The Role of Prosody in Spoken Question Answering

Anonymous ACL submission

Abstract

Spoken language understanding research to date has generally carried a heavy text perspective. Most datasets are derived from text, which is then subsequently synthesized into speech, and most models typically rely on automatic transcriptions of speech. This is to the detriment of prosody–additional information carried by the speech signal beyond the phonetics of the words themselves and difficult to recover from text alone. In this work, we investigate the role of prosody in Spoken Question Answering. By isolating prosodic and lexical information on the SLUE-SQA-5 dataset, which consists of natural speech, we demonstrate that models trained on prosodic information alone can perform reasonably well by utilizing prosodic cues. However, we find that when lexical information is available, models tend to predominantly rely on it. Our findings suggest that while prosodic cues provide valuable supplementary information, more effective integration methods are required to ensure prosody contributes more significantly alongside lexical features.

1 Introduction

Prosody, which is characterized by elements of speech beyond orthographic words, such as pitch, stress and rhythm, plays a critical role in both speech production and perception. It has been shown to impact how people perceive speech, with difficulties often arising when the natural variability in prosodic structure is limited, as is the case with synthetic speech (Winters and Pisoni, 2004; Wester et al., 2016). In human listening comprehension, prosodic cues are essential in guiding listeners through the process of interpreting spoken language (Buck, 2001; Keskin et al., 2019). The incorrectly-stressed elements in speech can also cause listeners to make incorrect inferences (Field, 2005). Motivated by these linguistic findings, researchers have explored how prosody can be leveraged in speech-related tasks computationally.

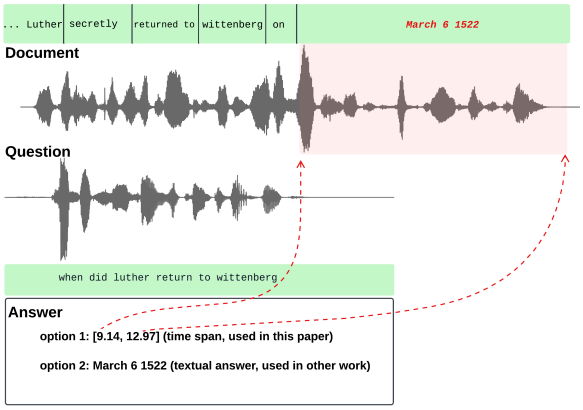


Figure 1: Illustration of the SQA format

One of the primary tasks in this area is Spoken Language Understanding (SLU), which focuses on extracting meaningful information from spoken language input. Unlike Natural Language Understanding (NLU), which primarily deals with text-based information, SLU incorporates the added complexity of processing signal made of prosodic features such as intonation, stress, and pauses. In this work, we use the term *lexical* information to refer to information that is also present in the text in its orthographic form. That is, info that encompasses phonetic information from the speech signal. Therefore, in our work, we categorize anything that is not lexical as *prosodic* information.

Recent advances in NLU research have significantly impacted SLU through the use of a cascade approach. This approach consists of two key components: an Automatic Speech Recognition (ASR) model that transcribes speech into text, followed by an NLU model that is fine-tuned for specific downstream tasks. Spoken Question Answering (SQA) is one of the challenging SLU tasks, and takes the form of listening comprehension. In SQA, the input consists of a spoken passage accompanied by a question about that passage, and the model is required to provide the correct answer. This answer

can take the form of timestamps in the passage, as used in this paper and previous works (Lee et al., 2018; Lin et al., 2022) , or it can be a textual output (Shon et al., 2024). Figure 1 illustrates the typical input-output structure of an SQA task. The evolution of SQA research has gradually shifted from synthetic speech datasets to those based on natural speech. In early studies, researchers utilized Text-To-Speech (TTS) systems to convert existing Textual Question Answering (TQA) datasets into large-scale SQA corpora (Lee et al., 2018; Lin et al., 2022; Ünlü Menevşe et al., 2022). However, since SQA means prosodic information is available in the input, the prosodic characteristics of synthetic speech may not accurately represent those found in natural speech (Wester et al., 2016; Clark et al., 2019; Chan and Kuang, 2024). This has led to concerns about the effectiveness of using synthetic speech for tasks where prosody plays a crucial role. To address these limitations, recent efforts have focused on integrating more natural speech into SQA datasets. Some studies have developed small test sets read by human speakers (Lin et al., 2022), while others have explored hybrid datasets where the questions are recorded by humans, but the passages are synthetically generated (Wu et al., 2024). Additionally, there has been work on creating training sets by sourcing spoken documents relevant to each question from external natural speech corpora (Shon et al., 2023).

Given the availability of the natural SQA training datasets, we aim to explore whether models can utilize prosody when comprehending speech, as humans do. The cascade approach, however, is not well-suited for this task, as prosodic information is typically lost after the transcription stage, and recovering prosody from text has been shown to be difficult (Talman et al., 2019). Although some research has attempted to explicitly incorporate word-level prosodic features into NLU models (Tran et al., 2018, 2019), the errors from ASR and alignments tend to propagate, resulting in ill-formed inputs and significantly impacting performance. Recent developments, particularly in Self-Supervised Learning (SSL) representations, have enabled researchers to bypass explicit transcription through end-to-end models(Chuang et al., 2020) or by using discrete units as pseudo-text(Lin et al., 2022), which latter is the framework adopted in this paper.

