PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES
COMPUTER SCIENCES OPEN ACCESS

Structure in conversation: Evidence for the vocabulary, semantics, and syntax of prosody

Nadav Matalon[a,b,1], Eyal Weinreb[a,1], Dominik Freche[a], Erez Volk[c], Tirza Biron[d], Elisha Moses[a,2], and David Biron[e,f,2]

Affiliations are included on p. 10.

Edited by David Weitz, Harvard University, Cambridge, MA; received February 16, 2024; accepted March 17, 2025

Prosody, the musical facet of speech, is pivotal in human communication, and its structure and meaning remain subjects of ongoing research. In this study, we introduce a data-driven model for English prosody, based on large-scale analysis of spontaneous conversations. As a first step, we identify approximately 200 discernible prosodic patterns—which we view as building blocks of the prosodic vocabulary—and outline their properties and range of meanings. Next, we reveal a Markovian logic, akin to a syntax, for concatenating these elementary building blocks into coherent utterances. We identify distinct compound functions associated with pairs of consecutive patterns and show that the Markovian syntax is more prevalent in spontaneous prosody, as compared to scripted speech. These findings offer invaluable insights into the underlying mechanisms of conversational prosody: They empirically inform and refine existing theoretical concepts. The methodology we present, combining unsupervised analysis of large datasets of spontaneous speech with manual sampling of the results, could guide future research aimed at refining our model and expanding it to other languages.

language | speech | prosody | unsupervised analysis

Significance

In conversation, prosody complements words, forming a structured communication system distinct from, yet connected to, text. By analyzing large datasets of spontaneous conversations and clustering similar snippets of speech, we identify the fundamental building blocks of this system. Our findings reveal a prosodic vocabulary of a few hundred patterns (far fewer than the thousands of words in a core verbal vocabulary), which fulfill interactional and attitudinal functions. Just as syntax governs word combinations, we observe recurring prosodic structures where certain patterns follow others more frequently than chance. Such ubiquitous pairs were not detected in scripted speech. These results provide data-driven support for the analogy of prosody to a linguistic system with its own vocabulary, semantics, and a simple syntax.

Prosody is the musical aspect of speech, conveying information on multiple levels, in parallel to the text (1, 2). It includes quantifiable vocal features like pitch, loudness and timing, as well as voice quality (3, 4). Theories of prosody vary in their views on its underlying structure and core functions (5–11). Moreover, the extent and means of conveying information through prosody are debated, with competing theories needing empirical support (12). The present study describes an empirical probing of the use of prosody in spontaneous speech.

Here, we take a data-driven approach and examine prosody as a linguistic system in its own right, as though it were an unknown language. We explore a large volume of conversational English, answering recent calls to move away from controlled stimuli toward natural data (13–16). The conversational data are compared to scripted speech to identify distinct features of spontaneous prosody.

Our analysis starts with parsing Intonation Units (IUs) (17)—the primary prosodic unit of spoken discourse (2, 6, 10, 18–21). Each IU is a snippet of speech that typically consists of 1 to 4 words (compared to about 15 to 20 words per typical written sentence). It is prosodically coherent and well-delimited (22–24), temporally structured (25), and considered a universal characteristic of speech (26). IU boundaries are characterized by prosodic cues such as pitch, intensity, and speech rate. Modulations of the rate are considered the most reliable cues for IU boundaries (27) and are used here for automatic parsing of IUs (28).

A conversation thus consists of a string of IUs and each IU is associated with a specific string of words. When stripped of the text, each IU consists of a pitch contour (22). This pitch variation over time can be viewed as a particular instance of an element of the prosodic vocabulary. The possibility of identifying such a vocabulary in conversational English is the first question addressed in this study.

Current approaches to identify a prosodic vocabulary are "top–down": They assume that a finite-state grammar that uses a predetermined set of basic constituents generates IU-sized pitch contours (6, 9, 12, 29–31). The basic constituents occupy specific locations within the IU, thus creating its inner structure. Estimates of the vocabulary size are combinatorial and range from 10 to 20 commonly used meaningful contours (6, 30) to over 1,000 "well-formed" contours (9, 12, 20, 32).

In our approach, each IU in the database is represented by a vector combining pitch and intensity. These vectors are encoded in a latent space and then grouped by unsupervised clustering. The centroid of each cluster is a pitch contour that represents a single building block, and the set of all building blocks is the prosodic vocabulary. Thus, we

Author contributions: N.M., E.W., D.F., E.V., T.B., E.M., and D.B. designed research; N.M., E.W., E.V., and D.B. performed research; E.W., D.F., and E.V. contributed new reagents/analytic tools; N.M., E.W., T.B., E.M., and D.B. analyzed data; and N.M., E.W., D.F., E.V., T.B., E.M., and D.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]N.M. and E.W. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: elisha.moses@weizmann.ac.il or dbiron@statistics.uchicago.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2403262122/-/DCSupplemental.

Published April 21, 2025.

build a vocabulary using a bottom–up approach without assuming a predefined collection of the structural components that constitute an IU.

Clustering of pitch contours was previously attempted for speech synthesis (33, 34), for validating predefined intonation categories (35, 36), to explore prosody in indigenous languages (37, 38), and to enhance computer understanding of speech (39). This study presents a large-scale unsupervised analysis of prosodic patterns in spontaneous conversation and a data-driven description of prosody in conversational English.

Ideally, clustering of pitch contours would result in homogenous and well-separated groups, representing the prosodic patterns in the data. However, such clusters were shown to be imperfect, with significant overlap in the relevant space (40). Ambiguous clustering partly results from unavoidable noise in naturalistic recordings. In addition, the boundaries between prosodic categories are inherently fuzzy (41, 42), and a given IU may present features characteristic of more than one cluster (43). Therefore, a clustering approach should be qualitatively correct but quantitatively approximate, offering order-of-magnitude estimates of cluster numbers and sizes, not precise measurements.

Nevertheless, an approximate prosodic vocabulary suffices to ask: Are there rules for combining these elements? Can they, or their combinations, convey linguistic meaning? The possibility to concatenate IUs to form meaningful sequences has been previously proposed (9, 10, 18). For example, ref. 30 describes a recipe for the construction of four frequently used IU pairs, and in refs. 31, 44, and 45 the accentual properties of multi-IU structures are investigated. Recently, analyses of sequences of chimpanzees' calls (46) and whale songs (47) revealed structural complexities. Our analysis of full conversations with minimal filtering empirically explores recurring prosodic patterns.

The meaning of prosodic elements has been previously investigated. Some theoreticians focus on the conveyance of emotion (7, 8) or attitude (30). Others disregard emotive meaning and highlight prosody's part in fulfilling text-related functions, such as signaling information status and propositional relationships (11, 29). Such claims often rely on introspection and on specific, at times contrived, examples. In contrast, we examine prosodic functions empirically, with no prior assumptions regarding their nature. In this we are most aligned with qualitative approaches for the analysis of prosody in conversation (48–50). Accordingly, our analysis follows standard practices in such approaches (14, 51), such as stratified sampling of the data for manual analysis.

We present two key findings: First, spontaneously produced IUs cluster meaningfully without considering their associated text. We estimate that 200 (within a factor of 2) distinct prosodic patterns are prevalent in conversational English. A typical cluster's prosodic pattern may serve various context-dependent linguistic functions yet convey a broader attitudinal meaning. These results support identifying a finite vocabulary of IU-sized prosodic patterns in a language.

Second, pairs of specific clusters—i.e., an IU from one cluster immediately followed in conversation by an IU from a second cluster—occur significantly more frequently than expected by chance. Often, such pairs serve a specific linguistic function. This result does not extend to triplets or longer sequences. It suggests that a syntax-like combinatorial system governs IU-sized prosodic pattern order, largely following a Markovian process where each pattern depends on the preceding one. Taken together, our findings provide a data-driven approximation for the prosodic vocabulary and syntax in conversational English.

Results

Clustering IUs Using Pitch and Intensity. We analyzed two representative datasets of natural spoken language—the CallHome Corpus (CH) (52) and the Santa Barbara Corpus (SBC) (53)—and three audiobooks representative of scripted speech—CT^{pro}, $SOIaF^{pro}$, and $SOIaF^{amt}$ (*Methods*). The SBC is manually segmented to IUs, and CH and the audiobooks were automatically segmented using the method we reported in ref. 28.

IUs were divided into deciles according to their duration, and each decile was clustered using the algorithm outlined in Fig. 1*A*. Clusters smaller than 1% of the decile population were discarded, and the corresponding IUs were used in repeated clustering cycles (*Methods*). In all datasets, more than 90% of IUs were successfully clustered. The resulting cluster statistics are described in Table 1. As Table 1 shows, the corresponding silhouette scores do not indicate well-separated clusters (40). Rather, they reflect the gradient transition between prosodic categories (41, 42). As expected (39), the prosodic markedness of a given cluster is inversely related to its size (*SI Appendix*, Fig. S1). Large clusters reflect the tendency of speakers to maintain relatively constant pitch, within their vocal comfort zone (54–57).

Fig. 1 *C* and *D* describe the clustering results of our main dataset (CH). Out of 43,086 IUs that are eligible for acoustic analysis, 39,612 (91.9%) were successfully clustered: 33,576 (77.9%) in the first clustering cycle, and 6,036 (14%) in repeated cycles. A preliminary quality check distinguished between "real clusters" (n = 159), depicting genuine prosodic patterns, and "clusters of noise" (n = 55), grouping IUs with severe F0 (pitch) extraction errors. It was found that, to a great extent, the first cycle of clustering (n = 168) filters out such erroneous IUs (Fig. 1*C*).

Next, we randomly sampled 20 "real" clusters, 2 from each decile (n = 3,484 IUs). Manual examination of this sample revealed that 70% ± 1 of IUs in a given cluster adhere to a distinct prosodic pattern. The manual analysis included an evaluation of the automatic segmentation and of the extraction of F0. It was found that 12% ± 1 of IUs in a given cluster involve segmentation errors and 9% ± 1 contain F0 extraction errors (see Fig. 1*D* and similar results for the SBC in *SI Appendix*, Fig. S2) (mean ± SE). Misclassification of IUs stemmed mostly from these two error types. Thus, the resulting clusters depict identifiable and distinct prosodic patterns.

