

Non-verbal information in spontaneous speech – towards a new framework of analysis

Tirza Biron^{1*}, Moshe Barboy¹, Eran Ben Artzy¹,
Alona Golubchik¹, Yanir Marmor¹, Smadar Szekely¹,
Yaron Winter¹, David Harel¹

^{1*}Faculty of Mathematics and Computer Science, Weizmann Institute of
Science.

*Corresponding author(s). E-mail(s): tirza.biron@weizmann.ac.il;

Non-verbal signals in speech are encoded by prosody and carry information that ranges from conversation action to attitude and emotion. Despite its importance, the principles that govern prosodic structure are not yet adequately understood.

This paper offers an analytical schema and a technological proof-of-concept for the categorization of prosodic signals and their association with meaning. The schema interprets surface-representations of multi-layered prosodic events.

As a first step towards implementation, we present a classification process that disentangles prosodic phenomena of three orders. It relies on fine-tuning a pre-trained speech recognition model, enabling the simultaneous multi-class/multi-label detection. It generalizes over a large variety of spontaneous data, performing on a par with, or superior to, human annotation.

In addition to a standardized formalization of prosody, disentangling prosodic patterns can direct a theory of communication and speech organization. A welcome by-product is an interpretation of prosody that will enhance speech- and language-related technologies.

Keywords

Context formalization, prosody, multi-layered information, computational linguistics, NLP

1 Introduction

1.1 A New Schema for Prosody Analysis

Non-verbal linguistic signals that are encoded by prosody and carry crucial information in speech. Prosodic messages range from conversation action (e.g., request, command) and discourse function (e.g., narration, parentheticals), to saliency of information (de/emphasis), attitude (e.g., sarcasm), and uninhibited emotion.

Written language registers some of the prosody’s many functions: punctuation denotes segmentation, certain speech-act types, and a few discourse functions. One also encounters the occasional orthographic *emphasis* or ‘misgivings’.

Despite its importance, the principles that govern prosodic structuring remain, by and large, unformulated; prosodic variability is a persistent source of debate (e.g., [1–4]). This appears to be due to a basic characteristic of prosodic signals – their simultaneity: speakers combine several messages at a time. Consider a potential breakdown for a surprised question “Really?” vs. its sarcastic counterpart “Really?...”. The latter exhibits at least two orders of non-verbal information: a rhetorical question and a mocking attitude. An analysis of prosodic structure must therefore account for its multidimensional nature. Recent developments in pattern recognition present a unique opportunity for use in such a context.

This article offers an analytical framework and a technological proof-of-concept for the categorization of prosodic signals and their association with meaning. At the core of our proposal is a schema that interprets the surface-representation of multi-layered prosodic events (cf. [5]). As a first step toward implementation, we present a prediction/classification process for the disentanglement of prosodic patterns that relies on a transformer-based architecture.

The primary objective of our experiment is to assess if and to what extent a model may simultaneously learn several prosodic messages of different non-verbal orders. The proposed method, then, enables the simultaneous training, followed by a one-pass multi-labeling. It generalizes well over a large variety of speakers, for several types of data, tagged by different annotators, and performs at 0.91/0.97 (Cohen’s Kappa/accuracy) for intonation unit (IU) detection, 0.55/0.81 for emphasis detection, and 0.45/0.70 for prosodic prototype detection (see section 3 below).

In addition to a standardized, careful explication of prosody, disentangling prosodic patterns can shed light on the organization of speech and expand theories of communication. It can enhance the pairing of prosodic form and function, help articulate the constraints that affect prosodic patterning (cf. [3]), and minimize the disparities in their acoustic description.

Furthermore, since prosody reflects much of the communicational context, a reliable analysis would be a gateway to an improved formalization of context. As a welcome by-product, speech technologies will be able to output exhaustive meaning, adding non-verbal conditioning to the recognised words. Speech analytics, natural language understanding, and speech synthesis are all expected to benefit from an accessible deciphering of prosody.

An additional contribution of our work involves simple means for adding prosodic labels to an aligned transcription. The transfer learning process presented here alters the model’s output labels to include new ones in the original, decoded series of tokens. The method may be applied to a variety of different domains. Lastly, we demonstrate the ability of re-training the STT WHISPER model [6] for prosodic disentanglement.

1.2 Linguistic Framework

Aiming at a broad approach to the analysis of prosody, two hypotheses underlie our proposal:

1. The germane unit and arena of prosodic events is the intonation unit (abbreviated IU; [7, 8]), also termed “Tone Group” [3], “intermediate intonational phrase” [9, 10], “prosodic intermediate phrases” [11], “turn construction unit” [12], or “minimal discourse unit” [13]; and cf. [14].
2. IUs exhibit semantically meaningful, sometimes grammaticalized, prosodic patterns [15, 16].

Our schema maintains that all prosodic phenomena may be analyzed as variations, either hierarchical or orthogonal, of a very small number of IU prototypes [7]. The variations include four basic layers: information structure, attitude, emotion, and 3-5 sub-categories of conversation action and/or discourse functions (see Figure 1).

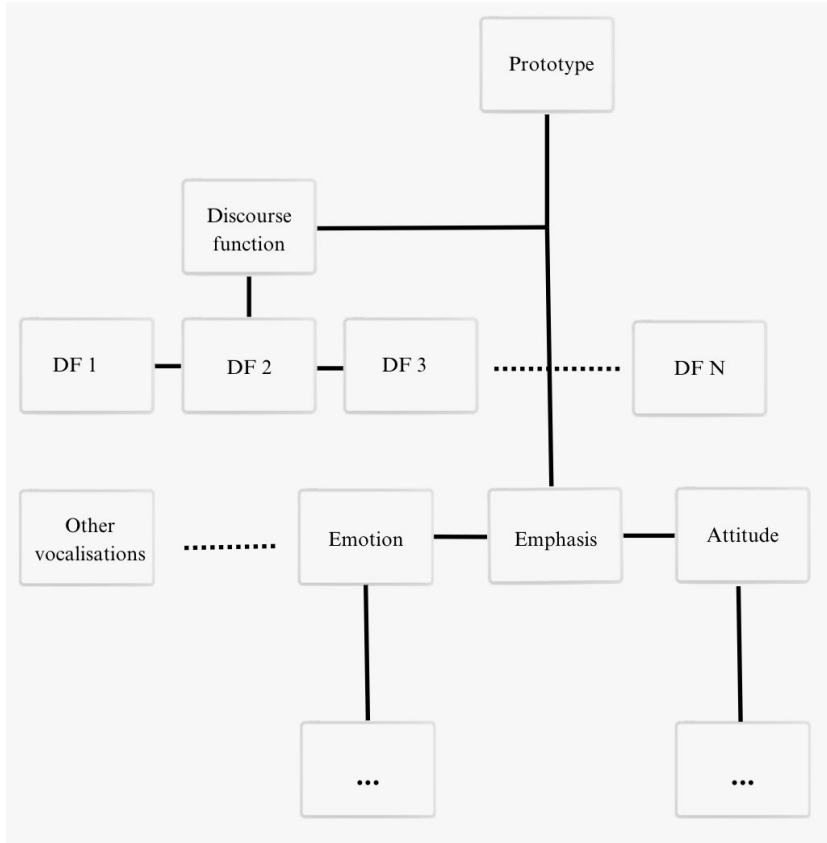


Fig. 1: An illustration of the analytical hierarchy for IUs. Note that emotion, emphasis and attitude are orthogonal to the prosodic prototype and discourse function hierarchy.

