JASA ARTICLE

Intonation contour similarity: *f*0 representations and distance measures compared to human perception in two languages

Constantijn Kaland[a)]
Institute of Linguistics, University of Cologne, Cologne, Germany

ABSTRACT:
Recently, cluster analysis on *f*0 contours has become a popular method in phonetic research. Cluster analysis provides an automated way of categorising *f*0 contours, which gives new insights into (phonological) categories of intonation that vary across languages. As cluster analysis can be performed in many different ways, it is important to understand the extent to which these analyses can capture human perception of *f*0. This study focuses on the way in which *f*0 contours and differences among them are represented numerically, i.e., a crucial methodological choice preceding cluster analysis. These representations are then compared to the way in which *f*0 contour differences are perceived by human listeners from two different languages. To this end, four time-series contour representations (equivalent rectangular bandwidth, standardisation, octave-median rescaling, first derivative) and three distance measures [Euclidean distance (L2 norm), Pearson correlation, and dynamic time warping) were tested. The perceived differences were obtained from listeners of German and Papuan Malay, two typologically different languages. Results show that computed contour differences reflect human perception moderately, with dynamic time warping applied to the first derivative of the contour performing best, and showing minimal differences between the languages. © *2023 Acoustical Society of America*. https://doi.org/10.1121/10.0019850

(Received 24 February 2023; revised 6 June 2023; accepted 7 June 2023; published online 6 July 2023)
[Editor: James F. Lynch] Pages: 95–107

I. INTRODUCTION

At the core of our understanding of prosody and intonation lies the categorisation of their components. Intonation has often been modelled by means of phonological categories of pitch accents and boundary tones (i.e., autosegmental-metrical models such as Tones and Break Indices, ToBI; e.g., Silverman *et al.*, 1992). The intonation of a considerable number of languages has been analysed using this model (e.g., Jun, 2005, 2014). A central assumption of these analyses is that intonation contours are composed of "building blocks" of high (H) and low (L) tones that combine into f0 movements with specific shapes (pitch accents and boundary tones), which themselves can be combined into tunes that have a particular meaning (Pierrehumbert and Hirschberg, 1990). In these approaches, languages are assumed to exhibit an inventory of pitch accents, comparable to the inventory of phonemes.

It is a challenge to understand how human speakers and listeners handle the assumed tonal categories that underlie the intonation of their language. It has been shown, for example, that two different pitch accents do not always have entirely different meanings and that one pitch accent might have multiple meanings (e.g., Féry and Stoel, 2006; Watson *et al.*, 2008). Yet, research has also shown that perceived meaning distinctions between pitch accent categories can be attributed to single phonetic cues (e.g., Ritter and Grice, 2015). To advance our understanding of how exactly phonetic differences between *f*0 contours relate to underlying phonological categories, research has adopted automatic classification techniques such as cluster analysis. Alternative approaches have used trained classifiers [e.g., Cole *et al.*, 2022 or Tonal Center of Gravity (TCoG); e.g., Barnes *et al.*, 2012; Albert *et al.*, 2018] as ways to distinguish pitch accents.

Clustering has been applied in intonation research to both the production and to the perception of *f*0 contours since the seventies (e.g., Collier, 1975). As clustering techniques find their applications in many fields of research, there is a large variety of methodological choices to make in order to tailor clustering to the specific data at hand (e.g., Kaufman and Rousseeuw, 1990). The cluster analyses that have been used in intonation research, therefore, vary to a large extent. We can distinguish two main aspects that affect the performance of the cluster analysis; (1) the representation of the data that needs to be clustered and (2) the type of cluster analysis. While the second aspect is planned for future studies, this paper covers a selected number of ways in which intonation contours can be represented *before* applying cluster analysis. As further outlined in the following sections, this aspect has not been covered in clustering approaches to intonation, although it provides a crucial step to understanding the extent to which cluster analysis is able to resemble human perception and therefore the extent to which cluster analysis is useful to explore intonation categories.

The next section provides a brief introduction to the main aspects of cluster analysis relevant to this study (Sec. I A). Thereafter, studies that applied cluster analysis to *f*0

[a)]Electronic mail: ckaland@uni-koeln.de

contours are discussed for the way they represented the contours (Sec. I B) and how differences between the contours have been calculated (Sec. I C). The final section summarises the aims and research questions of this study (Sec. I E).

A. Cluster analysis

Cluster analysis is a classification technique that groups numerically similar data(points). The desired outcome is a certain number of groups (clusters) between which there is dissimilarity and within which there is similarity. A key problem in cluster analysis is finding the ideal number of clusters. It is generally agreed upon that the optimum lies at the number of clusters for which the between-cluster variation is maximal and the within-cluster variation is minimal. The two most commonly applied clustering techniques that are relevant for the current overview are *k-means* clustering and *hierarchical* clustering. These types of clustering distinguish themselves in the way clusters are formed and in their need for deciding on the number of clusters. K-means clustering requires setting a number of clusters before clustering is performed. The algorithm first assigns each observation to a cluster (randomly) after which it updates the assignment until the means of each cluster (centroids) do not change anymore (convergence).

Hierarchical clustering can be performed without setting the number of clusters. The clustering is performed by iteratively merging (bottom-up; agglomerative) or splitting (top-down; divisive) clusters. The initial or final state of the clustering is when each observation forms a cluster or when all observations are in one cluster (depending on direction). The output of hierarchical clustering is a tree-structure (dendrogram) from which any number of clusters can be obtained by choosing a particular height in the tree. Hierarchical clustering requires a criterion on the basis of which the clusters are formed (linkage criterion). This criterion determines how the distance between sets of observations is calculated. Thus, for expressing the distance between individual observations, the *distance measure* is taken. However, between sets of observations distances are expressed by a calculation over all the distance measures in the sets. This can be done by taking the maximum (complete linkage), minimum (single linkage), (un)weighted average distance (UPGMA, WPGMA), centroid (UPGMC), median (WPGMC), or minimum increase in sum of squares (Ward). There are more types of cluster analysis and more linkage criteria, which will not be discussed here.

While the type of cluster analysis and choice of linkage criterion affects how the clustering is performed, the way the data is represented is equally crucial to the usefulness of the clustering outcome. In the case of intonation contours, two main representational aspects are covered in this study; contour representation and distance measures. The next section discusses how previous intonation research has dealt with these aspects.

B. Cluster analysis in intonation research: Contour representation

Intonation studies that applied cluster analysis (overview in Table I) can be roughly divided into two lines of research, depending on whether mainly production or mainly perception data were analysed. Note that some studies covered both, however, with primary focus on one of them (e.g., Cole and Steffman, 2021). Perception research performed cluster analysis on data obtained from tasks in which listeners had to group utterances with similar sounding intonation patterns (Collier, 1975; Collier, 1977; Odé, 1989). Listeners could decide themselves how many groups they formed. The distance measure was taken on the basis of how often the intonation patterns were grouped together (counts). The resulting distance matrix was then the input for hierarchical clusterings with either single or complete linkage as criteria. This group of studies is unique in that it does not perform cluster analysis directly on acoustic measures, as was done in the production studies. The clustering results were compared to acoustic properties of the contours in the utterances. It was found, for example, that in Dutch a timing continuum of a rising *f*0 movement over 12 items was grouped by listeners into three clusters, corresponding to three prototypical intonation contours in Dutch (Collier, 1975). As for Russian, the perceptual grouping of rising contours was done based on their excursion size, posttonic *f*0 level, and timing of the slope (Odé, 1989).

Production studies that applied cluster analysis to intonation can furthermore be divided into ones that represented *f*0 contours by (multiple) parameters derived from acoustic measures of the *f*0 contour and ones that represented intonation contours by time-series measures of *f*0 (Table I). As for the first, common parameters to describe an intonation contour are (starting/mean) level, range, and slope (e.g., Demenko and Wagner, 2006; Levow, 2006; Hirschberg and Rosenberg, 2007). More complex and integral approaches include Parametrisation of Intonation Events (PaIntE) (Möhler and Conkie, 1998, used in Calhoun and Schweitzer, 2012) and a model that combines a Contour-based, Parametric, and Superpositional approach to intonation (CoPaSul; Reichel, 2011). As it is beyond the scope of this overview to go into further detail concerning the models, only their contour representation methods are discussed here. Both models represent *f*0 contours using (a selection of) the previously mentioned parameters and additional temporal measures (alignment, domain) to reconstruct the original contour in a highly naturalistic way. This was confirmed by acceptability/natural ratings obtained in perception tests (Demenko and Wagner, 2006; Reichel, 2011).

Time-series *f*0 measures provide another way to represent the *f*0 contour. This method is adopted in the current study and is essentially different from the parametric one as the contour is represented by a vector of *f*0 measures in chronological order. For example, a contour with a length of one second can be represented by 20 values, taken each 50 ms throughout the contour, e.g., in Hertz with the

https://doi.org/10.1121/10.0019850

TABLE I. Studies that applied cluster analysis to intonation contours. Main aspects summarized: contour representation, *f*0 conversion, distance measure (Eucl: Euclidean), cluster analysis (H, hierarchical (linkage criterion); K-M, k-means; K-L, k-lines; N, network analysis) and language.