In this work we are motivated to gain a comprehensive understanding of how prosodic information, distinct from the lexical information used in NLU tasks, contributes to the SQA task. Specifically, two key research questions are investigated: 1) *Is Prosodic information sufficient for SQA tasks? as intonation, pitch, and pauses, signal important structural and emphatic aspects of speech.*, and 2) *Do SQA models utilize prosodic information when lexical information is also present?*

To address these questions, we carefully design two experimental conditions of our dataset: one that approximates prosodic information only, and another that approximates lexical information only. Directly disentangling prosodic and lexical content in speech is a complex challenge that remains unsolved (Quamer and Gutierrez-Osuna, 2024; Skerry-Ryan et al., 2018). Among the various approaches proposed to achieve delexicalization and isolate the contribution of prosody, applying a low-pass filter has been one of the most widely used techniques, in psycholinguistics but also in modelling (Goldman et al., 2014; Niebuhr et al., 2020; Audibert et al., 2023); it can preserve an approximation of the prosodic features while removing most of the discriminating information that phonetically delineates orthographic words (Mehler et al., 1988).

Hence, for the prosodic condition, we apply a low-pass filter to remove information above a certain cutoff frequency, ensuring that most lexical information is excluded from the speech signal. In contrast, for the lexical condition, we flatten both pitch and intensity to eliminate most prosodic variation. While prosodic and lexical information are not completely disentangled, our experiments show that the reduction of these elements is sufficient to prevent significant interference of one variable over the other in the results.

Through controlled experiments, we demonstrate that prosodic information alone can, to some extent, guide models to answer questions in SQA tasks. However, while prosody offers meaningful complementary cues, we find that models predominantly rely on lexical information when it is available. By providing a deeper understanding of prosody's role in SQA, we hope to pave the way for future work on developing more robust models capable of leveraging both lexical and prosodic information effectively, particularly in situations where lexical information is limited or degraded.

2 Related work

To date, SLU research has mostly involved first identifying word sequences. As such, there has been a focus on integrating prosodic information into ASR models. Previous research has used prosody by conditioning the acoustic and pronunciation modelling on prosodic features (Ostendorf et al., 2003; Chen et al., 2006), simultaneously predicting prosodic events (Chen et al., 2003; Hasegawa-Johnson et al., 2005), or incorporating prosodic information in N-best rescoring in hybrid ASR systems (Ananthakrishnan and Narayanan, 2007, 2009; Huang et al., 2010). However, current state-of-the-art ASR models do not model prosody explicitly.

Researchers have also explored using prosody to help other tasks. Assuming known time alignments, incorporating word-level prosodic features has yielded improvements in constituent parsing of conversational speech (Tran et al., 2018, 2019). Prosody has also been shown helpful in topic tracking (Guinaudeau and Hirschberg, 2011), dialogue act classification (Wei et al., 2022) speech to intent (Rajaa, 2023), and emotion recognition (Luengo et al., 2005; Naderi and Nasersharif, 2023). These studies modeled prosodic patterns either at the word or utterance level by averaging frame-level features such as pitch and intensity, or by using other hand-selected prosodic features. With the rise in popularity of neural networks, prosodic patterns can now be more effectively captured and modeled directly through CNNs, allowing for a more comprehensive representation of the prosodic features without the need for manual selection.

However, with the emergence of SSL models (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022), prosody is usually not explicitly modeled as a separate feature. Instead, it has been shown that these models capture prosodic information implicitly within their learned representations, alongside other linguistic features. SSL models are pre-trained on large amounts of unlabeled audio data, learning representations by predicting missing portions of the input signal or clustered latent speech units. There has been extensive research exploring the utility of SSL representations in prosody-related tasks, and it has been concluded that these representations encode prosodic information such as gender and speaker identity (de Seyssel et al., 2022; Liu et al., 2023; Mohamed et al., 2024). These representations have also been successfully applied to tasks such as emotion recognition, speaker identification, and intonation analysis (Lin et al., 2023). To leverage the capabilities of advanced language models, k-means clustering or other quantization approaches are typically used in conjunction with SSL representations to reduce the length of the input sequences. It has been observed that even within these discrete units, prosodic information is preserved, resulting in relatively low error rates for speaker and gender classification, particularly when more clusters are used (de Seyssel et al., 2022). This suggests that these discrete units retain key prosodic cues, which we use to represent speech in our study, allowing us to investigate the role of prosody in SQA tasks more effectively.

3 Methods

To investigate the role of prosody in SQA, in addition to the original datasets which combine both lexical and prosodic information (i.e., the dataset in its *natural condition*), we also designed a set of experiments that systematically examine the effects of prosodic information and lexical information individually. Our approach involves three main stages: data preparation, model training, and evaluation.