To address consistency, we combined our two conversational corpora and analyzed them as a single dataset. We note first that doubling the dataset size did not double the number of resulting clusters, nor did it change the overall distribution of cluster sizes (Fig. 1*B* and Table 1). We found that 35% ± 1 of the IUs found in a given cluster in one of the original datasets ended up in the same cluster in the joined dataset. A total of 60% ± 1 of the IUs found in a given cluster in one of the original datasets were found in three clusters from the joined dataset (mean ± SE). *SI Appendix*, Fig. S3 shows the mean pitch contours of the resulting clusters from the three datasets for an example decile (5th). As the figure shows, most clusters in the combined dataset have a clear equivalent in one or both of the original datasets.

Finally, when the clustering hyperparameters and algorithm were varied, the numbers of resulting clusters did not substantially change (*SI Appendix*, Tables S1 and S2). In addition, since the clustering process involves inherent randomness, we repeated it 100 times with fixed hyperparameter values and obtained consistent cluster-size distributions (*SI Appendix*, Fig. S4).

Clusters of IUs Exhibit Functional Polysemy and Share a Broad Linguistic Meaning. A meaningful clustering of prosodic patterns should group together IUs that exhibit a functional common

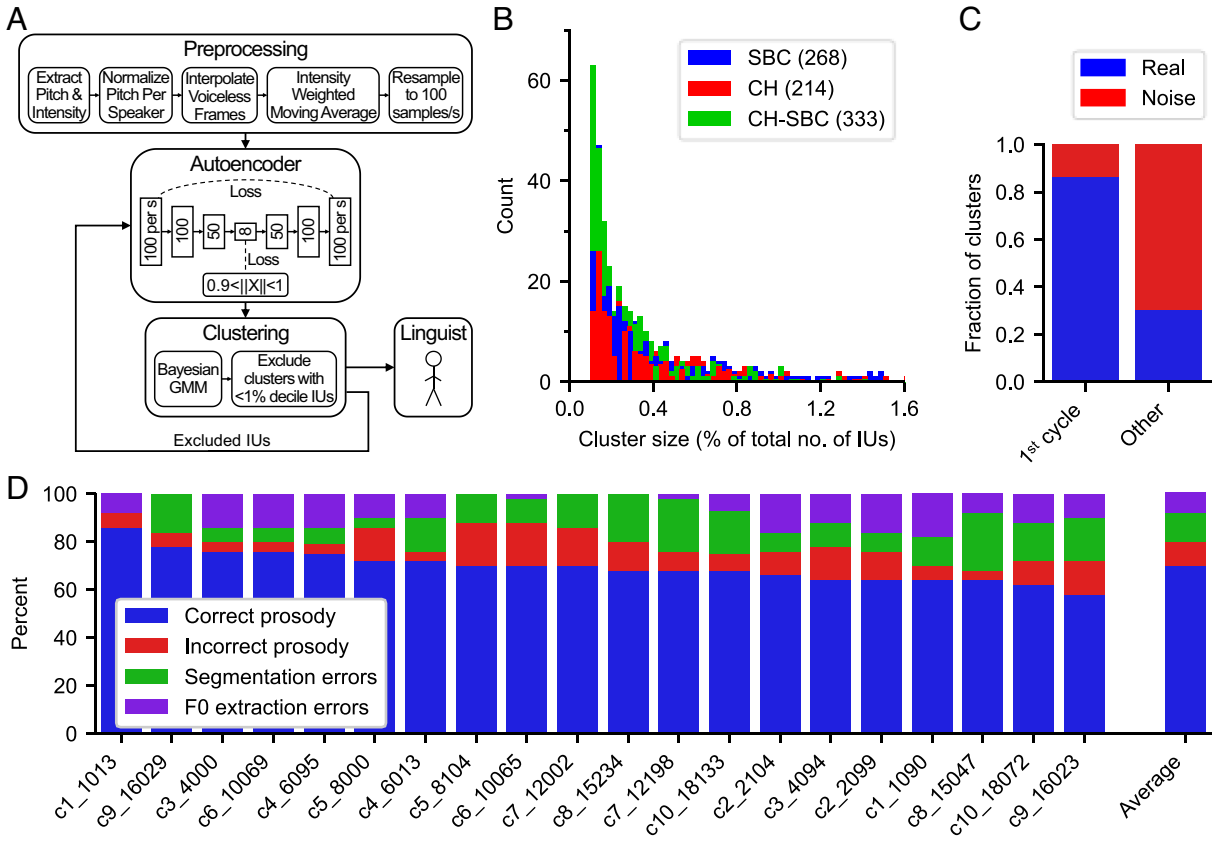

Fig. 1. Clustering of IUs. (*A*) A block diagram of the clustering process. (*B*) The distribution of the sizes of clusters in three spontaneous speech datasets, as a ratio of population size. (*C*) The fraction of "real" and "noise" clusters in the first and repeated clustering cycles (CH dataset). (*D*) The distribution of four categories—"correct prosody," "incorrect prosody" (IUs not adhering to the cluster's prosodic pattern), "segmentation error," and "F0 extraction error"—within 20 clusters (n = 3,484 IUs, CH dataset).

denominator, associated with at least one of the interactional functions achieved through prosody. A functional analysis of the CH dataset is summarized in Fig. 2. Within the random sample of 20 clusters described above, 78% ± 3 of the "correct prosody" IUs exhibit a recurring function (*Methods*) (mean ± SE). Importantly, all examined clusters exhibit multiple recurring functions, ranging between 2 to 5 (see Fig. 2*A* and similar results for the SBC in *SI Appendix*, Fig. S5). This functional polysemy, i.e., a single prosodic pattern fulfilling more than one linguistic function, is a significant observation. It is consistent with claims for the context-sensitive nature of prosody (49, 58) and its interrelatedness with lexis and syntax (59).

A post hoc control for text can reveal functional distinctions cued solely by prosody. As an example, Fig. 2*B* presents eight responsive IUs, consisting only of the word "yeah," taken from four different clusters. A mid-range small pitch fall (cluster c1_1013) accomplishes the function of simple *affirmation*; a mid-high-range small pitch rise followed by a large fall (cluster c3_4000) accomplishes the function of *weighty agreement*, i.e., with a preceding statement which is somewhat sensitive; A mid-range small pitch rise (cluster c2_2099) accomplishes the function of *newsmarking* (60), i.e., acknowledging the reception of new information; Finally, a high-range large pitch rise (c2_2104) functions as a *surprised newsmark*, i.e., in signaling the reception of new information that counters expectations (61). At the same time, our findings show that prosodic patterns can fulfill the same function when coupled with different segments of text. For example, IUs with texts other than yeah from cluster c2_2104 accomplish the aforementioned function *surprised newsmark*. Examples include "no," "really," or partial repetitions such as "was it" and "did it" (*SI Appendix*, Fig. S6).

Fig. 2*C* presents further support for the functional independence of prosody. Cluster c4_6095, which depicts a prosodic pattern characterized by a high-range and extremely large rise-fall pitch movement, exhibits three frequent recurring functions. The function

Table 1. Resulting clusters

Dataset	Number of clusters (first-cycle clusters)	Cluster size mean (IUs)	Cluster size range (IUs)	Coverage (% IUs)	Silhouette score
CH	214 (168)	185	43 to 1,226	91.9	0.11
SBC	268 (189)	134	40 to 618	90.9	0.11
CH-SBC	333 (220)	229	82 to 2,879	92.5	0.11
CT^{pro}	216 (160)	225	52 to 2,062	92.1	0.1
$SOIaF^{pro}$	226 (165)	307	74 to 2,731	93.4	0.1
$SOIaF^{amt}$	340 (228)	162	59 to 1,491	94.1	0.12

Spontaneous speech: CH = CallHome Corpus; SBC = Santa Barbara Corpus; CH-SBC = CH and SBC combined. Audiobooks: CT^{pro} = Century Trilogy book No. 3 (professional production); $SOIaF^{pro}$ = A Song of Ice and Fire book No. 5 (professional production); $SOIaF^{amt}$ = A Song of Ice and Fire book No. 5 (amateur production).

