F0 derivatives in the classification of meaningful tonal movements

Constantijn Kaland

Suggested Citation:
Kaland, C. (2025). F0 derivatives in the classification of meaningful tonal movements. Journal of Phonetics, 113, Article 101454. https://doi.org/10.1016/j.wocn.2025.101454

Journal of Phonetics 113 (2025) 101454

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

F0 derivatives in the classification of meaningful tonal movements

Constantijn Kaland *

Linguistics I – Phonetics and Phonology, University of Düsseldorf, Germany
Institute of Linguistics – Phonetics, University of Cologne, Germany

ARTICLE INFO

Article history:
Received 25 November 2024
Received in revised form 15 September 2025
Accepted 1 October 2025
Available online 4 November 2025

ABSTRACT

Recent work applied cluster analysis on f0 contours in order to find 'prototypical' or 'underlying' categories as assumed in intonational phonology. However, it remains to be answered to what extent meaningful f0 variation can indeed be captured using automatic classification of surface realizations. Studies on f0 dynamics have suggested that derivatives (e.g., f0 velocity, acceleration and jerk) closely approximate the meaningful components of f0. The question answered in this study is to what extent f0 derivatives are more informative for cluster analysis than other metrics, such as the (time series) f0 contour they are derived from, a static measure representing it, or other acoustic measures such as intensity and duration. This is tested across two clustering techniques (hierarchical and k-medoids) for three different meaningful features expressed in Dutch noun phrases (of the type 'blue sofa'): focus type (broad, narrow), focus position (adjective, noun) and phrase position (medial, final). Results show that derivatives are among the most informative acoustic measures, although the best performing cluster analyses are the ones based on multiple acoustic measures. Crucially, cluster analyses reveal that the different meaningful prosodic features each have their own characteristics in terms of acoustics and number of clusters.

1. Introduction

There has been a recent increase in interest in applying cluster analysis to automatically classify intonation patterns. The aim is to find linguistically meaningful aspects of f0 contours that correspond to phonological categories. Studies have used cluster analysis to either *explore* such categories in understudied languages (e.g., Kaland, 2021; Björklund, 2024; Hakim, 2024) or to *refine* them based on existing theory of well-studied languages (e.g., Laméris et al., 2023; Cole et al., 2023; Seeliger and Kaland, 2022). Cluster analysis on f0 contours offers researchers a promising tool to deal with the high variability that is generally found in intonation. That is, naturally produced f0 contours signal much more than just linguistic content. They also provide emotional, socio-cultural, and physical information about the speaker. Intonational phonology abstracts over this kind of non-linguistic variation of f0 and aims to find the underlying linguistically meaningful categories. Cluster analysis is appealing in this endeavour because it automatically classifies data into groups of numerically similar observations (Kaufman and Rousseeuw, 1990), revealing the most robust patterns and avoiding the need for researchers to stipulate categories based on potentially biased auditory impressions or individual interpretation (Kaland, 2021).

With the adoption of cluster analysis as a classification tool, a stronger pressure arises on our understanding of what constitutes a class. In other words, what are the phonological categories in intonation? It has been shown repeatedly that traditional categorical approaches to intonation do not explain certain gradient phenomena (e.g., Watson, 2010; Grice et al., 2017; Ladd, 2022). Recent work that adopted nonlinear dynamical systems theory offers promising insights into how variability and gradience may go hand-in–hand with stability and discreteness. Concretely, a growing body of studies indicates that meaningful aspects of tonal contours are found in f0 dynamics (most recently, e.g., Roessig et al., 2019; Iskarous et al., 2024). This insight is an important advancement of the traditional perspective centered around the shape of a contour, as represented by measures of fundamental frequency level. The current study further investigates the usefulness of dynamic f0 measures, in particular the derivatives (velocity, acceleration and jerk), in the application of clustering.

Clustered f0 contours, i.e., the outcomes of automatic classification, hardly ever correspond directly to intonation

* Address: Heinrich-Heine-Universität Düsseldorf, Linguistics I – Phonetics and Phonology, Universitätstraße 1 (Building 23.21), 40225 Düsseldorf, Germany.
E-mail address: kaland@hhu.de

https://doi.org/10.1016/j.wocn.2025.101454
0095-4470/

categories, but neither are they redundant in a meaningful classification of f0 variation in a given dataset (Kaland, 2021). Central to the current study is the question: To what extent is cluster analysis able to approximate meaningful differences in intonation using dynamic measures? Answering this question not only contributes to the theoretical integration of articulatory dynamics and intonation, it also advances our understanding of how cluster analysis may facilitate intonation research. The current study does so by investigating Dutch adjective-noun combinations (noun phrases comparable to, e.g., 'blue sofa' in English), which were produced in different focus conditions (adjective, noun, broad) and phrase positions (medial, final). The intonation of focus and phrasing has been well-studied across several Western-Germanic languages and is generally expressed by f0 differences (e.g., Ladd, 2008). The f0 contour measures taken from the noun phrases were submitted to a series of cluster analyses and evaluated for how well the semantic context in which the noun phrases were produced could be classified. To assess the robustness of the role of the dynamic f0 measures (derivatives) in the classification, the analyses were performed across two different clustering techniques (hierarchical and k-medoids) and compared to static f0 measures as well as other acoustic cues (intensity and duration).

1.1. Categoricality

Influential models of intonation, in particular the ones within the autosegmental-metrical (AM) framework, are based on the idea of inventories of tonal events (Silverman et al., 1992; Pierrehumbert and Hirschberg, 1990; Beckman et al., 2005; Ladd, 2022). The idea is that high (H) and low (L) tones mark specific syllables, words or phrases to indicate their meaning and/or structure. Languages differ as to which combinations of tones they use and what their specific function is (e.g., Jun, 2014). The combinatory possibilities are given by the intonational grammar and based on a small inventory of (single or combined H and L) tones. These tonal movements may have specific functions, for example, marking sentence modality (statement/question), signaling pragmatic meaning, indicating discourse status, and/or marking the edges of smaller (intermediate) or larger (intonation) phrases (Silverman et al., 1992; Pierrehumbert and Hirschberg, 1990). AM theory has been centered around the idea of categories of tonal events, in which specific tonal shapes correspond to specific meanings. A well-studied example in Western-Germanic languages concerns the difference between new and contrastive discourse information (i.e., new and contrastive *focus*). For example, in English, the response to the question "What did you buy yesterday?" provides new information that is acoustically prominent, as in: "I bought a SOFA" (prominent word in capitals). However, in response to the question "Did you buy a table yesterday?", the response is likely to be produced with more acoustic prominence on the word 'sofa', as in: "No, I bought a SOFA". In English, as well as in other Western-Germanic languages such as Dutch and German, the difference between the new reading and the contrastive reading is generally expressed by a difference in intonation. The new reading is marked by a high/rising tone on 'sofa', and the contrastive one is marked with a low tone, directly followed by a steep high/rising tone. AM accounts of intonation have labelled these patterns as H* and L + H*, respectively, with * indicating that the high tone is aligned to the stressed syllable (Silverman et al., 1992). Thus, under this view, the two patterns (termed *pitch accents*) are categorically different contours because they have different discourse meanings (Pierrehumbert and Hirschberg, 1990).

Research has shown repeatedly that the prosodic marking of new and contrastive information is not strictly categorically different (e.g., Watson et al., 2008; Grice et al., 2017; Ladd, 2022). For example, referents such as 'sofa' in the examples above may have a more or less prominent role in the discourse depending on the number of times they have been mentioned, their phrase position, etc. It has been shown that acoustic prominence correlates with the degree of discourse prominence in a gradient way (Watson, 2010). Thus, a representation in terms of either H* or L + H* might give the false belief that f0 cannot gradually vary in the degree of target height or steepness of the rise. It has also been questioned to what extent H* and L + H* are two different phonological categories, as they might constitute "gradient phonetic variability [. . .] within a single phonological category" (Ladd, 2022, p.253). Furthermore, apart from varying the type of pitch accent, studies have also varied their position, such as in the noun phrases "blue SOFA" and "BLUE sofa" (e.g., Krahmer and Swerts, 2001; Swerts et al., 2002). These works show that in Western-Germanic prosodic marking of contrastive focus, the acoustic difference depends not only on the shape of the pitch accent on the accented syllable, but also on the lack of any such accent (deaccentuation) of the word that is not in focus.

It is beyond the scope of this study to resolve the question of phonological categoricality of H* and L + H*. Instead, the current study aims to test alternative ways of representing these intonation contours acoustically to investigate the extent to which they improve automatic classification of their meaningful prosodic differences. This aim contributes to tackling a fundamental problem with AM theory, namely that it is unclear "which phonetic features of intonation are gradient and which categorical" (Ladd, 2022, p.253). Recent intonation studies that applied dynamical systems theory have shown promising results on this issue, as further discussed in the following section.

1.2. Articulatory phonology as dynamic system

The combination of phonological categories and phonetic gradience has long been seen as a central problem in speech research (e.g., Perkell et al., 1986). The underlying question is how we need to model human speech such that it accounts for both invariant and variable components (e.g., Fujimura, 1990; Browman and Goldstein, 1992; Iskarous, 2017). This question not only applies to suprasegmental phenomena such as intonation, but also to segmental ones (i.e. phonemes, see, e.g., Arvaniti et al., 2024). Nonlinear dynamical systems theory overcomes this problem to the extent that it is able to simultaneously capture the gradience of dynamic movements (i.e., articulation) as well as the attractors, or stable states (i.e., phonological categories), on which such a dynamic system may naturally converge (e.g., Pierrehumbert and Pierrehumbert, 1990; Tuller et al., 1994; Gafos and Benus,

2006). Speech research has used a dynamic systems approach to explain a variety of production and perception phenomena, such as distinguishing 'say' and 'stay' in an acoustically manipulated continuum (Tuller et al., 1994), syllabification at different speech rates (Tuller and Kelso, 1989; Tuller and Kelso, 1991), or voicing neutralization and vowel harmony (Gafos and Benus, 2006). As for prosody and intonation, a dynamic approach has been applied to, among other phenomena, final lengthening (Katsika, 2016), lexical tone (Karlin and Tilsen, 2015), and pitch accent selection (Roessig et al., 2019; Iskarous et al., 2024). These approaches generally incorporate the theory of *task dynamics* (e.g., Saltzman and Munhall, 1989; Browman and Goldstein, 1990), which models the articulation of phonological categories as gestures (*gestural primitives*). Gestures may be seen as a set of dynamic movements in the vocal tract (tasks), e.g. tongue body constriction, lip aperture, or glottal aperture. The activation of these movements depends on the gesture that needs to be made, and typically multiple movements need to be made to realize a gesture, e.g. for the phoneme/a/: vocal fold vibration, jaw lowering, back constriction, etc. The timing and activation of these movements in a sequence (of morphemes, words, etc.) is what makes them dynamic and are managed in a coordinative structure (*gestural score*).

The analysis of articulatory phonology as gestures in a dynamic system has shown promising results in the study of f0 in intonation. As for the timing of f0, it was shown that traditionally assumed turning points in the contour are less informative to model L + H* and L*+H pitch accent differences in American English than the so-called *tonal centre of gravity* (TCoG) (Barnes et al., 2012; Barnes et al., 2021). This measure takes a weighted average over the time-series f0 in order to find the location in time of the f0 event "that can serve as a reference location for that F0 event in perception" (Barnes et al., 2012, p.342). Crucially, TCoG abstracts over the f0 shape as such and depends less on its manifestation over time, whilst still able to capture timing differences in f0 shape (i.e., scooped vs. domed movements).

For German, a dynamic model of pitch accents signalling broad, narrow and contrastive focus has been proposed (Roessig et al., 2019). In this model, a single control parameter is needed to select the focus type in order to reach the distribution of pitch accents as produced by speakers in interaction with a computer (referential question–answer task). This distribution showed similar amounts of rising and falling accents for broad focus, and predominantly rising accents for narrow and contrastive focus. The measure representing the accents was *tonal onglide*, i.e., the difference between the f0 maximum and minimum on the accented syllable (a positive value for rises and a negative value for falls). Previous work showed that the onglide, when acoustically manipulated in a perception experiment, could serve as the sole cue to listeners to successfully distinguish new from contrastive interpretations in German (Ritter and Grice, 2015). Crucially, the dynamic model was able to capture qualitative (phonological) variability in the type of pitch accent (rising vs. falling) as well as quantitative (phonetic) variability (degree of rising).