3.1 Data preparation

SLUE-SQA-5 dataset

We use the SLUE-PHASE SQA dataset (Shon et al., 2023) for this work. Unlike earlier datasets, which often rely on synthetic speech generated from TTS systems, it features naturally occurring spoken data, allowing for the study of prosody in a more realistic context. The corpus is derived from five existing TQA datasets and the questions are collected from crowd-source workers. The documents are collected by retrieving relevant documents to each question from the Spoken Wikipedia dataset (Köhn et al., 2016). All audios are from natural speech, which in this paper, we refer to as the *natural condition* (as opposed to the *lexical* or *prosodic* conditions that we define shortly). Table 1 illustrates the statistics of the corpus. In addition to train, test, and dev sets, the dataset includes a verified test set consisting of hand-picked question-document pairs from the test set, in which the document provides sufficient clues to answer the question.

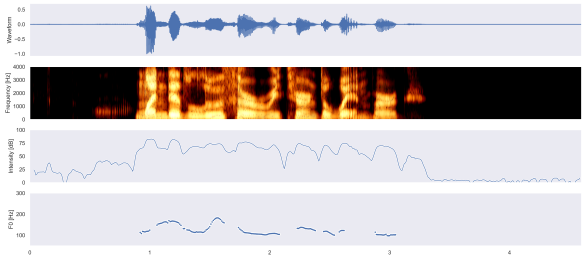
Dataset modification

We modify the audio in two different ways with Parselmouth (Jadoul et al., 2018; Boersma and Weenink, 2021), and examples of spectrograms

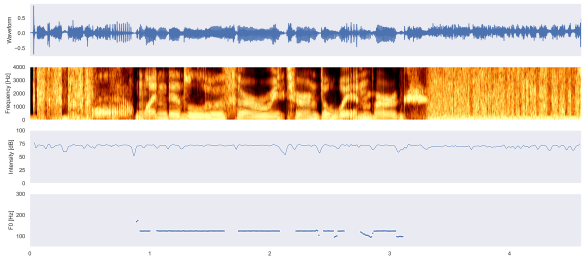
dataset	questions	documents	duration (hrs)
train	46186	15148	244
dev	1939	1624	21.2
test	2382	1969	25.8
verified test	408	322	4.2

Table 1: Statistics over the SLUE-SQA-5 dataset.

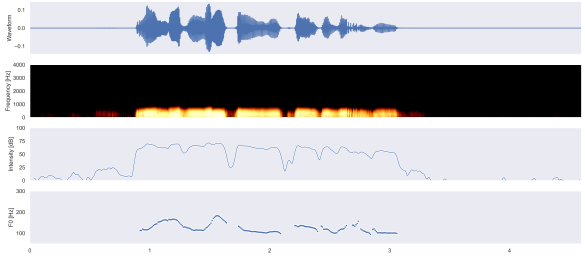
[1] for a same audio under different conditions are shown in Figure 2.



(a) Natural condition



(b) Lexical condition



(c) Prosodic condition

Figure 2: Spectrogram of the example speech under different conditions. In each sub-figure, the top plot is the waveform, the second plot is the spectrogram, the third plot is the intensity, and the bottom plot is the F0.

In the first setting, we remove the variations in both pitch and intensity, which we refer to as the *lexical condition*. This modification ensures that primarily lexical information remains, while that two of the main prosodic features, intonation and stress are considerably reduced. This results in a non-expressive, almost robotic-like quality to the sound. Specifically, we flatten both the fundamental frequency and intensity to the average value of each utterance. This approach helps us examine how models respond when only lexical information is present, with minimal prosodic variation. In Figure 2b, we can observe that the F0 and intensity contour is nearly flat. It should be noted that it is hard not to introduce the artifacts when flattening the intensity, for example, breath can become very loud and the gain inside smaller segments of silence is prominent (Ekstedt and Skantze, 2022). It should still be noted that rhythm (i.e duration) is not modified here.

In the second setting which we refer to as the *prosodic condition*, we apply a low-pass filter to the audio, removing high-frequency components, as shown in Figure 2c where the spectrogram shows a clear cut-off above the threshold. Filtering audio by frequency inevitably affects prosodic information as well, since prosody is embedded in various frequency bands. To mitigate this, we set the cut-off frequency to 300Hz, aiming to preserve as much prosodic information as possible while reducing high-frequency lexical cues. The choice of 300Hz is based on the distribution of speech energy: vowel sounds generally lie in the range of 250 to 2000Hz, voiced consonants between 250 and 4000Hz, and unvoiced or voiceless consonants primarily occupy the 2000 to 8000Hz range (Colatosti et al., 2024). By targeting the lower frequencies, we attempt to retain certain prosodic elements like pitch contour and rhythm, while removing higher-frequency lexical content. Section 4.1 explores the effects of applying different cut-off frequencies, allowing us to investigate how varying amounts of high-frequency information influence model performance.

It is important to note that we do not aim to fully disentangle prosodic and lexical information as this would be an extremely complex task given their intertwined nature in natural speech. Instead, our objective is to generate modified versions of the dataset that either keep lexical information by suppressing prosody or reduce lexical content while retaining some prosodic cues. We are not claiming that the remaining prosodic or lexical information is identical to its representation in natural speech, rather this research is carried out under the awareness of this is a limitation. These manipulations serve as controlled approximations, allowing us to systematically investigate the role of prosody and lexical information in SQA tasks.

[1]The corresponding audio files are provided in the supplementary materials.