Study	Contour representation	*f*0 conversion (scale)	Distance measure	Cluster analysis	Language
Collier (1975)	Perceptual Grouping	NA	grouping counts	H (single/complete)	Dutch
Collier (1977)	Perceptual grouping	NA	grouping counts	H (single/complete)	Dutch, (British) English
Odé (1989)	Perceptual grouping	NA	grouping counts	H (complete)	Russian
Klabbers and Van Santen (2004)	Time-series f0	None/single speaker (Hz)	Pearson	H (Ward)	(American) English
Demenko and Wagner (2006)	Acoustic parameters	None/single speaker (Hz)	NA	K-M	Polish
Levow (2006)	Acoustic parameters	Log-transformed, *z*-normalized (NA)	NA	K-L (spectral), K-M	Mandarin, American English
Hirschberg and Rosenberg (2007)	Acoustic parameters	*z*-normalized (NA)	NA	K-M	American English
Reichel (2011)	CoPaSul acoustic parameters	Range-normalized (ST)	Eucl.	K-M	German
Calhoun and Schweitzer (2012)	PaIntE acoustic parameters	Log-transformed, standardised (Hz)	Eucl., Mahalanobis	H (Ward), K-M	American English
Raškinis and Kazlauskienė (2013)	Time-series *f*0	Log-transformed, zero-centered (Hz)	DTW	K-M	Lithuanian
Zhang (2016)	Time-series *f*0	μ-normalized (Hz, Ct, Bark), D1	Eucl., DTW, MINDIST	N/clustering coeff.	Mandarin
Dockum (2017)	PCA loadings (2)	z-normalized (ERB)	NA	K-M	Chindwin Khamti
Kaland (2021a)	Time-series *f*0	Standardised (Hz)	Eucl.	H (complete)	Papuan Malay, Zhagawa
Cole and Steffman (2021)	Time-series *f*0	Scaled (ERB)	Eucl.	K-M	American English
Kaland *et al.* (2021b)	Time-series f0	Standardised (Hz)	Eucl.	H (complete)	Kera'a
Seeliger and Kaland (2022)	Time-series *f*0	OMe scaled (Hz)	Eucl.	H (complete)	German

measurement number in subscript: $\{120.70_1, 123.34_2, 127.94_3, \ldots, 91.78_{20}\}$. Time-series *f*0 measures overcome the challenge of finding the acoustic parameters that underlie its shape. That is, time-series values form a direct input to drawing the contour as measured acoustically and inherently provide all the information that can be captured by parameters. Studies have used normalized time-series *f*0 data to perform cluster analysis on. Normalization was done to account for speaker differences such as gender and *f*0 range.

Some studies using time-series *f*0 data had a primary interest in automatization, such as text-to-speech systems (Klabbers and Van Santen, 2004) or data mining (Zhang, 2016). These studies both *z*-normalized the *f*0 values. Results on American English (Klabbers and Van Santen, 2004) showed that a hierarchical cluster analysis assuming six clusters provided the most informative assessment of the variation of pitch accents in expressive speech taken from reading children's stories. It was also shown that clustering time-series data can be computationally reduced by a network analysis (Zhang, 2016). In such an analysis a clustering coefficient expresses how nodes in a network tend to group (e.g., Watts and Strogatz, 1998). This approach was applied previously to detect melodic patterns in Indian art music (Gulati *et al.*, 2016) and has so far only been applied to intonation to test the classification accuracy of Mandarin tones (Zhang, 2016). That study compared different representations of the time-series values, such as *f*0 in Hertz, Cent, and Bark, as well as the first derivative (D1, i.e., velocity approximation) of the *f*0 contour. In addition, *f*0 contour approximations based on intonation models were tested, i.e., by polynomial functions (Hirst *et al.*, 2000) and by syllable-wise pitch target approximations (Prom-on *et al.*, 2009). Furthermore, symbolic aggregate approximations (SAX) of *f*0 contours were included. SAX representation transforms the contour into a string of letters for computation reduction, where each letter corresponds to a part of the contour in a certain range. The division in ranges is based on equal probabilities for an *f*0 level to be in a certain range. Results showed that the highest classification accuracy was achieved by the Hertz, Bark, and D1 representations of the contours (all above 90%), outperforming the model-based and SAX representations.

Other studies that used time-series *f*0 data applied cluster analysis to find evidence for phonological categories of intonation. These either used k-means (Raškinis and Kazlauskienė, 2013; Cole and Steffman, 2021) or hierarchical clustering (Kaland, 2021a; Kaland *et al.*, 2021b; Seeliger and Kaland, 2022) on normalized *f*0 data. The normalization was done using zero-centered semitone values (Raškinis and Kazlauskienė, 2013), scaled equivalent rectangular bandwidth (ERB) values (Cole and Steffman, 2021), standardised values (Kaland, 2021a; Kaland *et al.*, 2021b, using the method in Rose, 1987), or octave-median (OMe) scaled values (Seeliger and Kaland, 2022, using the method in De Looze and Hirst, 2014). One of these studies (Cole and Steffman, 2021) compared the clustering output with human discrimination accuracy using eight American English

nuclear tunes (as proposed in Pierrehumbert, 1980). The perception task was setup as an AX-discrimination task and showed that listeners had difficulty distinguishing tune pairs that were acoustically similar (as expressed by root-mean-squared-error, RMSE, on the ERB values). Crucially, the clustering indicated an optimum of five instead of eight distinct tunes, merging exactly those tunes that were acoustically similar.

A somewhat different approach used the outcome of a principal component analysis (PCA) as a representation of the *f*0 contour used as lexical tone (Dockum, 2017). The analysis was performed on word corpora from Chindwin Khamti, spoken in northwestern Myanmar. The PCA was first performed on measures of *f*0 and phonation in a time-series fashion. The outcome showed that more than 95% of the variance in the data was explained by the first two components, corresponding to *f*0 level and slope respectively. The two outcome values (one for each component) represented the *f*0 contour and were submitted to a k-means cluster analysis. The clustering was able to reliably distinguish three of the four tones reported for this language.

C. Cluster analysis in intonation research: Distance measures

After choosing a contour representation, the next step required before applying cluster analysis is deciding on the way in which differences between contours are expressed, i.e., the *distance measure*. There is much less variation in the choice of distance measure in the studies that applied clustering to intonation compared to the choice of contour representation (Table I). Nevertheless, there are many distance measures available (e.g., Mori *et al*., 2016, for an overview of distance measures for time-series data). It is beyond the scope to review all of these. This section is therefore confined to three (common) distance measures that were used in previous intonation contour research; Euclidean distance, Pearson correlation, and dynamic time warping. The grouping counts taken from the perceptual evaluation of contours (Collier, 1975, 1977; Odé, 1989) are left out here as the primary focus is on the distance measure between contour representations based on acoustic measures (production).

The most commonly applied distance measure is Euclidean distance (Table I). However, Euclidean distance does not meet a number of properties that is desired for time-series distance measures, in particular, the need for insensitivity to outliers and the recognition of similar shapes among differences in scaling (Esling and Agon, 2012). It has been shown that these disadvantages affect small datasets more than larger ones (Ding *et al*., 2008), although this has not been tested for *f*0 contours. It is expected that the scaling problem could at least partially be overcome by expressing *f*0 on a scale that accounts for auditory perception of certain spectral characteristics (ST, ERB, Bark), which would make Euclidean distance more suitable to apply to *f*0 contours.

Scaling differences in the *f*0 domain can also be accounted for by taking the Pearson correlation coefficient (ρ) as a distance measure. Thus, two rise-fall *f*0 movements that are produced in a different range (e.g., due to gender differences) are recognized as similar by this measure. Pearson correlation distance was applied to *f*0 contours for this reason in one study (Klabbers and Van Santen, 2004). It is important to note that not all correlation-based distance measures are suitable for *f*0 contours. For example, the absolute Pearson correlation coefficient, i.e., $|\rho|$, would analyse a rising and a falling *f*0 with similar starting points and steepness as highly similar. A Spearman rank correlation coefficient would lose the chronological order from the time-series because measurements are ranked in its computation.

Scaling in the time-domain can also be accounted for. This is useful to detect similarities between intonation events that might have a different time-alignment, for example, early and late peaks. A common method to do so is dynamic time warping (DTW) and has been applied in two studies (Raškinis and Kazlauskienė, 2013; Zhang, 2016). It was used in time-series of single *f*0 measurements per syllable over varying phrase lengths (Raškinis and Kazlauskienė, 2013) and on an existing dataset of extracted Mandarin tones represented by a fixed number of points (30; Gauthier *et al*., 2007). Note that the desired effect of DTW is different in these studies. That is, for varying phrase lengths similar nuclear tunes might be revealed by the clustering when using DTW, rather than when using other distance measures, which is useful for exploring intonational phonology (Raškinis and Kazlauskienė, 2013). However, for testing different mining techniques, it is desired to study a well-described phenomenon such as Mandarin tone as represented by time-normalized and equidistant *f*0 measures from a "relatively clean dataset" (Zhang, 2016, p. 5). As (time-)alignment differences can be meaningful in intonation (e.g., Ladd, 2008), this might be a reason to not apply DTW. The decision largely depends on the type of data and the research question.

The other distance measures that were used in single studies concern Mahalanobis distance (in Calhoun and Schweitzer, 2012), which is scale-invariant (Mahalanobis, 1936) and computes faster than DTW in time-series (Prekopcsák and Lemire, 2012). In addition, MINDIST (Lin *et al*., 2003, as used in Zhang, 2016) was used as a distance measure that is tailored to SAX representations and outperformed Euclidean distance on normalized f0 (Hz) data in terms of clustering accuracy.

