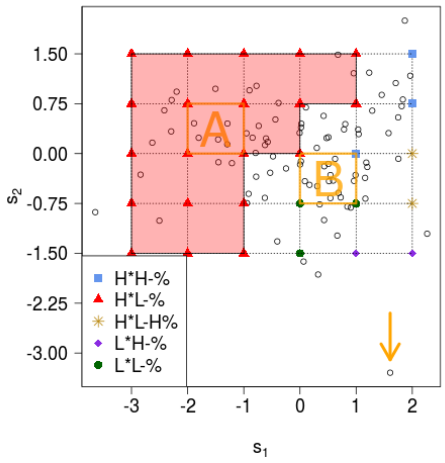
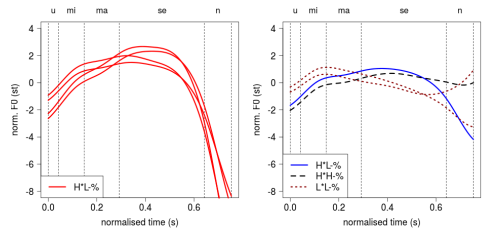


Figure 1: *PC scores scatterplot. Empty circles are located at (s_1, s_2) coordinates corresponding to the 89 input F_0 contours as modelled by Equation (1). Vertices on the grid are labelled with the ToBI category assigned by annotating the corresponding contours obtained from Equation (1). The red areas are sorrounded by vertices belonging to the same ToBI category, namely H* L-%.*

To make use of FPCA to predict where different pitch accents are located in the data set *before* starting data annotation, a regular grid of (s_1, s_2) points that covers most of the data is defined, as shown in Figure 1. Then, all F_0 contours corresponding to the grid points are annotated using the ToBI system. Note that the contours do not match any of the actual contours, but rather they are the result of applying Equation (1) to a set of (s_1, s_2) pairs defined in the study. The result of this operation is shown in Figure 1 by the symbols located at each grid crossing. At this point, PC scores modulate contours in a continuous and gradual way, i.e. points close to each other in Figure 1 correspond to F_0 contours that look similar. As a consequence, it can be assumed that rectangular (s_1, s_2) regions framed by vertices belonging to the same accent class contain points belonging to the same class. The area shaded red in Figure 1 shows all the regions where presumably all the actual F_0 contours, corresponding to empty circles, are realisations of H* L-% accents. Region A is one of these homogeneous regions, since it is framed by four vertices labelled H* L-%, which correspond to the four (artificial) F_0 contours in Figure 2(a). On the contrary, it is not possible to predict pitch accents in heterogeneous regions like B, since its vertices belong to different accents, as shown in Figure 2(b). In the next phase, one can choose to automatically assign H* L-% to the contours in the shaded homogeneous area and manually annotate a subset of the contours belonging to the heterogeneous regions.

Finally, FPCA can be used to detect outliers. The isolated

point indicated by the arrow in Figure 1 is considerably far from the rest of the data. Since its shape parameters (PC scores) are very different from all other contours, one may expect it either to belong to a ToBI category that is unusual or extreme for this particular data set or that it contains an error.



(a) (b)

Figure 2: *F_0 contours obtained from Equation (1) using s_1 and s_2 values corresponding to the vertices of rectangular region (a) A and (b) B in Figure 1. The contours have been manually labelled using the ToBI system.*

3.3. Self-Organising Maps (SOM)

The second method is a visual analytic system that divides the data into groups (clusters) based on the SOM algorithm. The result of the algorithm is a neural network that provides not only group information but also data regarding similarity among groups by means of a topological representation where similar cells (groups) are adjacent to each other. The data contained in each cell is represented by its cluster centroid, an average of all the data inside an individual cell.

The system consists of three components. The first is *Data Input*, which is flexible in accepting a great variety of input formats. In this study, the input consisted of the smoothed and landmark-registered F_0 contours that also formed the input for FPCA. The second component is *Machine Learning*, through which the SOM algorithm is carried out. Here the level of detail can be changed interactively by selecting the number of desired cells or by removing data in any iteration. After various attempts with different numbers of cells, 3 x 3 was selected as the desired configuration. The visualisation based on the SOM result is realised by the last component, *Interactive Visualisation*.