3.2 Model training

We use the Discrete Spoken Unit Adaptive Learning (DUAL) framework (Lin et al., 2022), which consists of two main components: a Speech Content Encoder (SCE) and a Pre-trained Language Model (PLM). Unlike conventional cascade models, DUAL bypasses the reliance on ASR transcripts, and thus also the associated ASR error propagation. The SCE leverages WavLM, a self-supervised pre-trained model known for strong performance in prosody-related tasks (Lin et al., 2023), to encode representations directly from raw audio waveforms. These representations are then processed using k-means clustering, converting them into discrete units that are deduplicated before being fed into the PLM.

Although deduplication could potentially discard duration information, which is an important aspect of prosody, the impact in our case is minimal. Using 1000 clusters, we observe very few repetitions, with only 8% of units showing more than three consecutive repetitions in the verified testset, which are likely due to silence. Furthermore, higher cluster counts have been shown to retain more prosodic information, as demonstrated by performance improvements in tasks like gender and speaker classification (Sicherman and Adi, 2023).

For our SCE, we use a pretrained SpeechBrain model (Ravanelli et al., 2021)[2], which is a WavLM Large model pretrained on the Librispeech 960-hour corpus. The representations from this model remain frozen for all our experiments. The PLM is responsible for predicting the answer span within the context passage by identifying the start and end positions, similar to a typical TQA model. Consistent with the DUAL paper, we use the Longformer-base model[3] as the PLM, a BERT-like model for long documents, pretrained on unlabeled long text documents (Beltagy et al., 2020).

For reproducibility, all configurations are detailed here. We utilize eight A100-80 GPUs with a total batch size of 128, training the models for up to 18 epochs. Following the original DUAL framework, the learning rate is warmed up over the first 500 steps. We conduct a learning rate search within the range of $[5e-6, 1e-5, 5e-5, 1e-4]$, starting with the natural condition. Once the optimal learning rate is identified, we evaluate the performance on other conditions to ensure that the selected rate is not biased toward the natural condition. Ultimately, the learning rate is fixed at $1e-5$ for all experiments. The model needs 160 GPU hrs when tracking all evaluations in Section 4.2. To account for variability, all models are trained using three different random seeds, and we report the mean and standard deviation across all results.

[2]We use the pre-trained WavLM representations available at https://huggingface.co/speechbrain/SSL_Quantization/tree/main/LibriSpeech960/wavlm/LibriSpeech_wavlm_k1000_L23.pt for our experiments. While it is possible to select representations from different layers, a thorough layer-wise analysis falls beyond the scope of this work.

[3]https://huggingface.co/allenai/longformer-base-4096

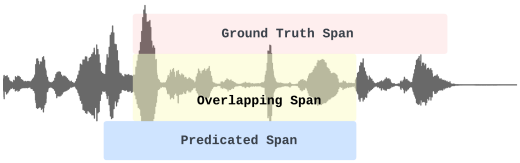


Figure 3: Illustration of ground truth span, predicted span and overlapping span for evaluation.

3.3 Evaluation

To evaluate the performance of our models, we employ two key metrics: Frame-level F1 (FF1) score (Chuang et al., 2020) and Audio Overlapping Score (AOS) (Lee et al., 2018), both of which are commonly used in SQA tasks to assess model accuracy with respect to time-based predictions.

FF1 score is an adaptation of the standard F1 score used in text-based question answering (TQA) tasks. While in TQA, the F1 score is calculated based on token-level matches between predicted and ground-truth answers, in SQA, the answers are temporal segments of audio rather than discrete tokens. Therefore, FF1 measures the precision and recall of frame-level matches between the predicted and actual answer spans, as shown in Figure 3. The equation is as follows

$$Precision = \frac{Overlapping\ Span}{Predicted\ Span}$$
$$Recall = \frac{Overlapping\ Span}{Ground\ Truth\ Span}$$
$$FF1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

AOS on the other hand, provides an additional evaluation by measuring the overlap between the predicted and ground-truth answer spans using the intersection-over-union ratio at the frame level.

	Natural		Lexical		Prosodic	
	Test	Verified-Test	Test	Verified-Test	Test	Verified-Test
	FF1					
Natural	**33.27 ± 1.20**	**36.01 ± 0.86**	26.12 ± 0.59	30.01 ± 1.23	10.04 ± 0.47	10.57 ± 0.87
Lexical	26.72 ± 0.41	29.31 ± 0.71	**32.37 ± 0.12**	**33.42 ± 0.55**	8.94 ± 0.61	8.19 ± 0.66
Prosodic	8.33 ± 0.51	9.90 ± 2.19	6.96 ± 1.25	9.01 ± 1.07	**18.49 ± 0.67**	**18.29 ± 1.14**
	AOS					
Natural	**29.67 ± 1.24**	**31.80 ± 0.78**	22.47 ± 0.68	26.04 ± 0.97	6.36 ± 0.41	6.75 ± 0.82
Lexical	23.13 ± 0.26	26.04 ± 0.97	**28.63 ± 0.02**	**29.30 ± 0.68**	5.95 ± 0.46	5.57 ± 0.32
Prosodic	5.39 ± 0.39	6.79 ± 1.65	4.37 ± 0.94	5.99 ± 0.94	**13.97 ± 0.73**	**13.86 ± 0.98**

Table 2: Results for different training and testing conditions (natural, lexical, and prosodic) on the test and verified test set. Bold diagonal cells indicate results when training and tesing conditions are the same.