D. Typological differences in intonation

As can be seen from Table I, cluster analysis has been applied to typologically different languages. Often, cluster analysis was chosen to explore under-researched languages (Raškinis and Kazlauskienė, 2013; Dockum, 2017; Kaland, 2021a; Kaland *et al*., 2021b). For other languages, cluster analysis provided a new perspective that could be compared to existing work on that language (e.g., Klabbers and Van Santen, 2004; Reichel, 2011; Calhoun and Schweitzer, 2012; Zhang, 2016; Cole and Steffman, 2021; Seeliger and

Kaland, 2022). Given the varying ways in which cluster analyses, in particular contour representations and distance measures, were applied, it is challenging to understand the effect of the chosen method on the outcomes. With a large variation in prosody across languages (e.g., Jun, 2005, 2014) it is important to assess at which stage in the cluster analyses language differences are to be expected. Thus, to reveal intonation patterns that are prototypical (i.e., nuclear) in a certain language, these would be expected to show from different clusters. We do not know, however, whether contour representation needs to be tailored to the language, given that contrasts between two different prototypical contours might be perceptually relevant in one language and not in another. To assess this, perception research is needed across languages and the outcomes need to be compared to the data before clustering. In the current study, this is done by comparing Papuan Malay (PMY) with German (DEU), two typologically different languages.

Papuan Malay is spoken in the Eastern Indonesian provinces Papua and West-Papua. The language is rather under researched, although its prosody has been studied in more depth in recent years. Studies have investigated word-level and phrase-level prosody. As for word-level prosody, research indicates that this language has word stress (Kaland, 2019, 2020, 2021b; Kaland *et al.*, 2021a). As for phrase-level prosody, it was found that Papuan Malay listeners agreed more on where prosodic boundaries (phrase-final) occurred than on where prosodic prominence (phrase-medial) occurred (Riesberg *et al.*, 2020). That study compared those transcriptions to the ones by German listeners, who showed a stronger agreement on the location of prominences. These results are in line with the finding that in Papuan Malay the largest *f*0 movements are found in pre-final and final syllables in the phrase (Kaland and Baumann, 2020) and that these movements facilitate word recognition (Kaland and Gordon, 2022). Focus marking has been studied for semantic contrasts in Papuan Malay and Dutch noun phrases (e.g., *red banana* vs *blue banana*; Kaland *et al.*, 2023). While these were typically marked with a pitch accent in Dutch, they were not in Papuan Malay. Irrespective of the contrastive focus, Papuan Malay speakers always produced a rise on the pre-final syllable and a rise or fall on the final syllable in the phrase. More research is needed to understand phrase(-final) prosody in Papuan Malay and in particular to what extent it is useful to distinguish pitch accents from boundary tones, as done by default in autosegmental metrical models (e.g., Jun, 2005, 2014). Thus, to date, no inventory of pitch accents and boundary tones is established for this language.

German prosody has been well studied and detailed transcriptions are available (e.g., Adriaens, 1991; Grice *et al.*, 2005). German has word level stress and uses phrase prosody for a variety of highlighting and phrasing functions. For example, clause-attachment, information structure (givenness, focus), speech acts, as well as paralinguistic functions such as signalling emotion (e.g., Grice and Baumann, 2007 for a general overview using German examples). The intonational phonology of German (e.g., Féry, 1993) is based on an inventory of six pitch accents (L*, H*, L*+H, L + H*, H + L*, and H+!H*) and eight boundary tones (intonational phrase: L-%, H-%, L-H%, H-ˆH%; intermediate phrase: L-, H-,! H-; initial boundary tone: %H). An overview of how these combine into *nuclear tunes* is available as online training material (Grice *et al.*, 2022). In a perceptual evaluation of the categoric nature of German intonation contours (i.e., acceptability ratings) it was concluded that a gradient grammar would account better for the results than a formal grammar (Féry and Stoel, 2006), which nuanced the mapping of form to function in German intonation.

E. Research aims

The studies that performed cluster analysis on *f*0 contours differ widely in their combinations of contour representation, distance measures and languages on which they were applied (n.b. leaving aside the types of cluster analysis). Although some studies explicitly compared representational methods (Zhang, 2016) or provided human perception data to compare the clustering output to (Demenko and Wagner, 2006; Reichel, 2011; Cole and Steffman, 2021), no study has systematically compared multiple contour (difference) representations to human perception across languages. Commonly, it has been taken for granted that perceptual scale and/or a normalization would approximate human pitch perception in speech. However, these *f*0 representations do not necessarily capture the way in which *f*0 contour *differences* are perceived. These aspects crucially precede the clustering, on which they have potentially a large effect. Thus, the *f*0 representation in terms of scale and normalization affects the calculation of distance measures, which in turn affects the clustering output. The effects of the representational choices in contour clustering are compared to the way human listeners perceive differences between *f*0 contours in the current study. As the ultimate goal of the cluster analysis is understanding the ways in which *f*0 movements can be meaningfully (i.e., phonologically) categorised, this study further compares the effect of native language on contour perception. It remains to be seen whether choices in contour representation for clustering depend on the language under investigation, or whether these choices rather correspond to more language-independent (psycho-)acoustic processes. The two main research questions are formulated as follows:

(RQ1) Which time-series contour representation, in terms of f0 values and distance measures, does reflect human perception best?
(RQ2) To what extent does the contour representation depend on the (prosody of the) language under investigation?

To answer these research questions, the current study compares the distance matrices obtained from all combinations of *f*0 representations and distance measures in Table II to distance matrices based on perceptual similarity judgments by Papuan Malay and German listeners. The

TABLE II. Overview of *f*0 representations and distance measures tested in this study.

*f*0 representation		Distance measure
Unconverted (ERB)		Euclidean
Standardised (Hz)	×	Pearson
OMe rescaled (Hz)		DTW
First derivative (ERB)		

similarity judgments were obtained from all combinations of nine carefully chosen contours. The investigation concerns time-series *f*0 data only, as this method provides fine-grained detail superseding parametric approaches. In order to test the difference between the languages, the *f*0 contours under investigation were naturally produced ones from Papuan Malay and presented to listeners using acoustically manipulated stimuli (Sec. II).

II. METHODOLOGY

This section describes the methodological choices made in this study. The contours investigated here are taken from a corpus of Papuan Malay recordings from speaker pairs who put together pieces that form a tangram figure in a collaborative way (Riesberg and Himmelmann, 2012). The recordings were unscripted, and are a representative and naturalistic type of speech for this language. Transcriptions were cross-checked with the help of native speakers.

A. Stimulus preparation

From the corpus of recordings (Riesberg and Himmelmann, 2012) phrase-final words were selected. This was done because previous studies found that phrase-final *f*0 movements are the largest in Papuan Malay (Kaland and Baumann, 2020). The goal was to select a set of *f*0 contours that showed a natural degree of variation. Previous studies reported rise-fall contours as the most common *f*0 pattern in Papuan Malay (Himmelmann and Kaufman, 2020). However, the data showed that other patterns were frequent as well. For this reason, three overall patterns were distinguished for the selection of the stimulus material: (rise-)falling, level, and rising. Note that these are coarse categorisations, as variation was found among falling contours in whether they were preceded by a small rise and among the rising contours in the steepness of the rise. In the former case, the rise always had a smaller range and shorter duration than the fall (i.e., the contour was overall more falling than rising).

The phrase-final words were chosen from three speakers (two females, S and T; one male, Y). These speakers were chosen because the duration variation among the (rise-)falling, level, and rising contours were smaller compared to other speakers (386 to 678 ms; see Table III). This was done to minimize the degree of duration manipulation (described in the next paragraph). For the same reason, only bisyllabic words were considered. The final set consisted of nine selected words (three speakers × three types of contours). See Table III and Fig. 1 for details on the selected contours.

TABLE III. Properties of the nine selected contours.

Speaker	# Contour	Word	Duration (ms)
S	1: (rise-)fall	*prahu* (boat)	530
S	2: level	*ekor* (tail)	502
S	3: rise	*kecil* (small)	523
T	1: (rise-)fall	*beda* (different)	396
T	2: level	*dengan* (with)	475
T	3: rise	*besar* (big)	678
Y	1: fall	*besar* (big)	589
Y	2: level	*bagyan* (part)	483
Y	3: rise	*baru* (only)	386

The selected words were prepared as stimuli for the perception experiment in several stages of acoustic processing and manipulation using Praat (Boersma and Weenink, 2022). First, their *f*0 contour was interpolated and smoothed. The resulting contour was used as input to generate a "hummed" version of the word. This step preserves the prosody (*f*0, duration, and intensity) while the individual segments become inaudible. This was done to avoid any effect of the originally produced word. To further ensure that from the prosodic cues only f0 differed between the stimuli, intensity and duration were both normalized. For intensity, this was done by flattening the sound pressure level to a constant of 70 dB throughout the stimulus. For duration, the stimuli were time-normalized such that they all lasted exactly 500 ms (without changing pitch). Although the naturalness of the stimuli was therefore degraded to a certain extent, any perceptual effect can be attributed to *f*0 only, which exactly followed the naturally produced contour (Fig. 1).