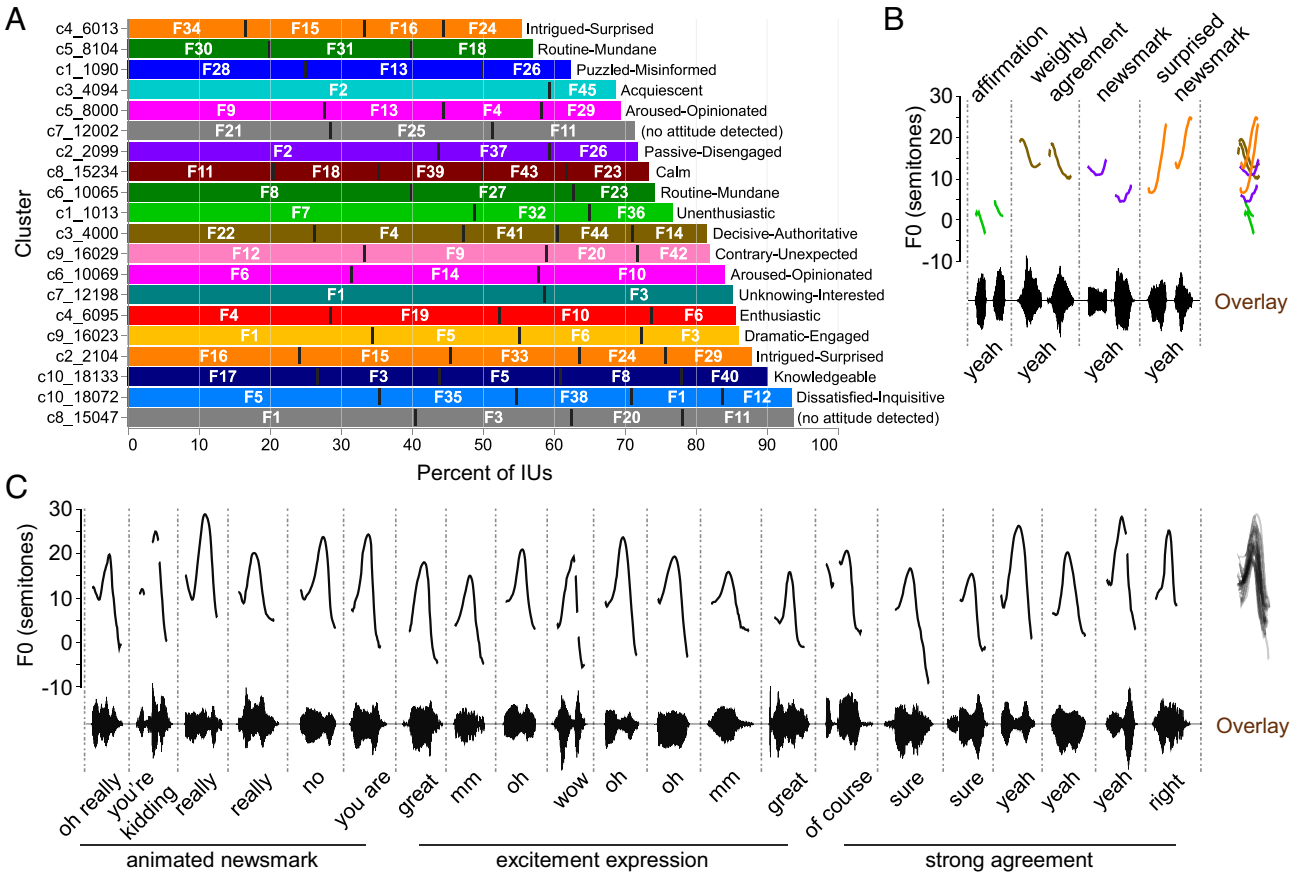

Fig. 2. Clusters and linguistic function. (*A*) Summary of the functions and attitudes for 20 clusters. Each bar represents a cluster, with the *x*-axis indicating the fraction of IUs exhibiting recurring functions. The stacked bars represent the frequencies of the functions in the cluster. Function numbers correspond to *SI Appendix*, Table S3. Attitudes associated with each cluster are distinguished by hues and specified on the *Right*. (*B*) Eight IUs consisting only of the word "yeah," taken from four different clusters and exhibiting prosodic form–function relations (attitudes marked by hues matching those in panel *A*, functions specified above). The overlay (*Right*) demonstrates the differences between the clusters. See Audio S1. (*C*) Representative sample of IUs from cluster c4_6095 with varying text, accomplishing three distinct *functions* and sharing one *attitude*—"enthusiastic." The overlay (*Right*) demonstrates the common form of the IUs within the cluster. See Audio S2. In panels *B* and *C*, presented pitch contours are not speaker normalized (0 semitones = 100 Hz). (CH dataset).

strong agreement was expressed, inter alia, using "sure," yeah, and "of course"; *animated newsmark*, evaluating received information as highly "remarkable" (62), was expressed, inter alia, using "oh really," "you're kidding," and "you are"; and *expression of excitement*, was performed, inter alia, using "mm," "oh" and "wow." The full list of single-cluster functions we identified is presented in *SI Appendix*, Table S3 (see *SI Appendix*, Table S4 for the SBC dataset).

Importantly, different functions of a given cluster are often related by a broader meaning, viz., the *attitude* of the speaker, defined as shared by at least 50% of the cluster's IUs. Here, *attitude* refers collectively to displays of epistemic, affiliative, affective, and deontic stances (63). Of the examined clusters, 18 out of 20 (90%) exhibited an attitude. Thus, attitude, as opposed to function, is less context sensitive, and relates more directly to prosody. In the case of cluster c4_6095 (Fig. 2*C*), 93% of IUs share the attitude *enthusiastic*. In general, our findings support the claim that dynamic prosody is indicative of the speaker's emotional involvement (64, 65). The full list of single-cluster attitudes we identified is presented in *SI Appendix*, Table S5.

Taken together, these results demonstrate that clusters do not only group prosodically similar IUs, but are also characterized by identifiable linguistic functions.

Markovian Dynamics of Prosodic Patterns. Following insights into the vocabulary and semantics of IU-sized prosodic patterns, we turn to investigate their syntactic behavior. Examining the cluster association of consecutive IUs revealed an excess of recurring cluster pairs as compared to randomized data. In contrast, recurring cluster triplets and higher-level structures were exceedingly rare. We note that more than half of the speaker turns in our data consist of only a single IU and that lengthy turns are somewhat rare in conversational English (*SI Appendix*, Fig. S7).

To explore whether common pairing of prosodic patterns could be serving interactional needs, a functional analysis was applied to a sample of 34 recurrent cluster pairs (*Methods*). In 17 of the 34 cluster pairs, a single linguistic function was shared by at least 50% of occurrences (Fig. 3*A*). For comparison, only 2 of the 20 analyzed single clusters exhibited such functional uniformity (Fig. 2*A*).

The two following examples demonstrate this effect. Cluster pair c9_16027-c6_10137, i.e., an IU from cluster c9_16027 immediately followed by an IU from cluster c6_10137, occurred 12 times. Both IUs are long and exhibit a flat pitch contour, slow speech rate, and nonemphatic voice quality. Eight of these 12 occurrences share a distinct interactional function: reporting a habitual behavior or continuous state, often as background information. Fig. 3*B* shows the pitch contour and text of two representative cases (Table 2).

Cluster pair c1_1042-c6_10069 occurred six times. It includes a combination of a short IU followed by a long one, both exhibit a high-range and extremely dynamic pitch contour. Three of these six occurrences function as a positive assessment given in response to informing (66). The first IU marks the reception of new information (60), and the second provides the positive assessment. This

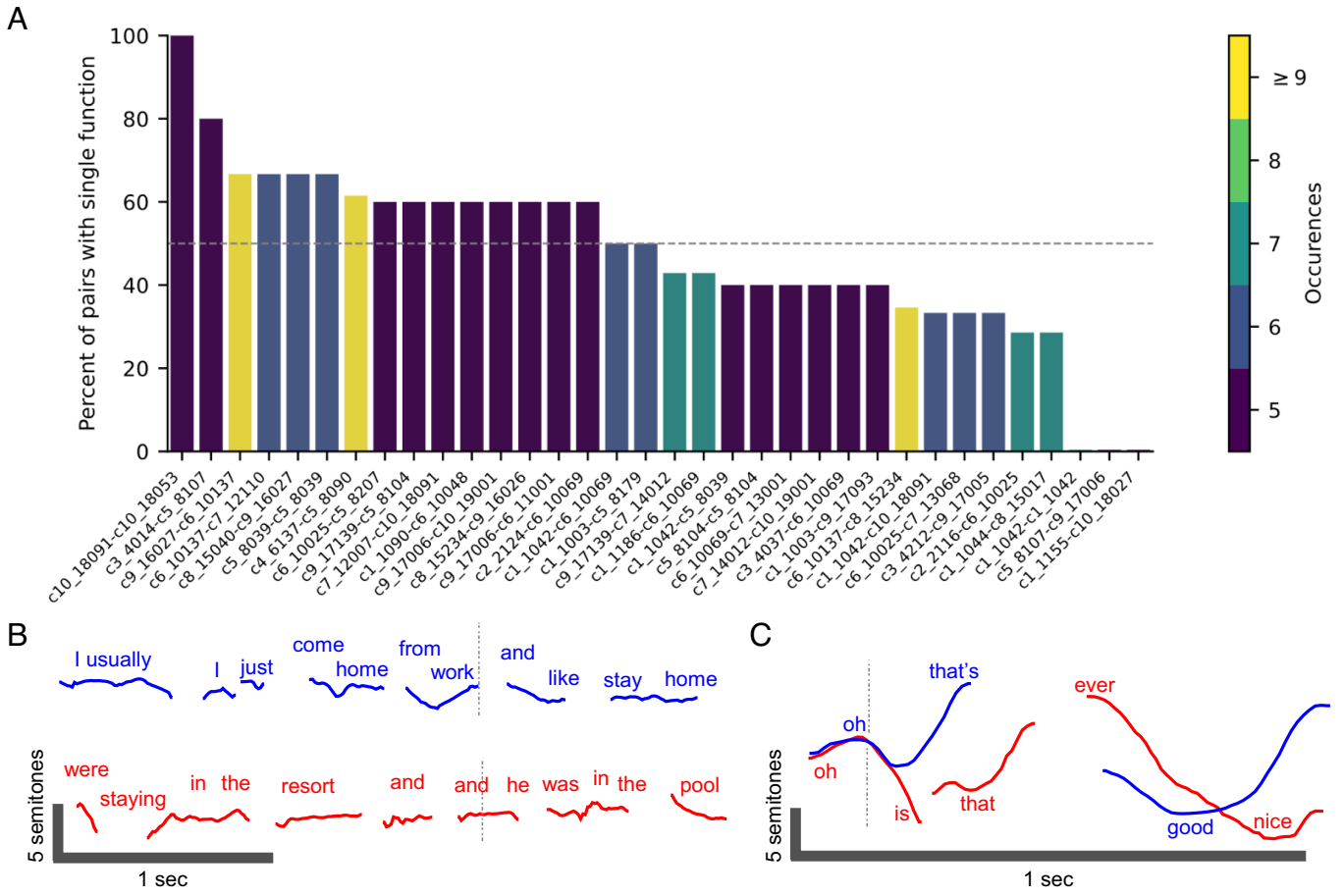

Fig. 3. Cluster pairs and linguistic function. (*A*) Fraction of IU pairs sharing a recurring function in 34 cluster pairs (the horizontal dashed line marks the 50% threshold of "functional uniformity"). (*B* and *C*) Text and pitch trace of two representative instances of cluster pair c9_16027-c6_10137 (*B*) See Audio S3, and cluster pair c1_1042-c6_10069 (*C*) See Audio S4. In panels *B* and *C*, vertical dashed lines represent boundaries between IUs.

exemplifies a well-defined conversational move accomplished by exactly two consecutive IUs. As in single clusters (Fig. 2*C*), Dynamic prosody signals the speakers' high emotive involvement. Fig. 3*C* shows the pitch contour and text of two representative cases (Table 3). The full list of cluster-pair functions we identified is presented in *SI Appendix*, Table S6.

The prevalence of recurring cluster pairs and the scarcity of higher-level structures suggest that the IU pair is a distinctive construct, regularly utilized in spontaneous conversation. The functional uniformity of cluster pairs and the identification of pairings which serve single interactional functions suggest that the IU pair is a locus for reciprocal contextualization, resulting in a disambiguation of prosodic meaning.

Statistical Analyses of Markovian Dynamics in IU Sequences. To quantitatively evaluate pairwise IU association, Fig. 4 *A* and *B* compare the observed distribution of the number of occurrences for each cluster pair in the CH and SBC datasets (blue) to a randomized sequence of IUs (gray) with the same structure of speaker turns (*Methods*). In both datasets, a clear signature emerges—an excess of highly recurring cluster pairs in the actual data compared to the randomized sequence. Using the Earth Mover's Distance (EMD) (67) as a measure of distance between the histograms we show that the excesses in both datasets are statistically significant ($P < 0.0001$, see *Methods*) (Fig. 4*E*).