The equation can be written as follows

$$AOS = \frac{Overlapping\ Span}{Predicted\ Span \cup Ground\ Truth\ Span}$$

4 Experiments and results

To explore the role of prosodic information in SQA tasks, we design two stages of experiments to answer the research questions stated in Section 1.

4.1 Is prosodic information sufficient for SQA tasks?

First, we train the model separately over our three different data conditions: *natural*, *lexical* and *prosodic* and test on the corresponding conditions. The results are shown in Table 2. We also establish a *chance-level* baseline by generating white noise speech using a normal distribution with the same duration as the documents in the verified-test set in order to simulate the model's performance in the absence of all meaningful information. Evaluating all models across different seeds for generating the white noise speech, the best FF1 and AOS are 6.03 ± 0.16 and 3.29 ± 0.09, respectively. This serves as the default result when random predictions are made, providing a baseline for comparison with the model's performance on actual data.

As expected, when the training and testing conditions are the same, the model performs best under the natural condition, followed by the lexical condition. But interestingly, the prosodic condition also performs reasonably well, far better than the chance level baseline [4].

The model performs the worst when trained and tested on some combination of lexical and prosodic conditions, as this combination has the least information overlap among all the tested configurations. Although the results for natural and lexical conditions are not identical, the relatively close performance suggests that the model relies heavily on lexical information, as both conditions include it. When trained on either natural or lexical condition, the model performs well on the other condition, indicating that the model can generalize effectively between natural and lexical information. However, when tested on the prosodic condition, performance drops significantly. This pattern is mirrored when the model is trained on prosodic information and tested on natural or lexical conditions. These results highlight that, although prosody alone provides meaningful information for SQA from the results, it cannot fully compensate for the absence of lexical content. This ability to generalize only occurs despite a domain mismatch in prosody between lexical and natural conditions, suggesting that the model prioritizes lexical cues over prosodic variations when both are available.

Can the prosodic condition scores stem from leftover lexical information? One possible concern is whether the performance observed in the prosodic condition could be purely influenced by *residual lexical information*. To address this, we select a cut-off frequency of 300Hz, which theoretically removes the majority of energy associated with both vowels and consonants, given that vowel sounds typically lie in the range of 250–2000Hz, and consonants span from 250Hz to as high as 8000Hz. To further investigate, we conduct ad-

[4]The DUAL baseline, using *wav2vec* as the encoder, achieved an FF1 score of 23.1 on the same natural condition for verified-test set, as reported in (Shon et al., 2023). We include their result here not for direct comparison of the models, but to highlight that the results obtained using prosodic information in our experiments are very close to a functional baseline using both prosodic and lexical information, significantly better than random performance.

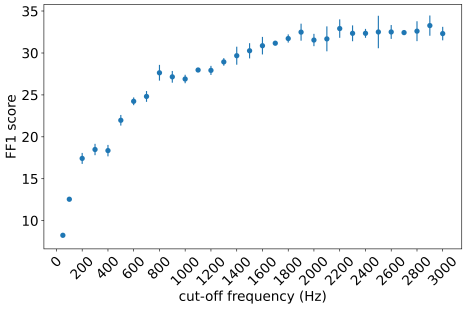


Figure 4: Performance on the test set with different cut-off frequencies

ditional experiments using different cut-off frequencies (from 50 to 3000Hz) to assess the impact of filtering on model performance. The result is presented in Figure 4. We observe there is no significant performance drop when the cut-off frequency is above 1800Hz and between the 200Hz and 400Hz. This suggests that frequencies above 1800Hz are primarily associated with high-frequency sounds such as fricatives and certain consonants, which are less critical for understanding the core content of speech. However, the model's performance gradually decreases as the cut-off frequency is lowered from 1800Hz to 400Hz, indicating that most lexical information is included within this range as it contains many formant frequencies of vowels and essential cues for consonants. The sharp performance drop below 200Hz can be attributed to the loss of crucial prosodic information, especially F0, which plays a key role in intonation, and stress patterns. Therefore, a cut-off frequency between 200Hz and 400Hz offers a good compromise, retaining enough prosodic information while effectively removing most lexical content.

Is the question relevant in the prosodic condition? In the context of SQA, prosodic information, such as intonation, pitch, and pauses, signal important structural and emphatic aspects of speech. These cues can highlight portions of the context that are more likely to contain relevant information. For example, changes in intonation may signal the introduction of key points, while pauses and shifts in pitch can emphasize certain phrases or concepts. As a result, prosodic information can help narrow down the likely locations of the answer within the context, even when lexical information is absent or reduced. However, one concern is whether prosodic information alone can meaningfully contribute to SQA if it can disregard the actual *question*. If prosody merely highlights key parts of the context without connecting them to the question, contribution of the question might be limited. To explore this, we conduct an experiment in which questions and contexts for the verified test set were randomly paired. In this setup, the model's performance dropped significantly, reaching levels similar to those observed when lexical information was present but not prosodic information. This drop in performance indicates that prosodic cues alone are insufficient to fully answer the question, as they do not directly convey the question's relationship to the relevant parts of the context.