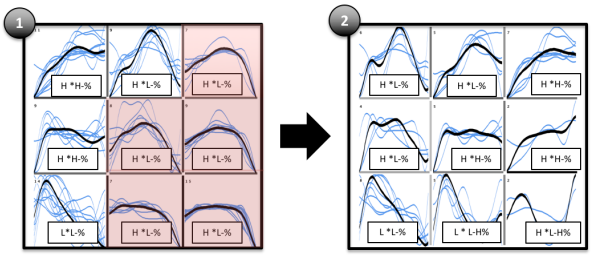


Figure 3: *The 9 clusters found in each training and ToBI annotation. (1) shows the output from the first run of SOM and (2) the second run after removing the red cells in (1). Black lines are centroid contours in each cluster. Blue thin lines are the F_0 contours within the cluster. Black borders between the cells show the cluster dis-similarity.*

Figure 3-(1) shows the output of SOM excuted on the whole contour data set. The cells are aligned according to their similarity. Each cell shows its centroid contour as a thick black curve, and all the contours belonging to the cell as thin blue curves. In the next step, the first author annotated the 9 cluster centroids according to ToBI. Five cells showed similar F_0 contours and were all coded as H* L-%, with small ToBI-within-categorical differences. Then, cells whose centroid contours were obvious to assign to a ToBI category and where member contours (blue curves) showed low variation around their centroids were selected. These homogeneous cells, coloured in red in Figure 3-(1), were excluded from further analysis, and all the contours belonging to these cells were automatically assigned to the ToBI category of their centroid. After having removed the red cells, the SOM algorithm was run for a second time, now including only heterogeneous cells. The output is shown in Figure 3-(2). The contours found originally in the white cells in Figure 3-(1) are more differentiated in Figure 3-(2). Finally, the 9 centroid contours obtained in the second run were annotated, and their member contours were assigned accordingly. Note that this interactive iteration between SOM-training and data-selecting can theoretically be repeated until only one contour remains.

4. Evaluation

4.1. Manual annotation

For the evaluation of FPCA and SOM, a manual annotation was carried out by a Japanese L1 speaker who was highly proficient in German. The types of accents and boundary tones were coded. Phonological form of pitch accent is restricted to H*+L in Japanese, but German learners of Japanese were expected to produce other accent types influenced by their L1, so the pitch accent categories of the German G-ToBI system [3] were used for both the Japanese and German data. The available accent types were six basic pitch accents (H*, L*, L*+H, L+H*, H+L*, H+!H*) [29]. Additionally, the H tones can be downstepped, which increases the inventory from 6 to 11 accents (ibid.). The available boundary tones were L-%, L-H%, H-% and H-^H%. The pitch form of the actual lexical pitch accent was always annotated on the mora [se], even though some L2 utterances showed a pitch fall in deviant positions, e.g. a pitch fall occurring earlier than in the mora [se], which was coded as H+L*. In this paper, the change between L* and H* pitch accents and L% and H% boundary tones is primarily reported, and other variations (e.g. downsteps or upsteps) are regarded as secondary modifications.

4.2. FPCA

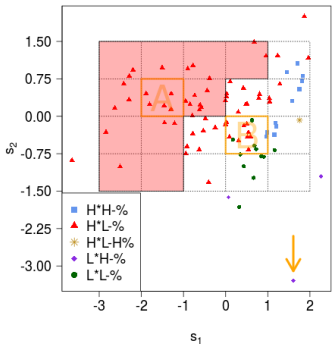


Figure 4: *PC scores scatterplot and ToBI annotation.*

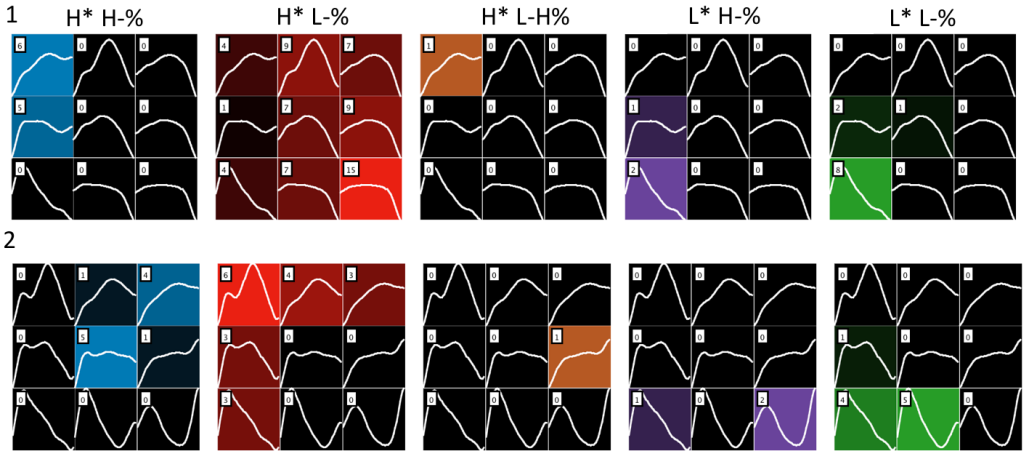


Figure 5: *Heatmaps of SOM sorted according to the ToBI categorisations. (1) is the output of the first run of SOM and (2) is the output of the second run. Numbers in each cell show the numbers of the data points in the cell.*

Figure 4 shows that the red areas predicted to contain only H* L-% (cf. Figure 1) do indeed contain the contours annotated as H* L-%. Moreover, rectangular areas on the grid in Figure 1, where vertices were labelled with mixed categories do indeed contain those categories (e.g. area B). Finally, note that the outlier (marked with an arrow) in Figure 1 does indeed belong to L* H-%, an outlier (minor) category for this data set.

4.3. SOM

Figure 5-(1) shows that contours with H* L-% were distributed throughout all 9 cells. H* H-% were categorised in the two cells that were predicted to be H* H-% in the data modelling (compare Figure 5-(1) and Figure 3-(1)). Other minor categories did not follow a pattern and were found crowded into only a small number of cells. Figure 5-(2) shows the output of the second SOM. The cells in the first and second rows matched the ToBI annotations (compare Figure 5-(2) and Figure 3-(2)). The cells in the third rows could be included for the third iteration in order to categorise the data more precisely.

5. Discussion

The data modelling techniques tested in this study proved to be successful tools for F_0 data exploration and preliminary annotation. Both FPCA and SOM allowed the identification of data subsets where automatic labelling could be safely applied. Note that, in this particular data set, there is very little to gain in terms of number of ToBI annotations to carry out since model-based annotation requires the annotation of only a few contours. However, the same procedure can be applied to data sets of arbitrary size, e.g. in the order of 10K utterances, by computing F_0 contours and providing the system with segmental boundary information, which would save a considerable amount of manual annotation. The current study has shown that, in general, it is difficult to provide automatic annotation for all data. This is partially solvable by increasing the resolution of the map, e.g. increasing the number of SOM clusters. However, it is difficult to rely entirely on automatic ToBI labelling, since the categorisation induced by FPCA and SOM does not always correspond to human categorical perception.

FPCA and SOM were able to automatically reveal underlying intonational categories and detect outliers before annotation was carried out. This was especially insightful in the current study because it contains L2 data, which are known to exhibit deviant variations [11] that are difficult to describe using the target language ToBI system. Data modelling provides us with an overview of the most typical contour shapes present in the data, which serves as a guide to selecting the five ToBI categories used in the analysis.

Even though FPCA and SOM are very different tools, both on the theoretical and practical levels, a comparison between them with regard to their value as tools for F_0 contour data exploration should be provided. FPCA proved to be highly valuable in providing an overview of the data and moderately successful (high precision, low recall) in providing a partial annotation of the data set. The results presented here could be improved by interactively changing the grid resolution in areas where contour categories are mixed, i.e. further subdividing the rectangles on the right side of the grid in Figure 1. However, the software tool used to carry out FPCA (the R package `fda` [30]) is not interactive. In order to make FPCA practical for interactive data labelling, a user interface specialised for this task must be created. SOM software, on the other hand, already comes with an interface for data visualisation that enables the user to tune the cluster resolution and examine results interactively. It should be noted that FPCA and SOM were tested on a small data set in this work, since a full manual annotation of all data was needed for evaluation. However, the results reported here could change when data sets grow by orders of magnitude. Time and memory costs of running the software tools in addition to quality of the obtained models (e.g. purity of the clusters, number of relevant PCs) are among the factors that certainly require further experimental investigation.

6. Conclusions

This study showed how to save costly manual work for the analysis of F_0 data through the use of automatic analysis methods. The methods provide an overview of a data set that detects homogeneous areas in which data points can be automatically assigned to the same intonational category without manually annotating them, while heterogeneous areas and outliers should be carefully analysed by a human. The steps shown in this study may be applied to larger data sets, providing in sizeable cost reduction.

7. References

[1] M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, "The original ToBI system and the evolution of the ToBI framework," in *Prosodic Typology – The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed. Oxford University Press, 2005.

[2] D. R. Ladd, *Intonational Phonology*. Cambridge: Cambridge University Press, 1996.