This statistical signature is distinct from and subtler than the cluster-pair functional uniformity identified by manual examination. For example, while cluster pairs occurring five times in the CH dataset show an excess of 19% compared to the corresponding random permutations (154 vs. 129 pairs), ~56% of them exhibit functional uniformity. Manual analysis of ten cluster pairs with fewer occurrences (*Methods*) assured that the effect of functional uniformity holds even when the excess of recurring pairs in real data as compared to randomized data is not statistically significant.

Consistent with a meaningful pairing of IU-sized prosodic patterns, real data differ from randomized data also in within-pair ordering. Fig. 4 *C* and *D* show the distribution of absolute differences between the number of instances of cluster pairs (cluster i followed by cluster j) and of their reverse pairs (cluster j followed by cluster i). The differences between the actual (blue) and randomized data (gray) are statistically significant ($P < 0.0001$, see *Methods*).

Given the differences between spontaneous and scripted speech (*Discussion*), we hypothesized that IU pairwise association may also differ according to data type. To address this question, we applied the same statistical analysis to two professionally produced audiobooks, and to an amateur production of one of them (*Methods*). Fig. 4*E* Summarizes this analysis. Both of the professionally produced audiobooks (Blue) exhibited cluster-pair occurrence distributions that were similar to their corresponding randomized sequences. Importantly, excluding IUs corresponding to "holes" in the spontaneous speech datasets (created by speaker changes and unanalyzable IUs) did not significantly affect the corresponding EMDs. Interestingly, the amateur audiobook production yielded intermediate statistics, significantly below both spontaneous datasets but significantly above both professional audiobooks. These results support the notion that rehearsals and

Table 2. Text of the 12 occurrences of cluster pair c9_16027-c6_10137 (Fig. 3*B*) and their corresponding function

First IU	Second IU	Function
usually I just come home from work	and like stay home	habitual/continuous
were staying in the resort and and	he was in the pool	habitual/continuous
you know during the time we just call	on a daily basis	habitual/continuous
you're ending one year and starting the next year and	you know doing a lot with the	habitual/continuous
running through the	the um	habitual/continuous
you know everybody in the lobby you know	before leaving	habitual/continuous
president sorenson was down hunting somewhere	down in mexico or	habitual/continuous
but the trucks were very careful	which was um	habitual/continuous
from that one time I met her she's not very uh	uh	grounds for assessment
when she was talking to me and what I know	even more now	grounds for assessment
I mean like three weeks later it was like falling down	falling down off of her	disappointing fact
report so now you can uh	fill me in on yours	N/A (seg. error)

The IU pairs at the eight top rows share a distinct function.

adherence to written text alter the prosodic patterning of IUs in a measurable manner.

In sum, the functional uniformity of cluster pairs (shown in the previous section) along with the pairwise associations of IUs suggest that prosodic patterns in spontaneous speech follow a basic Markovian syntax: The probability of a particular pattern to appear depends on the preceding one.

Discussion

Languages function as structured systems of communication in which signs or symbols convey meaning. These building blocks constitute the vocabulary while syntax and grammar provide the rules for combining them into coherent expressions. In this study, we identify building blocks represented by centroid pitch contours and define them as the conceptual elements of the prosodic vocabulary. While this remains an analogy, we believe it is a useful one. To further illustrate, consider the dictionary entry "chair." It is an idealized concept, representing all instances of chairs. Similarly, our method constructs a prosodic "dictionary entry" by grouping instances of similar prosodic patterns and distilling a single representative form; akin to representing multiple real-world occurrences by the word chair (*SI Appendix*, Texts S1 and S2).

Our findings provide empirical evidence for the existence of a finite vocabulary of prosodic patterns at the IU scale. Different English datasets yielded 200 to 350 clusters of pitch contours. These values were stable to variations in the clustering technique and hyperparameter settings (*SI Appendix*, Table S1). In all datasets, more than 90% of IUs were successfully clustered, and, as expected (54–57), the resulting clusters show an inverse correlation between prosodic markedness and cluster size (*SI Appendix*, Fig. S1).

Table 3. Text of the six occurrences of cluster pair c1_1042-c6_10069 (Fig. 3*C*) and their corresponding function

First IU	Second IU	Function
oh	that's good	Responsive positive assessment
oh	is that ever nice	Responsive positive assessment
oh	that's great	Responsive positive assessment
that's	it's not	N/A (F0 extraction error)
well	one of the things danny rice	N/A (other's talk between IUs)
they	had to go yeah	N/A (segmentation error)

The IU pairs at the three top rows share a distinct function and the ones at the three bottom rows suffer from either an F0 extraction error or a segmentation issue.

Our reported cluster numbers are not exact, likely due to the impromptu and inexact nature of conversational prosody (41, 42) and to noise that is intrinsic to it (68). We interpret our results as an estimate of 200 (within a factor of 2) distinct prosodic patterns—higher than (6, 30) and lower than (9, 12, 20, 32). In contrast to these previous studies, we do not assume a predetermined set of constituents or specific positions within the IU that they may occupy. Nevertheless, the internal structure of our a posteriori patterns can be systematically described using such frameworks (*SI Appendix*, Fig. S8 and *Text S3*). Additionally, we identify a hierarchical similarity between our prosodic patterns, which depends both on the internal structure and the register (*SI Appendix*, Figs. S9 and S10).

Semantically, identified prosodic patterns serve multiple functions but generally convey a broad attitude. Many of the identified functions (*SI Appendix*, Tables S3 and S5) are characteristic of talk-in-interaction—responsive in nature (Figs. 2 *B* and *C* and 3*C*) or inviting a response. A specific function is a result of a combination of prosody and text within a specific sequential position (59, 65). Thus, our study reveals one-to-many form–function relations, as the algorithm ignores text and context. However, speaker attitude depends more directly on prosody (8) and can be shared across IUs with different functions. Indeed, 90% of the manually analyzed clusters were associated with a distinct attitude.

IUs follow a Markovian-like structure, where a pattern's occurrence depends mainly on the preceding one. This aligns with linguistic theory emphasizing the IU's centrality in spoken language (17). First, theory suggests that completed IUs can be of two kinds: "substantive," i.e., conveying ideas of events, states, or referents, or "regulatory," which coincide to a large extent with Discourse Markers (69). Identified cluster pairs often combine a short regulatory IU with a longer substantive one, forming a full conversational move (Fig. 3*C*). Second, substantive IUs adhere to a "one-new-idea" constraint. Thus, information which is easily packed in one written sentence, for example, a basic grammatical nexus including a subject and a finite verb (70), often occupies an IU pair in spontaneous speech.

Our findings do not reveal frequently occurring triplets of IUs. This does not mean that longer strings of IUs are never utilized or have linguistic meaning. For example, list constructions

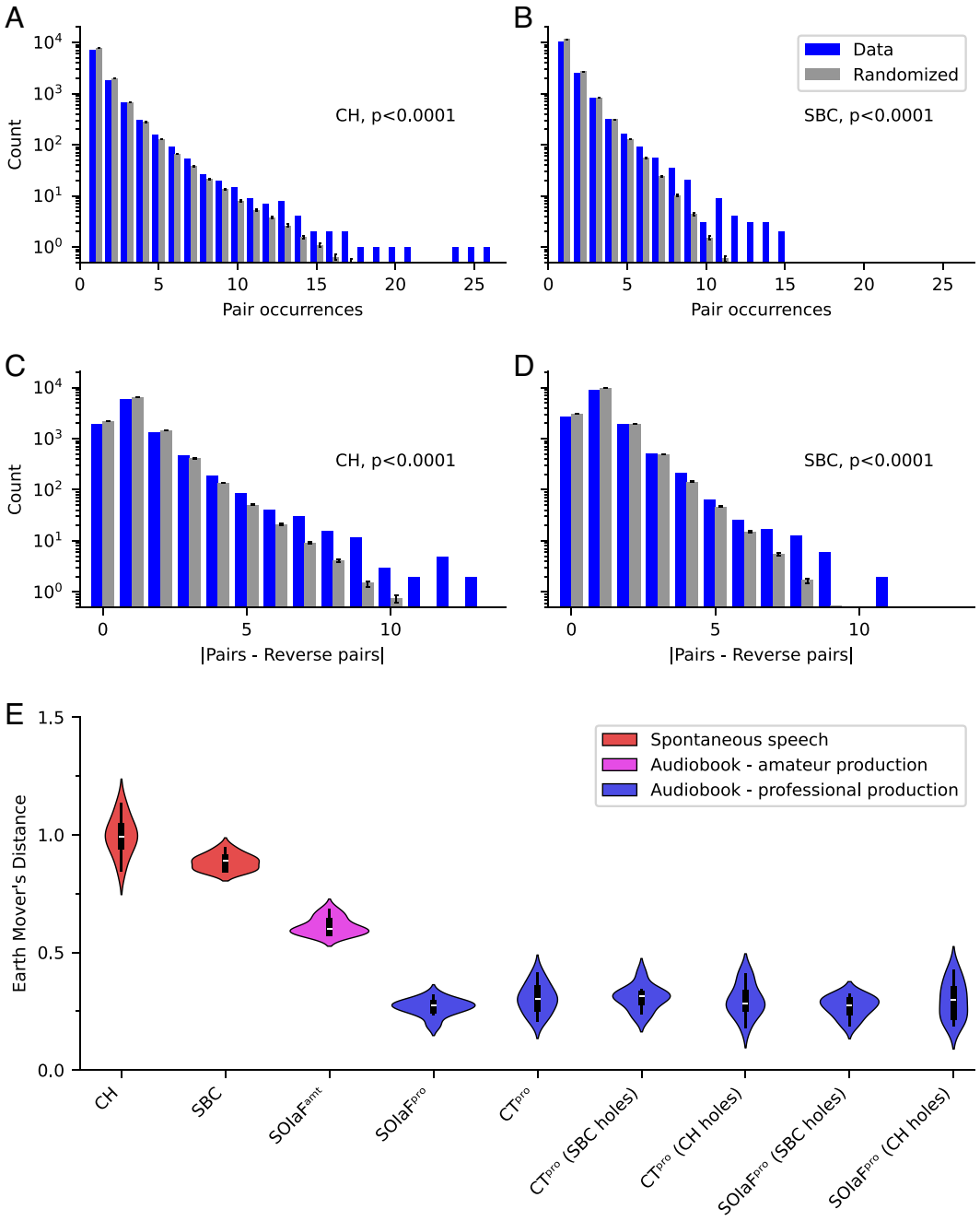

Fig. 4. Cluster-pair analysis. (*A* and *B*) Abundance of cluster pairs in real data compared to random sequences, in the CH (*A*) and SBC (*B*) datasets (semilog). (*C* and *D*) Absolute differences between the number of occurrences of cluster pairs and of their reverse pairs in real data compared to random sequences, in the CH (*C*) and SBC (*D*) datasets (semilog). In panels *A*–*D* error bars are shown for the random distributions. (*E*) EMD between real data and 100 random permutations in three data types. Each "violin" shows the median (white dot), interquartile range (dark box), min to max spread (dark line), and the estimated probability density of the data (a single density curve around the centerline rather than the baseline).