	FF1	AOS
Natural	17.09 ± 0.96	14.79 ± 0.89
Lexical	17.26 ± 0.33	14.74 ± 0.25
Prosodic	9.77 ± 0.54	7.05 ± 0.57

Table 3: Results for the same training and testing conditions on the random-paired verified test set.

Nevertheless, it is important to note that the model still performed better than random chance, suggesting that prosodic cues provide some utility even in the absence of a meaningful connection between the question and context. These cues likely highlight segments of the passage that are perceived as more important or emphasized, helping the model identify areas where relevant information might be located. This explains why the model outperforms a chance-level baseline even when the question-context alignment is disrupted. In summary, while prosody alone cannot entirely guide the model to the correct answer, it serves as a helpful supplementary signal that directs attention to key parts of the passage.

4.2 Do SQA models utilize prosodic information when lexical information is also present?

From the results presented in Table 2, we observe that the model tends to prioritize lexical information from the similar performance on both the natural and lexical sets across different configurations. To better understand this behavior, we conduct experiments for the prosodic condition when combining training with portions of 0%, 5%, and 100% of the training sets from other two conditions. We then track the evaluation loss on all conditions throughout the training process. As shown in Figure 5, the evaluation loss trends clearly demonstrate that the model predominantly relies on lexical in-

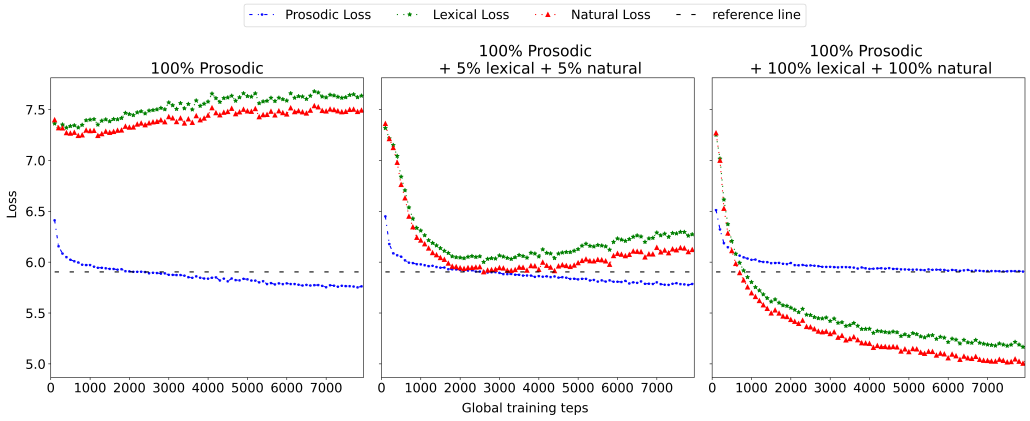


Figure 5: Evaluation loss across different conditions. From left to right, the model is trained on (1) the full prosodic training set, (2) a combination of the full prosodic training set and 5% from both the lexical and natural training sets, and (3) the full training sets for all conditions. The reference line indicates the lowest prosodic loss when the model is trained on all conditions..

formation. When this lexical information is absent during training, the model learns to utilize prosodic information as indicated by the decrease in prosodic evaluation loss. However, even when only 10% of the training data contains lexical information and the overwhelming majority consists of prosodic data, the model quickly learns from the lexical data. This is reflected by the rapid decrease in evaluation loss for both the lexical and natural sets, which soon approach the same level as the prosodic loss. When equal amounts of data from each condition are provided, we observe a more rapid decrease in loss for both the natural and lexical sets, whereas the prosodic set exhibits higher loss than the other two training conditions with less lexical information, as indicated by the reference line. This suggests that, when given access to both lexical and prosodic features, the model primarily uses lexical information, possibly because lexical features offer a more straightforward path to understanding and answering questions based on the content of the speech. In contrast, prosodic cues, though helpful, do not seem to be the model's primary source of information in these settings.

5 Conclusion

Through a series of controlled experiments, we explore the role of prosody in SQA tasks. Our findings demonstrate that while lexical information remains the dominant feature in models trained on both prosodic and lexical data, prosody still provides meaningful complementary cues. In experiments where prosodic information was isolated, the model performed reasonably well, indicating that prosody alone can guide the model toward identifying relevant segments in the context. This underscores the independent value of prosodic features such as intonation, stress, and pauses in guiding the model's understanding of spoken language.

However, when both prosodic and lexical information were available, the model predominantly rely on lexical cues, as they offer a more direct path to understanding the meaning of the speech. Even when the amount of lexical information in the training data was reduced to just 10%, the model continued to prioritize and learn from these features over prosodic cues. This suggests that while prosody is useful, it is often overshadowed by lexical content in tasks where both types of information are present.

Additionally, our analysis of random question-context pairings reveal that prosodic cues alone cannot fully guide the model to the correct answer without considering the relationship between the question and the relevant context. Nevertheless, the model still performed above random chance, suggesting that prosodic information highlights important parts of the context, even when it cannot provide the complete answer.