A recent study on American English proposed a somewhat similar model describing the f0 variation observed in produced imitations of intonation contours (Iskarous et al., 2024). The contours all consisted of a pitch accent (H*, L + H*, L*+H, L*) and an edge tone (HH, HL, LH, LL), totaling 16 different contours. Their basic shape distinctions concerned five measures: the extremes (minimum or maximum f0), velocity, maximum velocity, rise latency (time lag between rise onset and its maximum velocity) and span (f0 range). Instead of modeling these features explicitly, the model was built up in a minimal way such that all shape features follow from the model. This is a crucial difference with the Roessig et al. (2019) study, in which modeling was done directly on the onglide measures. Note that the acoustic properties of the f0 contours in both modeling studies are nevertheless *dynamic* in the sense that they go beyond the representation of f0 as expressed by a fundamental frequency level, but rather take non-static properties into account (e.g., tonal onglide, velocity, rise latency, span). Research has shown that *derivatives*, i.e., dynamic measures that are derived from the f0 contour, are often more informative representations than static measures such as mean f0, maximum/minimum f0, etc. Derivatives are particularly relevant in the study of the categorical nature of intonation, as further discussed in the next section.

1.3. The role of f0 derivatives

(1) $f0_{normalized}\{0.84, 0.91, 0.14, -0.76, -0.96, -0.28, 0.66, 0.99, 0.41, -0.54\}$

Mathematical curves can be described by their derivatives (e.g., Herman and Strang, 2016). Three derivatives are particularly relevant in the study of f0; the first (d1 or d′), second (d2 or d″) and third (d3 or d‴). For curves represented by time series such as the example in (1), their time derivatives are obtained by iteratively taking the difference between two adjacent points in the time series. Time derivatives express the *rate of change* of the original curve over time (Fig. 1). Note that in Fig. 1 the timescale is an arbitrary one for illustrative purposes. D1 is the result of this calculation applied to the f0 measures (i.e., the original f0 contour representing fundamental frequency), and expresses how fast the f0 increases or decreases over time (velocity or Δf0). D2 is obtained by rate-of-change calculations on d1, thus expressing the acceleration over time, or the concavity of the curve. Positive d2 values indicate that the f0 curve is concave up, negative d2 values indicate that the f0 curve is concave down (Fig. 1). D3 is obtained by rate-of-change calculations on d2, expressing 'jerk', or the rate of f0 acceleration, which has been argued to be minimal in human and other primates' motor control in order to make movements as smooth as possible (Hogan, 1984; see further discussion in Section 4). The example in Fig. 1 shows that with each derivative, the range of the values becomes smaller than that of the curve it was taken from. Note that with each derivative, the time series is N-1 shorter, which can be dealt with using extrapolation (see Section 2.2 and The MathWorks Inc., 2025 for further explanation). It can furthermore be seen that maximum and minimum jerk correspond to minimum and maximum velocity, respectively, and that maximum and minimum acceleration correspond to minimum and maximum f0, respectively. It also follows from the example in Fig. 1 that f0 rises such as the one observed between time points 5 and 8 correspond to positive velocity values, decreasing acceleration, and negative jerk values. F0 falls such as the

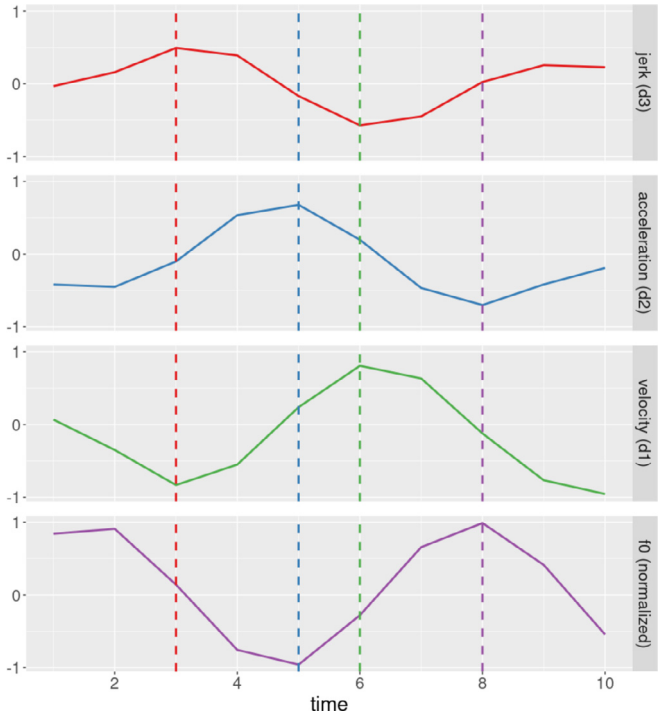

Fig. 1. Example of the normalized f0 contour in (1) in the bottom panel with its three interpolated time derivatives (d1, d2, d3) in the upper panels. For interpretation of their relation, dashed lines indicate the maximum value for each curve in the same color across all panels. Time is represented by an arbitrary series of increasing values for illustration purposes.

one observed between time points 2 and 5 correspond to negative velocity values, increasing acceleration and positive jerk values.

Although it is possible to take derivatives of a higher order, speech research has mainly focused on the first three in the study of f0 dynamics. Studies on speech synthesis (e.g., Narusawa et al., 2004; Gu et al., 2004; Sun et al., 2012; Comini et al., 2022) and automatic speech recognition (e.g., Liu et al., 1998; Le Roux et al., 2007; Hasan et al., 2014) have used them to improve the speech models that are used in such systems. A particular interest in derivatives in this respect has been shown in studies modeling affect in speech (e.g., Ross et al., 1986; Silva et al., 2016). It has also been shown that in the study of vocal fold dynamics the first derivative of the electroglottographic signal is highly informative (e.g., Henrich et al., 2004; Rahman and Shimamura, 2006). Work that has taken the syllable as a (phonological) unit in prosodic structure furthermore points to the importance of f0 derivatives. Multiple studies on Mandarin Chinese have shown that its lexical tones are best understood as dynamic targets that are aligned to syllables, as approximated by (among other acoustic values), the first and second derivative of f0 (Xu, 1998; Xu, 1999; Xu and Wang, 2009; Target Approximation Model; Xu and Wang, 2001). The first derivative of f0 also facilitates the detection of periodic energy, as measured in the ProPer toolbox (Albert et al., 2018), which bridges between acoustic and perceptual representations of intonation and is centered around syllable-based measures.

1.4. F0 derivatives in automatic classification

Particularly relevant to the current study is research in which the f0 derivatives were used in classification. This was done in a series of studies on Mandarin lexical tone, modeling how infants acquire lexical tone distinctions based on highly variable speech input (Gauthier et al., 2007; Gauthier et al., 2007; Gauthier and Shi, 2011). A neural network approach using self-organizing maps was used for unsupervised classification of 1800 instances of four Mandarin lexical tones (Gauthier et al., 2007). The tone contours were represented as time series of 30 f0 measurement points taken per syllable, obtained from the disyllabic sequence ‘mama’, taken from a carrier phrase as produced in a scripted, laboratory setting. The first derivatives were taken from these time series. After a training phase, the model was tested on classifying the four tones. The classification based on f0 time series showed an error rate of approximately 20%, and an error rate of 3% when based on d1. The tonal prototypes that the neural network model had learned were used in a subsequent simulation, which reduced the d1 error rate to zero. Note that no contextual information was included in any of the modeling, i.e., the classification was entirely based on the (derived) acoustic measures. Later studies replicated these results successfully using additional simulations (Gauthier et al., 2007), using infant-directed speech as model input (Gauthier and Shi, 2011), and using cluster analysis instead of neural networks (Zhang, 2016).

A recent study with a particular focus on cluster analysis investigated how (differences between) f0 contours should be represented such that the clustering outcome best reflects human perception (Kaland, 2023). This was tested in two typologically different languages, Papuan Malay and German, using different time-series f0 representations and different *distance measures* in hierarchical agglomerative clustering. Among the contour representations that correlated best with human perception was d1 in combination with Euclidean (time-warped) distances, outperforming traditional speaker-standardization methods and other distance metrics. These measures showed little difference in accuracy between the two languages, indicative of having approximated a low-auditory level of perceived contour differences rather than a phonological one, for which larger between-language differences would have been expected given their different prosodic systems. Results showed overall moderate to weak correlations between the contour dissimilarities as calculated based on acoustic measures, and those based on human perception. This was taken as an indication that despite d1 being a promising step towards modeling the perceived, and potentially meaningful, f0 contour differences, more work is needed to improve their correlation. This holds in particular for the approximation of linguistically meaningful (i.e., phonological) f0 contour differences, which is the goal of the current study. The question is to what extent cluster analysis is indeed able to capture these kinds of f0 differences, as further discussed in the following.

1.5. Clustering f0 variation

Cluster analysis, the grouping of observations based on their numerical similarity, is widely used across scientific disciplines (Scitovski et al., 2021). Recent research on prosody has shown an increased interest in using this technique to classify f0 variation. It appeared particularly useful to explore underresearched languages (e.g., Babinski, 2022; Björklund, 2024; Hakim, 2024; Zahrer, 2024) and to refine existing work on well-studied languages (Seeliger et al., 2023; Jeon et al., 2024; Cole et al., 2023; Steffman et al., 2024). The usefulness of cluster analysis in the study of linguistically meaningful f0 variation lies in several aspects related to this technique. First, cluster analysis avoids human interpretation of potential phonological categories based on researchers' own hearing, an aspect that has received little attention in the literature (Kaland, 2021). An additional benefit of hierarchical cluster analysis is the ability to perform analysis without having to set the number of clusters. In this way, the clustering process is represented by a tree structure (dendrogram), with the height in the tree corresponding to the number of clusters, which can then be chosen at a later moment. The latter option is unavailable for k-means and k-medoids clustering, which require setting a cluster number *prior to* the clustering. It is important to note that regardless of the clustering technique, *finding* the most suitable number of clusters is key in cluster analysis in general (Kaufman and Rousseeuw, 1990; Scitovski et al., 2021).

Second, cluster analysis can be applied to a wide range of suprasegmental phenomena, such as lexical tone (Laméris et al., 2023), phrase intonation (Kaland, 2021;Zahrer, 2024; Hakim, 2024), speech hesitations (Jabeen and Wagner, 2023), the expression of sarcasm (Tatár et al., 2024) and question intonation (Albert et al., 2024; Jeon et al., 2024). This wide applicability requires the further (re-) thinking of what constitutes a phonological category (see also Section 1). In particular, the classification of f0 contours signaling mutually exclusive meanings, such as Mandarin lexical tones (Zhang, 2016), lead to a different clustering and clustering evaluation than the classification of f0 contours that may be used with a considerable degree of overlap, such as focus marking in German (e.g., Roessig et al., 2019). A recent study on American English showed that clustering combined with other acoustic analyses and with perceptual data provides crucial new insights into AM accounts of intonation (Cole et al., 2023; Steffman et al., 2024). In particular, no support was found for certain assumed tonal contrasts found in phrase-final intonation contours ('nuclear tunes'). Analyses rather supported the idea that American English nuclear tonal contrasts are hierarchically organized such that some tunes are variants of each other, assuming a smaller 'inventory' and more subtle distinctions between what was previously assumed to be categorical distinctions.

Thus, third, cluster analysis as a classification tool, in particular when used with a flexible number of clusters, offers multiple ways of obtaining a more dynamic and integrative account of meaningful f0 variation. One way of improving its dynamic nature concerns the acoustic measures, such as the inclusion of f0 derivatives and multiple other acoustic cues. Another consists in the combination of cluster analysis with other classification techniques. Recent studies combined clustering with random forest analyses in the study of Papuan Malay and Korean f0 contours (Veilleux et al., 2023; Kaland and Grice, 2024; Jeon et al., 2025). These studies show that random forests can complement cluster analysis by means of a variable ranking of all theoretically possible factors explaining the classified f0 variation in the cluster analysis. A specific case was described for American English expressions of surprise (mirativity), which could be accurately classified into fillers and exclamatives using both clustering and random forests based on PoLaR annotated acoustic features (Ahn et al., 2019; Veilleux et al., 2023). For the spontaneous and interactive speech data of Papuan Malay, results showed that turn-taking (continuation or ending) explained most of the f0 variation. For Korean spontaneously produced questions, a series of cluster analyses with varying numbers of clusters was run, with a random forest analysis applied to each clustering. Different variable rankings were obtained for different numbers of clusters. That is, in addition to the high degree of speaker variation in this data, the type of question particle explained most of the f0 variation when assuming four clusters, whereas for higher numbers of clusters, the combination of language variety of the speaker and type of question explained most of the f0 variation (Jeon et al., 2025). Although these results seem to show a trivial outcome in that meaningful categories of f0 variation depend on how many categories one allows to be found, they are testimony of how classification can be applied along multiple dimensions simultaneously. Crucially, this kind of analysis gives an insight into the potentially complex interaction between multiple sources of f0 variation. This is an important advancement with respect to the well-studied H* vs. L + H* distinction, which is a simplified, stylized and abstract representation of the f0 variation found in naturally produced speech. While it is undisputed that H* and L + H* often represent a meaningful difference, there is only one level at which the meaning difference occurs (discourse level; Pierrehumbert and Hirschberg, 1990), ignoring other sources of f0 variation. The current study is centered around classification (cluster analysis) of meaningful f0 movements only. It is beyond the scope of this study to include f0 perception or additional classification techniques. The current study exclusively focuses on the role of dynamic measures of produced f0 contours in clustering in order to further test and exploit its advantages for prosodic modeling, which advances previous work in several important ways, as further outlined in the next section.

1.6. Research aims

Summing up the literature discussed in the previous sections, two main directions for further research become clear. First, f0 dynamics as represented by derivatives of time series seem more informative than static measures. This is particularly true for the categorization of meaningful f0 variation, as shown by dynamic models (Barnes et al., 2012; Roessig et al., 2019; Iskarous et al., 2024). Second, cluster analysis has shown promising results when f0 derivatives were included (Zhang, 2016; Kaland, 2023). This technique also has the advantage of integrating multiple variables and testing the number of clusters flexibly. These are crucial for our understanding of categoricality in intonation.

The research question addressed in this study is therefore how informative f0 derivatives are for automatic classification of f0 contours by means of cluster analysis. The way in which the current study answers this question resolves a number of issues that have not been (simultaneously) addressed in the previous literature.

Regarding the investigation of informativeness of f0 derivatives, the current study includes all derivatives that have been reported to be relevant for f0 (d1, d2 and d3). Previous work has included d1 in particular, with only a few studies including other ones or multiple derivatives (Xu, 1998; Xu, 1999; Xu and Wang, 2009; Target Approximation Model: Xu and Wang, 2001). Furthermore, the informativeness of the derivatives to cluster analysis in the current study is compared to the informativeness of other acoustic cues such as intensity and duration. Both traditional accounts (Pierrehumbert and Hirschberg, 1990) as well as recent dynamic modeling (Barnes et al., 2012; Roessig et al., 2019; Iskarous et al., 2024) have exclusively focused on modeling f0 cues. While f0 might be the most important aspect of the speech signal for intonation differences, in naturalistic speech it is never perceived in isolation from other cues (see Albert et al., 2018 for an integrative approach). With regard to the distinction between new and contrastive information, research has shown that cues such as intensity and duration are also affected (e.g., Kaland, 2014). Regarding clustering, the current study compares two different techniques: hierarchical agglomerative clustering and k-medoids clustering. This is done to assess to what extent the informativeness of f0 dynamics in clustering depends on the technique used.

The above mentioned issues are investigated using f0 contours obtained from Dutch adjective-noun combinations (noun phrases) in different focus conditions and phrase positions. Three linguistically meaningful features of these noun phrases are distinguished, as they have all been reported to be expressed by prosodic differences: focus type (narrow, broad), focus position (adjective, noun) and phrase position (medial, final). The two focus features are disentangled to account for the similarity between noun and broad focus in Dutch: they are both expressed on the noun despite being semantically different and showing some prosodic differences (e.g., Krahmer and Swerts, 2001). Thus, the two focus features are not entirely symmetrical, because broad focus on the adjective is not possible in Dutch. Consequently, the design to collect the focus features was not entirely symmetrical in the sense that it followed a complete 2x2 design. The three meaningful features furthermore differ in how local or global their domain of realization is. That is, focus type is investigated in this study within the word domain (the noun in the noun phrase), focus position concerns a variation between two words within the domain of the noun phrase, whereas phrase position concerns a more global feature varying within the domain of the phrase, with local variation expected particularly on the final syllable of the noun phrase (boundary tone). Teasing these three features apart is therefore particularly useful in the clustering approach in the current study, following previous work that showed that meaningful f0 differences are best captured by measures that incorporate both local and global aspects of f0 (e.g., Barnes et al., 2012). The implications of the local or global nature of f0 variation that is expected for each feature is further discussed in Section 2.2 and Section 3.2.

The contours, although produced in a laboratory setting and therefore lacking some degree of naturalness, still have a natural degree of speaker variation in the extent to which prosodic cues were used to mark linguistically meaningful features. It is furthermore important to note that the current study does not assume a particular phonological status of the expected meaningful f0 differences. The main aim is to perform an automatic categorization of f0 contours whilst knowing their linguistic context, and whilst expecting a considerable degree of variation, both within focus conditions/phrase positions and across speakers. Whether the 'best' outcome in the current study comprises two, three, four, or more clusters does not primarily depend on the assumption that there should be a fixed number, but rather on the quality of clustering, as assessed by multiple metrics, across different clusterings, acoustic cues and techniques. Such an approach further distinguishes the current study from recent dynamic modeling studies (Roessig et al., 2019; Iskarous et al., 2024), which both performed modeling on phonological f0 categories as assumed in AM accounts. That is, the German model was based on known shape differences between two pitch accent categories (onglides), and the American English model was based on imitated contours, which neatly followed traditionally assumed phonological categories and were lacking speakers' natural variation. The current approach is entirely bottom-up and (acoustic) data driven, and lacks a pre-specified categorization of f0 variation as the goal of the classification. In this way, the current approach is largely agnostic about the way intonational form and meaning are mapped. The minimal and careful assumption is that there is some degree of mapping expected. Thus, it is very well possible in the current study that the most informative clustering has more clusters than focus categories, just because the speakers' prosodic variation is better captured by that number of clusters.

The next section provides a detailed report of the methodology (Section 2), after which the results are reported (Section 3). The final section provides a general discussion and conclusion (Section 4).

2. Methodology

2.1. Data

Dutch noun phrases consisting of an adjective and a noun were elicited in a picture naming task. The data and collection procedures are described in Kaland et al. (2023), and only the relevant aspects of the collected data are (re-) reported here.

The noun phrases were produced by speakers in a task in which they described two pictures on a computer screen; one on the left side of the screen and one on the right side. The pictures differed in color and/or shape in such a way that the information status of the adjective and or the noun could be manipulated. Participants described the pictures from left to right using one of the two matrix phrases that differed in the position of the noun phrase referring to the pictures, either medial: "I see [LEFT PICTURE] on the left side, but I see [RIGHT PICTURE] on the right side.", or final: "On the left side I see [LEFT PICTURE], but on the right side I see [RIGHT PICTURE]." (In Dutch: "Ik zie een [LEFT PICTURE] aan de linkerkant, maar ik zie een [RIGHT PICTURE] aan de

rechterkant." and "Aan de linkerkant zie ik een [LEFT PICTURE], maar aan de rechterkant zie ik een [RIGHT PICTURE]."). The noun phrases that were produced to describe the right picture were selected for further analysis, as they differed either in color (narrow focus adjective), shape (narrow focus noun) or both (broad focus) from the left picture. Five colors and five shapes were used to elicit 25 right-picture-references (10 adjective focus, 10 noun focus and 5 broad). The Dutch words used in the noun phrases were all disyllabic with stress on the first syllable, colors: 'zwarte', 'witte', 'rode', 'groene', 'blauwe' - *black, white, red, green, blue*; shapes: 'kano', 'robot', 'satan', 'python', 'haring', 'radar', 'lichaam', 'sofa', 'tosti', 'limo' - *canoe, robot, satan, python, herring, radar, body, sofa, toast, lemonade*. A total of 23 native speakers of Dutch carried out the task; 2 males and 21 females (mean age: 20.4 years, age range: 18–23 years). In total, 1150 noun phrases were collected (25 items * 2 phrase positions * 23 participants).

2.2. Acoustic analysis and processing

Three acoustic measures were taken from the waveform recordings of the noun phrases in R using the `wrassp` package (R Studio Team, 2022a; R Studio Team, 2022b; Winkelmann et al., 2023; Mahr, 2020): f0 (Harmonic Sieve method; Scheffers, 1983), intensity (Root Mean Square amplitude) and duration. F0 (Hz) and intensity (dB) were taken as time-series measures using 20 equidistant measurement points per word in the noun phrase (points 1–20 for the adjective, points 21–40 for the noun). These time-series measures thus 'warp' the duration of the adjective and the noun into a 20-point timescale for each word, as required by the clustering technique (see Section 2.3). Although the original timescale gets lost in this way, the boundary between adjective and noun is retained between measurement point 20 and 21. The degree of warping can be read from Table 1, which gives an overview of the word durations for each meaningful feature and the average interval duration between adjacent measurement points. Duration was measured separately for the adjective and the noun as absolute duration in milliseconds from the start of the word until the end of the word, excluding any silences that did not belong to the production of the word's segments. Thus, silences preceding plosives were taken into account for the duration measurement.

Before turning to a detailed description of how the time-series data were further processed, it is important to note that two types of acoustic measures were used to represent them. One pertains to their nature as time series, i.e. a vector of values over the entire noun phrase (see Table 2). The other pertains to a measure taking into account the difference between adjective and noun within the noun phrase, i.e., a difference score subtracting a value representing the noun from a value representing the adjective. Note that the difference scores were calculated on the basis of the time-series data as described above. Difference scores are syntagmatic in the sense that they account for the fact that the noun phrase is composed of two words and for the fact that produced and perceived contrastive focus in Dutch noun phrases not only depends on the shape of the f0 contour, but in particular also on the acoustic difference between adjective and noun (Krahmer and Swerts, 2001). The difference in performance between the time-series measures and the differences scores sheds further light on how locally or globally informed f0 measures affect the automatic classification. For comparison to the derivative measures, the time series from which these were derived are still included, using different scales that either represent the way f0 has been traditionally measured (Hertz) or more informed ones based on perception (ERB). Also included in the current study is a measure of f0 range, which captures some aspect of the time-series dynamics in a more abstract way than the f0 time series itself. F0 range has been shown to be an important correlate of meaningful intonation differences in many languages (e.g., Ladd, 2008) and in the case of Dutch contrastive noun phrases might act as an informative measure for global aspects of the contour.

The voiceless segments in the noun phrases caused missing f0 values in the time series. Linear interpolation and constant extrapolation were applied to approximate the perceived continuation of the f0 contour (e.g., Mixdorff and Niebuhr, 2013). Any disfluencies caused in this process were smoothed using Nadaraya–Watson kernel regression as implemented in the `stats` package in R (bandwidth set at '20': R Studio Team, 2022a; Nadaraya, 1964; Watson, 1964). The smoothed time series were then checked for their velocity (steepness of the movement). That is, the velocity of f0 between any two adjacent measurement point was checked against the maximum rates of rising and falling f0 as reported in Xu and Sun (2002, Table X): rises: 72 semitones/s; falls: 96 semitones/s. Any contour for which either no f0 could be tracked accurately or for which the velocity at any point exceeded those thresholds was removed from the data (N = 151).

The remaining contours were then converted to equivalent rectangular bandwidth (ERB), a scale with a logarithmic component to account for perceived differences in pitch (Glasberg and Moore, 1990; Hermes and Van Gestel, 1991).

Table 1
Average durations (and standard deviations) and interval duration between adjacent measurement points (both in ms) per word in the noun phrase (adjective, noun) for each of the phrase positions (medial, final) and focus conditions (adjective, noun, broad).

Position	Focus	Adjective		Noun	
		Mean (sd)	Interval	Mean (sd)	Interval
Medial	Adjective	316.36 (57.08)	15.06	391.51 (75.74)	18.64
	Noun	319.73 (90.45)	15.23	421.05 (79.59)	20.05
	Broad	327.55 (115.25)	15.60	407.13 (80.94)	19.39
Final	Adjective	325.47 (67.70)	15.50	499.29 (79.65)	23.78
	Noun	301.59 (71.95)	14.36	517.67 (81.46)	24.65
	Broad	308.91 (94.47)	14.71	496.71 (72.59)	23.65

Table 2
Overview of the acoustic measures taken from the noun phrases, their type (time series or difference score), their calculation, and their abbreviation as used in Section 3. Hz = Hertz, z-ERB = standardized ERB, max = maximum, min = minimum, σ = standard deviation, μ = mean, m = median, Δ = delta (difference); subscripts: spk = speaker, A = adjective, N = noun.

Cue	Measure/scale	Type	Calculation	Abbr.
F0	Hertz	Time series	Harmonic sieve (Scheffers, 1983) - smoothed	Hz
	ERB	Time series	$16.6 * log10(1 + (Hz/165.4))$	erb
	z-ERB	Time series	$(ERB_{spk} - \mu ERB_{spk})/\sigma ERB_{spk}$	ste
	F0 range	Difference	$max(z - ERB)_A - min(z - ERB)_A - max(z - ERB)_N - min(z - ERB)_N$	rg
	Velocity	Time series	$\Delta z - ERB/\Delta time$	d1
	Acceleration	Time series	$\Delta d1/\Delta time$	d2
	Jerk	Time series	$\Delta d2/\Delta time$	d3
	Mean velocity	Difference	$\mu\|d1\|_A - \mu\|d1\|_N$	md1
	Mean acceleration	Difference	$\mu\|d2\|_A - \mu\|d2\|_N$	md2
	Mean jerk	Difference	$\mu\|d3\|_A - \mu\|d3\|_N$	md3
Intensity	dB	Difference	$m(dB)_A - m(dB)_N$	db
Duration	ms	Difference	$\|ms\|_A - \|ms\|_N$	dur

This scale was also used to standardize (z-score) the contours based on each speakers' f0 range, as calculated over these contours. In this way, the standardized ERB contours (z-ERB) were centered around zero (the mean of the speaker's f0 range). F0 range was also calculated for each contour (maximum-minimum z-ERB) and any contours with a range beyond 2 standard deviations (1.49 ERB, approximately 8 semitones) were removed (N = 136). Manual inspection showed that these contours were mostly reflecting erroneously tracked f0 and/or constituted outliers, given that a perceptual model of Dutch intonation has defaulted to f0 movements spanning a maximum of 6 semitones (Hart et al., 1990). The f0 range values were then converted to a difference score, subtracting the range of the noun from the range of the adjective as a single-value approximation of the excursion size of the f0 movement on either word.

The final set of f0 contours (N = 863) thus consisted of Hertz, ERB and z-ERB time series that were largely free from errors. The f0 derivatives were calculated based on the z-ERB time series. This was done to achieve a minimal degree of f0 range variation originating from physical differences between speakers. Note, however, that some degree of speaker variation in the contours was still expected in the degree to which they marked focus (Roessig et al., 2019). Velocity (d1), acceleration (d2) and jerk (d3) were measured by the first three derivatives using the `gradient()` function in the `pracma` package (Borchers, 2022). Note that this function extrapolates linearly to overcome the N-1 problem in derivation. That is, the difference between two subsequent values in a time series can only be taken N-1 times, where N stands for the length of the original time series. This is solved at the edges, by extrapolating an additional measurement point to obtain the difference that can be time-aligned (see The MathWorks Inc., 2025 and the resulting output example in Fig. 1). For each derivative, the mean value of the noun was subtracted from the mean value of the adjective in order to obtain their difference score.

After inspection of the intensity time series, it was found that these were affected by the segmental material to a large extent. For example, words with plosives or fricatives (e.g. 'zwarte', 'robot') showed intensity extremes (minima and maxima) at specific points in the time series, causing the intensity time series to reflect segmental differences to a much larger extent than prosodic ones. To overcome these effects, median dB values for the adjective and the noun were taken. Median values are less affected by outlying measures than mean values. Then, an intensity difference score was calculated by subtracting the median intensity of the noun from the median intensity of the adjective.

Duration measures were also expressed as a difference score to abstract over the segmental differences between the words in the noun phrases. Note that due to final lengthening of the noun (μ duration adjective: 316.56 ms., μ duration noun: 450.40 ms.), negative difference scores are expected.

Table 2 lists the acoustic measures taken from the noun phrases. They were all used as input for the cluster analyses, which are explained in the following.

2.3. Clustering

Two cluster analysis techniques were used: hierarchical agglomerative clustering (HAC) using complete linkage as linkage criterion and k-medoids clustering, also known as partitioning around medoids (PAM). PAM was chosen instead of the more popular k-means as PAM is less sensitive to outliers and performs – just like HAC – clustering over distance matrices. A distance matrix consists of the dissimilarities between all observations as expressed by a distance metric. Euclidean distance was used in the current study for all cluster analyses as it was shown to be among the closest approximations to perceived contour differences and used in the majority of previous studies applying cluster analysis to f0 contours (Kaland, 2023). Distance matrices furthermore allow for multivariate clustering, which is used in the current study by means of combining multiple acoustic cues. HAC was performed using the `hclust()` function and PAM was performed using the `pam()` function in R (R Studio Team, 2022a). Multiple clusterings were performed, ranging from 2 to 15 clusters. This was done to assess which clustering provided the best categorization according to the criteria reported in Section 2.4.

Distance matrices were computed for each of the 12 acoustic measures in Table 2. All of them were rescaled between 0 and 1 (`rescale()` function in the `scales` package; Wickham and Seidel, 2022) such that summing multiple distance matrices into a new combined one warrants equal

contribution of the individual matrices of which it is composed. This was done to test the informativeness of multiple acoustic cues. Thus, in addition to the single distance matrices, distance matrices were combined in ways that took most of the possible combinations into account, whilst discarding combinations of cues that were variants of each other. That is, it was estimated as redundant to combine the distance matrices of, for example, Hertz time series with ERB time series. F0 range, however, was taken as a separate f0 measure from the time-series and derivative difference scores due to its global nature and its important role in intonation research (e.g., Ladd, 2008). Thus, the 9 f0 contour measures (time-series and derivative difference scores) were each combined with all possible combinations of intensity, duration and/or f0 range (N = 7), the latter of which were also tested without the f0 time series and derivative difference scores (N = 4). In addition, a combination of all f0 derivative difference scores was included to test their combined informativeness and to account for the theoretical possibility that each of them targets a different dynamic aspect of the f0 contour. All tested (combined) distance matrices are listed in Appendix A (12 + (9*7) + 4 + 1 = 80). Each of them was run through 14 rounds of cluster analysis (2 to 15 clusters) for each of the clustering techniques (HAC and PAM), resulting in 2240 cluster analyses.

2.4. Cluster evaluation

The evaluation of the cluster analyses concerns three types of meaningful features of the noun phrases that were expressed prosodically: focus position (adjective or noun), focus type (broad or narrow) and phrase position (medial or final). Each meaningful feature was evaluated for two quality aspects in each cluster analysis; namely separation and grouping. That is, quality of separation was assessed using chi-square tests on the two clusters that showed the best separation of the levels of the meaningful feature. In the case of focus position, for example, the assessment ascertained which cluster had the maximal difference in proportions of the number of observations between adjective and noun focus (proportional contingency table). This was done by subtracting the proportions of the two levels from each other (e.g., proportions adjective focus minus proportions noun focus). Then the assessment ascertained which cluster had the largest difference (i.e., favouring adjective focus) and which cluster had the smallest (negative) difference (i.e., favouring noun focus). A chi-square test was done on the 2 x 2 table of absolute number of observations from the two maximally different clusters. This procedure was followed in the same way for the other meaningful features (focus type and focus position), such that focus type (broad vs. noun) was evaluated whilst ignoring adjective focus observations, and focus position (adjective vs. noun) was evaluated whilst ignoring broad focus observations. These separate procedures guaranteed that the overall similarity between broad and noun focus in the f0 of Dutch noun phrases was neither masked nor boosted by including a third category, as would be the case if all focus conditions were lumped together. Two values obtained from the chi-square test, *p*-value and the absolute standardized residuals, were then used to assess the quality of separation. Standardized residuals indicate the degree to which the observed counts differed from chance level distribution, with higher standardized residuals indicating better separation. The quality of grouping was assessed by assigning each cluster to a level of the meaningful feature. This was done by taking the level that had the highest absolute number of observations in a given cluster. For example, if cluster 1 had the most observations for noun focus, then cluster 1 was taken as a noun-focus cluster. After all clusters in the analyses were labelled in this way, it was calculated how many observations were clustered 'correctly' according to the observed majority, as a proportion of the total number of observations (N = 863).

After all cluster analyses, the separation quality values (standardized residuals) were rescaled between 0 and 1 for comparability with the proportions assessing the grouping quality, only if the residuals were obtained from a chi-square test with a *p*-value below.05. Then, the mean of the separation and grouping value was taken for each meaningful feature, resulting in three values per cluster analysis (focus position, focus type, phrase position). Subsequently, a single value expressing the overall quality of that clustering for all three meaningful features was calculated by taking the mean of the three values of the meaningful features. Note that the evaluations of both separation and grouping do not automatically favor clusterings in which the number of clusters matches the number of levels of the meaningful feature under assessment (i.e., two). That is, better separation and grouping quality may be obtained for analyses assuming 3 or more clusters, because ambiguous productions form their own cluster(s) and/or multiple prototypical productions for specific levels of meaningful features are identified. The evaluation procedure also does not favor higher numbers of clusters as optimal ones, given the use of proportional contingency tables that take the proportion of observations relative to the grand total into account.

3. Results

3.1. Acoustic measures

The results of the acoustic measures show some general tendencies in the f0 contour on the noun phrases (Fig. 2). For the description of the results, measurement points were coarsely mapped onto the syllables, whilst loosing some accuracy due to the time-warping: points 1–10 and 11–20 for the adjective and points 21–30 and 31–40 for the noun respectively. The f0 contours were highly affected by phrase position and are therefore discussed separately for medial and final noun phrases. Phrase-medial noun phrases were produced with a high f0 in the region of the final syllable (of the noun), at a minimum of 0.5 z-ERB, with noun focus adding additional height and showing a steeper rise in the region of the stressed syllable of the noun compared to broad focus. Adjective focus was realized with an f0 rise on the adjective. Phrase-final noun phrases were produced with a low f0, between 0.25 and 0.5 z-ERB. Broad and narrow focus on the noun showed almost identical contours, except for a minimally steeper fall in the region of the stressed syllable of the noun in broad focus, i.e. a more sustained f0 level on the noun for noun focus. Adjective focus was realized with a falling contour from the start of the adjective to the region of the stressed syllable of the noun,

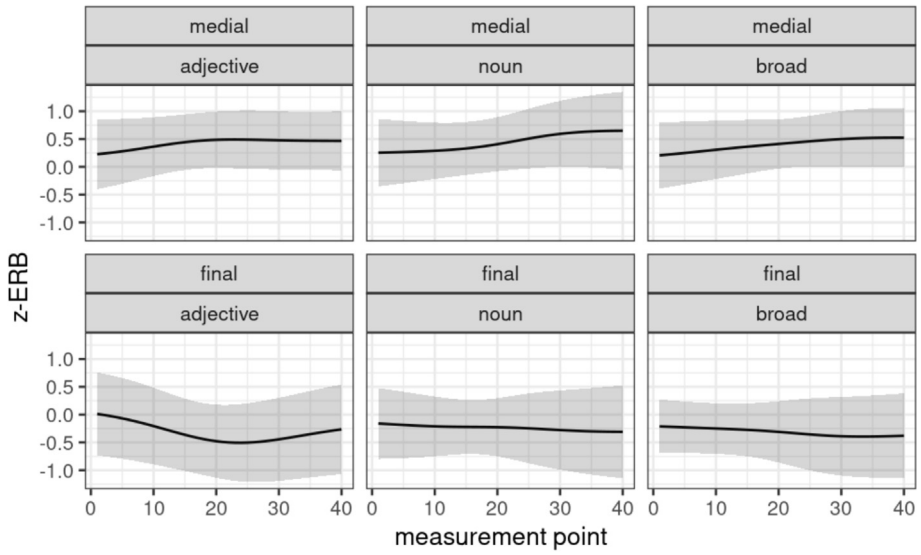

Fig. 2. Time-series measures of speaker standardized ERB (z-ERB) f0 values taken from noun phrases (adjective: points 1–20, noun: points 21–40) in each phrase position (top: medial, bottom: final) and focus condition (left: adjective, middle: noun, right: broad).

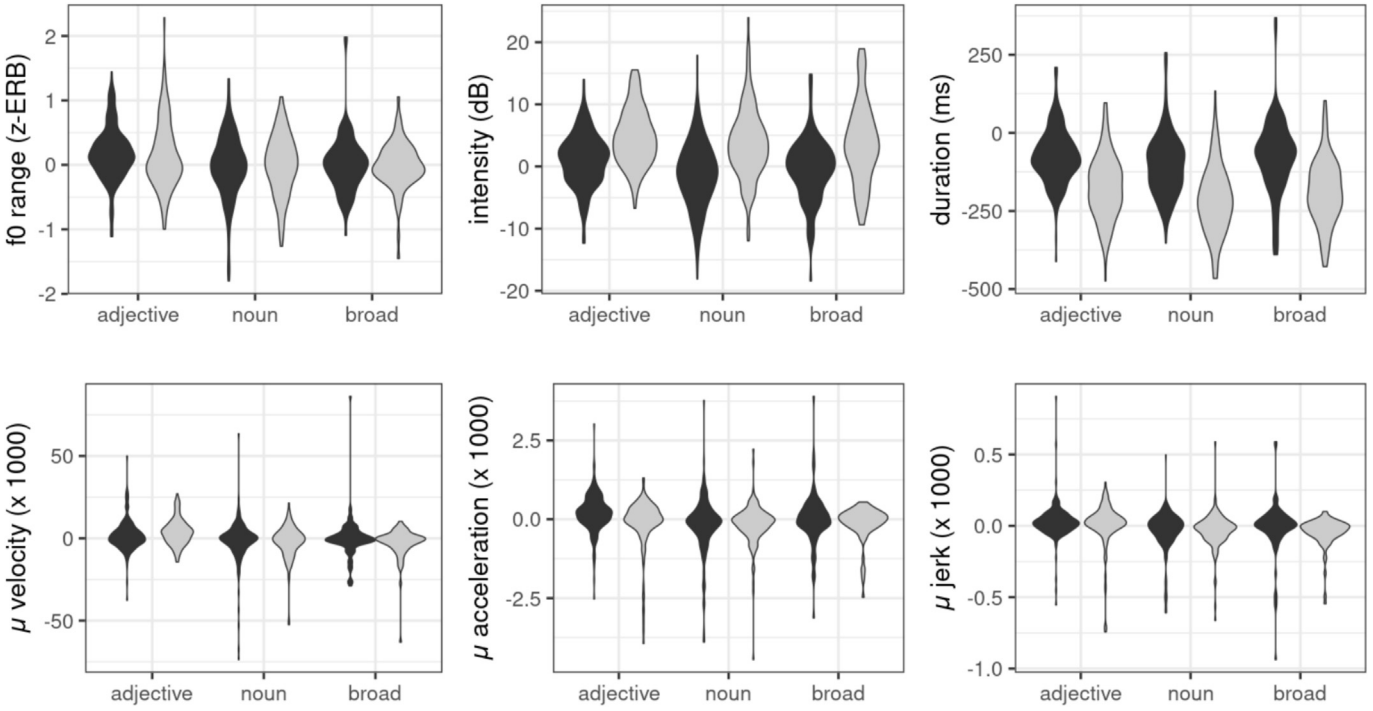

Fig. 3. Violin plots for each acoustic measure that was expressed as difference score (adjective minus noun) for each phrase position (black: medial, grey: final) and focus condition (adjective, noun, broad).

spanning between 0 and −0.5 z-ERB, with a slight rise towards the final syllable of the noun. (see (Fig. 3)).

The difference scores (Table 2) show a tendency for the values of adjective and noun focus to lie the farthest apart, with broad focus either patterning with noun focus (f0 range, intensity, and velocity and jerk in final phrase positions) or lying somewhat in between the two focus positions (acceleration, and velocity and jerk in medial phrase position). Duration showed similar values for adjective and broad focus, with noun focus standing out with more extreme values. The results indicate that f0 range was larger on the adjective when it was in focus, whereas little to no f0 range difference was found between adjective and noun for noun focus and broad focus. Intensity in medial phrase position was higher on the focused word than on the unfocused word, with the noun being slightly louder than the adjective in broad focus. Intensity in phrase-final position was overall higher on the adjective than on the noun, with the largest difference found for adjective focus. Duration indicated that the noun was generally longer than the adjective, hence the overall negative values, with the smallest difference observed for adjective focus in phrase-medial position and the largest difference observed for noun focus in phrase-final position. Velocity was the highest on the focused word, with broad focus showing a slightly higher veloc-

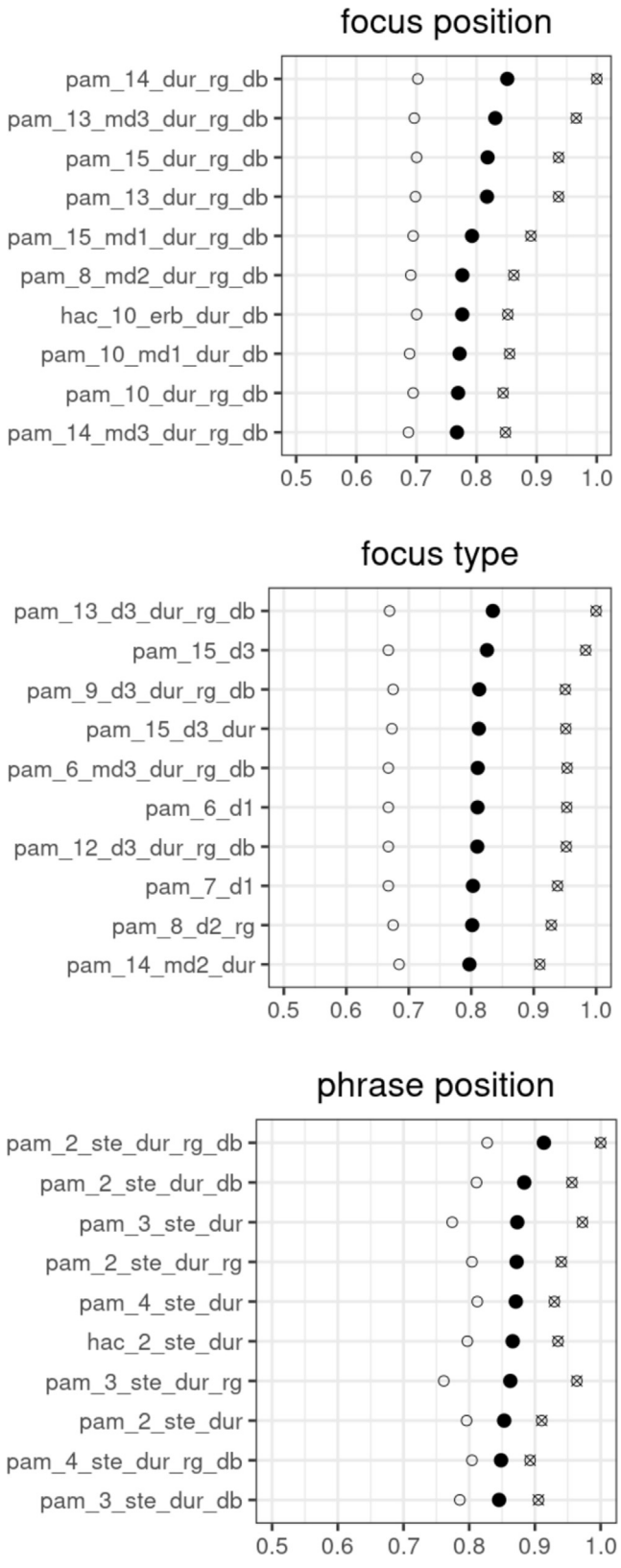

Fig. 4. Mean values for separation (crossed circles), grouping (unfilled circles) and their means (filled circles) for the 10 best performing cluster analyses referred to by their clustering method, number of clusters and combination of distance matrices in the format {method_Nclusters_mtx1_mtx2_mtx3_mtx4} for each of the meaningful features (top: focus position, mid: focus type, bottom: phrase position). See Appendix B for heatmaps of the best performing cluster analyses per meaningful feature.

ity on the noun than on the adjective in phrase-medial position (value close to zero), and a much higher velocity on the noun than on the adjective in phrase-final position (large negative value). Acceleration in phrase-medial position was the highest on the focused word, with broad focus showing almost equal acceleration rates. Acceleration in phrase-final position was always higher for nouns than for adjectives, with the smallest difference found for adjective focus. Jerk patterned with acceleration in that it was the highest for the focused word and almost equal for broad focus in phrase-medial position, whereas in phrase-final position, acceleration was almost equal for adjective and noun in adjective focus and higher for the noun in noun and broad focus.

3.2. Cluster evaluation

The separation and grouping values as well as their means are given in Fig. 4 and Fig. 5. Appendix B provides additional heatmaps of the best performing cluster analyses per meaningful feature. With regard to focus position (adjective or noun), the results show that among the best performing clusterings were the ones based on distance matrices that included duration, f0 range and intensity. Derivative difference scores had additional informativeness in some cases (e.g., jerk, velocity, and acceleration). The range of numbers of clusters for the best performing focus position analyses range from 8 to 15 (maximum obtained for 11 clusters). A similar outcome is found for focus type, with the difference that derivatives as time series appeared more informative. In particular jerk and velocity without other cues were among the best performing ones. Cluster numbers for the best performing focus type analyses range from 6 to 15 (maximum obtained for 15 clusters). Phrase position was best analysed with matrices including the standardized ERB time series and duration, as well as f0 range and intensity information. Derivatives, either as time series or as difference score did not appear in the best performing analyses. The number of clusters range from 2 to 4 for phrase position (maximum obtained for 3 clusters).

When comparing the best performing distance matrices by their overall score (Table 4), the matrices in the top 10 virtually always include a derivative, either as time series or as difference score. The only exception concerns the distance matrix with duration, f0 range and intensity, which had the third best overall performance, and performed particularly well for phrase position. Duration and intensity appeared in all of the best performing distance matrices, which were combinations of three or four measures in total.

A comparison of the performance across the acoustic measures (Table 5) shows that the best performing ones were the three derivatives as time series, duration, intensity and standardized ERB. The derivative or f0 range difference scores and the Hertz or ERB time series showed lower overall performance.

Among the two clustering techniques, PAM had a higher overall score (μ = 0.60, sd = 0.07) than HAC (μ = 0.59, sd = 0.07). Further inspection of these scores showed that PAM outperformed HAC for lower numbers of clusters (N $\leqslant$9; μPAM = 0.61, μHAC = 0.57), whereas the techniques were

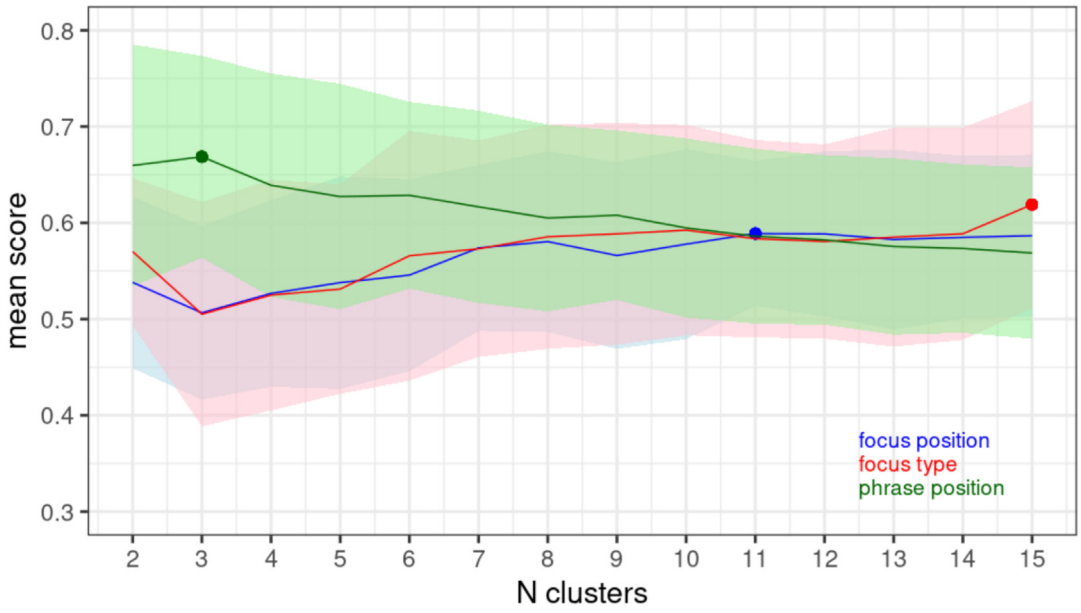

Fig. 5. Mean evaluation scores (averaged separation and grouping) for each meaningful feature (blue: focus position, red: focus type, green: phrase position) for all cluster rounds (2 to 15 clusters), light-colored shaded areas indicate standard deviations for each feature.

more similar for higher numbers of clusters ($N > 9$; μPAM = 0.59, μHAC = 0.60).

4. Discussion and conclusion

This study investigated the informativeness of f0 derivatives in automatic classification by means of cluster analyses on acoustic measures taken from Dutch noun phrases. The prosody of these noun phrases expressed three meaningful aspects: focus position, focus type and phrase position. These aspects will be discussed one by one in the following, after which a general discussion and conclusion is given.

4.1. Focus position

Focus position, the difference between adjective and noun in narrow focus, was expressed in the data by means of f0 (Fig. 2). This is done differently depending on phrase position. That is, in medial position, the focused word is marked with a rising movement, whereas in final positions, there is an interaction with the final low edge tone. This causes peak retraction to avoid tonal crowding, a phenomenon observed in Dutch (Caspers and Van Heuven, 1993; Schepman et al., 2006) as well as in other languages (e.g., Silverman and Pierrehumbert, 1987; Prieto et al., 1995; Myers, 2003; Gordon, 2008; Shue et al., 2010). The result of the retraction is an early peak to reach the high target in phrase-final adjective focus, such that the resulting f0 movement on the adjective is a fall. In fact, the fall continues until the first syllable of the noun, likely a result of deaccentuation of the noun (Krahmer and Swerts, 2001). For phrase-final noun focus, the accentuation of the noun appears from the sustained f0 on the adjective followed by a slight fall on the noun. The other acoustic measures generally show that adjective and noun focus pattern as opposites, with broad focus lying generally closer to noun focus (Table 3), as expected under the view that the position of the accent is the noun in both cases (e.g., Krahmer and Swerts, 2001).

The cluster results show that focus positions were most successfully separated and grouped by analyses that assumed a high number of clusters ($\geqslant 8$) using duration, f0 range and intensity in virtually all of the best performing matrices (Fig. 4, top). Although the derivative difference scores appeared in some of these matrices, they did not do so consistently. The difference in performance can be mainly ascribed to the separation quality as measured by the standardized residuals. The grouping quality did not differ much among the best performing matrices. The number of clusters at which the analyses performed best (N = 11) is indicative of a high degree of variation in the use of prosodic cues. This variation is best captured using difference scores from all three acoustic sources

Table 3
Overview of the mean values for each acoustic measure that was expressed as difference score (adjective minus noun) for each phrase position (medial, final) and focus condition (adjective, noun, broad). Derivative difference scores are multiplied by 1000 for readability.

Phrase pos.	Focus	f0 Range	Intensity	Duration	Velocity	Acceleration	Jerk
Medial	Adjective	0.18	1.06	−75.14	1.77	0.20	0.03
	Noun	−0.02	−1.25	−101.31	−1.63	−0.20	−0.03
	Broad	0.01	−0.56	−79.58	−0.41	−0.01	−0.01
Final	Adjective	0.16	4.76	−173.83	3.79	−0.12	0.01
	Noun	0.00	3.72	−216.08	−2.06	−0.16	−0.03
	Broad	−0.01	3.79	−187.80	−2.96	−0.13	−0.05

(f0, duration, intensity). Jerk (difference score) was ranked the highest among the best performing analyses that included derivative information.

4.2. Focus type

The acoustic differences signaling focus type are more subtle than those of focus position (Fig. 2). In phrase-final broad focus, the fall on the noun is minimally steeper than for narrow focus in that phrase position. This seems to be the only subtle difference in f0 between phrase-final broad and narrow noun focus. The same holds for phrase-medial broad and narrow noun focus, in which the marking is done with a steeper rise for noun focus than for broad focus. With the exception of duration difference (Table 3), all other acoustic measures showed that the two focus types (narrow and broad) are acoustically more similar to each other than the two focus positions.

Most likely because of the acoustic similarities between broad and noun focus, focus type appeared to be challenging for the cluster analysis. This can be seen from the trend that when assuming more clusters, the combined quality of separation and grouping keeps increasing (Fig. 5). This caused the best clustering for focus type to be found at 15 clusters, the maximum tested in this study. Due to computational cost, the maximum number of clusters was set to 15 in this study, although higher numbers are likely to have produced even higher clustering quality. This observation is indicative of a large amount of variation in the prosodic realization that needs to be accounted for by the clustering in order to obtain a high (er) quality separation and grouping for focus type. Fig. 4 (mid) furthermore supports this observation in that the best performing distance matrices varied considerably in the number of clusters. The only consistent acoustic measures found across all best performing matrices concerned f0 derivatives, i.e., jerk (time series and differences scores), velocity (time series) and acceleration (time series and difference scores). Note that relatively high and varying numbers of clusters could also be indicative of the subtlety of the effects of the derivatives among the other acoustic variation. F0 range, intensity and/or duration were among some of the best performing combined matrices, whereas time-series jerk and velocity appeared as single measures in the best performing matrices on three occasions (Fig. 4, mid).

4.3. Phrase position

Acoustic differences in phrase position mainly show up in the final syllable of the noun phrase. In medial positions, final f0 levels are well above the average f0 level of the speaker (between 0.25 and 0.50 z-ERB), whereas in final positions, final f0 levels are well below the speaker average (between −0.25 and −0.50 z-ERB), see Fig. 2. Phrase position furthermore appears to affect the overall f0 trajectory; rising in medial positions and falling in final positions (as discussed above). It seems, therefore, that phrase position not only has scope over a larger phonological domain (i.e., the phrase), but also affects a larger phonetic domain (i.e., overall f0 trend in the noun phrase), when compared to the effect of the respective focus conditions. This result is also supported in that the different focus conditions showed more extreme values for intensity, duration and velocity (Table 3) for phrase-final productions compared to phrase-medial ones. These cues together are indicative of domain-final prosodic phenomena related to articulatory muscle relaxation (e.g., Trouvain et al., 1998; Dubeda, 2006; Wagner and McAuliffe, 2019; Ots and Taremaa, 2023); that is, a drop in intensity, final lengthening and a falling f0.

Cluster analyses showed overall higher quality scores for phrase position, compared to the focus conditions (Fig. 4). There also appeared to be more variation in the individual scores of separation and grouping. In the best performing distance matrices, a consistent core set of acoustic measures can be observed. That is, all matrices consist at least of standardized ERB and duration difference measures. F0 range and intensity difference only appear in matrices with three or four measures, although matrices with only f0 range and intensity were among the tested ones (Appendix A). This final observation shows that f0 range and intensity only had additive informativeness to the clustering of phrase position differences.

4.4. Overall evaluations

Table 4 and Table 5 show the overall best performing distance matrices and acoustic cues respectively. As for the distance matrices, results show that derivative information was included in 9 out of 10 matrices (Table 4). Matrices with time-series measures of the derivatives showed better performance than the ones with derivative difference scores. This is expected from the idea that time series provide much more fine-grained f0 information. However, time series were largely blind to the division of noun phrases into units. This could have led to better performance of the difference scores, which took exactly the adjective-noun difference into account. Duration and intensity were included in all best performing matrices, indicating that these acoustic measures are highly informative. F0 range occurred in 6 out of 10 matrices. Note that the overall scores for the best performing distance matrices are the result of averaging over all three scores for the meaningful features. The separate scores for each meaningful feature in Fig. 4 therefore show different maxima than in Fig. 4. Nevertheless, similar trends can be observed for which (combined) matrices perform best. Except for focus position, all the best performing matrices of the individual features as well as the overall best performing one consisted of four acoustic measures, indicating that their combined informativeness generally exceeded the informativeness of matrices consisting of single measures. In addition, PAM clusterings were among the best performing ones in virtually all cases, suggesting that it generally outperforms HAC. HAC was found to perform similar to PAM for higher numbers of clusters only. It should also be noted that the overall scores of the two clustering techniques lie close to each other (Section 3.2).

As for the acoustic measures (Table 5), the highest overall score was obtained for jerk (time series), indicating that this acoustic measure was the most informative to the clustering of the meaningful features. The other time-series derivatives velocity and acceleration appeared among the top half of all the measures investigated in this study, indicating that they were also highly informative. Duration, intensity and standard-

Table 4
The 10 best performing distance matrices based on their mean score (overall and per meaningful feature) and standard deviation, in descending order of their overall mean score. Maximum scores for each column in boldface.

Distance matrix	μ Overall	μ f.Position	μ f.Type	μ phr.Position
d_d3_dur_rg_db	**0.69** (0.03)	0.70 (0.04)	**0.72** (0.08)	0.62 (0.06)
d_d1_dur_db	0.68 (0.02)	0.67 (0.05)	0.68 (0.05)	0.69 (0.03)
d_dur_rg_db	0.67 (0.02)	**0.73** (0.08)	0.61 (0.07)	0.65 (0.05)
d_d1_dur_rg_db	0.67 (0.02)	0.64 (0.05)	0.65 (0.13)	0.66 (0.03)
d_d2_dur_rg_db	0.67 (0.04)	0.61 (0.10)	0.67 (0.11)	0.62 (0.08)
d_d2_dur_db	0.67 (0.02)	0.61 (0.03)	0.69 (0.08)	**0.70** (0.05)
d_md3_dur_rg_db	0.66 (0.06)	0.70 (0.09)	0.55 (0.13)	0.67 (0.04)
d_md1_dur_db	0.66 (0.03)	0.64 (0.10)	0.60 (0.11)	0.66 (0.06)
d_md2_dur_db	0.65 (0.03)	0.64 (0.08)	0.64 (0.08)	0.66 (0.05)
d_md1_dur_rg_db	0.65 (0.03)	0.67 (0.08)	0.60 (0.09)	0.64 (0.07)

Table 5
Overview of the mean scores (and standard deviations) for each of the acoustic measures, overall (descending) and for each meaningful feature, based on scores for analyses in which the distance matrix consisted of the measure, either alone or in combination with other measures. Maximum scores for each column in boldface.

Measure	μ Overall	μ f.Position	μ f.Type	μ phr.Position
Jerk	**0.63** (0.07)	0.59 (0.10)	0.62 (0.11)	0.59 (0.09)
Duration	0.63 (0.05)	0.58 (0.09)	0.62 (0.10)	0.65 (0.06)
Velocity	0.62 (0.07)	0.57 (0.09)	0.61 (0.12)	0.61 (0.08)
Intensity	0.62 (0.06	**0.60** (0.09)	0.57 (0.11)	0.62 (0.08)
z-ERB	0.62 (0.05)	0.52 (0.09)	0.60 (0.10)	**0.73** (0.06)
Acceleration	0.61 (0.07)	0.59 (0.10)	**0.62** (0.11)	0.60 (0.09)
F0 range	0.60 (0.08)	0.57 (0.09)	0.57 (0.12)	0.59 (0.11)
Mean acceleration	0.60 (0.07)	0.58 (0.10)	0.56 (0.12)	0.59 (0.10)
Mean velocity	0.59 (0.08)	0.60 (0.09)	0.55 (0.11)	0.58 (0.11)
Mean jerk	0.58 (0.08)	0.58 (0.09)	0.55 (0.13)	0.57 (0.11)
Hertz	0.57 (0.05)	0.53 (0.08)	0.55 (0.10)	0.59 (0.07)
ERB	0.56 (0.05)	0.54 (0.09)	0.53 (0.10)	0.57 (0.08)

ized ERB were among the best performing ones too, outperforming the derivative difference scores as well as the Hertz and ERB time series.

4.5. General conclusion

This study showed that the inclusion of f0 dynamics is overall highly informative in the automatic classification of meaningful intonation. This was shown by an evaluation of focus position, focus type and phrase position, features that are commonly signaled by prosodic cues in Dutch. The results show that f0 derivatives are particularly informative for focus type, somewhat informative for focus position and minimally informative for phrase position. This nuanced result not only provides an insight into how automatic classification by means of cluster analysis may benefit from specific acoustic measures, but also into how speakers encode these meaningful features acoustically. That is, derivatives, in particular jerk, are abstract with regard to the variation found in the f0 contour they are derived from and express subtle acceleration rate differences. It appears that focus type, which indeed showed subtle f0 differences in the standardized ERB contours (Fig. 2), is signaled mostly by these jerk differences. As an analogy, jerk in train rides is comparable to abrupt acceleration and braking, which is generally avoided to ensure passenger comfort (i.e., jerk is kept within certain limits, e.g., Sharma and Chaturvedi, 2016). Jerk caused by a train's sudden braking typically attracts the attention of passengers, much more so than the gradually varying speeds during a long train ride. As for prosodic focus type differences expressed in the same phrase location such as the noun in the current study, jerk seems a particularly suitable signaler to attract perceptual attention, as jerk can be achieved in a short time-span and does not require movements elsewhere. In other words, speakers may use jerk if meaningful f0 differences are bound to a small time-window. While this conclusion explains the outcomes of the current study, it is in need of further production and perception testing (the latter is elaborated on below). It needs to be repeated here that the high number of clusters needed for better clustering performance is taken as indicative of the large amount of acoustic variation the analysis needs to deal with in order to arrive at one in which jerk is informative. In other words, jerk appears to be informative given that other prosodic variation is dealt with. It is not entirely clear what kind of other variation there could be in the current dataset. A likely option is that it originates from speaker differences in the degree to which focus type was marked, as these where not abstracted over in the standardized ERB contours.

It is furthermore interesting to observe that derivatives were not found in any of the best performing matrices for the clustering of phrase position. From all the time-series measures representing f0, standardized ERB performed best for phrase position. Thus, the clustering results reveal that medial and final phrase positions are signaled differently compared to the focus conditions, for which the derivatives played a more important role. Note that this observation is unsurprising given the articulatory dynamics at these different positions. Sustaining motor control when articulation continues (medial) or relaxing muscles when articulation stops (final) seem to involve different acoustic parameters then the focus marking on specific syllables. It appears that for phrase position, the f0 level relative to the speaker's range (z-ERB) matters most, whereas for focus marking, the acceleration rate at which a tonal target is reached is more telling. These conclusions fit with accounts

that have approached focus marking by means of f0 dynamics (e.g., Roessig et al., 2019; Iskarous et al., 2024) and ones that have modeled these processes as gestures (e.g., Saltzman and Munhall, 1989; Browman and Goldstein, 1990). Thus, it is too simplistic to maintain the view that a high tone that marks focus and a high tone that marks an edge are both intonational targets of the same kind. It seems that only for edge marking the actual f0 *level* plays a role, whereas focus marking is much more signaled by f0 *dynamics*, regardless of the level at which they occur (i.e., *gestures*). Thus, 'target' in the literal (level) sense only holds for edge tones, but not so much for focus marking. An important contribution of the current study lies therefore in the observation that the clustering of three meaningful features each had their own characteristics (i.e., optimal number of clusters and selection of acoustic cues) that are likely revealing of the intricate interplay of prosodic meaning and form. Given that the current dataset of Dutch noun phrases concerned scripted speech, more variation would be expected in a similar analysis of spontaneous data. Such spontaneous speech might pose a challenge for successful clustering that requires additional standardization/normalization steps that are yet to be investigated.

This study also shows that it is particularly helpful to break down the semantics of prosody into 'meaningful features'. That is, some of these features lend themselves better to clustering into a small number of groups (i.e., phrase position) than others (focus position/type). With respect to the question of categoricality (Section 1.1), these results seem to point out that some aspects of meaning are easier to put into categories than others. Above all, the quality of the categorization depends on the informativeness of the acoustic measures. It is likely that other acoustic features, or conversions of the measures taken into account in this study, improve classification even further. The underlying question for future research in this area is whether it is possible for cluster analysis to arrive at a (near-) perfect distinguishing of meaningful features just by adding more informative acoustic measures. It is still a possibility that the nature of some aspects of prosodic meaning is such that speakers and listeners do not maintain a strict separation as implied by the intonation categories that researchers have imposed to study them. Note that the role of perception research in this line of future research is invaluable. One way of verifying the outcomes of the current study would be to conduct a perception experiment with acoustically manipulated speech. The manipulated cues in such an experiment would be the best performing ones as found in this study, i.e., in order to test listeners' performance in distinguishing the meaningful features such as focus position, focus type and phrase position.

This study has mainly focused on the role of f0 derivatives, in accordance with the goals formulated in Section 1.6. However, it is important to note that non-f0 cues (duration and intensity in this study) are highly informative, too, and the best performing cluster analyses were based on combined acoustic measures. This result shows again that in the study of intonation, often exclusively related to f0 movements, much more can be gained from taking into account other aspects of the speech signal (see Arvaniti et al., 2024 for a similar conclusion). In this regard, it also needs to be acknowledged that the cluster analyses in this study had maximum grouping scores of around 0.70 (focus conditions) or 0.80 (phrase position), see Fig. 4, unfilled circles), showing that 20–30% of the data could not be clustered accurately. This shows an interesting difference with regard to classification studies on lexical tone, which had (near) perfect performance (e.g., Gauthier et al., 2007). It appears that the kind of meaningful feature under analysis is crucial for predicting the classification performance. If freedom in f0 variation is more likely to be harmful for communicating meaning, as in the case of Mandarin Chinese lexical tone, speakers will produce more distinct prosody and automatic classification is likely to be optimal. In the case of focus marking in the current study, f0 variation is less harmful for communication, possibly because other acoustic cues play an important role, too. In this sense, form and meaning are more loosely related in Dutch focus marking than in Mandarin lexical tone. In a similar vein, higher accuracy observed for phrase position in the current study confirms its importance in speech communication. Thus, speakers are generally clearer in signaling phrase ends, with less ambiguous cases in between, than in signaling focus. This could explain the higher clustering performance rates for phrase position in the current study.

CRediT authorship contribution statement

Constantijn Kaland: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was funded by the German Research Foundation (DFG) – Project-ID 281511265 and 559664412. The author thanks Katherine Walker for proofreading.

Appendices A and B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.wocn.2025.101454.

References

Ahn, B., Veilleux, N., & Shattuck-Hufnagel, S. (2019). Annotating prosody with PoLaR: Conventions for a decompositional annotation system. In Sasha Calhoun, Paola Escudero, & Marija Tabain (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 1302–1306). Melbourne, Australia: Australasian Speech Science and Technology Association Inc..

Albert, A., Cangemi, F., & Grice, M. (2018). Using periodic energy to enrich acoustic representations of pitch in speech: A demonstration. In Speech Prosody 2018 (pp. 804–808). ISCA. doi:10.21437/SpeechProsody.2018- 162.

Albert, A., Kaland, C., Ellison, T. M., Cangemi, F., Winter, B., & Grice, M. (2024). Harvesting spontaneous speech data from digital reservoirs to study prosody. In *In 19th Conference on Laboratory Phonology (LabPhon 19). Seoul, Korea.* .

Arvaniti, A., Katsika, A., & Hu, N. (2024). Variability, overlap, and cue trading in intonation. *Language, 100*(2), 265–307. https://doi.org/10.1353/lan.2024.a929737.

Babinski, S. (2022). *Archival Phonetics & Prosodic Typology in Sixteen Australian Languages* PhD thesis. New Haven, Connecticut, USA: Yale University.

Barnes, J., Brugos, A., Veilleux, N., & Shattuck-Hufnagel, S. (2021). On (and off) ramps in intonational phonology: Rises, falls, and the Tonal Center of Gravity. *Journal of Phonetics, 85*, 101020. https://doi.org/10.1016/j.wocn.2020.101020.

Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2012). Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology, 3*(2), 337–383. https://doi.org/10.1515/lp- 2012-0017.

Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The Original ToBi System and the Evolution of the ToBi Framework. *In Prosodic Typology. Oxford: Oxford University Press.*. https://doi.org/10.1093/acprof:oso/9780199249633.003.0002.

Björklund, A. (2024). Automatic intonational contour clustering in Patwin. *Proceedings of the Linguistic Society of America, 9*(1), 5713. https://doi.org/10.3765/plsa.v9i1.5713.

Borchers, H.W. (2022). Pracma: Practical numerical math functions. retrieved from https://CRAN.R- pro ject.org/package=pracma.

Browman, C. P., & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics, 18*(3), 299–320. https://doi.org/10.1016/S0095- 4470(19)30376- 6.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica, 49*(3–4), 155–180. https://doi.org/10.1159/000261913.

Caspers, J., & Van Heuven, V. (1993). Effects of Time Pressure on the Phonetic Realization of the Dutch Accent-Lending Pitch Rise and Fall. *Phonetica, 50*(3), 161–171. https://doi.org/10.1159/000261936.

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), 1–51. https://doi.org/10.16995/labphon.9437.

Comini, G., Huybrechts, G., Ribeiro, M.S., Gabryś, A., & Lorenzo-Trueba, J. (2022). Low-data? No problem: Low-resource, language-agnostic conversational text-to-speech via F0-conditioned data augmentation. In Interspeech 2022 (pp. 1946–1950). ISCA. doi:10.21437/Interspeech.2022- 10338.

Dubeda, T. (2006). Intensity as a macroprosodic variable in Czech. Speech Prosody 2006, paper 016–. https://doi.org/10.21437/SpeechProsody.2006-59.

Fujimura, O. (1990). Toward a model of articulatory control: Comments on Browman and Goldstein's paper. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology* ((1st ed.,, pp. 377–381). Cambridge University Press. https://doi.org/10.1017/CBO9780511627736.020.

Gafos, A. I., & Benus, S. (2006). Dynamics of phonological cognition. *Cognitive Science, 30*(5), 905–943. https://doi.org/10.1207/s15516709cog0000_80.

Gauthier, B., & Shi, R. (2011). A connectionist study on the role of pitch in infant-directed speech. *The Journal of the Acoustical Society of America, 130*(6). https://doi.org/10.1121/1.3653546. EL380-EL386.

Gauthier, B., Shi, R., & Xu, Y. (2007). Learning phonetic categories by tracking movements. *Cognition, 103*(1), 80–106. https://doi.org/10.1016/j.cognition.2006.03.002.

Gauthier, B., Shi, R., & Xu, Y. (2007). Simulating the acquisition of lexical tones from continuous dynamic input. *The Journal of the Acoustical Society of America, 121*(5). https://doi.org/10.1121/1.2716160. EL190-EL195.

Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research, 47*(1–2), 103–138. https://doi.org/10.1016/0378-5955(90)90170- T.

Gordon, M. (2008). Pitch accent timing and scaling in Chickasaw. *Journal of Phonetics, 36*(3), 521–535. https://doi.org/10.1016/j.wocn.2006.10.003.

Grice, M., Ritter, S., Niemann, H., & Roettger, T. B. (2017). Integrating the discreteness and continuity of intonational categories. *Journal of Phonetics, 64*, 90–107. https://doi.org/10.1016/j.wocn.2017.03.003.

Gu, W., Hirose, K., & Fujisaki, H. (2004). Automatic extraction of tone command parameters for the model of F 0 contour generation for standard Chinese. IEICE Transactions on Information and Systems 87 (5), 1079–1085.

Hakim, J. (2024). Using role-playing tasks to document intonational tune prototypes in Nasal, an endangered language of Sumatra. In Speech Prosody 2024 (pp. 1175–1179). ISCA. doi:10.21437/SpeechProsody.2024- 237.

Hart, J. '., Collier, R., & Cohen, A. (1990). A perceptual study of intonation: An experimental-phonetic approach to speech melody. OCLC: 708567537. Cambridge, UK: Cambridge University Press.

Hasan, M., Doddipatla, R., & Hain, T. (2014). Multi-pass sentence-end detection of lecture speech. In Interspeech 2014 (pp. 2902–2906). ISCA. doi:10.21437/Interspeech.2014- 602.

Henrich, N., d'Alessandro, C., Doval, B., & Castellengo, M. (2004). On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *The Journal of the Acoustical Society of America, 115*(3), 1321–1332. https://doi.org/10.1121/1.1646401.

Herman, E., & Strang, G. (2016). *Calculus:* (volume 1) Houston, Texas: OpenStax, Rice University.

Hermes, D. J., & Van Gestel, J. C. (1991). The frequency scale of speech intonation. *The Journal of the Acoustical Society of America, 90*(1), 97–102. https://doi.org/10.1121/1.402397.

Hogan, N. (1984). An organizing principle for a class of voluntary movements. *The Journal of Neuroscience, 4*(11), 2745–2754. https://doi.org/10.1523/JNEUROSCI.04-11-02745.1984.

Iskarous, K. (2017). The relation between the continuous and the discrete: A note on the first principles of speech dynamics. *Journal of Phonetics, 64*, 8–20. https://doi.org/10.1016/j.wocn.2017.05.003.

Iskarous, K., Cole, J., & Steffman, J. (2024). A minimal dynamical model of Intonation: Tone contrast, alignment, and scaling of American English pitch accents as emergent properties. *Journal of Phonetics, 104*, 101309. https://doi.org/10.1016/j.wocn.2024.101309.

Jabeen, F. & Wagner, P. (2023). Variability in hesitations in Punjabi semi-spontaneous narrative speech: An automatic clustering based analysis. In Disuency in Spontaneous Speech (DiSS) Workshop 2023 (pp. 71–75). ISCA. doi:10.21437/DiSS.2023- 15.

Jeon, H.-S., Kaland, C., & Grice, M. (2024). Cluster analysis of Korean IP-final intonation. In Speech Prosody 2024 (pp. 1025–1029). ISCA. doi:10.21437/SpeechProsody.2024- 207.

Jeon, H.-S., Kaland, C., & Grice, M. (2025). Question intonation in conversational speech: Chungcheong and gyeongsang varieties of korean. *The Journal of the Acoustical Society of America, 158*(1), 684–696. https://doi.org/10.1121/10.0037191.

Jun, S.-A. (Ed.). (2014). Prosodic typology II: The phonology of intonation and phrasing. Oxford linguistics. Oxford: Oxford University Press. retrieved from https: // academic. oup.com/book/27198

Kaland, C. (2014). Prosodic marking of semantic contrasts: Do speakers adapt to addressees? LOT Dissertation Series. LOT Dissertation Series 366 - Tilburg University.

Kaland, C. (2021). Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours. *Journal of the International Phonetic Association, 53*(1), 159–188. https://doi.org/10.1017/S0025100321000049.

Kaland, C. (2023). Intonation contour similarity: f0 representations and distance measures compared to human perception in two languages. *The Journal of the Acoustical Society of America, 154*(1), 95–107. https://doi.org/10.1121/10.0019850.

Kaland, C., & Grice, M. (2024). Exploring and explaining variation in phrase-final f0 movements in spontaneous Papuan Malay. *Phonetica, 81*(3). https://doi.org/10.1515/phon- 2023- 0031.

Kaland, C., Swerts, M., & Himmelmann, N. P. (2023). Red and blue bananas: Time-series f0 analysis of contrastively focused noun phrases in Papuan Malay and Dutch. *Journal of Phonetics, 96*, 101200. https://doi.org/10.1016/j.wocn.2022.101200.

Karlin, R., & Tilsen, S. (2015). *The articulatory tone-bearing unit: Gestural coordination of lexical tone in Thai. Providence, Rhode Island*. https://doi.org/10.1121/2.0000089.

Katsika, A. (2016). The role of prominence in determining the scope of boundary-related lengthening in Greek. *Journal of Phonetics, 55*, 149–181. https://doi.org/10.1016/j.wocn.2015.12.003.

Kaufman, L. & Rousseeuw, P.J. (Eds.). (1990). Finding Groups in Data. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/9780470316801.

Krahmer, E., & Swerts, M. (2001). On the alleged existence of contrastive accents. *Speech Communication, 34*(4), 391–405. https://doi.org/10.1016/S0167- 6393(00)00058- 3.

Ladd, D. R. (2022). The Trouble with ToBI. In J. Barnes & S. Shattuck-Hufnagel (Eds.), *Prosodic Theory and Practice* (pp. 247–258). The MIT Press.

Ladd, D.R. (2008). Intonational phonology (2nd ed). Cambridge studies in linguistics. Cambridge; New York: Cambridge University Press.

Laméris, T. J., Li, K. K., & Post, B. (2023). Phonetic and phono-lexical accuracy of non-native tone production by English-L1 and Mandarin-L1 Speakers. *Language and Speech, 66*(4), 974–1006. https://doi.org/10.1177/00238309221143719.

Le Roux, J., Kameoka, H., Ono, N., De Cheveigne, A., & Sagayama, S. (2007). Harmonic-temporal clustering of speech for single and multiple F0 contour estimation in noisy environments. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, IV–1053–IV–1056. https: // doi. org/ 10. 1109/ ICASSP. 2007.367254.

Liu, S., Doyle, S., Morris, A., & Ehsani, F. (1998). The effect of fundamental frequency on Mandarin speech recognition. 5th International Conference on Spoken Language Processing (ICSLP 1998), paper 0847–. https://doi.org/10.21437/ICSLP.1998- 761.

Mahr, T. (2020). Readtextgrid: Read in a 'Praat' 'TextGrid' File. *Institution: Comprehensive R Archive Network Pages, 0.1.2*. https://doi.org/10.32614/CRAN.package.readtextgrid.

Mixdorff, H. & Niebuhr, O. (2013). The inuence of F0 contour continuity on prominence perception. In Interspeech 2013 (pp. 230–234). ISCA. doi:10.21437/Interspeech.2013- 73.

Myers, S. (2003). F 0 Timing in Kinyarwanda. *Phonetica, 60*(2), 71–97. https://doi.org/10.1159/000071448.

Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability & Its Applications, 9*(1), 141–142. https://doi.org/10.1137/1109020.

Narusawa, S., Minematsu, N., Hirose, K., & Fujisaki, H. (2004). Evaluation of an improved method for automatic extraction of model parameters from fundamental frequency contours of speech. In *In Speech Prosody 2004. ISCA* (pp. 443–446). https://doi.org/10.21437/SpeechProsody.2004-101.

Ots, N., & Taremaa, P. (2023). Chunking an unfamiliar language: Results from a perception study of German listeners. In F. Schubö, S. Zerbian, S. Hanne, & I. Wartenburger (Eds.), *Prosodic boundary phenomena* (pp. 87–117). Berlin: Language Science Press. https://doi.org/10.5281/zenodo.7777530.

Perkell, J.S., Klatt, D.H., Stevens, K.N., & Massachusetts Institute of Technology (Eds.). (1986). Invariance and variability in speech processes. Proceedings of the Symposium on Invariance and Variability of Speech Processes met at M.I.T. Meeting Name: Symposium on Invariance and Variability of Speech Processes. Hillsdale, N.J.: Erlbaum.

Pierrehumbert, J., & Hirschberg, J. (1990). The Meaning of Intonational Contours in the Interpretation of Discourse. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in Communication*. https://doi.org/10.7916/d8kd24fp.

Pierrehumbert, J., & Pierrehumbert, R. (1990). On attributing grammars to dynamical systems. *Journal of Phonetics, 18*(3), 465–477. https://doi.org/10.1016/S0095- 4470(19)30374- 2.

Prieto, P., Van Santen, J., & Hirschberg, J. (1995). Tonal alignment patterns in Spanish. *Journal of Phonetics, 23*(4), 429–451. https://doi.org/10.1006/jpho.1995.0032.

R Core Team. (2022). R: the R pro ject for statistical computing. Version 4.2.1. retrieved from https://www.r- pro ject.org/.
R Studio Team. (2022). RStudio: Integrated Development for R. RStudio, Inc. retrieved urlyearurlmonthurlday, from https://www.rstudio.com/.
Rahman, M.S. & Shimamura, T. (2006). Speech Analysis Based on Modeling the Effective Voice Source. IEICE Transactions on Information and Systems, E89-D (3), 1107–1115. doi:10.1093/ietisy/e89- d.3.1107.
Ritter, S., & Grice, M. (2015). The role of tonal onglides in German Nuclear Pitch Accents. *Language and Speech, 58*(1), 114–128. https://doi.org/10.1177/0023830914565688.
Roessig, S., Mücke, D., & Grice, M. (2019). The dynamics of intonation: Categorical and continuous variation in an attractor-based model. *PLOS ONE, 14*(5), e0216859. https://doi.org/10.1371/journal.pone.0216859.
Ross, E. D., Edmondson, J. A., & Seibert, G. B. (1986). The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice. *Journal of Phonetics, 14*(2), 283–302. https://doi.org/10.1016/S0095- 4470(19)30669-2.
Saltzman, E. L., & Munhall, K. G. (1989). A Dynamical Approach to Gestural Patterning in Speech Production. *Ecological Psychology, 1*(4), 333–382. https://doi.org/10.1207/s15326969eco0104_2.
Scheffers, M. T. M. (1983). Simulation of auditory analysis of pitch: An elaboration on the DWS pitch meter. *The Journal of the Acoustical Society of America, 74*(6), 1716–1725. https://doi.org/10.1121/1.390280.
Schepman, A., Lickley, R., & Ladd, D. R. (2006). Effects of vowel length and right context on the alignment of Dutch nuclear accents. *Journal of Phonetics, 34*(1), 1–28. https://doi.org/10.1016/j.wocn.2005.01.004.
Scitovski, R., Sabo, K., Martínez-Álvarez, F., & Ungar, S. (2021). *Cluster analysis and applications*. Cham, Switzerland: Springer Nature. https://doi.org/10.1007/978- 3-030- 74552- 3.
Seeliger, H. & Kaland, C. (2022). Boundary tones in German wh-questions and wh-exclamatives - a cluster-based approach. In S. Frota & M. Vigario (Eds.), Proceedings of the 11th International Conference on Speech Prosody 2022 (pp. 27–31). Lisbon, Portugal. doi:10.21437/SpeechProsody.2022- 6.
Seeliger, H., Lützeler, A., & Kaland, C. (2023). The perception of German wh-phrase-final intonation: A contour clustering evaluation. In The 2nd International Conference on Tone and Intonation (TAI 2023) (pp. 10–14). Singapore. doi:10.21437/TAI.2023- 3.
Sharma, S. K., & Chaturvedi, S. (2016). Jerk analysis in rail vehicle dynamics. *Perspectives in Science, 8*, 648–650. https://doi.org/10.1016/j.pisc.2016.06.047.
Shue, Y.-L., Shattuck-Hufnagel, S., Iseli, M., Jun, S.-A., Veilleux, N., & Alwan, A. (2010). On the acoustic correlates of high and low nuclear pitch accents in American English. *Speech Communication, 52*(2), 106–122. https://doi.org/10.1016/j.specom.2009.08.005.
Silva, W.D., Barbosa, P.A., & Abelin, Å. (2016). Cross-cultural and cross-linguistic perception of authentic emotions through speech: An acoustic-phonetic study with Brazilian and Swedish listeners. DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada, 32 (2), 449–480. doi:10.1590/0102- 445003263701432483.
Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. *In Second international conference on spoken language processing. Banff: ISCA.*. https://doi.org/10.21437/ICSLP.1992- 260.
Silverman, K., & Pierrehumbert, J. (1987). The timing of prenuclear high accents in English. *The Journal of the Acoustical Society of America, 82*(S1). https://doi.org/10.1121/1.2024693. S19-S19.
Steffman, J., Cole, J., & Shattuck-Hufnagel, S. (2024). Intonational categories and continua in American English rising nuclear tunes. *Journal of Phonetics, 104*, 101310. https://doi.org/10.1016/j.wocn.2024.101310.
Sun, Q., Hirose, K., & Minematsu, N. (2012). A method for generation of Mandarin F0 contours based on tone nucleus model and superpositional model. *Speech Communication, 54*(8), 932–945. https://doi.org/10.1016/j.specom.2012.03.005.
Swerts, M., Krahmer, E., & Avesani, C. (2002). Prosodic marking of information status in Dutch and Italian: a comparative analysis. *Journal of Phonetics, 30*(4), 629–654. https://doi.org/10.1006/jpho.2002.0178.
Tatár, C., Brennan, J., Krivokapić, J., & Keshet, E. (2024). Examining melodiousness in sarcasm: Wiggliness, spaciousness, and contour clustering. In Speech Prosody 2024 (pp. 677–681). ISCA. doi:10.21437/SpeechProsody.2024- 137.
The MathWorks Inc. (2025). MATLAB documentation - gradient. Natick, Massachusetts. retrieved from https://www.mathworks.com/help/matlab/ref/gradient.html
Trouvain, J., Barry, W.J., Nielsen, C., & Andersen, O.K. (1998). Implications of energy declination for speech synthesis. In E. (ed.) Mike (Ed.), Speech synthesis: Proceedings of the 3rd ESCA/COCOSDA workshop on speech synthesis, jenolan caves, australia, november 1998 (pp. 47–52).
Tuller, B., Case, P., Ding, M., & Kelso, J. A. S. (1994). The nonlinear dynamics of speech categorization. *Journal of Experimental Psychology: Human Perception and Performance, 20*(1), 3–16. https://doi.org/10.1037/0096- 1523.20.1.3.
Tuller, B., & Kelso, J. A. S. (1991). The Production and Perception of Syllable Structure. *Journal of Speech, Language, and Hearing Research, 34*(3), 501–508. https://doi.org/10.1044/jshr.3403.501.
Tuller, B., & Kelso, J. A. S. (1989). Phase transitions in speech production and their perceptual consequences. *The Journal of the Acoustical Society of America, 86*(S1). https://doi.org/10.1121/1.2027310. S114-S114.
Veilleux, N., Ahn, B., Brugos, A., Jeong, S., & Shattuck-Hufnagel, S. (2023). Methods for PoLaR Exploration with Machine Learning: Grammatical Analysis of Intonation without Grammatical Labels. In R. Skarnitzl (Ed.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 1648–1652). Prague (CZ): Guarant International.
Wagner, M., & McAuliffe, M. (2019). The effect of focus prominence on phrasing. *Journal of Phonetics, 77*, 100930. https://doi.org/10.1016/j.wocn.2019.100930.
Watson, D. (2010). The Many Roads to Prominence. In Psychology of Learning and Motivation (Vol. 52, pp. 163–183). Elsevier. doi:10.1016/S0079- 7421(10)52004- 8.
Watson, D., Tanenhaus, M., & Gunlogson, C. (2008). Interpreting Pitch Accents in Online Comprehension: H* vs. L+H*. *Cognitive Science: A Multidisciplinary Journal, 32*(7), 1232–1244. https://doi.org/10.1080/03640210802138755.
Watson, G. S. (1964). Smooth regression analysis. *Sankhy: The Indian Journal of Statistics, Series A*, 359–372.
Wickham, H. & Seidel, D. (2022). Scales: Scale Functions for Visualization. retrieved from https://CRAN.R- pro ject.org/package=scales.
Winkelmann, R., Bombien, L., Scheffers, M., & Jochim, M. (2023). Wrassp: Interface to the 'ASSP' library. retrieved from https://CRAN.R- pro ject.org/package=wrassp.
Xu, Y. (1998). Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica, 55*(4), 179–203. https://doi.org/10.1159/000028432.
Xu, Y. (1999). F0 peak delay: When, where, and why it occurs. In *In The 14th International Congress of Phonetic Sciences. San Francisco* (pp. 1881–1884).
Xu, Y., & Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *The Journal of the Acoustical Society of America, 111*(3), 1399–1413. https://doi.org/10.1121/1.1445789.
Xu, Y., & Wang, M. (2009). Organizing syllables into groups—Evidence from F0 and duration patterns in Mandarin. *Journal of Phonetics, 37*(4), 502–520. https://doi.org/10.1016/j.wocn.2009.08.003.
Xu, Y., & Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication, 33*(4), 319–337. https://doi.org/10.1016/S0167-6393(00)00063- 7.
Zahrer, A. (2024). Exploring natural speech intonation of an under-researched Papuan language. In Speech Prosody 2024 (pp. 1095–1099). ISCA. doi:10.21437/SpeechProsody.2024- 221.
Zhang, S. (2016). Mining linguistic tone patterns with symbolic representation. In Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology (pp. 1–9). Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/W16- 2001.