(71, 72) may include a concatenation of three IUs or more (73). However, such patterns are overall rare and do not readily appear in our analysis.

Despite the ubiquity and functional uniformity of IU-pairs, our findings support treating the IU, not the IU-pair, as the core unit of prosodic vocabulary. For one, over 50% of speaker turns in our conversational data consists of a single IU (*SI Appendix*, Fig. S7). IUs that appear in such singleton turns cluster well and appear in all clusters. Additionally, a prosodic pattern associated with a given cluster may appear in several pair combinations and can occupy either one of the two positions in a pair (Fig. 3*A*). These observations suggest that the meaningful pairing of IUs is evidence of a simple syntax.

Scripted (or read) speech is the basis for much of spoken language research. However, scripted speech is based on written forms (74), and thus preplanned and often rehearsed. It tends to be polished, stylized, and grammatically regular. In contrast, talk-in-interaction is colloquial and produced on-the-fly. It is typically beset with repairs and other types of discourse deviations, exhibiting "emergent" and flexible grammar (75). Furthermore, scripted speech is typically monologic and acted, while spontaneous conversation requires coordination between participants and carries social consequences. These discrepancies imply formal and functional differences in prosodic design which are not yet fully understood (76, 77). Comparing these data types showed that the statistical signature of cluster pairwise correlation in

Table 4. The number of words and IUs, total recording time, and segmentation type of the five datasets

Source	Number of words	Total number of IUs	Number of IUs (eligible for analysis)	Total recording duration (~hours)	Type of segmentation
CH	242,622	47,764	43,086	20	Automatic
SBC	249,445	70,069	39,509	23	Manual
CT^{pro}	368,490	61,631	52,687	37	Automatic
$SOIaF^{pro}$	416,474	81,886	74,257	49	Automatic
$SOIaF^{amt}$	416,537	69,912	58,515	49	Automatic

spontaneous speech was reduced in scripted speech (Fig. 4*E*). This measurable difference and the unique functions of talk-in-interaction support using spontaneous conversation data in spoken language research.

Our description of prosodic behavior in communication was not a foregone conclusion. A typical speaker utters thousands of IUs daily across various conversations, using thousands of words. Yet, their prosody largely draws from just a few hundred patterns, meaning that conversational English's prosodic vocabulary is far smaller than its theoretical limit.

Conversational messaging may be variable or imprecise, yet our vocabulary choices show regularity and meaning. They convey attitude and enhance linguistic function. As we string elements together, prosodic memory extends only to the previous IU. Unlike a composer envisioning an entire tune, we plan just two units ahead, spanning only a few seconds.

Limitations and Future Work. Understanding noise and its sources was crucial to our analysis. Conversational data present challenges beyond technical issues like background noise and recording quality. Intrinsic noise arises because humans are not "perfect prosody machines," leading to IU truncation from self-repair, overlapping speech, poor articulation, or inattentiveness. This creates correctly segmented IUs that cannot be meaningfully grouped, partly explaining our observed silhouette scores, similar to those in ref. 40. Human listeners use context to recognize transitional or incomplete IUs, while the automated analysis did not. As a result, intrinsic noise was prominent in the 11 ± 1% of IUs that were correctly assigned to clusters but did not exhibit a linguistic function (Fig. 2*A*). Since technology cannot eliminate this noise, understanding its frequency and impact is crucial for improving natural language processing and human–machine interactions.

Another goal would be curating larger, higher-quality datasets for robust analysis. A clean, sizable dataset may enable generalizability, predicting the cluster of unseen IUs. The CH-SBC dataset suggests some generalizability already. A definitive vocabulary set could support a data-driven dictionary, mapping prosodic patterns to their conversational functions in English.

More broadly, it would be highly interesting to incorporate text and context into a comprehensive model that could disambiguate prosodic meaning and directly associate an IU with its interactional functions.

The methodology we presented here could be applicable for exploring the structure of prosody in additional languages, cultural contexts, or social settings. For instance, our methodology could be used to characterize elements of language acquisition during normal or atypical development (78–80), effects of conditions related to aging on speech (81, 82), or speech within various institutional settings such as emergency calls (83), courtroom questioning (84), or classroom discourse (85). This incentivizes the construction of specialized yet extensive databases.

Conclusions. Unsupervised analysis of large datasets of spontaneous speech combined with manual sampling of the results identified signs of vocabulary, semantics, and syntax in the use of prosody on the IU scale. IUs can be meaningfully clustered based on their duration and the shape of their pitch curve and its register, i.e., its relative height normalized per speaker. The results yield an estimate of 200 (within a factor of 2) distinct prosodic patterns. A given IU-sized prosodic pattern typically exhibits several context-sensitive interactional functions, and a broader meaning, viz. speaker's attitude. Sequences of two IUs, but typically not longer, are frequently utilized by speakers to accomplish distinct compound functions, a behavior that is more characteristic of spontaneous speech.

Methods

Datasets. We analyzed the CallHome corpus (labeled CH) (52) and the SBC of Spoken American English (labeled SBC) (53). These corpora document spontaneous speech produced by a diverse group of speakers and in a variety of linguistic contexts and communicative situations. Both datasets are widely used in research on prosody in spoken communication, speech perception, and spoken language processing (86–88). The CH corpus consists of 120 unscripted dyadic telephone conversations between native speakers of American English. From each conversation, a contiguous 5- or 10-min segment is manually transcribed. The SBC consists of 60 audio files that record spontaneous speech of various genres, from multiparty kitchen conversations and couples' dialogues to child tutoring, guided tours, sermons, and university classes. A transcript accompanies each speech file, where manually identified IUs are time stamped. For means of validation, we also analyzed a joined version of CH and SBC, labeled CH-SBC. For scripted speech, we analyzed two professionally produced audiobooks: "Edge of Eternity," the third book of the "Century Trilogy" by Ken Follett, read by John Lee [text: (89), audio: (90)] (CT^{pro}) and "A Dance with Dragons," the fifth book of the series "A Song of Ice and Fire" by George R. R. Martin, read by Roy Dotrice [text: (91), audio: (92)] ($SOIaF^{pro}$). In addition, we analyzed an amateur production of "A Dance with Dragons" read by @DavidReadsASoIaF (93) ($SOIaF^{amt}$). Table 4 specifies the size of these datasets.

Alignment and Segmentation. Prior to clustering, speech was segmented into IUs. The SBC was manually segmented previously (53). CH, CT^{pro}, $SOIaF^{pro}$, and $SOIaF^{amt}$ were segmented using our previously published automatic segmentation method (28). In brief, we first automatically align speech at the phone level using the Montreal Forced Alignment (94). Speech is then segmented at interword pauses of at least 0.3 s, and then further segmented at interword points where the mean phone speed decreases by at least 80%. The mean phone speed is calculated from the 0.3 s preceding the end of each word. In sum, our segmentation method relies on modulations in speech timing (slowing down and stopping) which are considered to be the most reliable cues for IU boundaries (22, 27, 95).

Feature Extraction. Speech pitch and intensity were extracted from source audio using the parselmouth interface (96) to the Praat software package (97). The pitch was extracted at 100 samples/s. The intensity was extracted at 125 samples/s and then interpolated to 100 samples/s to match the pitch.

Pitch was converted from Hertz to a semitone scale, which is closer to human perception of pitch (98, 99). The pitch was then speaker-normalized by subtracting the median pitch of the specific speaker, as calculated over the entire dataset (100).

Calculating Pitch Dynamics. For each cluster, we calculate a measure of *pitch dynamics* to quantitatively estimate prosodic markedness. This measure was calculated based on the mean pitch contour of the cluster. It is defined as the product of pitch change, i.e., the min-max range of the pitch contour; and the pitch register, i.e., the absolute value of the average deviation of the pitch contour from the speaker's median.

Filtering. IUs containing under five samples of defined pitch, IUs over 3 s in duration, and IUs with no words were excluded from the dataset prior to processing. Note that nonlexical elements (e.g., *oh*, *wow*, and *m-hm*) are considered "words" and IUs consisting only of such elements were not filtered out. In the SBC dataset, where word-level manual annotations are available, we also removed IUs which were marked as containing environmental noise, overlapping speech of multiple speakers, unknown speakers, unintelligible speech, and whispers. Whispers and overlapping speech were removed only if over a third of the words in the IU were marked as such.

Splitting Data into Deciles. The IUs were split into 10 equally sized groups ("deciles"), such that each decile contains IUs of similar durations. The entire analysis, including data preparation, autoencoder (AE) training, and the subsequent clustering, is applied to each of these deciles separately.

Low-Dimensional Data Representation.

Input data preparation. For each IU, sample points where pitch is undefined are linearly interpolated. The pitch is then smoothed using a weighted moving average with a 10-point sample-window length. The weights used for averaging are the intensity at each sample point, normalized to a [0, 1] scale within each IU. The resulting pitch vectors are then resampled to a common length according to the median length of the IUs in the decile, at 100 samples per second. Thus, for the shortest decile across the analyzed datasets, a length of 12 points per IU was used, and for the longest 271.

AE architecture. We trained a multilayer AE network (101) on the IU pitch vectors in order to create the lower-dimensional latent space in which the clustering of these vectors will later occur.

The AE network layer architecture is shown schematically in Fig. 1*A*. The layers (and layer sizes in parentheses) were input (100/s), dropout, hidden (100), dropout, hidden (50), latent (8), hidden (50), hidden (100), output (100/s). The hidden and latent layers are fully connected layers that use the Scaled Exponential Linear Unit activation function (102). The output layer is a fully connected layer with no activation function. We used dropout layers to prevent overfitting (103), with each dropout layer zeroing out 20% of its input.

The Keras software package was used to construct and train the AE (104). Training was done with the following hyperparameters: 1,000 training epochs with data shuffling, a batch size of 500, a train-test split of 80 to 20%, the mean-squared-error loss function, and the Adam optimizer (105).

In addition, the Sliced-Wasserstein (SW) technique (106) was used to impose an approximate distribution on the vectors in the latent space. In our case, we chose a uniform distribution over a "thick-walled spherical shell," defined as $0.9 \leq \|\bar{x}\| < 1$, where $\bar{x}$ is the eight-dimensional latent vector. For each training batch, we sample 50 random slices and add the SW loss component with a weight of 100 to the overall loss of the AE.

Clustering. After training the AE, we used it to encode all of the IUs (in the decile), i.e., to find the eight-dimensional vector corresponding to each IU in the latent space of the trained AE. Clusters of IUs were then identified in this latent space.

Clustering was done using the variational Bayesian Gaussian mixture model implemented by the scikit-learn software package (107). The following parameters were used for clustering: 1,000 maximum iterations, 300 mixture components, and a weight-concentration-prior of 1/300. If there is a failure to converge, the clustering process is repeated up to three times with different random seeds.

Clusters containing less than 1% of the IUs in the decile are considered invalid. All the IUs in these invalid clusters are taken (without the valid IUs) and passed again through the process of AE training and clustering from scratch. This process is repeated until there are no additional valid clusters found, or until there are less than 600 IUs remaining in invalid clusters. The 1% threshold limits the minimal size of a cluster and the maximum number of clusters in each decile. The number of clusters we identify remains well below this limit (*SI Appendix*, Fig. S11).

Silhouette Scores. Silhouette scores were generated using the scikit-learn software package (107). Scores were calculated for each IU, then averaged into a score for each cluster (from all the deciles analyzed), and then averaged to get an overall score for the entire dataset.

Cluster and Cluster-Pair Manual Evaluation. Manual evaluation of clusters and cluster-pairs was performed on the primary run of the CH dataset with the aim of assessing the prosodic resemblance of IUs within clusters and their functional uniformity. In the following, we specify the different stages in this process and explain its rationale in light of related work.

Cluster prosodic resemblance. First, we examined a visualization of the extracted F0 contours of all IUs in each of the clusters. We identified clusters that include a majority of IUs with severe extraction errors. This resulted in a categorization of clusters to real clusters and "noise" clusters.

Second, we randomly sampled each decile to obtain a representative sample of 20 real clusters. The stratified sampling scheme (where the strata are the duration deciles) aimed at revealing prevalent trends or systematic problems. Clusters in the sample were of varying sizes, ranging from 56 to 647 IUs (in total n = 3,484 IUs). The sampled clusters were manually analyzed by repeated listening supported by acoustic analysis. The goals were to determine the degree to which the units in a given cluster adhere to a common prosodic pattern and to validate the automatic segmentation to IUs and the extraction of F0. This analysis provided a four-way categorization of IUs within a single cluster: "correct prosody"–the IU matches the prosodic pattern of the cluster; "incorrect prosody"–the IU does not match the cluster prosody; "segmentation error"–the segment of speech is not a valid IU; and "F0 extraction error"–the IU was assigned to the cluster due to an F0 extraction error.

Cluster function and attitude. The third step included a functional analysis of the same 20 clusters. From each cluster, a sample of 50 IUs was examined in its conversational context, in search for recurring *functions*. A function was required to be observed at least three times in the 50 IUs sample for the observation to be considered noncoincidental. In clusters with more than 35 "correct prosody" IUs, we required a function to occur at least four times in order to be considered recurring.

Additionally, when possible, clusters were labeled for the speaker's *attitude*. Attitude is taken here as a broader linguistic meaning than function, referring collectively to displays of epistemic, affiliative, affective, and deontic stances (108). A cluster was assigned an attitude only when at least 50% of its IUs share it.

Rather than assuming a closed set of predetermined categories, we allow for an open-ended set of functions and attitudes. This is consistent with current approaches to the study of spoken language (14, 51), which advocate a data-driven delineation of linguistic categories. The labels for single-cluster function and attitude are listed in *SI Appendix*, Tables S3 and S5. These lists do not purport to include all possible functions and attitudes. Rather, they represent the categories which emerged from the data analyzed. See *SI Appendix, Text S4* for a detailed explanation of our functional analysis and its relation to existing literature.

To validate our functional analysis, we performed a control experiment. The experiment involved annotation of a random sample of 80 IUs from our main dataset (CH), by three naive annotators, with no access to the clustering solution. Similar to ref. 109, the annotators achieved an average agreement of 71 ± 3% with our original analysis when considering specific functions, and an average agreement rate of 80 ± 2% when considering function categories (*SI Appendix*, Table S7), i.e., when allowing confusion with 1 to 3 semantically related functions (*SI Appendix*, Table S8). See *SI Appendix, Text S5* for a detailed description of the control experiment.

The *functions* we identify are closely related to the concept of Dialogue Acts (DA), whose roots are in the theory of Speech Acts (110). Over the years, the taxonomy of five basic speech-act types (111) has been significantly extended and refined, resulting in lists which range between several tens to a few hundred DA labels (112–115). *SI Appendix*, Tables S9 and S10 compare the function labels that we used with two central DA sets: SWBD-DAMSL (113) and ISO 24617-2 (112, 116).

The comparison underscores the strong alignment of our labeling system with existing research. 32 out of the 45 (73%) functions we have associated with the identified prosodic patterns have an equivalent in at least one of the

corresponding lists. Additionally, the overlap between our list and each of the two others is similar to that between the two DA sets themselves, all ranging between 49 to 60%. This point highlights a lack of consensus regarding the classification of functions.

The observed differences between our classification and that of the two DA label sets originate in recent developments in the study of conversational actions (117) within the fields of Conversation Analysis (51) and Interactional Linguistics (14), and in our focus on prosodic functions rather than on a primarily text-based approach. As a result, functions that are predominantly text-based, such as greeting, thanking, and apologizing, which appear in both sets, are absent from our inventory. Additionally, the granularity of our labels varies, depending on whether the differentiation relies more on prosody or on textual content (*SI Appendix*, Tables S9 and S10 and *Text S4*).

Our notion of *attitude* is related to emotion (118). Psychological theory has suggested more than a dozen different sets of basic emotions, each one includes between 2 and 11 labels (119). Most familiar are the set of six emotions, proposed by (120, 121), commonly used in speech emotion recognition studies (122, 123), and that of eight emotions, proposed by ref. 124.

Such basic taxonomies were often found to be insufficient for describing actual data, and analysts expanded and refined them (125) or, like us, allowed for an open-ended set of labels (126, 127). In addition, given our broad definition of attitude (see above), attitude labels may be related to other known domains, such as the politeness-rudeness continuum (128, 129) and the epistemic gradient (130).

Cluster-pair function. A functional analysis of cluster pairs was performed independently of the results of the single clusters functional analysis. We analyzed a sample of 34 cluster pairs, which collectively included 225 IU pairs. This sample included cluster pairings with at least five occurrences (ranging between 5 and 26, mean of 6.6, SD 3.8) and exhibited the highest occurrence rate, after normalizing for cluster size (cluster sizes range between 91 and 916, mean of 334, SD 201). The reason to focus on cluster pairs of at least five occurrences, despite comprising only a small fraction of the data (Fig. 4*A*), was that our functional analysis relies on the identification of recurring phenomena. The list of labels for cluster-pair function can be found in *SI Appendix*, Table S6. For means of validation, however, an additional sample of 10 cluster pairs was examined. The validation sample included five pairs with three occurrences and five with four occurrences.

Randomized Assignment of IUs to Clusters. Randomized datasets were generated by randomly permuting the cluster ID labels of the corresponding dataset. Thus, label i appeared S_i times, where S_i was the size of the ith cluster and locations with missing data and/or speaker changes were preserved. Pairs (or triplets, quadruplets, etc.) of IUs were then counted in a randomized sequence that resembled the original data in the frequencies of labels and in the positions of holes in the sequence. For each of the datasets, this procedure was repeated 100 times and the results were averaged to produce histograms.

Statistical Tests. To compare the histograms in Fig. 4, *P*-values were calculated as follows. The null and alternative hypotheses were concerned with IU pairs that appeared between 3 and 13 times (4 *A*-*B*) or 3 to 9 times (4 *C*-*D*). H0 stated that the numbers of occurrences in the data were drawn from the same distributions as those for the randomized data. Alternatively, they were drawn in any order from distributions with higher means.

For each bin corresponding to ≥3 occurrences, the empirical distribution was calculated using 100 randomized sequences of clusters (see above) and a kernel-density estimate with Gaussian kernels. These empirical distributions were integrated to obtain the probability of a count at least as large as the observed one. The resulting probabilities were multiplied (counts in separate bins were only weakly correlated as the majority of pairs contributed to the first two bins in each histogram). Since the order of the counts was arbitrary, the probabilities were further multiplied by the factorial of the number of bins tested. The resulting *P*-value estimated the probability, under the null hypothesis, of observing counts at least as large as the data in any order.

To compare the means of the groups of EMDs in Fig. 4*E*, *P*-values were obtained using Tukey's range test (honestly significant difference). In brief, this test controls the family-wise error rate, i.e., the probability of making at least a single type I error (false positive), across multiple pairwise comparisons.

Data, Materials, and Software Availability. Code and data are available at https://github.com/EyalWeinreb/Structure-in-conversation (131). Previously published data were used for this work (52, 53, 89–93).

ACKNOWLEDGMENTS. We would like to thank John Du-Bois, David Harel, Maya Inbar, Asaf Marron, and Tsvi Tlusty for useful conversations.

Author affiliations: [a]Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 7610001, Israel; [b]Department of Linguistics, Hebrew University of Jerusalem, Jerusalem 9190501, Israel; [c]NeuraLight Inc., Tel Aviv 6713818, Israel; [d]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 7610001, Israel; [e]The Data Science Institute, The University of Chicago, Chicago, IL 60637; and [f]The Department of Statistics, The University of Chicago, Chicago, IL 60637

1. J. K. Burgoon, V. Manusov, L. K. Guerrero, *Nonverbal Communication* (Routledge, 2021).
2. A. Wichmann, *Intonation in Text and Discourse* (Routledge, 2014).
3. B. Szczepek Reed, *Prosodic Orientation in English Conversation* (Palgrave Macmillan UK, 2007).
4. P. Ladefoged, *Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques* (Blackwell Publishing, 2003).
5. J. R. Firth, "Sounds and prosodies" in *Prosodic Analysis*, F. R. Palmer, Eds. (Oxford University Press, 1970), pp. 1–26.
6. M. A. K. Halliday, *Intonation and grammar in British English* (De Gruyter, 1967).
7. D. Bolinger, *Intonation and its Parts. Melody in Spoken English* (Stanford University Press, 1986).
8. D. Bolinger, *Intonation and Its Uses. Melody in Grammar and Discourse* (Stanford University Press, 1989).
9. J. Pierrehumbert, *The Phonology and Phonetics of English Intonation* (MIT, 1980).
10. A. Cruttenden, *Intonation* (Cambridge University Press, 1997).
11. D. R. Ladd, *Intonational Phonology* (Cambridge University Press, ed. 2, 2008).
12. J. Pierrehumbert, "Tonal elements and their alignment" in *Prosody: Theory and Experiment*, M. Horne, Ed. (Springer, Dordrecht, 2000), pp. 11–36.
13. E. A. Schegloff, "Turn organization: One intersection of grammar and interaction" in *Interaction and Grammar*, E. Ochs, E. A. Schegloff, S. A. Thompson, Eds. (Cambridge University Press, 1996), pp. 52–133.
14. E. Couper-Kuhlen, M. Selting, *Interactional Linguistics: Studying Language in Social Interaction* (Cambridge University Press, 2018).
15. S. A. Nastase, A. Goldstein, U. Hasson, Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *Neuroimage* **222**, 117254 (2020).
16. L. S. Hamilton, A. G. Huth, The revolution will not be controlled: Natural stimuli in speech neuroscience. *Lang. Cogn. Neurosci.* **35**, 573–582 (2020).
17. W. L. Chafe, *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing* (University of Chicago Press, 1994).
18. E. O. Selkirk, *Phonology and Syntax: The Relation between Sound and Structure* (The MIT Press, 1984).
19. M. Nespor, I. Vogel, *Prosodic Phonology* (De Gruyter, 2007).
20. M. E. Beckman, J. Hirschberg, S. Shattuck-Hufnagel, "The original tobi system and the evolution of the ToBi framework" in *Prosodic Typology* (Oxford University Press, Oxford, 2005), pp. 9–54.
21. D. Crystal, *Prosodic Systems and Intonation in English* (Cambridge University Press, 1969).
22. J. W. Du Bois, S. Schuetze-Coburn, S. Cumming, D. Paolino, "Outline of discourse transcription" in *Talking Data: Transcription and Coding in Discourse Research*, J. A. Edwards, M. D. Lampert, Eds. (Psychology Press, 1993), pp. 45–89.
23. S. Izre'el, H. Mello, A. Panunzi, T. Raso, Eds., *In Search of Basic Units of Spoken Language* (John Benjamins Publishing Company, 2020).
24. S. Shattuck-Hufnagel, A. E. Turk, A prosody tutorial for investigators of auditory sentence processing. *J. Psycholinguist Res.* **25**, 193–247 (1996).
25. M. Inbar, E. Grossman, A. N. Landau, Sequences of Intonation Units form a ~1 Hz rhythm. *Sci. Rep.* **10**, 15846 (2020).
26. N. P. Himmelmann, M. Sandler, J. Strunk, V. Unterladstetter, On the universality of intonational phrases: A cross-linguistic interrater study. *Phonology* **35**, 207–245 (2018).
27. F. Seifart *et al.*, The extent and degree of utterance-final word lengthening in spontaneous speech from 10 languages. *Linguistics Vanguard* **7**, 0190063 (2021).
28. T. Biron *et al.*, Automatic detection of prosodic boundaries in spontaneous speech. *PLoS One* **16**, e0250969 (2021).
29. J. Pierrehumbert, J. Hirschberg, "The meaning of intonational contours in the interpretation of discourse" in *Intentions in Communication*, P. R. Cohen, J. Morgan, M. E. Pollack, Eds. (MIT Press, 1990), pp. 271–311.
30. J. D. O'Connor, G. F. Arnold, *Intonation of Colloquial English* (Longaman, London, 1973).
31. M. E. Beckman, J. B. Pierrehumbert, Intonational structure in Japanese and English. *Phonology Yearbook* **3**, 255–309 (1986).
32. K. Silverman *et al.*, "ToBI: A Standard for Labelling English Prosody" in *the 1992 International Conference on Spoken Language Processing* (ISCA Archive, University of Alberta Press, Edmonton, 1992), vol. 92, pp. 867–870.
33. E. Klabbers, J. V. Van Santen, "Clustering of foot-based pitch contours in expressive speech" in *5th ISCA Speech Synthesis Workshop* (2004), pp. 73–78.
34. G. Demenko, A. Wagner, "The stylization of intonation contours" in *Proceedings of the 3rd International Conference on Speech Prosody*, R. Hoffmann, H. Mixdorf, Eds. (TUDpress, Dresden, 2006), p. 254.
35. G.-A. Levow, "Unsupervised and semi-supervised learning of tone and pitch accent" in *Proceedings of the Human Language Technology Conference of the NAACL*, R. C. Moore, J. Bilmes, J. Chu-Carroll, M. Sanderson, Eds. (Association for Computational Linguistics, 2006), pp. 224–231.

36. A. Rosenberg, J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure" in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, J. Eisner, Ed. (Association for Computational Linguistics, 2007), pp. 410–420.
37. C. Kaland, Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours. *J. Int. Phon. Assoc.* **53**, 159–188 (2023).
38. S. Babinski, C. Bowern, Automatic categorization of prosodic contours in Bardi. *Proc. Linguistic Soc. Am.* **7**, 5218 (2022).
39. T. Yoshimura, S. Hayamizu, H. Ohmura, K. Tanaka, "Pitch pattern clustering of user utterances in human-machine dialogue" in *Proceeding of Fourth International Conference on Spoken Language Processing*, H. T. Bunnell, W. Idsardi, Eds. (IEEE, New Castle, DE, 1996), pp. 837–840.
40. S. Calhoun, A. Schweitzer, "Can intonation contours be lexicalised? Implications for discourse meanings" in *Prosody and Meaning*, G. Elordieta, P. Prieto, Ed. (De Gruyter Mouton, 2012), pp. 271–327.
41. D. Barth-Weingarten, *Intonation Units Revisited: Cesuras in Talk-in-Interaction* (John Benjamins Publishing Company, 2016).
42. D. Escudero-Mancebo, C. González-Ferreras, C. Vivaracho-Pascual, V. Cardeñoso-Payo, A fuzzy classifier to deal with similarity between labels on automatic prosodic labeling. *Comput. Speech Lang.* **28**, 326–341 (2014).
43. N. Matalon, *Prosodic Variation in Yes-No Questions and its Functional Correlates* (Hebrew University of Jerusalem, Jerusalem, 2024).
44. S. Schuetze-Coburn, M. Shapley, E. G. Weber, Units of intonation in discourse: A comparison of acoustic and auditory analyses. *Lang. Speech* **34**, 207–234 (1991).
45. M. Liberman, J. Pierrehumbert, "Intonational invariance under changes in pitch rangeand length" in *Language Sound Structure*, M. Aronoff, R. T. Oehrle, Eds. (MIT Press, 1984), pp. 155–233.
46. C. Girard-Buttoz *et al.*, Chimpanzees produce diverse vocal sequences with ordered and recombinatorial properties. *Commun. Biol.* **5**, 410 (2022).
47. I. Arnon *et al.*, Whale song shows language-like statistical structure. *Science* **1979**, 649–653 (2025).
48. J. Kelly, J. Local, *Doing Phonology: Observing, Recording, Interpreting* (Manchester University Press, 1989).
49. E. Couper-Kuhlen, M. Selting, *Prosody in Conversation* (Cambridge University Press, 1996).
50. E. Couper-Kuhlen, C. E. Ford, Eds., *Sound Patterns in Interaction* (John Benjamins Publishing Company, 2004).
51. E. A. Schegloff, *Sequence organization in interaction: A primer in conversation analysis I* (Cambridge University Press, 2007), **vol. 1**.
52. A. Canavan, D. Graff, G. Zipperlen, CALLHOME American English Speech. LDC catalog. https://catalog.ldc.upenn.edu/LDC97S42. Accessed June 2020.
53. J. W. Du Bois *et al.*, Santa Barbara corpus of spoken American English, Parts 1-4 (Linguistic Data Consortium, Philadelphia, PA, 2005). https://www.linguistics.ucsb.edu/research/santa-barbara-corpus#access. Accessed June 2020.
54. M. Lennes, M. Stevanovic, D. Aalto, P. Palo, Comparing pitch distributions using Praat and R. *Phonetician* **111–112**, 35–53 (2015).
55. J. Gorisch, B. Wells, G. J. Brown, Pitch Contour matching and interactional alignment across turns: An acoustic investigation. *Lang. Speech* **55**, 57–76 (2012).
56. J. Laver, *Principles of Phonetics* (Cambridge University Press, 1994).
57. Y. Horii, Some statistical characteristics of voice fundamental frequency. *J. Acoust. Soc. Am.* **52**, 146–146 (1972).
58. J. Local, G. Walker, Methodological imperatives for investigating the phonetic organization and phonological structures of spontaneous speech. *Phonetica* **62**, 120–130 (2005).
59. J. Local, "Conversational phonetics: Some aspects of news receipts in everyday talk" in *Prosody in Conversation*, E. Couper-Kuhlen, M. Selting, Eds. (Cambridge University Press, 1996), pp. 177–230.
60. J. Heritage, "A change-of-state token and aspects of its sequential placement" in *Structures of Social Action: Studies in Conversation Analysis*, J. M. Atkinson, J. Heritage, Eds. (Cambridge University Press, 1984), pp. 299–345.
61. M. Selting, "Prosody as an activity-type distinctive cue in conversation: The case of so-called 'astonished' questions in repair initiation" in *Prosody in Conversation*, E. Couper-Kuhlen, M. Selting, Eds. (Cambridge University Press, 1996), pp. 231–271.
62. M. Marmorstein, B. Szczepek Reed, Newsmarks as an Interactional Resource for Indexing Remarkability: A Qualitative Analysis of Arabic wal̪lāhi and English really. *Contrast. Pragmatics* **5**, 238–273 (2023), 10.1163/26660393-bja10091.
63. J. W. Du Bois, "The stance triangle" in *Stancetaking in Discourse*, R. Englebretson, Ed. (2007), pp. 139–182.
64. M. Selting, Affectivity in conversational storytelling. *Pragmatics* **20**, 229–277 (2010).
65. J. Local, G. Walker, Stance and affect in conversation: On the interplay of sequential and phonetic resources. *Text & Talk* **28**, 723–747 (2008).
66. S. A. Thompson, B. A. Fox, E. Couper-Kuhlen, *Grammar in Everyday Talk: Building Responsive Actions* (Cambridge University Press, 2015).
67. E. Levina, P. Bickel, "The Earth Mover's distance is the Mallows distance: Some insights from statistics" in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, G. Medioni, S. B. Kang, Eds. (IEEE Computer Society, McFarland, WI, 2001), pp. 251–256.
68. J. T. Hart, R. Collier, A. Cohen, *A Perceptual Study of Intonation* (Cambridge University Press, 1990).
69. D. Schiffrin, *Discourse Markers* (Cambridge University Press, 1987).
70. O. Jespesen, *The Philosophy of Grammar* (Allen & Unwin, 1924).
71. M. Selting, Lists as embedded structures and the prosody of list construction as an interactional resource. *J. Pragmat.* **39**, 483–526 (2007).
72. N. Matalon, "The Camel Humps prosodic pattern: Listing for Disaffiliating in Spoken Hebrew" in *Building Categories in Interaction - Linguistic Resources at Work*, C. Mauri, E. Goria, I. Fiorentini, Eds. (John Benjamins, 2021), pp. 155–186.
73. G. Jefferson, "List construction as a task and resource" in *Interaction Competence*, G. Psathas, Ed. (University Press of America, 1990), pp. 63–92.
74. W. Chafe, D. Tannen, The relation between written and spoken language. *Annu. Rev. Anthropol.* **16**, 383–407 (1987).
75. P. J. Hopper, "Emergent grammar and temporality in interactional linguistics" in *Constructions: Emerging and Emergent* (De Gruyter, 2011), pp. 22–44.
76. Yang Liu *et al.*, Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Trans. Audio Speech Lang. Process* **14**, 1526–1540 (2006).
77. E. Shriberg *et al.*, Can prosody aid the automatic classification of dialog acts in conversational speech? *Lang. Speech* **41**, 443–492 (1998).
78. L. Goffman, Prosodic influences on speech production in children with specific language impairment and speech deficits. *J. Speech Lang. Hearing Res.* **42**, 1499–1517 (1999).
79. L. Goffman, Kinematic Differentiation of Prosodic Categories in Normal and Disordered Language Development. *J. Speech, Language, and Hearing Res.* **47**, 1088–1102 (2004).
80. L. Goffman, L. Heisler, R. Chakraborty, Mapping of prosodic structure onto words and phrases in children's and adults' speech production. *Lang. Cogn. Process* **21**, 25–47 (2006).
81. S. Misiewicz, A. M. Brickman, G. Tosto, Prosodic impairment in dementia: Review of the literature. *Curr. Alzheimer Res.* **15**, 157–163 (2018).
82. H. Baglione, V. Coulombe, V. Martel-Sauvageau, L. Monetta, The impacts of aging on the comprehension of affective prosody: A systematic review. *Appl. Neuropsychol. Adult* 1–16 (2023), 10.1080/23279095.2023.2245940.
83. D. H. Zimmerman, "The interactional organization of calls for emergency" in *Talk at Work: Interaction in Institutional Settings*, P. Drew, J. Heritage, Eds. (Cambridge University Press, 1992), pp. 418–469.
84. M. Komter, "Conversation analysis in the courtroom" in *The Handbook of Conversation Analysis*, J. Sidnell, T. Stivers, Eds. (Wiley, 2012), pp. 612–629.
85. R. Gardner, "Conversation analysis in the classroom" in *The Handbook of Conversation Analysis*, L. Sidnell, T. Stivers, Eds. (Wiley, 2012), pp. 593–611.
86. J. Raymond, B. Szczepek Reed, *Units of Talk–Units of Action*, B. Szczepek Reed, G. Raymond, Eds. (John Benjamins Publishing Company, 2013).
87. D. Barth Weingarten, M. Selting, E. Reber, *Prosody in Interaction* (John Benjamins Publishing Company, 2010).
88. E. Couper-Kuhlen, C. E. Ford, Ed. *Sound Patterns in Interaction* (John Benjamins Publishing Company, 2004).
89. K. Follett, *Edge of Eternity* (Penguin Books, 2014).
90. K. Follett, *Edge of Eternity* (Downpour, 2014).
91. G. R. R. Martin, *A Dance with Dragons* (Bantam, 2012).
92. G. R. R. Martin, *A Dance with Dragons* (Random House Audio, 2011).
93. David Reads ASoIaF, A dance with dragons. (2021). https://www.youtube.com/playlist?list=PLMLTM7CoBZvJqZfq8fTdfVohvoPnh98ve. Accessed 1 September 2022.
94. M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, M. Sonderegger, "Montreal forced aligner: trainable text-speech alignment using kaldi" in *Interspeech 2017* (ISCA, Red Hook, NY, 2017), pp. 498–502.
95. Y. Mo, "Duration and intensity as perceptual cues for naïve listeners' prominence and boundary perception" in *Proceedings of the Fourth International Conference on Speech Prosody*, P. A. Barbosa, S. Madureira, C. Reis, Eds. (2008), pp. 739–742.
96. Y. Jadoul, B. Thompson, B. de Boer, Introducing parselmouth: A python interface to praat. *J. Phon.* **71**, 1–15 (2018).
97. P. Boersma, D. Weenink, Praat: Doing phonetics by computer. (2020). http://www.praat.org/.
98. R. W. Young, Terminology for logarithmic frequency units. *J. Acoust. Soc. Am.* **11**, 166–166 (1939).
99. G. Walker, Visual representations of acoustic data: A survey and suggestions. *Res. Lang. Soc. Interact.* **50**, 363–387 (2017).
100. C. De Looze, D. J. Hirst, "The OMe (Octave-Median) scale: A natural scale for speech melody" in *Proceedings of the Seventh International Conference on Speech Prosody*, N. Campbell, D. Gibbon, D. Hirst, Eds. (International Speech Communication Association, Dublin, 2014), pp. 20–23.
101. D. E. Rumelhart, J. L. McClelland, *Parallel Distributed Processing* (The MIT Press, 1986).
102. G. Klambauer, T. Unterthiner, A. Mayr, "Self-normalizing neural networks" in *31st Conference on Neural Information Processing Systems*, I. Guyon *et al.*, Eds. (Curran Associates, Inc., 2017), pp. 971–980.
103. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting. *The J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
104. F. Chollet, Keras [Computer software]. GitHub. https://github.com/keras-team/keras. Accessed March 2020.
105. D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization" in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)* (2015), pp. 14–16.
106. S. Kolouri, P. E. Pope, C. E. Martin, G. K. Rohde, "Sliced Wasserstein auto-encoders" in *International Conference on Learning Representations* (New Orleans, LA, 2019).
107. F. Pedregosa *et al.*, Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
108. J. W. Du Bois, "The stance triangle" in *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*, R. Englebretson Ed. (John Benjamins Publishing Company, 2007), pp. 139–182.
109. N. Duran, S. Battle, J. Smith, Inter-annotator agreement using the conversation analysis modelling schema, for dialogue. *Commun. Methods Meas.* **16**, 182–214 (2022).
110. J. L. Austin, *How to Do Things with Words* (Oxford University Press, 1962).
111. J. R. Searle, A classification of illocutionary acts. *Lang. Soc.* **5**, 1–23 (1976).
112. H. Bunt *et al.*, "Towards an ISO standard for dialogue act annotation" in *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, N. Calzolari *et al.*, Eds. (European Language Resources Association, 2010).
113. A. Stolcke *et al.*, Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist.* **26**, 339–373 (2000).
114. D. Jurafsky, S. Elizabeth, D. Biasca, "Switchboard SWBD-DAMSLshallow-discoursefunction annotation coders manual, draft 13" (Tech. Rep. 97-02, University of Colorado at Boulder, CO, 1997).
115. S. Jekat *et al.*, "Dialogue acts in VERBMOBIL" (VM-Report 65, Universität Hamburg, DFKI GmbH, Universität Erlangen, and TU Berlin, 1995).
116. H. Bunt, V. Petukhova, A. Malchanau, A. Fang, K. Wijnhoven, The DialogBank: Dialogues with interoperable annotations. *Lang. Resour. Eval.* **53**, 213–249 (2019).
117. A. Deppermann, M. Haugh, "Action ascription in social interaction" in *Action Ascription in Interaction*, A. Deppermann, M. Haugh, Eds. (Cambridge University Press, 2022), pp. 3–28.
118. A. Wichmann, "The attitudinal effects of prosody, and how they relate to emotion" in *Proceedings of the ISCA Workshop on Speech and Emotion* (Textflow, Belfast, 2000), pp. 143–148.
119. A. Ortony, T. J. Turner, What's basic about basic emotions? *Psychol. Rev.* **97**, 315–331 (1990).
120. P. Ekman, H. Oster, Facial expressions of emotion. *Annu. Rev. Psychol.* **30**, 527–554 (1979).
121. P. Ekman, W. V. Friesen, P. Ellsworth, *Emotion in the Human Face: Guidelines for Research and An Integration of Findings* (Elsevier, 2013).
122. P. van Rijn, P. Larrouy-Maestri, Modelling individual and cross-cultural variation in the mapping of emotions to speech prosody. *Nat. Hum. Behav.* **7**, 386–396 (2023).
123. M. B. Akçay, K. Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* **116**, 56–76 (2020).
124. R. Plutchik, "A general psychevolutionary theory of emotion" in *Theories of Emotion*, R. Plutchik, H. Kellerman, Eds. (Elsevier, 1980), pp. 3–33.

125. H. Rashkin, E. M. Smith, M. Li, Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, L. Màrquez, Eds. (Association for Computational Linguistics, 2019), pp. 5370–5381.
126. J. A. Brooks *et al.*, Deep learning reveals what vocal bursts express in different cultures. *Nat. Hum. Behav.* **7**, 240–250 (2022).
127. A. S. Cowen, P. Laukka, H. A. Elfenbein, R. Liu, D. Keltner, The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nat. Hum. Behav.* **3**, 369–382 (2019).
128. P. Brown, S. C. Levinson, *Politeness: Some Universals in Language Usage* (Cambridge University Press, 1987).
129. T. Niu, M. Bansal, Polite dialogue generation without parallel data. *Trans. Assoc. Comput. Linguist.* **6**, 373–389 (2018).
130. J. Heritage, Epistemics in action: Action formation and territories of knowledge. *Res. Lang. Soc. Interact.* **45**, 1–29 (2012).
131. N. Matalon *et al.*, Structure in conversation. GitHub. https://github.com/EyalWeinreb/Structure-in-conversation. Deposited 10 April 2025.