B. Experimental design and procedure

Similarity judgments were elicited on a five-point scale in a perception experiment that presented the stimuli in pairwise fashion. The nine stimuli were combined into 45 pairs, such that each of the nine stimuli was combined with all others and with itself. The pairs in which both stimuli were identical were included to assess participants' performance (see the following) and excluded from the similarity analysis. The resulting similarity matrix of the participants' judgments is based on the pairs in which the stimuli were not identical ($N = 36$), as provided in the Appendix.

The experiment was designed using PsyToolkit (Stoet, 2010, 2017), which allows for online participation and collection of results. The experiment was run through a web browser. For each stimulus pair a screen was generated displaying (from top to bottom) a percentage counter showing task progress, the phrase "The word melodies are…," two play buttons, a five-point scale with the words "identical" and "different" on the left and right side of the scale respectively, and a button to proceed to the next stimulus (Fig. 2). To ensure that participants followed the intended procedure, the scale was only displayed after they had clicked on (and listened to) both stimuli and the proceed button was only displayed after they had made a judgment on the scale. The

https://doi.org/10.1121/10.0019850

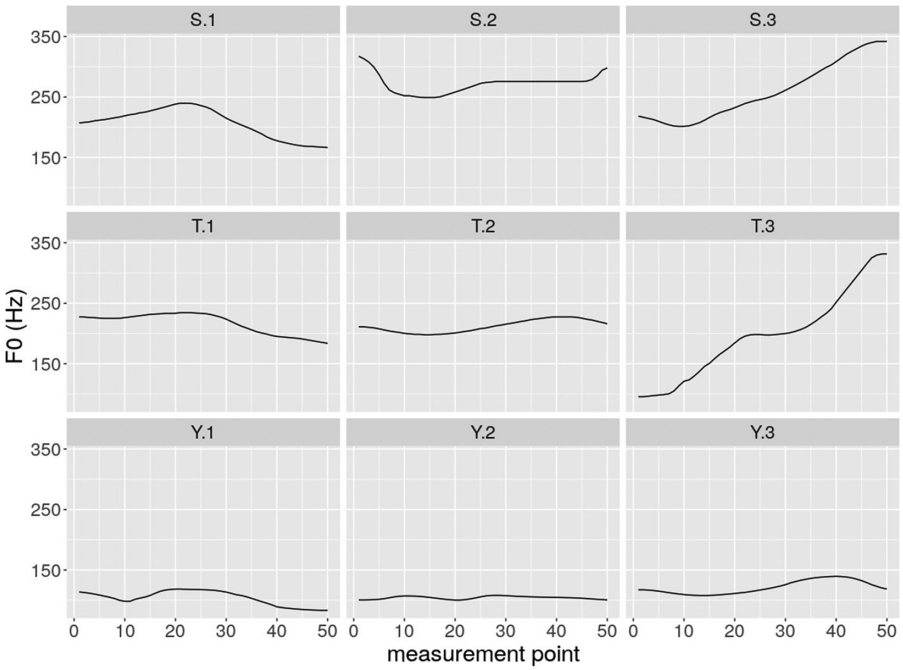

FIG. 1. Nine contours as used in the stimuli and produced by three different speakers (rows; S, T, Y) in three overall shapes [columns; 1: (rise-)fall, 2: level, 3: rising].

stimulus pairs were presented in random order, different for each participant. The position of the two play buttons corresponded to the two members of a stimulus pair in random fashion (either A:left–B:right or A:right–B:left). This was done to counterbalance any potential presentational bias. Participants could listen to the stimuli as much as needed and change their judgment until they had clicked the proceed button. After clicking the proceed button their final judgment was recorded and the next stimulus pair was shown.

Participants were instructed that they would listen to pairs of words that were made unrecognizable and that their task was to judge how similar/different the melodies of the words in the pair are. They were told that the stimuli were taken from different speakers and that speaker differences were still audible in the stimuli. Participants were instructed to ignore differences between speakers (i.e., in overall *f*0 range) as much as possible and only judge the melody of the words (i.e., contour shape). This instruction did not guarantee the desired auditory focus of participants, however, was taken as the optimal way to draw attention to the contour shape differences. Participants were further instructed to use the entire scale to express their judgments. After instructions, participants completed a training round consisting of five randomly chosen stimulus pairs. This training round was meant to familiarize them with the experimental procedure. During a pilot test without a training round, participants reported having difficulties judging different sounding stimuli as to their degree of difference. That is, it was hard for them to judge how different they were as they did not have anything to compare the difference to. Only after having heard multiple stimuli, this difficulty would decrease. Thus, the training round ensured that participants had a realistic impression of the degree of possible differences between the stimuli before starting the actual experiment. After having received the instructions and having completed the training round the actual experiment was done. The experiment lasted on average 40 min, with PMY participants taking longer ($\mu = 54$ min) than the DEU participants ($\mu = 26$ min). The latter difference could be ascribed to the experience participants had with doing experiments. That is, the DEU participants all had experience in participating in an experiment, whereas hardly any of the PMY participants had done an experiment before.

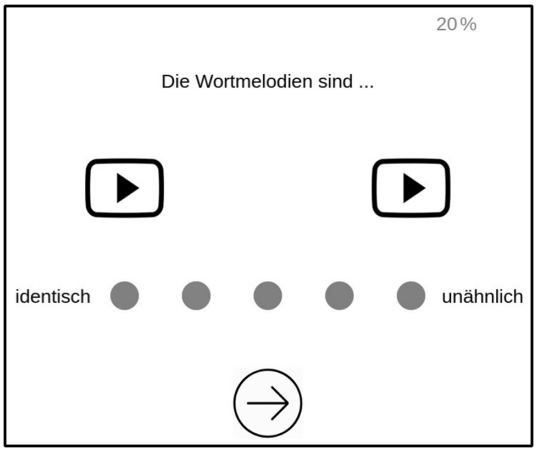

FIG. 2. Screen capture of the German version of the experiment with the phrase "The word melodies are…," two play buttons, a five-point scale with the words "identical" and "different" on the left and right side of the scale, respectively.

C. Participants

All participants were native speakers of the language without hearing problems. Although some of them spoke a local variety of the language as well, all of them were fluent in the standard language (PMY or DEU). Responses of participants were discarded if they failed to identify any of the

pairs in which the stimuli were identical (PMY, 9; DEU, 8). The judgments of these participants were considered unreliable. The remaining number of participants for analysis was as follows. PMY, 32 (26 F/6 M), *M* age, 26, age range: 19–41; DEU, 33 (23 F/10 M), *M* age, 31, age range, 24–41.

D. Contour representations

Time-series *f*0 measures were taken from the contours in the stimuli using 50 measurement points (i.e., every 10 ms). In this way, each *f*0 contour was represented by a vector of 50 values. They were measured in Hertz and then converted into four different *f*0 representations (Table II). First, they were converted to ERB using the formula from Glasberg and Moore (1990) given in Eq. (1). ERB is based on a logarithmic scale and therefore accounts for pitch perception across different *f*0 ranges. Second, standardisation/*z*-normalisation [Eq. (2)] was applied following previous work on intonation contour clustering (e.g., Levow, 2006; Hirschberg and Rosenberg, 2007; Calhoun and Schweitzer, 2012; Dockum, 2017; Kaland, 2021a; Kaland *et al.*, 2021b). Standardisation was proposed specifically to account for speaker differences in the production of lexical tones (Rose, 1987). Although the contours used in this study were phrase-final ones of a non-tonal language, their duration (0.5 s) matches the length of a word and standardisation might therefore be a suitable method for the stimuli. The input *f*0 values for standardisation were expressed in Hertz. Third, octave-median rescaling from Eq. (3) was applied to the Hertz values. This conversion was proposed to account for speaker (range) differences whilst taking into account the melodic nature of intonation (i.e., using octaves; De Looze and Hirst, 2014). It was applied in previous clustering research (Seeliger and Kaland, 2022). Note that for both standardisation and octave-median rescaling, the central tendency measures (mean, median, standard deviation) were calculated on the basis of more than the three contours per speaker (S, 17; T, 7; Y, 6) such that their estimates were more representative for the speakers. These additional contours used for estimation were also taken from phrase-final bisyllabic words in the same corpus. Note that the Hertz scale was chosen for standardisation and octave-median rescaling as these methods were proposed for this scale specifically. Fourth, the first derivative Eq. (4) of the *f*0 contour was taken from the time series values using the `gradient()` function from the R `pracma` package (Borchers, 2022). This method essentially computes the rate of change between two successive points such that the resulting velocity curve is scale-invariant, i.e., it only expresses the shape of the *f*0 movement, preserving range, direction, and slope information. This representation was tested on *f*0 contours in previous studies on Mandarin tone, outperforming a representation by direct *f*0 values (Zhang, 2016; Gauthier *et al.*, 2007). In the current study, ERB values were chosen to compute the first derivative, in order to match the velocity curves with the logarithmic nature of pitch perception,

$$ERB \quad f0_{\mathrm{ERB}} = 21.4 * \log 10(0.00437 * f0_{\mathrm{Hz}} + 1), \tag{1}$$

$$Standardisation(z\text{-}norm) \quad f0_{Std} = \frac{f0_{\mathrm{Hz}} - \overline{f0}_{\mathrm{Hz}}}{\sigma}, \tag{2}$$

$$Octave\text{-}median\ rescaling \quad f0_{\mathrm{OMe}} = \log 2\left(\frac{f0_{\mathrm{Hz}}}{\tilde{f0}_{\mathrm{Hz}}}\right), \tag{3}$$

$$First\ derivative \quad \nabla f(a_1, a_n) = \left(\frac{\partial f}{\partial \mathrm{x}_1}(a_1, \ldots, a_n)\right), \ldots, \frac{\partial f}{\partial \mathrm{x}_\mathrm{n}}(a_1, \ldots, a_n). \tag{4}$$