In conclusion, while prosodic information plays a valuable role in SQA, its contribution is secondary to lexical content when both are present. Future work should focus on developing models that can better integrate prosodic and lexical information to fully leverage the richness of spoken language, especially in scenarios where lexical information is degraded or limited.

6 Limitations

We have ensured full reproducibility of our results by using both an open-source model and original dataset, and providing detailed instructions (including hyperparameters) for replicating our experimental conditions and results. We acknowledge limitations in our work, with the primary challenge being the difficulty in making the prosodic and lexical information fully independent when designing our conditions. Indeed, in the prosodic condition, while we tried to minimize lexical information, there remains the possibility of some residual lexical cues contributing to the model's performance. Moreover, by applying a low-pass filter, we also degrade the quality of the prosodic information, potentially artifically lowering the scores related to prosodic only information. Future work could explore more sophisticated methods of explicitly modelling prosodic features separately from lexical ones.

Another limitation relies in the choice of layer used to extract the representations. While we used one of the deeper layers of the model to extract our discrete units, as it has been suggested that is where semantic information is the strongest (Pasad et al., 2021), it is possible that prosodic information is more heavily encoded in earlier layers. Further exploration of the different representations could bring more light on the role of prosody on SQA.

Furthermore, our study used SLUE-PHASE2, an extractive SQA dataset where answers are specific spans of audio within a passage. This approach limited our investigation to tasks requiring literal comprehension, such as identifying places, names, or dates. While this provides insight into how prosody helps in locating specific information, it would be valuable to extend this research to open-ended SQA tasks, where prosodic information may play a more significant role in guiding models to generate nuanced and contextually appropriate responses.

Finally, future work should explore how prosody influences inferential comprehension, where emotions, thoughts, and intentions are inferred from the speech. In these tasks, prosody could offer important cues that go beyond the lexical content, enriching the model's understanding of more abstract or emotional aspects of the spoken language.

References

S. Ananthakrishnan and S. Narayanan. 2009. Unsupervised adaptation of categorical prosody models for prosody labeling and speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):138–149.

Sankaranarayanan Ananthakrishnan and Shrikanth Narayanan. 2007. Improved speech recognition using acoustic and lexical correlates of pitch accent in a n-best rescoring framework. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pages IV–873–IV–876.

Nicolas Audibert, Francesca Carbone, Maud Champagne-Lavau, Aurélien Said Housseini, and Caterina Petrone. 2023. Evaluation of delexicalization methods for research on emotional speech. In *INTERSPEECH 2023*, pages 2618–2622.

Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *ArXiv*, abs/2006.11477.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Paul Boersma and David Weenink. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.38, retrieved 2 January 2021 http://www.praat.org/.

Gary Buck. 2001. *Assessing Listening*. Cambridge Language Assessment. Cambridge University Press.

Cedric Chan and Jianjing Kuang. 2024. Exploring the accuracy of prosodic encodings in state-of-the-art text-to-speech models. In *Speech Prosody 2024*, pages 27–31.

K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, Sung-Suk Kim, J. Cole, and Jeung-Yoon Choi. 2006. Prosody dependent speech recognition on radio news corpus of american english. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):232–245.

Ken Chen, Sarah Borys, Mark Hasegawa-Johnson, and Jennifer Cole. 2003. Prosody dependent speech recognition with explicit duration modelling at intonational phrase boundaries. In *INTERSPEECH*, pages 393–396.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yung-Sung Chuang, Chi-Liang Liu, Hung yi Lee, and Lin shan Lee. 2020. SpeechBERT: An Audio-and-Text Jointly Learned Language Model for End-to-End Spoken Question Answering. In *Proc. Interspeech 2020*, pages 4168–4172.

Robert A. J. Clark, Hanna Silén, Tom Kenter, and Ralph Leith. 2019. Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs. *ArXiv*, abs/1909.03965.

Adriana Colatosti, Ignacio Gil, Antonio Morant-Ventura, Emilia Monteagudo, Lucía Aranda, and Jaime Marco. 2024. Normal hearing and verbal discrimination in real sounds environments. *Acta otorrinolaringologica espanola*.

Maureen de Seyssel, Marvin Lavechin, Yossi Adi, Emmanuel Dupoux, and Guillaume Wisniewski. 2022. Probing phoneme, language and speaker information in unsupervised speech representations. In *Interspeech 2022*, pages 1402–1406.

Erik Ekstedt and Gabriel Skantze. 2022. How much does prosody help turn-taking? investigations using voice activity projection models. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 541–551, Edinburgh, UK. Association for Computational Linguistics.

John Field. 2005. Intelligibility and the listener: The role of lexical stress. *TESOL quarterly*, 39(3):399–423.

Jean-Philippe Goldman, Tea Pršir, George Christodoulides, Anne-Catherine Simon, and Antoine Auchlin. 2014. Phonogenre identification: A perceptual experiment with 8 delexicalised speaking styles. *Nouveaux cahiers de linguistique française*, pages 51–62.

Camille Guinaudeau and Julia Hirschberg. 2011. Accounting for prosodic information to improve asr-based topic tracking for tv broadcast news. In *Interspeech 2011*, pages 1401–1404.

Mark Hasegawa-Johnson, Ken Chen, Jennifer Cole, Sarah Borys, Sung-Suk Kim, Aaron Cohen, Tong Zhang, Jeung-Yoon Choi, Heejin Kim, and Tae-Jin Yoon. 2005. Simultaneous recognition of words and prosody in the boston university radio speech corpus. *Speech Communication*, 46:418–439.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Jui-Ting Huang, Po-Sen Huang, Yoonsook Mo, Mark Hasegawa-Johnson, and Jennifer Cole. 2010. Prosody-dependent acoustic modeling using variable-parameter hidden markov models. In *Speech Prosody 2010*, page paper 623.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

H Kagan Keskin, Gökhan Ari, and Muhammet Bastug. 2019. Role of prosodic reading in listening comprehension. *International Journal of Education and Literacy Studies*, 7(1):59–65.

Arne Köhn, Florian Stegen, and Timo Baumann. 2016. Mining the spoken wikipedia for speech data and beyond. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *Proc. Interspeech 2018*, pages 3459–3463.

Guan-Ting Lin, Yung-Sung Chuang, Ho-Lam Chung, Shu wen Yang, Hsuan-Jui Chen, Shuyan Annie Dong, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Lin shan Lee. 2022. DUAL: Discrete Spoken Unit Adaptive Learning for Textless Spoken Question Answering. In *Proc. Interspeech 2022*, pages 5165–5169.

Guan-Ting Lin, Chi-Luen Feng, Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li, Hung-yi Lee, and Nigel G. Ward. 2023. On the utility of self-supervised models for prosody-related tasks. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1104–1111.

Oli Danyi Liu, Hao Tang, and Sharon Goldwater. 2023. Self-supervised predictive coding models encode speaker and phonetic information in orthogonal subspaces. In *Proc. INTERSPEECH 2023*, Interspeech, pages 2968–2972. International Speech Communication Association. Funding Information: This work was supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology Language Sciences. Publisher Copyright: © 2023 International Speech Communication Association. All rights reserved.; Interspeech 2023 ; Conference date: 20-08-2023 Through 24-08-2023.

Iker Luengo, Eva Navas, Inma Hernáez, and Jon Sánchez. 2005. Automatic emotion recognition using prosodic parameters. In *Interspeech*.

Jacques Mehler, Peter Jusczyk, Ghislaine Lambertz, Nilofar Halsted, Josiane Bertoncini, and Claudine Amiel-Tison. 1988. A precursor of language acquisition in young infants. *Cognition*, 29(2):143–178.

Mukhtar Mohamed, Oli Danyi Liu, Hao Tang, and Sharon Goldwater. 2024. Orthogonality and isotropy of speaker and phonetic information in

self-supervised speech representations. In *INTERSPEECH 2024*. ISCA. The 25th Interspeech Conference, Interspeech 2024 ; Conference date: 01-09-2024 Through 05-09-2024.

Navid Naderi and Babak Nasersharif. 2023. Cross corpus speech emotion recognition using transfer learning and attention-based fusion of wav2vec2 and prosody features. *Knowledge-Based Systems*, 277:110814.

Oliver Niebuhr, Alexander Brem, Jan Michalsky, and Jana Neitsch. 2020. What makes business speakers sound charismatic? a contrastive acoustic-melodic analysis of steve jobs and mark zuckerberg. *Cadernos de Linguistica e Teoria da Literatura*, 1(1).

Mari Ostendorf, Izhak Shafran, and Rebecca Bates. 2003. Prosody models for conversational speech recognition.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921.

Waris Quamer and Ricardo Gutierrez-Osuna. 2024. Disentangling segmental and prosodic factors to non-native speech comprehensibility. *arXiv preprint arXiv:2408.10997*.

Shangeth Rajaa. 2023. Improving end-to-end slu performance with prosodic attention and distillation. In *INTERSPEECH 2023*, pages 1114–1118.

Mirco Ravanelli, Titouan Parcollet, Aku Rouhe, Peter Plantinga, Elena Rastorgueva, Loren Lugosch, Nauman Dawalatabad, Chou Ju-Chieh, Abdel Heba, Francois Grondin, William Aris, Chien-Feng Liao, Samuele Cornell, Sung-Lin Yeh, Hwidong Na, Yan Gao, Szu-Wei Fu, Cem Subakan, Renato De Mori, and Yoshua Bengio. 2021. Speechbrain.

urlhttps://github.com/speechbrain/speechbrain.

Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan Sharma, Wei-Lun Wu, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. 2023. SLUE phase-2: A benchmark suite of diverse spoken language understanding tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8906–8937, Toronto, Canada. Association for Computational Linguistics.

Suwon Shon, Kwangyoun Kim, Yi-Te Hsu, Prashant Sridhar, Shinji Watanabe, and Karen Livescu. 2024. Discreteslu: A large language model with self-supervised discrete speech units for spoken language understanding. *Preprint*, arXiv:2406.09345.

Amitay Sicherman and Yossi Adi. 2023. Analysing discrete self supervised speech representation for spoken language modeling. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 9, page 1–5. IEEE.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR.

Aarne Talman, Antti Suni, