PSST! Prosodic Speech Segmentation with Transformers

Nathan Roll
University of California Santa Barbara
nroll@ucsb.edu

Calbert Graham
University of Cambridge
crg29@cam.ac.uk

Simon Todd
University of California Santa Barbara
sjtodd@ucsb.edu

Abstract

We develop and probe a model for detecting the boundaries of prosodic chunks in untranscribed conversational English speech. The model is obtained by fine-tuning a Transformer-based speech-to-text (STT) model to integrate the identification of Intonation Unit (IU) boundaries with the STT task. The model shows robust performance, both on held-out data and on out-of-distribution data representing different dialects and transcription protocols. By evaluating the model on degraded speech data, and comparing it with alternatives, we establish that it relies heavily on lexico-syntactic information inferred from audio, and not solely on acoustic information typically understood to cue prosodic structure. We release our model[1] as both a transcription tool and a baseline for further improvements in prosodic segmentation.

1 Introduction

A growing body of research in phonetics, phonology, and speech processing focuses on prosody: the encoding of prominence and phrasal organization (Pierrehumbert, 1999; Ladd, 2008) through interconnected suprasegmental cues (intonation, stress, rhythm, etc.) (Arvaniti, 2020). One reason for this focus is that prosodic phrasing groups words into chunks that can facilitate the generation and processing of naturalistic running speech for both speakers and listeners. For example, in English, the presence of detectable boundaries between chunks enhances speech intelligibility (Cooper and Sorensen, 1981; Selkirk, 1984) and helps listeners correctly discern the syntactic structure of the utterance (Streeter, 1978; Wingfield et al., 1984; Beach, 1991; Crystal, 1986; Warren, 1996).

In this paper, we generate, evaluate, and probe machine-learned models for detecting the boundaries of prosodic chunks in untranscribed conversational English speech. We focus on boundaries of the Intonation Unit (IU), which delineate "chunks" of speech that reflect cognitive and prosodic cohesion (Chafe, 1994; Du Bois et al., 1992). Developing a robust boundary detector for conversational speech would have important implications for linguistics. Methodologically, it would open the door to automated systems for fine-grained discourse transcription, and theoretically, it would facilitate exploration of the way that suprasegmentals interact to cue prosodic structure (Du Bois et al., 1992). Given the utility of prosodic boundaries for human speech perception, it may also contribute to the robustness of Automatic Speech Recognition (ASR) generally for conversational speech. Robust conversational ASR is made difficult by the fact that cues to segmental information are often reduced in conversation, may be masked by significant interspeaker variation, and often do not correspond precisely to the rigid syntactic structures of written language, among other challenges.

The detection of prosodic boundaries via automated methods has a rich history in work that aims to segment transcriptions of speech. However prior works have largely taken a pipeline approach: first creating textual transcriptions (either manually or via ASR) and subsequently applying boundary detection methods to the generated transcript. In addition, they have not typically focused on identifying IU boundaries in everyday conversations. Many works (e.g., Stolcke and Shriberg, 1996, Wang and Narayanan, 2004, and Liu et al., 2006) use the Switchboard corpus to identify syntactically-based prosodic boundaries in telephone conversations between strangers, using orthographic inputs and/or manually crafted acoustic features. Xu et al. (2014) applies pause, pitch, energy, and duration information to a similar task in spoken Mandarin. More recent work has pursued integrated approaches that consider Speech-To-Text (STT) transcription and segmentation simultaneously, but still have not focused on IU boundaries in conversational speech.

[1] https://github.com/Nathan-Roll1/PSST

Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), pages 476–487

Sarkar et al. (2018) introduced a model to perform ASR, segmentation, and diarization concurrently on the LibriSpeech corpus of read speech. Similarly, Hou et al. (2020) detected phone- and word-level timestamps while performing ASR on the TIMIT and WSJ corpora of read speech.

Here, we follow this more recent work in taking an integrated approach, which we use to detect IU boundaries in everyday conversations. We develop an end-to-end model that incorporates IU boundary detection into a Transformer-based (Vaswani et al., 2017) STT task. Specifically, we fine-tune Whisper (Radford et al., 2023), a highly successful STT model, to generate IU boundaries as it processes audio and generates a transcription. The incorporation of IU boundary detection into STT transcription allows for counterfactual considerations of lexico-syntactic probabilities, and allows our model to recognize the strong correspondences and interactions between syntax and prosody that are fundamental to linguistic theory (Bennett and Elfner, 2019).

Studies on automatic boundary predictions in the prosodic domain have primarily concentrated on two key areas: feature engineering and modeling methods. Feature engineering (e.g., Ananthakrishnan and Narayanan, 2005) involves identifying and operationalizing acoustic features such as pitch as pause that correlate with prosodic boundaries. Modeling methods involves comparing various statistical machine learning frameworks – such as memory-based learning (Busser et al., 2001), maximum entropy (Sridhar et al., 2008), and deep neural networks (Rosenberg et al., 2015) – that use these features in different ways to identify prosodic boundaries in unlabeled data.

The Transformer architecture obviates the distinction between these areas by allowing the model to discover useful acoustic features itself, based on self-attention mechanisms applied to positionally-encoded audio data. The model therefore efficiently discovers and leverages rich features present in input audio, without enforcing strong assumptions about what those features are or how they are structured in the time or frequency domains. In doing so, it exhibits similarity to human IU boundary detection by considering a myriad of fine-grained cues, including those that are difficult to operationalize with direct feature engineering (Du Bois et al., 1992). This represents a significant departure from previous attempts to detect prosodic phrase boundaries, which have typically used either simple durational cues and pauses (Yang, 2003; Salomon et al., 2004) or a combination of other predetermined suprasegmental cues (Mandal et al., 2007; Peters, 2003), and/or have isolated the task of prosodic boundary detection from that of STT transcription (Biron et al., 2021; Stehwien and Vu, 2017).

We investigate whether fine-tuning on a small, high quality dataset can "teach" a pretrained Transformer-based STT model to segment conversational speech audio into IUs, by detecting IU boundaries in the course of transcription. We perform two experiments, with the following research objectives:

1. To fine-tune an ASR-optimized Transformer model to perform reliable IU boundary detection integrated with STT transcription, and test its robustness to variation in acoustics and transcription protocol by evaluating it on out-of-distribution data.

2. To explore the factors that contribute to the model's performance, by evaluating it on degraded speech data and comparing it with alternatives that do not integrate IU boundary detection with STT transcription.

2 Experiment 1: reliable IU detection

In Experiment 1, we fine-tune Whisper (Radford et al., 2023), a Transformer-based STT model, to identify IU boundaries as it processes and transcribes audio. Our goal is not to improve the basic word recognition rate of Whisper, but rather to investigate whether its capabilities can be leveraged to recognize intonation unit boundaries, in a generalizable way. The model is fine-tuned on a corpus of conversational American English, and we establish its performance on held-out data from the same corpus. Then, we assess its robustness to naturalistic acoustic variation and differences in prosodic transcription protocol, by evaluating it on out-of-distribution speech data (i.e., non-American English data not used in the training of the model) from a corpus of conversational British English that uses distinct criteria to determine IU boundaries.

2.1 Methods

2.1.1 Data and preprocessing

Our training and within-distribution testing data come from the Santa Barbara Corpus of Spoken

American English (SBCSAE) (Du Bois et al., 2000–2005), which contains 60 prosodically transcribed naturalistic conversations between 210 speakers, spanning a total of ∼20 hours. The speakers, who represent 30 U.S. states, exhibit variation in age, race/ethnicity, and educational background. The corpus is roughly gender balanced, with 55% of speakers identifying as female and 44% as male (1 unknown).

The transcripts include words, IU boundaries, and a variety of other features, with high inter-transcriber agreement. Disagreements between transcribers are resolved by experts[2] (Du Bois et al., 2000–2005).

IU-boundary timestamps are precise to 0.1 seconds. Each conversation is recorded on a single-channel 22,050 Hz `.wav` file. Each file contains the entire conversation, except for personal identifiers and sensitive information, which were masked using a 400 Hz low-pass filter. We hold out the first five conversations in the corpus (∼10% of the overall data, comprised of ∼2 hours of speech) for testing, and use the remainder for training.

To preprocess the data, we identify contiguous stretches of non-overlapping speech. We extract the word tokens for each stretch from the transcript, including filled pauses and disfluencies ("um","uh", "unhuh", etc.), and add a token of a symbol that is otherwise not used in the corpus to designate each IU boundary. To meet the input requirements of Whisper (Radford et al., 2023), we resample the audio from 22,050 Hz to 16,000 Hz and split it into 30-second chunks, padding with zeros as required. The model then converts each chunk to a log-Mel spectrogram with 80 channels, 25 ms windows, and 10 ms strides, globally rescaled to the interval $[-1, 1]$.

For out-of-distribution testing, we use the Intonational Variation in English (IViE) corpus (Grabe et al., 2001). IViE is different from the SBCSAE in two key ways: first, it contains conversations from speakers of different dialects (British English as opposed to American English); and second, it is transcribed with a distinct intonational phrase methodology, adapted from the ToBI framework (Silverman et al., 1992; Beckman and Ayers Elam, 1997). We use the spontaneous portion of the corpus, preprocessed in the same way as described above.

[2]Our version of the corpus presents a single authoritative transcription per file, with no information about the precise cases where there was transcriber disagreement.

We chose the SBCSAE and IViE corpus for our investigation because they are composed of conversational speech and have been subjected to detailed transcription that identifies IUs through multifaceted consideration of prosodic structure. This is a substantial difference from past work that has heavily focused on corpora of read speech (e.g., TIMIT and WSJ) and corpora that have been segmented shallowly according to syntactic structure, punctuation, and/or simple phonetic factors such as silence detection (e.g., Switchboard). Using prosodically transcribed corpora of conversational speech lets us investigate the rich structured variation inherent in natural speech, in which prosody reflects dynamic discourse and cognitive factors as well as more stable phonological and syntactic factors. Furthermore, using two corpora that represent different varieties of the same language, with generally similar lexico-syntactic systems but different intonational systems, lets us assess the extent to which the model's learning is based on IU boundary features and not merely the performance of the ASR system it incorporates.

2.1.2 Model and fine-tuning

Our Prosodic Speech Segmentation with Transformers (PSST) model is fine-tuned from the largest English-specific version of Whisper, with 764 million parameters and a size of 3.06 GB.[3] The architecture of PSST, based on (Radford et al., 2023), is shown in Figure 1. The fine-tuned model takes raw audio as input and produces a transcript, which includes both words and – crucially – IU boundaries.

We obtained PSST by fine-tuning Whisper in a supervised fashion[4], using manually generated transcripts as the ground truth. In fine-tuning, the model was trained using the same hyperparameters as the original Whisper model, except for batch size (number of samples per train iteration) and gradient accumulation steps (number of batches per effective train iteration), both of which were changed (from 256 to 32, and from 1 to 2) due to limitations of computational resources. We trained

[3]This distribution is trained on a non-public corpus of audio and accompanying (non-prosodically-annotated) transcripts, where heuristics were used to ascertain that the transcription was human-made. The 480,000-hour English subset was aggregated from web sources and represents a diverse range of speakers and situations, according to Radford et al. (2023).

[4]Fine-tuning used a single NVIDIA V100 Tensor Core GPU with 32 GB of VRAM.

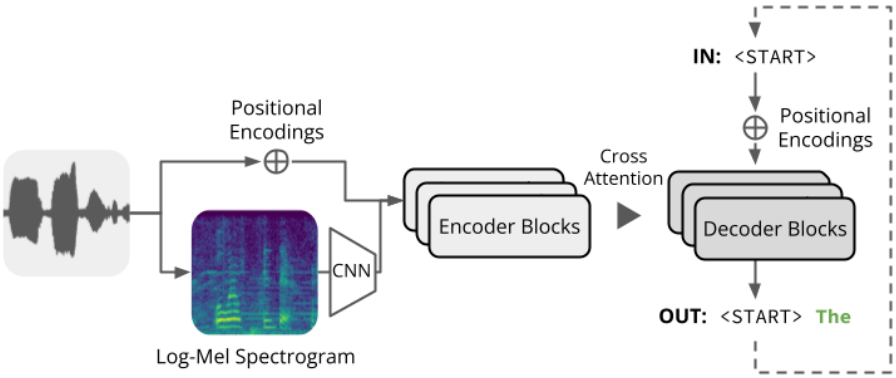

Figure 1: *PSST Architecture: Two convolutional layers activated by a Gaussian Error Linear Unit (GELU) convert a log-Mel spectrogram of each 30-second chunk of input into a linear vector, which is combined with a sinusoidal positional encoding. The array is passed through a series of encoder and decoder blocks, each composed of attention and multi-layer perceptron (MLP) components.*

the model for 400 steps (2 full passes of the training data). The learning rate hyperparameter was depressed for the first 50 steps to avoid early overfitting, increasingly linearly to reach 10^{-5}.

The trained model is highly efficient, requiring only four seconds to process a 30-second input using a consumer-grade GPU (and just over a minute using our CPU)[5]. Conversely, detailed and accurate manual discourse transcription by humans can take significantly longer (Du Bois et al., 1992).

2.1.3 Evaluation

The model outputs a transcript consisting of a stream of words and IU boundaries. We evaluate this output based not on the words it contains, but rather on the extent to which its boundaries are located in the correct temporal positions in the audio stream. To perform this evaluation, we generate timestamps for the output transcript by force-aligning it to the audio stream, using the Charsiu neural forced aligner[6] (Zhu et al., 2022). A generated IU boundary is deemed correct if it is force-aligned to within 20ms of the timestamp of a hand-transcribed boundary in the gold-standard SBCSAE data. Due to the use of forced alignment, successful IU boundary detection does not require perfect ASR performance, as incorrect tokens may still be placed in the correct location temporally.

Our primary metric for evaluating model performance is *F-score*, the harmonic mean of precision and recall. We calculate precision and recall based on boundary placement in the audio stream: precision is the proportion of boundaries in the model output that are force-aligned to within 20ms of a boundary in the hand-transcribed data, and recall is the proportion of boundaries in the hand-transcribed data that are within 20ms of a force-aligned boundary in the model output.

Generating IU boundaries in the right place is a difficult task: the model must both determine that a boundary occurs within a stream of words, and localize it with temporal precision. Even determining that a boundary occurs, independent of temporal alignment, is subject to significant ambiguity (Moore et al., 2016). Inter-labeler agreement for detecting intonational phrase boundaries in specific locations, for example, is 93.4% (Pitrelli et al., 1994).

Because F-score is based on the temporal placement of boundaries, it is affected by the dual difficulty of the task. To focus in on boundary occurrence, minimizing influences of temporal precision, we also report on word-level *accuracy*. Accuracy takes inspiration from word error rate in STT evaluation: it is based on the correct placement of boundary tokens in the transcript, independent of timestamps. We calculate it by considering the potential

[5] An 8-bit integer quantized version of our model is available as well, with nearly identical performance and a significantly faster inference speed.

[6] Charsiu uses convolutional layers built on top of a speech audio encoder (from wav2vec) and a phone sequence encoder (from BERT). It is trained to leverage phone sequence embeddings to reconstruct (quantized embeddings of) speech audio that has been masked through spectral augmentation in both the temporal and feature domains, based on both a reconstruction loss and a forward-sum loss. In this way, it learns a monotonic diagonal attention matrix that uniquely aligns the embeddings from the speech audio encoder and the phone sequence encoder in the temporal domain. We use the pre-trained `W2V2-FS-10ms` Charsiu model, which provides alignments for each 10ms window. This model has comparable performance to standard HMM-based forced aligners (such as the Montreal Forced Aligner and the Penn Forced Aligner) in the benchmarks reported by Zhu et al. (2022).

Table 1: *IU boundary detection performance on held-out data. PSST outperforms out-of-the-box Whisper and a baseline model that predicts no boundaries on the same test data, and seems to also outperform past models trained/tested on different data.*

Method	*F-Score*	*Acc.*
PSST (This Work)*	**0.87**	**0.96**
Rosenberg (2009)	0.81	0.93
Rosenberg (2010)	0.77	0.89
Hirschberg and Nakatani (1998)	0.70	0.83
Biron et al. (2021)*	0.66	0.86
Klejch et al. (2016)	0.63	0.87
Whisper (Radford et al., 2023)*	0.48	0.85
Baseline (No Boundaries)*	0.00	0.83

**Evaluated on the SBCSAE.*

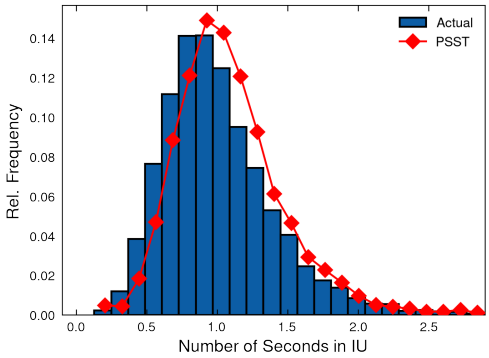

Figure 2: *Distributions of IU length (seconds) based on actual (blue bars) and model-generated (red dots/line) IU boundaries. IUs based on model-generated boundaries tend to be longer than expected, even though they typically contain the expected number of words.*

boundary sites in the output and gold-standard transcripts, which fall between every pair of words in each transcript. We align the two transcripts to each other, based on their separate alignments with the audio, and calculate accuracy as the proportion of aligned potential boundary sites that agree on whether or not a boundary occurs in that site. Accuracy is diminished by ASR failures, where a potential boundary site in one transcript is aligned to a word in the other transcript, and by boundary detection failures, where a site is labeled as containing a boundary in one transcript but not in the other.

2.2 Results

2.2.1 Performance on held-out test data

The results are shown in Table 1. PSST exhibits excellent IU boundary detection on held-out portions of the SBCSAE, in terms of both accuracy and F-score. Its performance is well above the baseline of a model that predicts no boundaries, and far exceeds that of out-of-the-box Whisper[7] on the same test set. Its performance also seems to exceed that of English-based models that have been previously reported in the literature; however, as these models all use different training and test data, it is difficult to make comparisons that are not affected by variation in aspects such as corpus content (number of speakers, dialect, scripted or unscripted, etc.) and transcription protocol.

[7]Whisper is trained to identify "phrase boundaries" (without a specific explanation of how they are defined). We assess the correspondence of these phrase boundaries to IU boundaries as a baseline of Whisper's performance on the IU segmentation task.

In order to get an overview of model outputs, we compare the distributions of IU length between the predicted and actual transcripts. When measured in terms of number of words, the predicted and actual distributions of IU lengths are highly similar, and show no significant differences in a Kolmogorov-Smirnov test ($p = 0.72$). When measured in terms of time, the distributions are qualitatively similar as seen in Figure 2 but significantly different ($p = 3.2 \times 10^{-9}$). We believe this effect to reflect shortcomings of the forced aligner rather than the transcription system: even when the model transcribes an IU correctly, the aligner may not place its boundaries in surrounding regions of silence in the same way as a human would.

After replacing boundary tokens with new lines, the PSST output can be compared with the human-annotated transcript. A successful sample transcription is shown in Table 2.

2.2.2 Performance on out-of-distribution data

Even on out-of-distribution data from the IViE corpus, PSST performs well, as shown in Table 3. Notably, it sees an improvement in performance relative to a baseline model that predicts no boundaries, whereas out-of-the-box Whisper does not. This indicates that the information PSST has learned from SBCSAE provides generalizable advantages for IU boundary detection. However, the fact that performance on IViE appears worse than performance on SBCSAE suggests that the reliability of PSST can be affected by variation in acoustics (e.g., across speakers of different dialects) and transcription protocol.

Table 2: *Sample Successful Transcription (SBCSAE04 8:33 to 8:50). Line breaks indicate IU boundaries.*

Actual Transcription	**PSST Transcription**
I'm the only teacher who's not experienced	I'm the only teacher who's not experienced
who's not certified	who's not certified
who just started teaching	who just started teaching
All these other teachers are old hands	All these other teachers are old hands
I mean they've all been at it for at	I mean they've all been at it for at
Well Chris is the least experienced besides me	Well Chris is the least experienced besides me
but still he's	but still he's
you know	you know
he's had his certification	he's had his certification
and he's had a year and stuff	and he's had a year and stuff
he's real good at it	he's real good at it

Table 3: *IU boundary detection performance on out-of-distribution test data from the IViE Corpus. PSST shows strong performance despite differences in dialect and transcription protocol compared to its training set.*

Method	*F-Score*	*Acc.*
PSST	**0.73**	**0.93**
Baseline	0.00	0.88
Whisper (Radford et al., 2023)	0.35	0.87

Table 4: *Confusion matrix for PSST IU boundary detection on held-out data from the SBCSAE.*

	Predicted	
Actual	Boundary	No Boundary
Boundary	**1,931**	371
No Boundary	378	**11,241**

2.2.3 Error Analysis

At the level of the transcript (i.e., not considering errors in temporal placement), PSST makes very few errors. As shown in Table 4, these errors include both false positives (boundaries predicted where they don't occur) and false negatives (boundaries missed). Inspection showed that errors in boundary detection are correlated with errors in word transcription, but not strongly: boundary errors also occur when all words are correctly transcribed, and there are many cases where boundaries are correctly detected in spite of errors in word transcription. This suggests that errors in PSST have two main causes: ASR-related inaccuracies and prosodic inaccuracies.

ASR-related inaccuracies refer to cases where the STT model either generates too many words, too few words, or the wrong words. The implications of ASR-related inaccuracies for joint or downstream boundary prediction have been well established in classic work (e.g. Liu et al., 2006). It is easy to imagine how poor STT transcription may limit IU boundary detection performance. Generating too many words can lead to false positives because the output transcript contains additional potential boundary sites, while generating too few words can lead to false negatives because the output transcript does not contain the required boundary sites. Generating the wrong words can lead to false positives or false negatives because the generated words may not fit in the same syntactic frames as the actual words, and IUs tend to be syntactically coherent, as demonstrated by the unsuccessful transcription in Table 5. However, because STT transcription and IU boundary detection are integrated in PSST, it is not possible to definitively say that poor transcription limiting boundary detection is the cause of the correlation between word error rate and boundary error rate; the reverse is also possible. We explore this issue further in Section 3.2.

Prosodic inaccuracies refer to cases where the model's word-level transcription is correct (or near enough to be accurately aligned with the gold-standard transcript), but an IU boundary prediction is nevertheless incorrect. Listening to such cases indicates that they often exhibit ambiguous prosodic cues to segmentation. Navigating this ambiguity requires weighting prosodic factors in a specific way; for human transcription, such weighting is codified in a transcription protocol. It is likely that PSST's weighting of prosodic factors does not precisely

Table 5: *Sample Unsuccessful Transcription (SBCSAE02 14:01 to 14:10). Line breaks indicate IU boundaries, with additional vertical space added for visual consistency.*

Actual Transcription	**PSST Transcription**
cause I've heard em for the past three months	cause I burned em for the past three months
I didn't think anything of it	I didn't think anything of it
but then	but then
this guy played songs for a whole hour	this guy played songs for a whole hour
and it was like	and it was like
eighty per cent of those songs I'd	eighty percent of those songs out
that band had sung that very night	that band
	his son
	that very night
Mhm	mhm

match that of the SBCSAE protocol.

3 Experiment 2: understanding the model

In Experiment 2, we explore factors that contribute to PSST's excellent results. In Experiment 2A, we explore the kind of acoustic features that the model may be relying upon, by evaluating performance on acoustically degraded stimuli. In Experiment 2B, we explore the extent to which the model integrates acoustic and lexico-syntactic information, by comparing its performance with that of alternatives that have limited integration.

3.1 Experiment 2A: use of acoustic features

As a STT model, PSST uses acoustic features to infer the identity of words. The error analysis in Section 2.2.3 suggests that inaccuracies in lexical inference can cause cascading errors in IU boundary detection, yet also reveals that the model can still struggle to detect acoustically-cued IU boundaries even when word-level inference is correct. Does this imply that the acoustic features PSST uses are primarily those that cue lexical identity?

To address this question, we analyze model performance on acoustically-degraded inputs via frequency-based filtering. In humans, it has been shown that vowel formants are particularly important for correct lexical inference and intelligibility in running speech (Kewley-Port et al., 2007; Fogerty and Humes, 2012), while pitch contours captured by fundamental frequency (F0) are a salient cue to prosodic boundaries (Streeter, 1978; Pierrehumbert, 1980; Jusczyk et al., 1992). If PSST uses acoustic features primarily to cue lexical identity, then filtering out frequencies in the range that represent F1–F3 vowel formants for American English (∼200–3200Hz) (Peterson and Barney, 1952; Hillenbrand et al., 1995; Kent and Vorperian, 2018) should reduce performance to near-baseline levels, while filtering out frequencies in the F0 range (less than ∼200Hz) should not dramatically impair performance.[8]

We applied a series of low-pass and high-pass Butterworth filters (Figure 3) to the audio in the held-out test set (Butterworth, 1930). We crossed the choice of low- or high-pass filter with the choice of a threshold frequencies of 200 Hz, 400 Hz, 800 Hz, 1.6 kHz, or 3.2 kHz, yielding 10 different versions of degraded test data. We applied the model described in Section 2 to each version of the test set. The model was unable to generate any word tokens for the 200 Hz low-pass filtered data, so we do not report its boundary prediction performance in what follows.

The results are shown in Figure 4. Generally, PSST's performance declines as larger acoustic ranges are filtered out, for both low- and high-pass filters. When crucial frequencies representing F1–F3 are removed (400 Hz low-pass and 3.2 kHz high-pass), performance is notably poor, but still better than performance of the baseline or out-of-the-box Whisper model on undegraded test data (cf. Table 1). Conversely, performance under a 200Hz high-pass filter that removes F0 but leaves F1–F3 intact shows little change relative to performance

[8] Other acoustic features such as duration and intensity have also been identified as relevant to prosodic boundary detection in humans. We do not explore these features here because they are less strongly linked to lexical inference than frequency; however, a more explicit investigation of the impact of acoustic features on our model is worth considering in a future study.