The common, unmarked prototypes (see [17] for markedness) are analyzed as modulated into stacked variants (cf. [18]). The resulting signal is realized as an integration of the above layers with additional constraints, such as syllable structure and unit length (cf. e.g., [19]).

To illustrate the principle of multi-layering, consider Figure 2c, which shows emphasis production for the prosodic prototype “comma”/“continuation”. Note the difference in pitch maxima at the beginning vs. the ending of the IU, echoing its latent pitch template (cf. [20]).

This view of prosodic template-variation is inspired by Semitic word-formation (see [22, 23] and the *supplementary material*). Non-concatenative morphology - that is, composites of morphemes of different orders - makes a useful metaphor for a layered, integrated patterning. When applied to prosody, this organizing principle enables a substantial reduction in complexity: from seemingly infinite variation to a hierarchical system. It thus facilitates the distinction between different non-verbal messages, readily accounting for the simultaneity of prosodic events.

1.3 Related Work

In the interest of smooth reading, and since the article touches upon a number of fields, a more detailed description of related work has been relegated to section 2 in the *supplementary material*. Here we provide a broad description only.

For overviews of the prevalent linguistic approaches to the study of prosody, see [1, 16, 24, 25]. As pointed out by [1] and [5], a predictive, general framework for associating prosodic form and function is yet to be put forward.

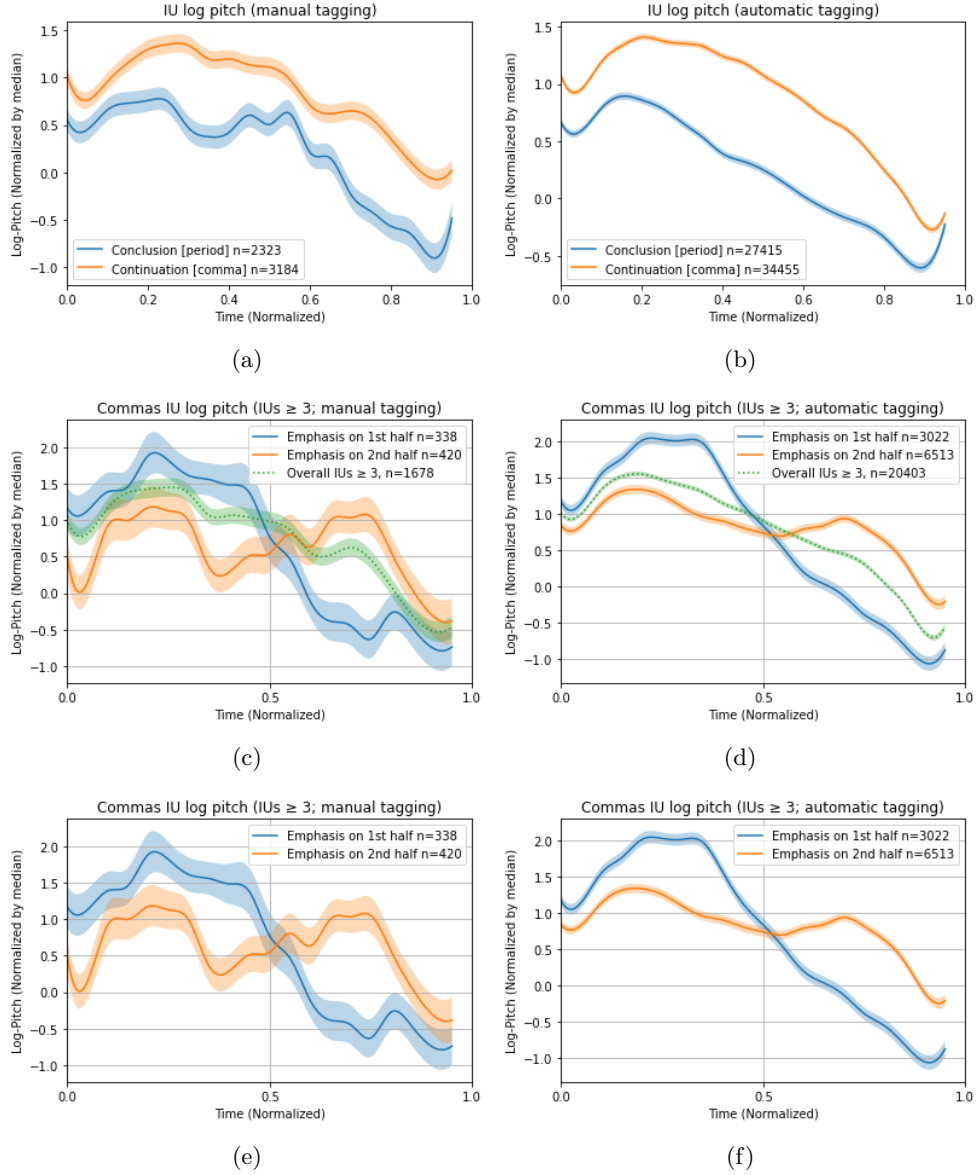


Fig. 2: Log pitch course, median-normalized and time-normalized, of manually (a, c, e) and automatically (b, d, f) annotated IUs, “This American Life” corpus [21]. **2(a-b)** Log pitch course of the prototypes “continuation” (“comma”; **2a** $n=3,184$; **2b** $n=34,455$) and “conclusion” (“period”; **2a** $n=2,323$; **2b** $n=27,415$) for manual and automatic annotation, respectively. **2(c-d)** Log pitch course of “continuation” IUs that bear emphasis in their first half (blue), second half (orange), and all “continuation” IUs (green), for manual and automatic annotation, respectively. **2(e-f)** Log pitch course of “continuation” for IUs that bear emphasis in their first half (blue) or their second half (orange), for manual and automatic annotation, respectively. Note the influence of the underlying “comma” pitch pattern on the production of emphasis, and the resemblance between manually annotated and automatically obtained IUs.

In the domain of computational prosody, improving speech synthesis has been a subject of significant research (e.g., [26]). As for automated analyses of prosody, most have been aimed at detecting single phenomena, such as unit boundaries (e.g., [27, 28]), prominence/saliency [28–30], or specific dialogue-acts (e.g., [31]). Crucially, none of the above tackle the multi-layered nature of many prosodic events.

A method for fine-tuning WHISPER to predict IU boundaries is described in [32]. Our proposal is similar, in that it enriches a transcription with prosodic tags. However, to the best of our knowledge, the multi-class/multi-label transfer learning that we employ has not been used for prosody analysis.

Machine learning methods have been used for semantic disentanglement in a large variety of domains, mainly in image processing (e.g., [33]). As far as we are aware, the disentanglement of non-verbal prosodic layers has not been the focus of such efforts.

2 Motivation: The Challenge of Context Formalization

Semiotic studies define context as that which accords meaning to a sign (cf. [34]). Verbal contextuality is traditionally viewed as the relationship – and indeed the contrast – between a sign and its fellow signs, with which it can be either joined or replaced [35]. Yet, despite its obvious contribution to contextual meaning, non-verbal information is rarely considered in descriptions of phonetics, phonology, and morpho-syntax. In response, Austin [36] stresses that “what we have to study is not the sentence but the issuing of an utterance in a speech situation” (p. 138). Consider example no. 1:

“There is a bull in the field.” (ibid., p.32) (1)

This statement is ordinarily either a description or a warning, the distinction relying on the speaker’s identity and motivation. To the discerning ear, prosody reflects, remarkably and accurately, such speech situations and their contextual meaning: the performance of a speech act, or imparting feelings, conveying epistemological information and other speaker intentions.

Language technologies have been wrestling with contextualization for several decades. Early mathematical representations of linguistic entities [37] instituted syntactic analysis as the base for natural language processing (NLP) (e.g., [38]), treating words as discrete, atomic units. With the introduction of robust word conversions, words and phrases were represented as continuous vectors (e.g., [39]), relying on an element’s immediate environment (often referred to as “context” in related domains as well (e.g., [40])). Continuous vectors that represent less immediate neighbors (n-grams) [41, 42] were later fed into convolutional neural networks (CNNs) and long-short term memory networks (LSTMs). LSTMs have been using the reciprocal “attention” of words in a text [43]; that is, an output that is affected by each element/word in the input series, by considering both their relative and absolute positions. LSTMs eventually culminated in the large, flexible models that stem from transformers [6]. Those excel at modeling contextual information through statistical learning, and have recently been augmented with visual and audio data, embedded in their input [44–46].

However, unformalised contextual information results in obvious weaknesses of linguistic accounts, whether heuristic or statistical. A more formal solution would require a systematic inclusion of non-verbal conditioning to verbal output. Apparently, a simple rule (paraphrasing [47]) would suffice: “A feature F is contextual for an action (or meaning) A if F constrains A, and may affect the outcome of A, but is not a constituent of A”. The prosodic output for example no. 1 would therefore be either “There is a bull in the field (warning, urgent)” or “There is a bull in the field (description, narrative, neutral)”, or, for that matter, any combination of speech act, intention and attitude/emotion with which the text was produced.

2.1 Contextuality and Scope

Contextuality typically relies on scope. Relationships between linguistic signs, either when joined to- or when replaced with one another, vary according to the size of the unit at hand. These range from a short retort (“Yes!”) to entire genres.

The effect of the wider context on meaning includes multi-unit prosodic patterns. Simple examples are appositions, list patterns [48], and prosodic bi-partites (such as if-then constructions; see samples [here](#)). Paragraphs, narratives, and formal addresses encompass a larger scale, which often presents nested structures (e.g., a list of events within a narrative). Those form the syntax of prosodic patterns (cf. [49]).

Analyses of larger-than-sentence entities are central to the domains of text linguistics (e.g., [50–52]), discourse analysis (DA), and conversation analysis (CA) (e.g., [53–55]), all of which provide valuable tools for our work.

3 The schema

The schema we propose here posits that communicative intentions, encoded by prosody, may be ordered and tracked hierarchically. Surface-representations of patterns within IUs may be interpreted through a layered classification procedure. Thus, the overwhelming diversity of prosody, often referred to as its ‘elusive’ nature (e.g., [56, 57]), may be broken down beneficially.

Similarly to [1], our proposal concurs with the idea of stacking, and follows the functional-contrastive approach, whereby a sign-function draws its systemic value from the contrast with other sign-functions. Conversely, our schema stipulates a different discrete unit and hence a different scope of patterning, as well as a different view on stacking (cf. [58]).

We offer a framework that will eventually become predictive, in that it will describe the underlying structures and constraints that form a consequent pattern. We propose between 3 and 8 (but no more) classes of variation on 3 to 5 basic labels (see Figure 3).

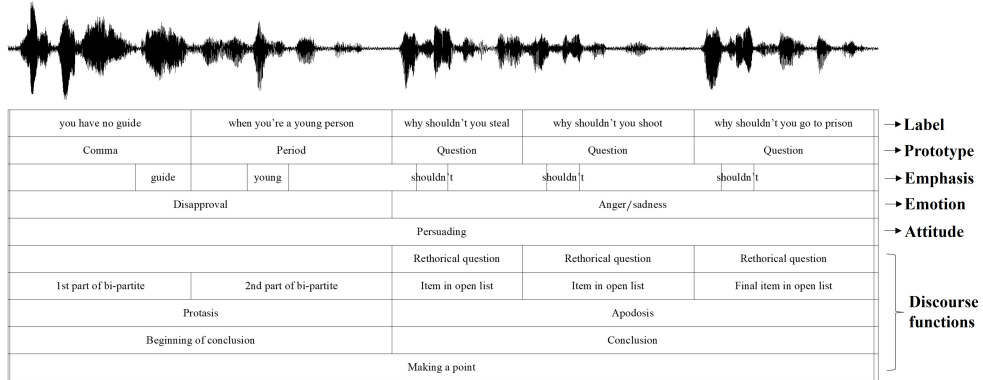


Fig. 3: An example of the categories and labels that constitute the prosodic messages in [audio](#). Excerpt drawn from [59].

Once identified, IUs are classified into the following categories:

1. *Para-syntactic modality, termed here prosodic prototype* – The prototypes that we identify are: (,) “continuation”; (.) “conclusion”; and (?) “request for response” (cf. [7]).
2. *Discourse function and/or conversation action* – These patterns signal the organization of discourse. It is a category that covers a wide scope, from syntax to

rhetoric, and its labels include, for example, “circumstantial unit”, “title of discourse”, “background of narrative”, “narrative event” and so on (see Table *SM3* in the *supplementary material*). The sub-category of conversation action refers to speech acts that are designed to affect the interlocutor’s behavior; for example, questions that serve as requests, warnings, commands, etc.

3. *Information structure* – The prosodic signaling of saliency of information (de/emphasis).
4. *Express sentiment/attitude* – Irony, feigned anger and calculated indifference are examples of overt attitudes.
5. *Unintentional/unplanned emotion* – This category includes, for example, delight, disgust, reserve, fear, pain and other emotions and feelings that can change one’s prosody (see overview in [60]).

Our theory of patterning posits a prosodic prototype (category (1)) that is interpreted within a set of pre-established alterations (categories (2)-(5)). In other words, the global signal can lead the listener to infer the speaker’s intentions based on their prior knowledge of the prototypical template and its available variations. The text in example no. 2, below, should be read as a disapproving rhetorical question with an emphasis on the last word:

“You want to go *home*?!” (2)

When the underlying patterns of an IU are identified, other prosodic messages may be disentangled. Thus, the question pattern in example no. 2, can be extricated for differential flagging and distinguished from the pattern of disapproval and the emphasis on “home”.

The resulting classification outlines an inventory of variations that are projected, or ‘grafted’, onto a prototype-pattern (see Tables *SM1*, *SM2* and *SM3* in the *supplementary material*). As stated above, in a technological context, the disentanglement enables an enhanced detection of prosodic semantics. For a detailed presentation of the schema, see the *supplementary material*.

Some prosodic layers are more subtly marked, while others are more clear cut (e.g., discourse function vs. information structure). Still, when analyzing speech, our description strives to be as detailed as possible, in as much as the details may be perceived. An advanced prototype-classification tree would define what constitutes a distinctive feature for prosodic patterning on the scale of IUs.

In the following sections, we report upon the methods for, and results of, a successful disentanglement procedure of three prosodic categories, as detected simultaneously through fine-tuning the WHISPER speech language model.

4 Methods

4.1 Datasets and data preparation

The problem of multi-layered prosodic classification has received little attention in the ML community. Moreover, existing datasets and benchmarks do not match our analytical framework. Therefore, a substantial part of our work is dedicated to creating designated datasets.

Our principal set is drawn from the “This American Life” podcast (abbreviated TAL, [21]). As an auxiliary set, we compiled a collection of 24 interviews, each recording less than 30 seconds long, totalling 7 minutes of tagged speech. Among the speakers are Oprah, Will Smith, Frances Arnold and Connan O’Brian ([interviews dataset](#)). This set differs from TAL, in that it contains spontaneous speech only, with no narrated parts. Created for validation purposes, it was annotated by a different expert and was not represented in the training set. Both sets have partially timestamped transcriptions.

4.1.1 Manual annotation

A primary automatic segmentation was carried out using the TAL transcript: word sequences between punctuation marks were regarded as IU-proxies, and a preliminary classification into prosodic prototypes was done using that same punctuation: (,) (“continuation”), (.) (“conclusion”), or (?) (“request for response”). Of those, 80% of ≤ 7 -word units were found to correspond to IUs, and were therefore included as a suggestion for manual tagging – their labels to be confirmed or corrected.

The annotation was added manually, per word, using INCEpTION [61]. For the experiments presented here, word labels included IU boundary information, IU prototype, and a class of saliency (primary or secondary emphasis, and de-emphasis). The result of the annotation process is a table of time-aligned, tagged words. See Table 1 for the statistics of the annotation.

(a) Main speaker vs. interviewees	
Speaker	Number (Fraction)
Narrator	1,385 (23.33%)
Interviewee	4,551 (76.67%)
Total	5,936
(b) Prosodic prototypes	
Prototype	Number (Fraction)
Continuation (comma)	3,246 (54.99%)
Conclusion (period)	2,362 (39.79%)
Request for response (question mark)	310 (5.22%)
Total	5,936
(c) Emphasis tags	
Emphasis	Number (Fraction)
Primary	5,320 (26.34%)
Secondary	2,726 (12.99%)
Non-emphasized words	12,946 (61.67%)
Total	20,992

Table 1: *The annotated data.* 1a. Number and fraction of main speaker data vs. interviewees (n=82); 1b. Number and fraction of prosodic prototypes; 1c. Number and fraction of emphasis tags (= the number of words annotated).

4.1.2 Preprocessing for labeling and training

TAL transcripts were normalized as follows:

The text was converted into lower case; abbreviations (e.g., Dr., Ms.) and transcribed digits were replaced by their long forms using [62]; for the purposes of our analysis, transcribed (–) was replaced by (.), and (!) by (.).

To remove background music, the audio was processed using SPLEETER [63]. The transcription of TAL and Interviews were force-aligned using the Montreal Forced Aligner [64], in order to produce timestamps for each word and phone.

4.1.3 Turn compilation

In conversation analysis (CA), continued speech by a single speaker is termed a “turn” [65]. Turns are typically constructed of at least one IU, and may extend to entire communications. In our experiments, however, “turn” is the audio unit that is input to the model for analysis.

Our considerations for obtaining optimal turns in this context included:

1. Turns should contain at least two IUs by the same speaker, so that the model may learn IU switches;

2. Turns should not contain long pauses, both for efficiency of computation and in order to avoid IU switches that are too obvious;
3. Multiple speakers in a turn may be beneficial, as they better reflect real-life speech situations;
4. Turns should not exceed 30 seconds or 448 tokens, as per the WHISPER constraints.

Preliminary tests for optimizing turn generation considered three parameters: avoid/use multiple speakers; determine maximal speech pause; and determine the minimal number of IUs. These considerations led to a dataset in which most “natural” turns measure less than 10 sec.; 88% feature one speaker, 11.5% feature two, and 0.5% three speakers.

In order to determine the best turn-compilation strategy, we chose the WHISPER-Small model, one of six sub-models published along with the WHISPER paper [6]. This choice stemmed from its performance, which is close to the best obtained result (see section 1 in the *supplementary material*; For further details, see Figure 5 and Table 2).

4.2 Experiment objectives and setup

As mentioned above, the primary objective of our experiment was to assess if and to what extent a model may simultaneously learn several prosodic messages of different non-verbal orders. Another objective was to predict these labels simultaneously. To this end, we applied transfer learning and fine-tuning to the WHISPER model, the backbone of our experiments (Figure 4).

4.2.1 Training

We used the HuggingFace WHISPER implementation to fine-tune the various models. The default optimization procedure is described in [66], and the learning rate was fixed at 10^{-5} . We applied an early stop mechanism, using 5% of the training set for evaluation, which induced 5-15 epochs of training. For efficiency, turns were sorted by length (i.e., the number of words), and the generated batches included 256 tokens, inducing mini-batch sizes of 1-20.

Each turn of speech was treated as a single instance, the training input consisting of its audio and transcription. The transcription was enriched with prosodic labels per word. Those were inserted alternately, as single strings, with text-words preceded by their prosodic label-combination (see Figure 4). As far as we are aware, this method of multi-class/multi-label transfer learning has not yet been used for prosody analysis (cf. [32]).

In addition to training for the triple, simultaneous recognition of prosodic events, we trained for three distinct recognition tasks of the same events. This required replacing the complex labels (that represent a combination of phenomena) by simplex labels (that denote just one).

4.2.2 Prediction

The WHISPER base building block is a transformer, whose input is a spectrogram and a sequence of tokens that represent the audio and the text, respectively. The transformer then generates predictions for the next token to be concatenated to the sequence. Our challenge was to predict correct prosodic labels only, excluding textual ones. The prediction proceeded as described in Algorithm 1.

On each iteration, word tokens are concatenated together with the prosodic token predictions that have accumulated so far. The prosodic token with the highest probability is picked, then inserted between the accumulated word-token predictions. Since prosodic label-combinations are defined per word, the output string alternates the generated prosodic labels and words in the odd and even positions.

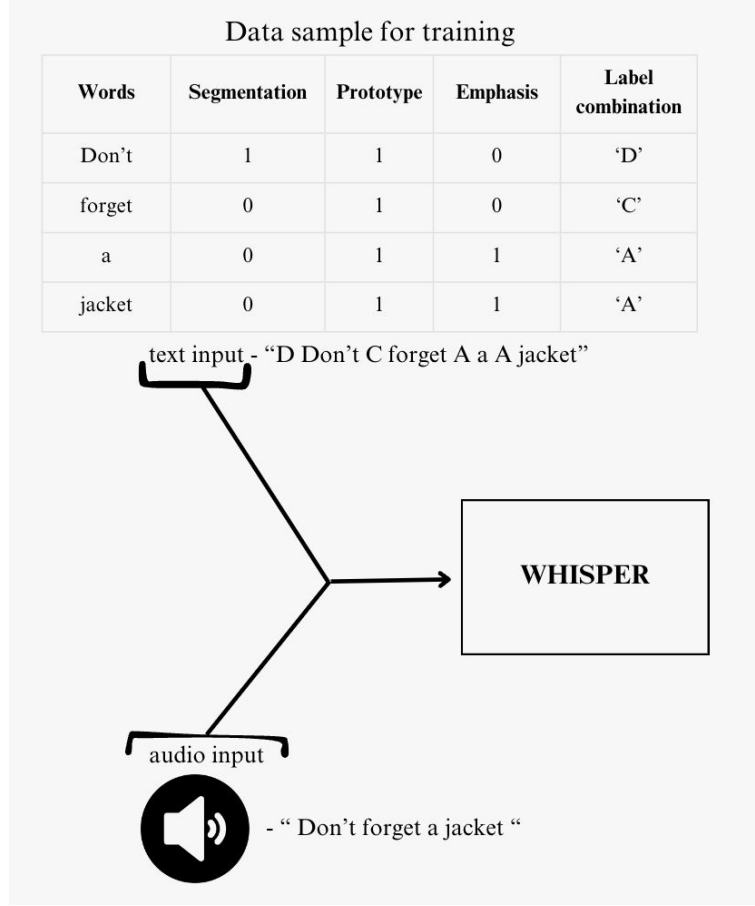


Fig. 4: *Training scheme.* The backbone of our method is the fine-tuned WHISPER [6]. Its input includes speech audio, its corresponding text and prosodic labels; output predicts label combinations for each word of the input text.

Note the differences vis-a-vis the regular WHISPER prediction scheme: when trained on a language task, the WHISPER inference is not required to distinguish transcription-related tokens from non-transcription ones. Conversely, our method requires that only prosodic labels be drawn at the inference stage (for a manually tagged text vs. the output of the trained model see Figure *SM2* in the *supplementary material*).

4.2.3 Validation/Evaluation

Metrics

To evaluate the capabilities of the model, we used Cohen’s Kappa (CK) metric of inter-annotator agreement (see [8] for IU boundaries and [67] citing scores for two of our three labels).

Two CK metrics were used for IU boundary recognition/segmentation: the first considered the prediction for the first uttered word in a turn, and the second did not. Since the beginning of a turn is a predetermined IU boundary, the classification for the first word carries no predictive power.

Prototype performance was calculated per IU, and only for the well-identified IUs ($\sim 94\%$ of the units). The evaluation was based on the predicted prototype label for the first and last words of an IU, assuming that this is a match which best represents the prosodic information required for the task.

Algorithm 1 *Pseudo code of the inference procedure.* This method enables prosodic labeling only. Note that *next_label* holds the predicted label of a multiclass-multilabel combination

Require: Model: the re-trained model (based on WHISPER), which consists of an audio encoder and a text decoder.

Require: Tokenizer: converts text into the model’s known tokens.

Require: Audio_spectrogram: audio in the format suited for the model’s input (of the currently handled turn).

Require: Word_list: the words in the transcription, sorted by order of utterance.

Ensure: Label_list: the tags corresponding to the word list and aligned with them.

```
1: label_list  $\leftarrow$  empty list
2: token_list  $\leftarrow$  model’s starting tokens
3: audio_features  $\leftarrow$  model.audio_encoder(audio_spectrogram)
4: for word in word_list do
5:   label_logits  $\leftarrow$  model.text_decoder(token_list, audio_features)
6:   next_label  $\leftarrow$  label with highest probability in label_logits
7:   append next_label to label_list
8:   append next_label to token_list
9:   append tokenizer.encode(word) to token_list
10: end for
11: return label_list
```

Experiment Setup

First, we explored the effect of several pre-trained WHISPER architectures. Whereas fine-tuning the largest model yielded the best results, it required roughly three hours of training on a single GPU. To balance training speed and performance, we tested smaller models, including the “Tiny”, “Small”, “Base” and “Medium” variants. By eliminating gradient accumulation and using a larger batch size, training time for the smaller model was reduced to a half an hour on a smaller GPU.

As mentioned above (section 4.2.1), we trained for single recognition tasks in order to compare the performance on a single task vs. the triple one. In addition, we tested three different representation methods of prosodic labels: (1) ‘raw’, which refers to special ‘words’ that were generated for this process; (2) ‘compact’, which refers to twelve labels that stand for the twelve combinations of prosodic tags; and (3) ‘bits’, which is similar to ‘raw’, and represents each prosodic feature by a single token (see *supplementary material*).

5 Results

Fine-tuning the WHISPER models for simultaneous detection of prosodic phenomena proved very successful. This is especially true for predicting IU boundaries, whereas simultaneous detection of prosodic prototypes and emphases were more demanding tasks. Notably, the outcome indicates that the fine-tuned model is on par with human annotators (when they tag individual tasks). In the task of prototype recognition, the rare prototype “?” (“request for response”) was best recognised when employing the WHISPER-Large V2 (see Table 2).

Table 3 shows that the model generalizes well across datasets and genres. It was more successful when employed on the TAL data than on the Interviews data (which were excluded from the training material), and specifically so in regard to IU boundary recognition.

Table 4 reports slight differences in performance for TAL interviewer (Ira Glass) vs. his interviewees (n=49 in the test set; n=81 in the train set). The difference may

(a)					
Metric	Segmentation	Emphasis	Question	Period	Comma
Cohen’s Kappa	0.932	0.588	0.664	0.453	0.442
Recall	0.958	0.713	0.594	0.644	0.789
Precision	0.941	0.7	0.784	0.724	0.708
F1-score	0.949	0.7	0.676	0.682	0.746
Accuracy	0.974	0.831	0.978	0.733	0.722

(b)					
Metric	Segmentation	Emphasis	Question	Period	Comma
Cohen’s Kappa	0.914	0.552	0.497	0.443	0.419
Recall	0.938	0.738	0.391	0.626	0.797
Precision	0.936	0.639	0.735	0.726	0.672
F1-score	0.937	0.685	0.510	0.672	0.741
Accuracy	0.968	0.808	0.971	0.729	0.711

Table 2: Comparison of various metrics on TAL dataset. Results for main split, re-trained WHISPER-Large V2 (2a) and WHISPER-Small (2b), using the “Compact” labels.

Dataset	Model	Segmentation	Segmentation (wos)	Emphasis
TAL	Small	0.914	0.895	0.552
Interviews	Small	0.680	0.593	0.456
Interviews	Large-V2	0.711	0.629	0.519

Table 3: Cohen’s Kappa scores on TAL and Interviews datasets. These tests employed the Large version of the model on the main split of TAL dataset, using the “Compact” labels.

Test Set	#Turns	#Speakers	Segmentation	Segmentation (wos)	Emphasis	Prototype
All	192	50	0.914	0.895	0.552	0.447
Ira Glass	47	1	0.915	0.895	0.574	0.419
Others	145	49	0.914	0.895	0.547	0.451

Table 4: Ira Glass - the show host - vs. other speakers (n=49 in test set), results for WHISPER-Small.

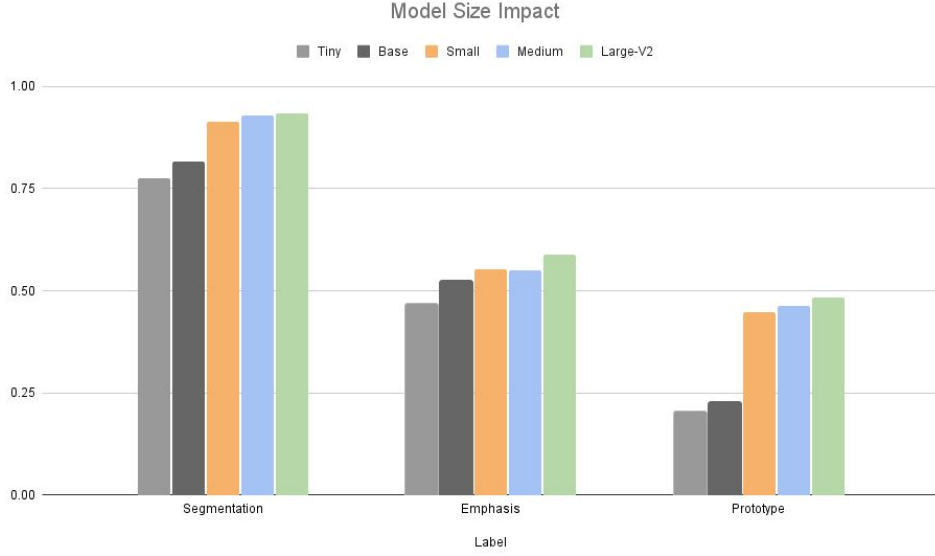
be attributed to genre: the interviewer’s speech may be scripted/ narrated, whereas the interviewees are spontaneous speakers.

As for model size, unsurprisingly and generally speaking, the larger the model, the better the performance (Figure 5). The Large V2 model performed significantly better on Prototype detection. However, over the majority of tasks, the improvement was not dramatic.

The triple detection task begs the question of how well the fine-tuned model would fare when trained to detect a single prosodic phenomenon. Table 5 shows that the results are not all that different: the performance is stable and somewhat weaker for single tasks.

Another finding is the robustness of the models, regardless of the differences in turn generation method (Table 3, section 4). Note the difference in number of turns vis-a-vis the stability of performance.

The results in Tables 2 to 4 and Figure 5b indicate that the re-trained WHISPER models separate well three different prosodic simultaneous messages. They generalize over a large variety of speakers, for several types of data, spontaneous and scripted, and for different expert annotators.



(a)

Model	Segmentation	Segmentation (wos)	Emphasis	Prototype
Tiny	0.776	0.718	0.469	0.205
Base	0.815	0.771	0.524	0.228
Small	0.914	0.895	0.552	0.447
Medium	0.929	0.914	0.551	0.462
Large-V2	0.933	0.918	0.588	0.484

(b)

Fig. 5: Impact of model size on performance of re-trained WHISPER for three simultaneous tasks. (5b) Tests on TAL dataset, with/out considering the first word of a turn.

(a)				
Full Train Set				
Model	IU Detect	IU (wos)	Emphasis	Prototype
Small	0.941	0.928	0.561	0.471
Large-V2	0.931	0.916	0.563	0.506
<i>Best Multi-Label</i>	<i>0.946</i>	<i>0.934</i>	<i>0.588</i>	<i>0.503</i>

(b)				
8% Train Set				
Model	IU Detect	IU (wos)	Emphasis	Prototype
Small	0.850	0.817	0.428	0.183
Large-V2	0.887	0.864	0.489	0.325
<i>Best Multi-Label</i>	<i>0.915</i>	<i>0.896</i>	<i>0.504</i>	<i>0.274</i>

Table 5: Performance of re-trained WHISPER models for three single tasks vs. the triple task (in italics). Tests on TAL dataset, with/out considering the first word of a turn. Models were trained either on the entire set (5a) or on 8% (5b) of it, to rule out data loss due to label encoding.

6 Discussion

6.1 Summary

We have shown that simultaneous prosodic messages of different non-verbal orders may be disentangled and detected, independently and simultaneously. This is an encouraging validation of the layered approach to prosodic patterning that this article proposes. The fact that the triple detection task outperforms the single detection ones further corroborates our decision to use the IU as the central arena of prosodic events, and are key to their successful recognition.

In addition, we presented a new method for multi-label, multi-class transfer learning, which enriches the sequence of ASR training with prosodic labels. This ‘dynamic tokenizer’ – that is, a fine-tuning that uses existing WHISPER tokens for a new task – seems to draw out information that already exists within the weights of the original model. The performance of this method is just as encouraging. Despite the difficulty in training for various detection tasks at a time, on diverse data, labeled by different annotators, it is either on par with, or superior to, that of average human annotation. As discussed in [8] and [67], the agreement for annotating prosodic boundary and emphasis (separately) is estimated at 0.52-0.78 Cohen’s Kappa. Therefore, our model can be considered an expert annotator for the prosodic phenomena learned.

6.2 Future work

Future challenges abound, and encompass many of the domains that this multidisciplinary work touches upon. They may be divided into four principal veins:

1. Extending the repertoire of reliably recognised prosodic patterns of all non-verbal orders, including emotions and speaker attitudes. This includes exploring prosodic universals vs. language- or community-specific phenomena, as well as other socio-linguistic factors and fine-grained analyses.
2. Applying our transfer learning method to additional fields: computer vision, NLP, etc. The method of intertwining new labels with known tokens enables the labeling of “extra” information, which exists in the model’s weights side-by-side with already-formalized data. It can also enhance the measuring of the “new token” recognition.
3. Exploring the model and its internal representations, in order to determine, and better make use of, the distinctive features of its prosodic classification (cf. [68]).
4. Studying the relationships of prosodic patterns with other linguistic components, and developing a new tool for context formalization.

This article offers a first attempt at the disentanglement of prosodic messages, based on IU analysis. Through the systematic recognition of non-verbal messages, it can expand the horizons of speech and language descriptions, and support the long-standing effort on context elucidation. As our framework differentiates between non-verbal signals, it can also set apart emotional from non-emotional patterns. Thus, it might just produce the holy grail of speech analytics – reliable emotion and sentiment recognition for spontaneous speech.

Acknowledgements

We wish to thank Amir Becker, David Biron, Sharon Fireman and Assaf Marron for valuable suggestions throughout our research. Parts of the work were funded by a research grant to DH from the Israel Science Foundation, and by BINA – the Translational Research and Innovation unit of the Weizmann Institute of Science. Some of the methods and techniques presented in the article have been submitted for US Provisional patent protection (Application No. 63/614,588, filed on 12.24.23).

References

- [1] Xu, Y.: Speech prosody: A methodological review. *Journal of Speech Sciences* **1**(1), 85–115 (2011)
- [2] Thein, M.L.: *Die Informationelle Struktur Im Englischen: Syntax und Information Als Mittel der Hervorhebung* vol. 323. Walter de Gruyter GmbH & Co KG, ??? (2017)
- [3] Xu, Y., Lee, A., Prom-On, S., Liu, F.: Explaining the penta model: a reply to arvaniti and ladd. *Phonology* **32**(3), 505–535 (2015)
- [4] Cole, J.: Prosody in context: A review. *Language, Cognition and Neuroscience* **30**(1-2), 1–31 (2015)
- [5] Ladd, D.R.: *Simultaneous Structure in Phonology* vol. 28. OUP Oxford, ??? (2014)
- [6] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: *International Conference on Machine Learning*, pp. 28492–28518 (2023). PMLR
- [7] Du Bois, J.W., Schuetze-Coburn, S., Cumming, S., Paolino, D.: Outline of discourse transcription. In: *Talking Data*, pp. 45–89. Psychology Press, ??? (2014)
- [8] Himmelmann, N.P., Sandler, M., Strunk, J., Unterladstetter, V.: On the universality of intonational phrases: A cross-linguistic interrater study. *Phonology* **35**(2), 207–245 (2018)
- [9] Halliday, M.A.K.: *Intonation and Grammar in British English* vol. 48. Walter de Gruyter GmbH & Co KG, ??? (2015)
- [10] Beckman, M.E., Pierrehumbert, J.B.: Intonational structure in japanese and english. *Phonology* **3**, 255–309 (1986)
- [11] Silverman, K.E., Beckman, M.E., Pitrelli, J.F., Ostendorf, M., Wightman, C.W., Price, P., Pierrehumbert, J.B., Hirschberg, J.: Tobit: A standard for labeling english prosody. In: *ICSLP*, vol. 2, pp. 867–870 (1992)
- [12] Reed, B.S.: Units of interaction: “intonation phrases” or “turn constructional phrases”. *Actes/Proceedings from IDP (Interface Discours & Prosodie)*, 351–363 (2009)
- [13] Degand, L., Simon, A.C.: Minimal discourse units: Can we define them, and why should we. *Proceedings of SEM-05. Connectors, discourse framing and discourse structure: from corpus-based and experimental analyses to discourse theories, Biarritz*, 14–15 (2005)
- [14] Su, C.-y., Tseng, C.-y.: Perceivable information structure in discourse prosody-detecting prominent prosodic words in spoken discourse using f0 contour. In: *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 424–428 (2018). IEEE
- [15] Hannay, M., Kroon, C.: Acts and the relationship between discourse and grammar. *Functions of language* **12**(1), 87–124 (2005)
- [16] Couper-Kuhlen, E., Selting, M.: *Interactional Linguistics: Studying Language in*

Social Interaction. Cambridge University Press, ??? (2017)

- [17] Jakobson, R.: Russian and Slavic Grammar: Studies, 1931-1981 vol. 106. Walter de Gruyter, ??? (1984)
- [18] Hockett, C.F.: The origin of speech. *Scientific American* **203**(3), 88–97 (1960)
- [19] Jacobs, C.L., Yiu, L.K., Watson, D.G., Dell, G.S.: Why are repeated words produced with reduced durations? evidence from inner speech and homophone production. *Journal of Memory and Language* **84**, 37–48 (2015)
- [20] Hirose, K., Fujisaki, H., Yamaguchi, M.: Synthesis by rule of voice fundamental frequency contours of spoken japanese from linguistic information. In: ICASSP’84. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 9, pp. 597–600 (1984). IEEE
- [21] Glass, I.: This American Life. Chicago Public Media. [Online]. Available: <https://www.thisamericanlife.org/archive> (1995-present)
- [22] Owens, J.: The arabic grammatical tradition. *The Semitic Languages* **46** (2013)
- [23] Schippers, A.: The hebrew grammatical tradition. *The Semitic Languages*, 59–65 (1997)
- [24] Wagner, M., Watson, D.G.: Experimental and theoretical advances in prosody: A review. *Language and cognitive processes* **25**(7-9), 905–945 (2010)
- [25] Wennerstrom, A.: *The Music of Everyday Speech: Prosody and Discourse Analysis*. Oxford University Press, ??? (2001)
- [26] Triantafyllopoulos, A., Schuller, B.W., İymen, G., Sezgin, M., He, X., Yang, Z., Tzirakis, P., Liu, S., Mertes, S., André, E., et al.: An overview of affective speech synthesis and conversion in the deep learning era. *Proceedings of the IEEE* (2023)
- [27] Biron, T., Baum, D., Freche, D., Matalon, N., Ehrmann, N., Weinreb, E., Biron, D., Moses, E.: Automatic detection of prosodic boundaries in spontaneous speech. *PloS one* **16**(5), 0250969 (2021)
- [28] Rosenberg, A.: Classification of prosodic events using quantized contour modeling. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 721–724 (2010)
- [29] Barbosa, P.A.: Prominence-and boundary-related acoustic correlations in brazilian portuguese read and spontaneous speech. In: *Proceedings of the Speech Prosody 2008 Conference*, pp. 257–260 (2008). Citeseer
- [30] Calhoun, S., Yan, M., Salanoa, H., Taupi, F., Kruse Va’ai, E.: Focus effects on immediate and delayed recognition of referents in samoan. *Language and Speech* **66**(1), 175–201 (2023)
- [31] Sridhar, V.K.R., Bangalore, S., Narayanan, S.S.: Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. *IEEE transactions on audio, speech, and language processing* **16**(4), 797–811 (2008)
- [32] Roll, N., Graham, C., Todd, S.: Psst! prosodic speech segmentation with

- transformers. arXiv preprint arXiv:2302.01984 (2023)
- [33] Wu, Q., Liu, Y., Zhao, H., Kale, A., Bui, T., Yu, T., Lin, Z., Zhang, Y., Chang, S.: Uncovering the disentanglement capability in text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1900–1910 (2023)
 - [34] De Saussure, F.: Cours de Linguistique Générale vol. 1. Otto Harrassowitz Verlag, ??? (1989)
 - [35] Hjelmslev, L., Whitfield, F.J.: Prolegomena to a theory of language (1953)
 - [36] Austin, J.L.: How to do Things with Words vol. 88. Oxford university press, ??? (1975)
 - [37] Chomsky, N.: Syntactic Structures. Mouton de Gruyter, ??? (2002)
 - [38] Carroll, J., Briscoe, T., Sanfilippo, A.: Parser evaluation: a survey and a new proposal. In: LREC, vol. 998, pp. 447–454 (1998)
 - [39] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
 - [40] Behre, P., Tan, S., Varadharajan, P., Chang, S.: Streaming punctuation: A novel punctuation technique leveraging bidirectional context for continuous speech recognition. arXiv preprint arXiv:2301.03819 (2023)
 - [41] Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014). Association for Computational Linguistics
 - [42] Yin, W., Kann, K., Yu, M., Schütze, H.: Comparative study of cnn and rnn for natural language processing. arXiv preprint arXiv:1702.01923 (2017)
 - [43] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
 - [44] Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
 - [45] Peng, C., Chen, K., Shou, L., Chen, G.: Carat: Contrastive feature reconstruction and aggregation for multi-modal multi-label emotion recognition. arXiv preprint arXiv:2312.10201 (2023)
 - [46] Li, Y., Du, H., Ni, Y., Zhao, P., Guo, Q., Yuan, F., Zhou, X.: Multi-modality is all you need for transferable recommender systems. arXiv preprint arXiv:2312.09602 (2023)
 - [47] Devlin, K.: Confronting context effects in intelligence analysis: How can mathematics help. Center for the Study of Language and Information, Stanford University (2005)
 - [48] Matalon, N.: The camel humps prosodic pattern. Building categories in interaction: Linguistic resources at work **220**, 155 (2021)

- [49] Matalon, N., Weinreb, E., Freche, D., Volk, E., Biron, T., Moses, E., Biron, D.: Structure in Conversational Prosody: Evidence for Vocabulary, Semantics and Syntax of Intonation Units. Under revision (Under revision)
- [50] Weinrich, H.: *Tempus: The World of Discussion and the World of Narration*. Fordham Univ Press, ??? (2024)
- [51] Shisha-Halevy, A.: Epistolary grammar: Syntactical highlights in kate roberts’s correspondence with saunders lewis. *Journal of Celtic Linguistics* **9**(1), 83–103 (2005)
- [52] Shisha-Halevy, A.: Converbs in welsh and irish. In: 13th International Congress of Celtic Studies, Bonn (2007). Conference Paper
- [53] Couper-Kuhlen, E.: Intonation and discourse. *The handbook of discourse analysis*, 82–104 (2015)
- [54] Couper-Kuhlen, E.: *An Introduction to English Prosody*. TUEBINGEN, ??? (1986)
- [55] Selting, M., Barth-Weingarten, D., Reber, E., Selting, M.: Prosody in interaction. *Prosody in Interaction*, Amsterdam/Philadelphia, John Benjamins, 3–40 (2010)
- [56] Dogil, G.: Understanding prosody. In: Rickheit, G., Herrmann, T., Deutsch, W. (eds.) *Psycholinguistics: Ein Internationales Handbuch*, pp. 544–565. De Gruyter Mouton, Berlin • New York (2003). <https://doi.org/10.1515/9783110114249.4.544>
- [57] Cornille, T., Wang, F., Bekker, J.: Interactive multi-level prosody control for expressive speech synthesis. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8312–8316 (2022). IEEE
- [58] Cenceschi, S., Sbattella, L., Tedesco, R.: Calliope: A multi-dimensional model for the prosodic characterization of information units. *Estudios de fonética experimental*, 227–245 (2021)
- [59] Du Bois, J.W., Chafe, W.L., Meyer, C., Thompson, S.A., Martey, N.: *Santa barbara corpus of spoken american english*. CD-ROM. Philadelphia: Linguistic Data Consortium (2000)
- [60] Hashem, A., Arif, M., Alghamdi, M.: Speech emotion recognition approaches: A systematic review. *Speech Communication*, 102974 (2023)
- [61] Klie, J.-C., Bugert, M., Boullosa, B., Castilho, R.E., Gurevych, I.: The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pp. 5–9 (2018)
- [62] Honnibal, M., Montani, I., Landeghem, S.V., Boyd, A.: *spaCy: Industrial-strength Natural Language Processing in Python*. Available online. Accessed: 2023-01-14 (2020)
- [63] Hennequin, R., Khelif, A., Voituret, F., Moussallam, M.: Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software* **5**(50), 2154 (2020)

- [64] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M.: Montreal forced aligner: Trainable text-speech alignment using kaldi. In: Interspeech, vol. 2017, pp. 498–502 (2017)
- [65] Goodwin, C., Heritage, J.: Conversation analysis. *Annual review of anthropology* **19**(1), 283–307 (1990)
- [66] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019)
- [67] Breen, M., Dille, L.C., Kraemer, J., Gibson, E.: Inter-transcriber reliability for two systems of prosodic annotation: Tobi (tones and break indices) and rap (rhythm and pitch). *Corpus linguistics and linguistic theory* **8**(2), 277–312 (2012)
- [68] Belinkov, Y., Glass, J.: Analyzing hidden representations in end-to-end automatic speech recognition systems. *Advances in Neural Information Processing Systems* **30** (2017)