[3] M. Grice, S. Baumann, and R. Benzmüller, "German intonation in autosegmental-metrical phonology," in *Prosodic Typology. The Phonology of Intonation and Phrasing*, J. Sun-Ah, Ed. Oxford: Oxford University Press, 2005, pp. 55–83.

[4] J. Venditti, "Japanese ToBI labelling guidelines," *Ohio State University Working Papers in Linguistics*, vol. 50, pp. 127–62, 1997.

[5] M. Jilka, "The contribution of intonation to the perception of foreign accent," Ph.D. dissertation, University of Stuttgart, 2000.

[6] I. Mennen, "Bi-directional interference in the intonation of Dutch speakers of Greek," *Journal of Phonetics*, vol. 32, no. 4, pp. 543–563, 2004.

[7] T. A.-T. Nguyen, C. L. J. Ingram, and J. R. Pensalfini, "Prosodic transfer in Vietnamese acquisition of English contrastive stress patterns," *Journal of Phonetics*, vol. 36, no. 1, pp. 158–190, 2008.

[8] M. Ueyama and S.-A. Jun, *Focus realisation in Japanese English and Korean English intonation*. CSLI/Stanford University Press, 1998, vol. 7.

[9] S.-A. Jun and M. Oh, "Acquisition of second language intonation," in *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, 2000.

[10] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Physiological Measurements*, vol. 20, pp. 37–46, 1960.

[11] L. Selinker, "Interlanguage," *International Review of Applied Linguistics*, vol. 10, pp. 209–241, 1972.

[12] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis - 2nd Ed.* Springer, 2005.

[13] T. Kohonen, *Self-organising maps*. Springer, 2001, vol. 30.

[14] T. Schreck, *Visual-Interactive Analysis With Self- Organising Maps - Advances and Research Challenges*. Intech, 2010, pp. 83–96.

[15] T. Schreck, J. Bernard, T. Tekušová, and J. Kohlhammer, "Visual cluster analysis of trajectory data with interactive Kohonen maps," *Palgrave Macmillan Information Visualization*, vol. 8, pp. 14–29, 2009.

[16] H. Fujisaki, "Modeling the process of fundamental frequency contour generation," in *Speech Perception, Production and Linguistic Structure*, Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka, Eds. Ohmsha, 1992, pp. 313–328.

[17] E. Grabe, G. Kochanski, and J. Coleman, "Connecting intonation labels to mathematical descriptions of fundamental frequency," *Language and Speech*, vol. 50, no. 3, pp. 281–310, 2007.

[18] D. Hirst and R. Espesser, "Automatic modelling of fundamental frequency using a quadratic spline function." 1993, pp. 75–85.

[19] A. Rosenberg, "AuToBI - a tool for automatic ToBI annotation," in *Interspeech 2010*, 2010.

[20] A. Schweitzer, "Experiments on automatic prosodic labeling," in *Interspeech 2009*, Brighton, UK, 2009, pp. 2515–2518.

[21] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *The Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1697–1714, 2000.

[22] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program] version 5.2.20," 2011.

[23] O. Turk, M. Schöder, B. Bozkurt, and L. Arslan, "Voice quality interpolation for emotional text-to-speech synthesis," in *Proceedings of the 7th Interspeech*, Lisbon, 2005, pp. 797–800.

[24] M. Gubian, F. Torreira, and L. Boves, "Using functional data analysis for investigating multidimensional dynamic phonetic contrasts," *Journal of Phonetics*, vol. 49, pp. 16–40, 2015.

[25] J. Jackson, *A User's Guide to Principal Components*. Hoboken, NJ: Wiley, 1991.

[26] R. H. Baayen, D. J. Davidson, and D. M. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *Journal of Memory and Language*, vol. 59, no. 4, pp. 390–412, 11 2008.

[27] G. Turco and M. Gubian, "L1 prosodic transfer and priming effects: A quantitative study on semi-spontaneous dialogues," in *Proceedings of Speech Prosody 2012*, Shanghai, China, 2012.

[28] M. Zellers, M. Gubian, and B. Post, "Redescribing intonational categories with functional data analysis," in *Proceedings of INTERSPEECH 2010*, Chiba, Japan, 2010, pp. 1141 – 1144.

[29] S. Baumann, M. Grice, and R. Benzmüller, "GToBI – a phonological system for the transcription of German intonation," in *Prosody 2000: Speech recognition and synthesis*, ser. 21–28, S. Puppel and G. Demenko, Eds. Adam Mickiewicz University, 2001.

[30] J. O. Ramsay, G. Hookers, and S. Graves, *Functional Data Analysis with R and MATLAB*. New York, NY: Springer Verlag, 2009.