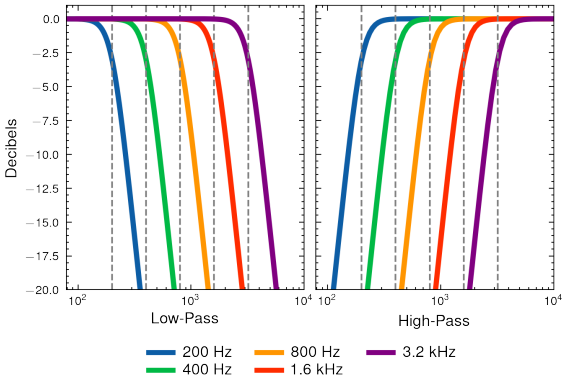

Figure 3: *Low-pass (left) and high-pass (right) Butterworth filters applied to audio input. These filters have a soft cut-off, which smoothly attentuates frequencies above (for low-pass) or below (for high-pass) the threshold frequency.*

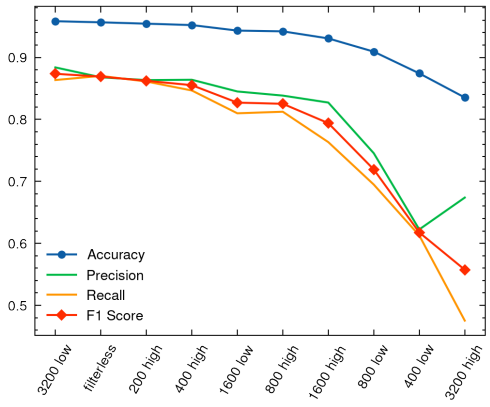

Figure 4: *IU boundary detection performance on acoustically-degraded audio, by filter. Performance decreases as frequencies from the F1–F3 range are filtered out, but shows little decrease when F0 is filtered out.*

on unfiltered data. Taken together, these results suggest that PSST does indeed primarily use acoustic features to cue lexical identity, and not, for example, to track pitch contours. Nevertheless, given that performance decreases slightly (∼0.8%) when F0 is filtered out, it remains possible that PSST uses pitch (and other acoustic features) to a secondary extent for IU boundary detection.

3.2 Experiment 2B: integration of acoustic and lexico-syntactic information

The results of Section 3.1 imply that the IU boundaries that PSST detects are primarily cued by lexico-syntactic information, rather than acoustics. At the same time, however, the results of Section 2.2.3 show that PSST can identify boundaries even when lexical identity is obscured, suggesting a broader role for acoustics. Does this mean that the success of PSST is affected by its integration of IU boundary detection with STT transcription, allowing it to jointly leverage acoustic and lexico-syntactic information?

Table 6: *IU boundary detection on held-out SBCSAE data: comparison of models from Experiments 1-2. Lexical and Masked models that dissociate IU boundary detection from STT transcription perform worse than PSST models that integrate them, even when inputs are degraded.*

Method	*F-Score*	*Acc.*
PSST	**0.87**	**0.96**
PSST (1.6 kHz high-pass)	0.79	0.93
Lexical	0.77	0.93
Masked	0.71	0.87
Whiper (Radford et al., 2023)	0.48	0.85
Baseline	0.00	0.83

To address this question, we construct two alternative models that dissociate STT transcription from IU boundary detection: a Lexical model and a Masked model. The Lexical model represents the best boundary predictions a model could make without direct access to acoustics. It takes Whisper-generated text as input and predicts (force-aligned) IU boundaries in it, based on fine-tuning of the 1.2 billion parameter (5.36 GB) distribution of GPT-NEO (Black et al., 2021). The Masked model represents an attempt to downplay lexical identification in the IU boundary detection task, by replacing all words in the test and training data with a common mask token. It is otherwise identical to PSST; thus, even though it is not required to output distinct lexical items, it likely maintains latent lexico-syntactic representations. Both models are trained and tested using the SBCSAE data described in Section 2.1.1.

The results are shown in Table 6, together with previously-described models for context. Both the Lexical and the Masked model perform better than the baseline and out-of-the-box Whisper models, indicating that IU boundary detection can draw upon lexico-syntactic and acoustic information separately. However, both models perform worse than PSST, even when the input is substantially acoustically degraded. This suggests that at least some of the success of PSST is due to the interaction of acoustic and lexico-syntactic information, which arises due to its integration of IU boundary detection with STT transcription.

4 Discussion & Conclusion

This study had two research objectives, as stated in Section 1. In relation to Objective 1, we successfully fine-tuned Whisper (Radford et al., 2023) to segment conversational speech into IUs. We achieved F-scores of 0.87 on held-out test data and 0.72 on out-of-distribution data, indicating strong reliability. Whisper was originally trained on the simple objective of discerning words from audio, yet the fact that we were able to repurpose it successfully using few-shot learning holds significant promise for other NLP studies that rely on smaller datasets.

In relation to Objective 2, we explored the potential factors influencing the model's performance. Our findings suggest that the model uses acoustic information primarily for lexical identification. Interestingly, the model also appears to benefit from the interactions between acoustic and lexico-syntactic information that are made possible through the integration of IU boundary detection with STT transcription. These results may be surprising from an expectation that prosodic boundaries would be reflected primarily by acoustic cues, but they reinforce the understanding from linguistic theory that prosody involves complex interactions between syntax and phonology (Bennett and Elfner, 2019).

Given these results, there are two clear next steps. First, though our model was able to perform reliable IU boundary detection, its performance was hindered in out-of-distribution contexts involving different dialects and transcription protocols. Expanding the training set to be more representative of such variation would further improve its reliability and adaptability. Second, though we observed a benefit from integrating acoustic and lexico-syntactic information, it appears that the acoustic information was relatively underweighted. This is likely a reflection of the fact that fine-tuning the integrated model represents a very small amount of training relative to training the original STT model, in which acoustic cues to prosodic boundaries have limited relevance. Fine-tuning for longer, or on more data, may help increase the weight of acoustic cues to prosodic boundaries. In addition, experiments with acoustically enhanced rather than degraded stimuli may help to illuminate the circumstances under which acoustic cues to prosodic boundaries can override biases from lexico-syntactic information.

Our results suggest that STT transcription and prosodic boundary identification should not be approached as independent challenges, but rather as interacting components of a unified speech processing objective. Simply requiring prosodic features to be represented in the desired output transcriptions unlocks a seemingly latent ability for STT models to identify them. Overall, our results suggest that such STT models implicitly represent prosodically-relevant information, given their success in a few-shot context. Furthermore, the robustness of segmentation performance when exposed to moderate frequency-based signal tampering, or even complete F0 masking, strengthens the case for prosody-syntax interplay at the "heart" of high-performance ASR models. By following a similar process to what we have shown here, there is strong potential for STT models to be extended to detect other speech phenomena as well – such as prosodic accents, vocal quality changes, or even environmental contexts – which would put us one step closer to a fully automated discourse transcription system.

5 Acknowledgments

The authors would like to thank John W. Du Bois for useful discussions about the SBCSAE. This project was supported jointly by grants from the the Office of Undergraduate Research & Creative Activities (URCA) at UC Santa Barbara and the Cambridge Language Sciences Incubator Fund.

References

S. Ananthakrishnan and S. S. Narayanan. 2005. An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model. In *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 269–272.

Amalia Arvaniti. 2020. The phonetics of prosody. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Cheryl M. Beach. 1991. The interpretations of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. *Journal of Memory and Language*, 30(6):644–663.

Mary E Beckman and Gayle Ayers Elam. 1997. Guidelines for ToBI labelling. Technical report, The Ohio State University.

Ryan Bennett and Emily Elfner. 2019. The syntax-prosody interface. *Annual Review of Linguistics*, 5(1):151–171.

Tirza Biron, Daniel Baum, Dominik Freche, Nadav Matalon, Netanel Ehrmann, Eyal Weinreb, David Biron, and Elisha Moses. 2021. Automatic detection of prosodic boundaries in spontaneous speech. *PLOS One*, 16(5):e0250969.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow.

Bertjan Busser, Walter Daelemans, and Antal van den Bosch. 2001. Predicting phrase breaks with memory-based learning. In *Proceedings of the 4th ISCA Tutorial & Research Workshop on Speech Synthesis*, page 125.

Stephen Butterworth. 1930. On the theory of filter amplifiers. *Wireless Engineer*, 7(6):536–541.

Wallace L. Chafe. 1994. *Discourse, Consciousness, and Time*. The University of Chicago Press, Chicago, IL.

William E. Cooper and John M. Sorensen. 1981. *Fundamental Frequency in Sentence Production*. Springer, New York, NY.

David Crystal. 1986. Prosodic development. In Paul Fletcher and Michael Garman, editors, *Language Acquisition: Studies in First Language Development*, pages 174–197. Cambridge University Press, New York, NY.

John W. Du Bois, Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000–2005. Santa Barbara corpus of spoken American English, parts 1–4. Philadelphia, PA: Linguistic Data Consortium.

John W. Du Bois, Susanna Cumming, Stephen Schuetze-Coburn, and Danae Paolino. 1992. Discourse transcription. *Santa Barbara Papers in Linguistics*, 4.

Daniel Fogerty and Larry E Humes. 2012. The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *The Journal of the Acoustical Society of America*, 131(2):1490–1501.

Esther Grabe, B Post, and F Nolan. 2001. Modelling intonational variation in English: The IViE system. In *Proceedings of Speech Prosody*, pages 51–57.

James Hillenbrand, Laura A. Getty, Michael J. Clark, and Kimberlee Wheeler. 1995. Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, 97(5):3099–3111.

Julia Hirschberg and Christine H. Nakatani. 1998. Acoustic indicators of topic segmentation. In *Proceedings of the 5th International Conference on Spoken Language Processing*.

Junfeng Hou, Wu Guo, Yan Song, and Li-Rong Dai. 2020. Segment boundary detection directed attention for online end-to-end speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2020:3.

Peter W Jusczyk, Kathy Hirsh-Pasek, Deborah G Kemler Nelson, Lori J Kennedy, Amanda Woodward, and Julie Piwoz. 1992. Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology*, 24(2):252–293.

Raymond D. Kent and Houri K. Vorperian. 2018. Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of Communication Disorders*, 74:74–97.

Diane Kewley-Port, T Zachary Burkle, and Jae Hee Lee. 2007. Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 122(4):2365–2375.

Ondrej Klejch, Peter Bell, and Steve Renals. 2016. Punctuated transcription of multi-genre broadcasts using acoustic and lexical approaches. In *Proceedings of the 2016 IEEE Spoken Language Technology Workshop*, pages 433–440.

D. Robert Ladd. 2008. *Intonational Phonology*, 2 edition. Cambridge University Press.

Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1526–1540.

Shyamal Kr. Das Mandal, Bhaskar Gupta, and Asoke Kumar Datta. 2007. Word boundary detection based on suprasegmental features: A case study on Bangla speech. *International Journal of Speech Technology*, 9:17–28.

Russell Moore, Andrew Caines, Calbert Graham, and Paula Buttery. 2016. Automated speech-unit delimitation in spoken learner English. In *Proceedings of COLING*, pages 782–793.

Benno Peters. 2003. Multiple cues for phonetic phrase boundaries in German spontaneous speech. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 1795–1798.

Gordon E. Peterson and Harold L. Barney. 1952. Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2):175–184.

Janet Pierrehumbert. 1999. Prosody and intonation. In Robert A. Wilson and Frank C. Keil, editors, *The MIT Encyclopedia of Cognitive Sciences*, pages 479–482. MIT Press, Cambridge, MA.

Janet Breckenridge Pierrehumbert. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. thesis, Massachusetts Institute of Technology.

John F. Pitrelli, Mary E. Beckman, and Julia Hirschberg. 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the 3rd International Conference on Spoken Language Processing*, pages 123–126.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518.

Andrew Rosenberg. 2009. *Automatic Detection and Classification of Prosodic Events*. Phd dissertation, Columbia University.

Andrew Rosenberg. 2010. Classification of prosodic events using quantized contour modeling. In *Proceedings of NAACL*, pages 721–724.

Andrew Rosenberg, Raul Fernandez, and Bhuvana Ramabhadran. 2015. Modeling phrasing and prominence using deep recurrent learning. In *Proceedings of INTERSPEECH 2015*, pages 3066–3070.

Ariel Salomon, Carol Y. Espy-Wilson, and Om Deshmukh. 2004. Detection of speech landmarks: Use of temporal information. *Journal of the Acoustical Society of America*, 115:1296–1305.

Amitrajit Sarkar, Surajit Dasgupta, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2018. Says who? Deep learning models for joint speech recognition, segmentation and diarization. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5229–5233.

Elisabeth Selkirk. 1984. *Phonology and Syntax: The Relation between Sound and Structure*. MIT Press, Cambridge, MA.

Kim E A Silverman, Mary E Beckman, John F Pitrelli, Mari Ostendorf, Colin W Wightman, Patti Price, Janet B Pierrehumbert, and Julia Hirschberg. 1992. ToBI: A standard for labeling English prosody. In *Proceedings of the Second International Conference on Spoken Language Processing*, pages 867–870.

Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Shrikanth S. Narayanan. 2008. 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.

Sabrina Stehwien and Ngoc Thang Vu. 2017. Prosodic event recognition using convolutional neural networks with context information. In *Proceedings of INTERSPEECH 2017*, pages 2326–2330.

A. Stolcke and E. Shriberg. 1996. Automatic linguistic segmentation of conversational speech. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, volume 2, pages 1005–1008.

Lynn A. Streeter. 1978. Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America*, 64:1582–1592.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*.

Dagen Wang and S.S. Narayanan. 2004. A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues. In *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages I–525.

Paul Warren. 1996. Prosody and parsing: An introduction. *Language and Cognitive Processes*, 11:1–16.

Arthur Wingfield, Linda Lombardi, and Scott Sokol. 1984. Prosodic features and the intelligibility of accelerated speech: Syntactic versus periodic segmentation. *Journal of Speech and Hearing Research*, 27:128–134.

Chenglin Xu, Lei Xie, and Zhonghua Fu. 2014. Sentence boundary detection in Chinese broadcast news using conditional random fields and prosodic features. In *Proceedings of the 2014 IEEE China Summit & International Conference on Signal and Information Processing*, pages 37–41.

Li-Chiung Yang. 2003. Duration and pauses as phrase and boundary marking indicators in speech. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 1791–1794.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Phone-to-audio alignment without text: A semi-supervised approach. In *Proceedings of the 47th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8167–8171.

A Limitations

Our approach to prosodic boundary detection is not without limitations. Firstly, as with any automatic evaluation procedure, the challenge of quantifying performance is a significant hurdle. Due to the strong dependence of the gold-standard hand-annotated data on human perception and nuanced transcription protocols, which together raise the potential for variation and inter-annotator disagreement, our evaluations are only as good as our ability to create effective and reliable performance metrics.

Secondly, our model is designed to operate in an end-to-end manner: it detects prosodic boundaries based on the processing of raw audio data, without explicitly generating intermediate (human-accessible) levels of representation. This approach

obscures the contribution of the specific features (acoustic and otherwise) that are implicitly learned by the model as cues to prosodic boundaries. The inherent lack of interpretability of the model's decisions makes it challenging to assign importance to specific prosodic elements. While we work to tease apart the contributing factors through acoustic degradation and lexical/acoustic masking, the interconnectedness of prosody at times presents ill-posed problems for such analyses. This both provides an opportunity for future projects and maintains the relevance of the many previous works which address factors individually.