E. Distance measures

Each of the contour representations were the input for the computation of distance matrices using three distance measures; Euclidean distance (L2 norm), Pearson correlation and dynamic time warping (DTW). The computations were carried out in R (R Core Team, 2022) and R Studio (R Studio Team, 2022) using the package `TSDist` (Mori *et al.*, 2016). Their common formulas are given in the following, in which x and y are the two time-series (vectors) of *f*0 values, ρ is the Pearson correlation coefficient between x and y, and df_ϕ is the average accumulated deformation due to warping the time-indices of x and y. As for Euclidian distance in Eq. (5), the square root of the sum of the squares of the differences between x and y are taken. This method is highly similar to the one based on RMSD adopted in Hermes (1998) as successful quantifier of perceived differences between *f*0 contours. Pearson correlation [Eq. (6)] as applied here expresses the distance between x and y on scale between 0 and 2, such that negative values are avoided. Thus, strongly negative correlations end up close to 2, whereas strongly positive correlations end up close to 0. In this way, all distance measures applied here share that identical contours are expressed using zero distance and that increasing distance values reflect increasing dissimilarity. Dynamic time warping [Eq. (7)] first remaps the time-axis of x and y (all formulas in Giorgino, 2009). The deformation resulting from this remapping (warping) that aligns x and y as close as possible is taken as the distance measure. Distances between warped x and y are Euclidean distances. Without further constraining the maximum allowed path (window) of warping (ϕ), DTW is highly similar to Euclidean distance. Therefore, in the current analysis the window constraint was set to 5 (measurement points), thus allowing a maximum of ±50 ms misalignment, which equals a maximum of 10% misalignment given the total duration of the contours (0.5 s). The window calculation was done using the Sakoe–Chiba method (Sakoe and Chiba, 1978), which was shown to outperform other common uses of DTW on 85 different time-series datasets (Geler *et al.*, 2019),

$$Euclidean \quad \mathrm{d}(\mathrm{x}, \mathrm{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \tag{5}$$

$$Pearson \quad \mathrm{d}(\mathrm{x}, \mathrm{y}) = \sqrt{2(1 - \rho(\mathrm{x}, \mathrm{y}))}, \tag{6}$$

$$Dynamic\ Time\ Warping \quad d(x, y) = \min_{\phi} df_\phi(x, y). \tag{7}$$

https://doi.org/10.1121/10.0019850

Participants' judgments on the five-point scale were recorded as a value between 0 and 4, where 0 corresponded to two contours judged as identical and 4 to two contours judged as maximally dissimilar. The mean of these values was computed for each stimulus pair and formed the input values for a distance matrix, one for each language.

F. Data analysis

The distance matrix computation generated production-based matrices (all combinations of the four contour representations and three distance measures, Table I, $N = 12$) and two perception-based matrices (PMY and DEU). The crucial comparisons for the current study concern the ones between the production-based ones on the one hand and the perception-based ones on the other hand (2×12). Nevertheless, correlation coefficients were computed for all combinations of distance matrices ($N = 91$) to further understand how they differ among each other. The correlation coefficient (Kendall τ) was chosen as a measure of similarity between the distances matrices (Dietz, 1983) and testing was carried out using the `cor.test()` function in R (R Core Team, 2022) for all complete pairwise observations (36 pairs of values for each comparison of two distance matrices). The resulting correlation matrix thus consisted of 91 coefficients (see Sec. III).

III. RESULTS

The distances as computed for each stimulus pair according to each tested method and their perceptual rating are given in the Appendix. Before turning to a comparison of the production-based distances and the perception-based distances, they are briefly discussed separately one by one.

Among the production-based distances, some combinations of contour representation and distance measure show a near perfect correlation (>0.97); ERB-Pρ with Standardisation-Pρ and with OMe rescaling-Pρ (Table IV). Very strong correlations (>0.80) are found within each contour representation between the distance measures Euclidean distance and DTW. Across all tested combinations, ERB-Pρ, Standardisation-Pρ, OMe-rescaling-Pρ, and First derivative-DTW show significant correlations with most of the other production-based distances.

As for the perception-based distances, it can be seen that Papuan Malay and German overall show equally low or equally high similarity ratings for the contour pairs (Appendix). Figure 3 shows the perceived distances in graphs for each language. For example, contour pair #16 (S.3–T.1) and #36 (Y.2–Y.3) both show minimal differences in rating across the languages (<0.01). Exceptions are contour pair #14 (S.2–Y.2) and #24 (T.1–Y.1), which both showed more than one point difference on the five-point rating scale. Some contour pair ratings varied less than others, as can be seen from their standard deviations (SD). These were unsurprisingly found at either end of the rating scale, i.e., rated as highly similar or highly different contours, respectively. For Papuan Malay, pair #3 (S.1–T-1) had the smallest SD for a high similarity rating and pair #7 (S.1–Y.2) had the smallest SD for a highly different rating. For German, pair #3 (S.1–T-1) had the smallest SD for a high similarity rating and pair #12 (S.2–T.3) had the smallest SD for a highly different rating. The largest SDs were found for pair #30 (T.2–Y.3) for Papuan Malay and for pair #14 (S.2-Y.2) for German. It furthermore becomes clear from the Table IV that the perception ratings of both languages correlated moderately to strongly.

When comparing the perceived distances with the computed (production-based) ones (Table IV), correlations are mostly found to be moderate in strength. The strongest correlations are found for Standardisation-DTW (PMY, 0.38; DEU, 0.42) and First derivative-DTW (PMY, 037; DEU,

TABLE IV. Correlation coefficients (Kendall τ) as calculated for all combinations of distance matrices. Coefficients for the perception-based matrices occur in bold face when the correlation test had a p-value below 0.05.

		ERB			Standardisation			OMe rescaling			First derivative			Perception	
		EU	Pρ	DTW	EU	Pρ	DTW	EU	Pρ	DTW	EU	Pρ	DTW	PMY	DEU
ERB	EU														
	Pρ	0													
	DTW	**0.90**	−0.01												
Standardisation	EU	0.14	**0.25**	0.04											
	Pρ	0.01	**0.99**	−0.01	**0.26**										
	DTW	0.12	**0.29**	0.07	**0.87**	**0.30**									
OMe rescaling	EU	0.09	**0.27**	0	**0.86**	**0.28**	**0.79**								
	Pρ	0	**0.99**	−0.01	**0.25**	**0.98**	**0.29**	**0.27**							
	DTW	0.06	**0.31**	0	**0.80**	**0.32**	**0.84**	**0.84**	**0.31**						
First derivative	EU	0.13	**0.24**	0.05	**0.77**	**0.25**	**0.74**	**0.71**	**0.24**	**0.66**					
	Pρ	−0.02	**0.47**	−0.03	0.18	**0.46**	0.22	0.18	**0.47**	0.21	0.21				
	DTW	0.06	**0.34**	−0.02	**0.72**	**0.35**	**0.66**	**0.71**	**0.34**	**0.63**	**0.80**	**0.27**			
Perception	PMY	**0.33**	**0.24**	**0.29**	**0.35**	**0.25**	**0.38**	**0.37**	0.23	**0.34**	**0.37**	0.19	**0.37**		
	DEU	**0.25**	**0.33**	0.20	**0.42**	**0.33**	**0.42**	**0.40**	**0.33**	**0.35**	**0.42**	0.15	**0.49**	**0.65**	

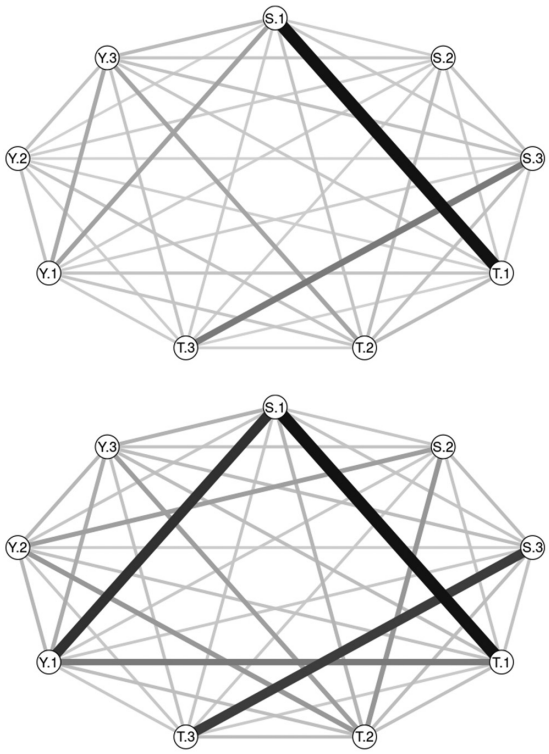

FIG. 3. Perceived distances between all contours (see Fig. 1) by Papuan Malay (top) and German (bottom) listeners. Visualization in a circular network graph based on the distance matrix of perceptual scores (values in the Appendix) for each language using `qgraph` package in R (Epskamp *et al.*, 2012). Thicker and darker lines correspond to smaller perceived distances.

0.49). Overall, weaker correlations were found for Papuan Malay than for German. Across the languages (Table IV bottom two rows only) and among the different contour representations, standardisation always resulted in a significant correlation. Among the different distance measures, Euclidean distance and DTW tend to show stronger correlations than Pearson ρ. In this respect, the contour representations standardisation and first derivative are highly similar when taking into account only Euclidean distance and DTW. OMe rescaling is similar to Standardisation and First derivate in that it has the strongest correlations for Euclidean distance and is different in that DTW correlations are weaker. Contour representations in ERB led to the overall weakest correlations across languages and across distance measures. Pearson ρ applied to the first derivative showed the lowest τ values across the languages.

IV. DISCUSSION AND CONCLUSION

The aim of this study was to find the combination of contour representation and distance measure that best captures perceived differences between intonation contours across two languages. The results have shown that the majority of the combinations work moderately well. The strongest correlations between computed and perceived distance were found for DTW, when applied to either standardized *f*0 values or the first derivative (RQ1). Although these were the strongest correlations in numeric terms, their strength generally did not differ much from the other contour representations and distance measures. Standardisation had the best result across distance measures and Pearson correlation had the worst result when applied to the first derivative. Thus, the outcomes rather indicate which combinations should be avoided when performing cluster analysis on *f*0 contours.

The results do not clearly indicate language differences in the way the contour differences were perceived. The largest difference between Papuan Malay and German perception was found for the first derivative-DTW (0.12). This difference indicates that the first derivative-DTW better reflects German contour difference perception than it does for Papuan Malay. It is however unlikely that this difference is a reflection of true linguistic differences, for two reasons. First, the correlations were overall weaker for Papuan Malay than for German. Second, since the contours were taken from Papuan Malay data, a true linguistic difference would likely have shown in the other direction; i.e., a perceptual advantage of hearing intonation contrasts for the Papuan Malay listeners as the contours were taken from their language. In the context of overall moderate correlations between production-based and perception-based similarities, it is likely that true language differences did not exist in the current setup. In addition, it should be noted that the two contour pairs for which the largest language differences in the ratings were found, were similar in shape and different in register (#14 and #24). The language difference for both pairs was such that Papuan Malay listeners perceived the contours are less similar than the German listeners. The perception-based distances largely confirm the original division of contours in terms of (rise-)fall, level, and rise (cf. Table III and Fig. 3). The smallest perceived distances were indeed found within these shape categories across speakers and across languages, in particular (rise-) falls (X.1) and rises (X.3). Speakers S and T (female) tend to be perceived as more similar than each of them compared to speaker Y (male), indicating that participants were not entirely able to abstract over speaker gender. It should be noted that some informal reports from participants indicated a challenge they had with contour pairs that only differed in register. Although they were explicitly instructed to not rate differences between male and female voices, not all participants managed to do so equally well. In this respect, there was a difference between the Papuan Malay and the German listeners, in that participants from the latter group had more often a background in linguistics and were therefore potentially better able to focus on contour shape differences only, as potentially reflected in shorter completion times (see Sec. II C).

Related to the difficulty of the task, it should also be noted that none of the combinations of contour representations and distance measures was able to correlate more than moderately with the perception ratings. Table IV indicates that the best reflection of participants' ratings were the ones obtained from the other language, not from the computed ones. It can be expected that the task difficulty participants experienced or the decreased naturalness of the stimuli

https://doi.org/10.1121/10.0019850

contributed to these results. Although the methodological choices for the stimulus material were careful, it could not be avoided that they deviated from natural speech perception. That is, listeners do generally not perceive speech in stretches of the same length and intensity. Crucially, no information content was present in any of the stimuli. These factors are likely to reduce the correlation strength to some extent. In addition, it could be the case that the production-based distances are not optimal approximations of perceived contour differences. More research is needed to fully understand the perception of f0 and how they can be approximated by acoustic measures.

Despite the shortcomings of the current study, the results provide a useful basis for the application of cluster analysis to *f*0 contours. This is particularly true given the lack of notable language differences, indicating that the current experiments are likely to reflect a level of auditory perception that is shared across languages (RQ2). The results support the choices for standardisation and Euclidean distance in a considerable number of previous studies (Table I), as these methods were shown to be among the best-performing ones in the current study. The results also show that human perception is even better reflected in methods that were less commonly applied. This holds in particular for the first derivative as contour representation and DTW as a distance measure. Their combination led to (one of) the best results in both languages tested in this study. Only a few studies applied these methods so far (DTW, Raškinis and Kazlauskienė, 2013; D1-DTW, Zhang, 2016) and the current results indicate that they are worth applying in future work. This outcome is not entirely surprising, given that both methods are improved versions of others. That is,

TABLE V. *Numbered stimulus pair list showing all combinations of the nine contours (Table III and Fig. 1) and all distances between them as measured by the different combination of contour representations (ERB, Standardisation, OMe rescaling, First derivative) and distance measures (EU, Pρ, DTW), and the mean perceived distances (with standard deviation) by PMY and German DEU listeners.

Pair	Contour		ERB			Standardisation			OMe rescaling			First derivative			Perception	
#	A	B	EU	Pρ	DTW	EU	Pρ	DTW	EU	Pρ	DTW	EU	Pρ	DTW	PMY	DEU
1	S.1	S.2	10.48	1.72	114.16	14.50	1.72	158.58	3.25	1.72	35.22	0.70	1.58	5.11	3.28 (1.14)	3.42 (0.83)
2	S.1	S.3	11.88	1.92	105.99	16.84	1.92	148.14	3.62	1.91	32.53	0.75	1.73	6.95	3.00 (1.32)	3.30 (1.13)
3	S.1	T.1	1.95	0.21	12.38	5.64	0.22	51.74	1.56	0.20	15.46	0.20	0.51	1.02	0.56 (0.95)	0.67 (0.74)
4	S.1	T.2	5.08	1.92	50.96	8.47	1.92	79.13	2.10	1.91	18.70	0.50	1.76	4.62	2.75 (1.32)	2.61 (1.25)
5	S.1	T.3	13.72	1.83	137.58	23.49	1.84	221.16	4.69	1.80	45.09	1.01	1.23	8.41	3.16 (1.30)	3.36 (0.90)
6	S.1	Y.1	17.82	0.40	230.86	2.10	0.41	8.81	0.56	0.38	2.10	0.31	0.91	1.70	1.72 (1.49)	0.82 (1.07)
7	S.1	Y.2	18.02	1.33	235.99	4.82	1.33	41.95	1.28	1.32	10.69	0.40	1.41	2.79	3.62 (0.83)	3.42 (0.97)
8	S.1	Y.3	15.33	1.86	191.76	10.54	1.86	94.23	2.76	1.85	24.94	0.55	1.70	4.99	2.38 (1.45)	2.48 (1.33)
9	S.2	S.3	6.15	1.12	50.66	9.05	1.13	74.25	1.81	1.11	14.88	0.55	0.89	4.08	2.72 (1.28)	2.94 (1.09)
10	S.2	T.1	8.65	1.67	96.33	9.90	1.66	86.97	1.79	1.67	15.99	0.59	1.51	3.99	3.09 (1.23)	3.33 (0.96)
11	S.2	T.2	8.59	0.95	104.86	9.08	0.97	99.35	1.62	0.93	17.76	0.49	1.17	2.75	2.50 (1.39)	1.79 (1.47)
12	S.2	T.3	15.83	1.31	151.57	23.64	1.28	222.35	4.88	1.35	42.59	1.03	1.32	8.12	3.34 (1.07)	3.61 (0.56)
13	S.2	Y.1	27.14	1.54	361.75	13.14	1.53	141.43	2.91	1.54	29.36	0.52	1.18	3.23	3.28 (1.22)	3.55 (0.83)
14	S.2	Y.2	26.94	1.65	366.33	12.33	1.65	150.70	2.54	1.65	31.09	0.59	1.66	2.79	3.28 (1.37)	1.97 (1.59)
15	S.2	Y.3	23.66	1.08	315.55	6.98	1.10	57.82	1.23	1.05	9.61	0.56	1.33	3.29	2.84 (1.48)	3.27 (1.04)
16	S.3	T.1	10.30	1.95	94.92	14.58	1.96	146.94	2.66	1.95	27.66	0.64	1.62	6.15	3.19 (1.28)	3.18 (1.10)
17	S.3	T.2	8.03	0.51	73.62	9.06	0.52	70.86	1.65	0.51	13.81	0.41	0.88	3.11	2.62 (1.45)	2.94 (1.03)
18	S.3	T.3	10.94	0.31	97.14	15.43	0.28	127.78	3.26	0.38	23.46	0.62	1.05	3.27	1.16 (1.27)	0.88 (0.89)
19	S.3	Y.1	26.12	1.85	334.91	15.58	1.86	131.42	3.36	1.85	28.25	0.62	1.31	5.22	3.34 (1.12)	3.55 (0.71)
20	S.3	Y.2	25.54	1.46	333.91	13.06	1.47	118.09	2.55	1.45	23.82	0.53	1.73	4.98	3.59 (1.04)	3.79 (0.74)
21	S.3	Y.3	22.00	0.64	283.14	7.44	0.67	51.85	1.24	0.63	8.90	0.46	1.07	3.15	2.69 (1.47)	2.91 (1.26)
22	T.1	T.2	3.98	1.92	45.48	7.64	1.92	87.45	1.29	1.92	14.63	0.35	1.65	2.69	2.41 (1.27)	3.12 (0.99)
23	T.1	T.3	13.46	1.89	134.45	24.78	1.90	249.42	4.79	1.86	46.28	0.98	1.16	7.96	3.53 (0.88)	3.64 (0.74)
24	T.1	Y.1	19.22	0.41	253.99	5.06	0.42	42.39	1.31	0.40	10.87	0.25	0.86	1.41	2.72 (1.61)	1.36 (1.29)
25	T.1	Y.2	19.28	1.34	259.94	6.67	1.35	69.21	1.38	1.34	13.52	0.27	1.31	1.62	3.41 (1.07)	3.15 (1.00)
26	T.1	Y.3	16.40	1.87	214.25	8.23	1.87	84.90	1.67	1.87	16.57	0.42	1.63	3.25	2.97 (1.33)	2.91 (1.16)
27	T.2	T.3	10.30	0.77	85.08	18.55	0.76	155.04	3.83	0.81	30.29	0.91	1.54	7.13	3.41 (1.04)	2.94 (1.12)
28	T.2	Y.1	18.86	1.81	249.46	7.19	1.81	57.71	1.85	1.81	14.50	0.40	1.46	3.35	2.84 (1.35)	3.00 (0.90)
29	T.2	Y.2	18.50	1.36	252.31	4.38	1.36	38.58	1.04	1.35	10.39	0.23	1.38	1.53	2.72 (1.35)	1.82 (1.38)
30	T.2	Y.3	15.17	0.25	199.60	3.01	0.25	17.47	0.77	0.24	4.37	0.13	0.39	0.55	1.78 (1.66)	1.97 (1.31)
31	T.3	Y.1	19.60	1.80	199.72	22.87	1.82	216.10	4.66	1.77	44.77	1.01	1.29	7.24	3.22 (1.10)	3.55 (0.79)
32	T.3	Y.2	18.58	1.43	193.72	20.05	1.46	170.26	3.88	1.39	32.86	0.93	1.68	7.54	3.34 (1.00)	3.48 (0.76)
33	T.3	Y.3	15.25	0.85	155.74	19.27	0.86	171.73	3.95	0.86	32.91	0.98	1.63	7.64	2.78 (1.39)	3.18 (0.88)
34	Y.1	Y.2	2.40	1.39	20.31	4.19	1.39	35.55	1.23	1.38	10.28	0.41	1.65	3.04	2.62 (1.29)	2.70 (1.16)
35	Y.1	Y.3	5.10	1.74	41.29	9.18	1.74	74.45	2.45	1.74	19.64	0.45	1.46	3.83	1.88 (1.50)	2.30 (1.47)
36	Y.2	Y.3	3.83	1.26	37.87	7.03	1.26	68.81	1.76	1.26	17.68	0.30	1.34	2.06	2.69 (1.45)	2.70 (1.19)

taking the first derivative of an *f*0 contour automatically abstracts over register differences without the need of calculating a speaker's range first (as done in standardisation and OMe rescaling). DTW is essentially Euclidean distance with the added capability of allowing for misalignments in time. In this regard it is interesting that Pearson correlation turned out to be the least-performing distance measure. It appeared that although this measure is able to abstract over register differences it does not reflect the way listeners do this. Pearson correlation was shown to perform best to express contour differences among different representations of the *produced* contours, most likely because they were variants of each other based on the same set of f0 measures. To conclude, the current results showed that *perceived* contour distances are best expressed by Euclidean distance or DTW.

Now that some common contour representations and distance measures are compared to human perception, the next step is applying the outcomes to cluster analyses. Given the variety of clustering techniques (Table I), future work should compare them in the same structural way as done in the current study. One way of doing so would be to perform several clustering methods on the same dataset, which has known groupings of data (see, e.g., Cole and Steffman, 2021 for such a set of American English contours). In this way, our understanding of contour clustering and its potential implications for prosodic theory can be improved. The current study shows that dynamic time warping on the first derivative of ERB converted *f*0 contours would be a promising way to express the contour differences in such a study.

ACKNOWLEDGMENTS

Research for this paper was funded by the German Research Foundation (DFG,) – Project-ID 281511265 – SFB 1252. The author thanks the Papuan Malay Bible Translation Team (Tim Penerjema Alkitab Melayu Papua) for their help with participant recruitment and experiment facilitation, Janne Lorenzen and Timo Buchholz for translations, Jennifer Cole and Jeremy Steffman for sharing their work and for inspiring conversations, all participants for their effort and two reviewers for their valuable comments. The author has no conflicts of interest to declare. The experiments reported in this paper have been conducted following protocols and informed consent practices in compliance with the Helsinki Declaration, with approval of the Papuan Malay Bible Translation Team and the Faculty of Arts and Humanities of the University of Cologne. Informed consent was obtained from each individual participant prior to participation. Stimuli, data, and supplementary material are available at https://doi.org/10.17605/OSF.IO/AGNM5.

APPENDIX

See Table V for the stimulus pair list.

Adriaens, L. L. (**1991**). "Ein Modell deutscher Intonation: Eine experimentell-phonetische Untersuchung nach den perzeptiv relevanten Grundfrequenzänderungen in vorgelesenem Text" ("A model of German intonation: Experimental-phonetic investigation of the perceptually relevant fundamental frequency changes in text read aloud") (Technische Universiteit Eindhoven Eindhoven, the Netherlands).

Albert, A., Cangemi, F., and Grice, M. (**2018**). "Using periodic energy to enrich acoustic representations of pitch in speech: A demonstration," in *Speech Prosody 2018*, June 13–16, Poznan, Poland, pp. 804–808.

Barnes, J., Veilleux, N., Brugos, A., and Shattuck-Hufnagel, S. (**2012**). "Tonal center of gravity: A global approach to tonal implementation in a level-based intonational phonology," Lab. Phonol. **3**(2), 337–383.

Boersma, P., and Weenink, D. (**2022**). "Praat: Doing phonetics by computer," http://www.praat.org/ (Last viewed May 5, 2022).

Borchers, H. W. (**2022**). "pracma: Practical Numerical Math Functions," https://CRAN.R-project.org/package=pracma (Last viewed November 22, 2022).

Calhoun, S., and Schweitzer, A. (**2012**). "Can intonation contours be lexicalised? Implications for discourse meanings," in *Prosody and Meaning*, edited by G. Elordieta and P. Prieto (De Gruyter, Berlin), pp. 271–327.

Cole, J., and Steffman, J. (**2021**). "The primacy of the rising/non-rising dichotomy in American English intonational tunes," in *Proceedings of the 1st International Conference on Tone and Intonation (TAI)*, March 28–30, Beijing, China, pp. 122–126.

Cole, J., Steffman, J., and Tilsen, S. (**2022**). "Shape matters: Machine classification and listeners' perceptual discrimination of American English intonational tunes," in *Proceedings of Speech Prosody 2022*, May 23–26, Lisbon, Portugal, pp. 297–301.

Collier, R. (**1975**). "Perceptual and linguistic tolerance in intonation," IRAL Int. Rev. Appl. Ling. Lang. Teach. **13**(1-4), 293–308.

Collier, R. (**1977**). "The perception of English intonation by Dutch and English listeners," IPO Annu. Prog. Rep. **12**, 69–73.

De Looze, C., and Hirst, D. (**2014**). "The OMe (Octave-Median) scale: A natural scale for speech melody," in *Proceedings of the 7th International Conference on Speech Prosody 2014*, May 20–23, Dublin, Ireland, pp. 910–914.

Demenko, G., and Wagner, A. (**2006**). "The stylization of intonation contours," in *Proceedings of Speech Prosody 2006*, May 2–5, Dresden, Germany, p. 254.

Dietz, E. J. (**1983**). "Permutation tests for association between two distance matrices," Syst. Biol. **32**(1), 21–26.

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., and Keogh, E. (**2008**). "Querying and mining of time series data: Experimental comparison of representations and distance measures," Proc. VLDB Endow. **1**(2), 1542–1552.

Dockum, R. (**2017**). "Computational modeling of tone in language documentation: Citation tones vs. running speech in Chindwin Khamti," in *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, Vol. 43, pp. 43–73.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., and Borsboom, D. (**2012**). "qgraph: Network visualizations of relationships in psychometric data," J. Stat. Softw. **48**(4), 1–18.

Esling, P., and Agon, C. (**2012**). "Time-series data mining," ACM Comput. Surv. **45**(1), 1–34.

Féry, C. (**1993**). *German Intonational Patterns* (De Gruyter, Berlin).

Féry, C., and Stoel, R. (**2006**). "Gradient perception of intonation," in *Gradience in Grammar*, 1st ed., edited by G. Fanselow, C. Féry, M. Schlesewsky, and R. Vogel (Oxford University Press, Oxford), pp. 145–166.

Gauthier, B., Shi, R., and Xu, Y. (**2007**). "Learning phonetic categories by tracking movements," Cognition **103**(1), 80–106.

Geler, Z., Kurbalija, V., Ivanovic, M., Radovanovic, M., and Dai, W. (**2019**). "Dynamic Time Warping: Itakura vs Sakoe-Chiba," in *Proceedings of the 2019 IEEE International Symposium on Innovations In Intelligent Systems And Applications (INISTA)*, July 3–5, Sofia, Bulgaria, pp. 1–6,.

Giorgino, T. (**2009**). "Computing and visualizing dynamic time warping alignments in R: The DTW package," J. Stat. Softw. **31**(7), 1–24.

Glasberg, B. R., and Moore, B. C. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**(1–2), 103–138.

Grice, M., and Baumann, S. (**2007**). "An introduction to intonation – functions and models," in *Phonetic Description and Teaching Practice*, edited by J. Trouvain and U. Gut (De Gruyter Mouton, Berlin), pp. 25–52.

Grice, M., Baumann, S., and Benzmüller, R. (**2005**). "German intonation in Autosegmental-Metrical phonology," in *Prosodic Typology: The*

Phonology of Intonation and Phrasing, edited by S.-A. Jun (Oxford University Press, Oxford, UK), pp. 55–83.
Grice, M., Baumann, S., Rössig, S., and Röhr, C. (**2022**). "GToBI: Übungsmaterialien zur deutschen Intonation" ("GToBI: Training materials for German intonation"), http://www.gtobi. uni-koeln.de/index.html (Last viewed November 21, 2022).
Gulati, S., Serra, J., Ishwar, V., and Serra, X. (**2016**). "Discovering rāga motifs by characterizing communities in networks of melodic patterns," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 20–25, Shanghai, China, pp. 286–290.
Hermes, D. J. (**1998**). "Measuring the perceptual similarity of pitch contours," J. Speech. Lang. Hear. Res. **41**(1), 73–82.
Himmelmann, N. P., and Kaufman, D. (**2020**). "Austronesia," in *The Oxford Handbook of Language Prosody*, edited by C. Gussenhoven and A. Chen (Oxford University Press, Oxford, UK), pp. 369–383.
Hirschberg, J. B., and Rosenberg, A. (**2007**). "V-Measure: A conditional entropy-based external cluster evaluation," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, June 28–30, Prague, Czechia, pp. 410–420.
Hirst, D., Di Cristo, A., and Espesser, R. (**2000**). "Levels of representation and levels of analysis for the description of intonation systems," in *Prosody: Theory and Experiment*, edited by N. Ide, J. Véronis, and M. Horne (Springer, Dordrecht, the Netherlands), Vol. 14, pp. 51–87.
Jun, S.-A. (**2005**). *Oxford Linguistics Prosodic Typology: The Phonology of Intonation and Phrasing* (Oxford University Press, New York).
Jun, S.-A. (**2014**). *Oxford Linguistics Prosodic Typology II: The Phonology of Intonation and Phrasing* (Oxford University Press, New York).
Kaland, C. (**2019**). "Acoustic correlates of word stress in Papuan Malay," J. Phon. **74**, 55–74.
Kaland, C. (**2020**). "Offline and online processing of acoustic cues to word stress in Papuan Malay," J. Acoust. Soc. Am. **147**(2), 731–747.
Kaland, C. (**2021a**). "Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours," J. Int. Phonetic Assoc. **53**, 159–188.
Kaland, C. (**2021b**). "The perception of word stress cues in Papuan Malay: A typological perspective and experimental investigation," Lab. Phonol. **12**(1), 1–33.
Kaland, C., and Baumann, S. (**2020**). "Demarcating and highlighting in Papuan Malay phrase prosody," J. Acoust. Soc. Am. **147**(4), 2974–2988.
Kaland, C., and Gordon, M. K. (**2022**). "The role of f0 shape and phrasal position in Papuan Malay and American English word identification," Phonetica **79**(3), 219–245.
Kaland, C., Kluge, A., and Van Heuven, V. J. (**2021a**). "Lexical analyses of the function and phonology of Papuan Malay word stress," Phonetica **78**(2), 141–168.
Kaland, C., Peck, N., Ellison, T. M., and Reinöhl, U. (**2021b**). "An initial exploration of the interaction of tone and intonation in Kera'a," in *Proceedings of the 1st International Conference on Tone and Intonation (TAI)*, December 6–9, Sonderborg, Denmark, pp. 132–136.
Kaland, C., Swerts, M., and Himmelmann, N. P. (**2023**). "Red and blue bananas: Time-series f0 analysis of contrastively focused noun phrases in Papuan Malay and Dutch," J. Phon. **96**, 101200.
Kaufman, L., and Rousseeuw, P. J. (**1990**). *Wiley Series in Probability and Statistics Finding Groups in Data* (John Wiley & Sons, Inc., Hoboken, NJ).
Klabbers, E., and Van Santen, J. P. (**2004**). "Clustering of Foot-Based pitch contours in expressive speech," in *Proceedings of the 5th ISCA Speech Synthesis Workshop*, June 14–16, Pittsburgh, PA.
Ladd, D. R. (**2008**). *Cambridge Studies in Linguistics Intonational Phonology*, 2nd ed. (Cambridge University Press, Cambridge, UK).
Levow, G.-A. (**2006**). "Unsupervised and semi-supervised learning of tone and pitch accent," in *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, June 4–9, New York, pp. 224–231.
Lin, J., Keogh, E., Lonardi, S., and Chiu, B. (**2003**). "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge discovery—DMKD '03*, June 13, San Diego, CA, p. 2.
Mahalanobis, P. C. (**1936**). "On the generalised distance in statistics," Proc. Nat. Inst. Sci. India **2**(1), 49–55.
Möhler, G., and Conkie, A. (**1998**). "Parametric modeling of intonation using vector quantization," in *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, November 26–29, Blue Mountains, Australia, pp. 311–316.
Mori, U., Mendiburu, A., and , Lozano, A. J. (**2016**). "Distance measures for time series in R: The TSdist Package," The R J. **8**(2), 451–459.
Odé, C. (**1989**). *Russian Intonation: A Perceptual Description* (Rodopi, Amsterdam).
Pierrehumbert, J. (**1980**). "The phonology and phonetics of English intonation," Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.
Pierrehumbert, J., and Hirschberg, J. (**1990**). "The meaning of intonational contours in the interpretation of discourse," in *Intentions in Communication*, edited by P. R. Cohen, J. Morgan, and M. E. Pollack (MIT Press, Cambridge, MA).
Prekopcsák, Z., and Lemire, D. (**2012**). "Time series classification by class-specific Mahalanobis distance measures," Adv. Data Anal. Class. **6**(3), 185–200.
Prom-on, S., Xu, Y., and Thipakorn, B. (**2009**). "Modeling tone and intonation in Mandarin and English as a process of target approximation," J. Acoust. Soc. Am. **125**(1), 405–424.
Raškinis, G., and Kazlauskienė, A. (**2013**). "From speech corpus to intonation corpus: Clustering phrase pitch contours of Lithuanian," in *Proceedings of the 19th Nordic Conference of Computational Linguistics*, May 22–24, Oslo, Norway, pp. 353–363.
R Core Team (**2022**). "R: The R project for statistical computing," https://www.r-project.org/ (Last viewed November 4, 2022).
Reichel, U. D. (**2011**). "The CoPaSul intonation model," in *Studientexte Zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2011* (*Study Texts on Speech Communication: Electronic Signal Processing 2011*), edited by B. J. Kröger and P. Birkholz (TUDpress, Dresden), pp. 341–348.
Riesberg, S., and Himmelmann, N. P. (**2012**). "The DoBeS Summits-PAGE Collection of Papuan Malay," https://hdl.handle.net/1839/00-0000-0000-0019-FF78-5 (Last viewed July 11, 2019).
Riesberg, S., Kalbertodt, J., Baumann, S., and Himmelmann, N. P. (**2020**). "Using rapid prosody transcription to probe little-known prosodic systems: The case of Papuan Malay," Lab. Phonol J. Assoc. Lab. Phonol. **11**(1), 8.
Ritter, S., and Grice, M. (**2015**). "The role of tonal onglides in german nuclear pitch accents," Lang. Speech **58**(1), 114–128.
Rose, P. (**1987**). "Considerations in the normalisation of the fundamental frequency of linguistic tone," Speech Commun. **6**(4), 343–352.
R Studio Team (**2022**). "RStudio: Integrated Development for R," https://www.rstudio.com/ (Last viewed November 4, 2022).
Sakoe, H., and Chiba, S. (**1978**). "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoust, Speech, Signal Process. **26**(1), 43–49.
Seeliger, H., and Kaland, C. (**2022**). "Boundary tones in German wh-questions and wh-exclamatives—A cluster-based approach," in *Proceedings of the 11th International Conference on Speech Prosody 2022*, May 23–26, Lisbon, Portugal, pp. 27–31.
Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (**1992**). "ToBI: A standard for labeling English prosody," in *Proceedings of the Second International Conference on Spoken Language Processing*, October 12–16, Banff, Canada.
Stoet, G. (**2010**). "PsyToolkit: A software package for programming psychological experiments using Linux," Behav. Res. Methods **42**(4), 1096–1104.
Stoet, G. (**2017**). "PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments," Teaching Psychol. **44**(1), 24–31.
Watson, D., Tanenhaus, M., and Gunlogson, C. (**2008**). "Interpreting pitch accents in online comprehension: H* vs. L+H*," Cogn. Sci.: A Multidiscip. J. **32**(7), 1232–1244.
Watts, D. J., and Strogatz, S. H. (**1998**). "Collective dynamics of 'small-world' networks," Nature **393**(6684), 440–442.
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*, August 11, Berlin, Germany, pp. 1–9.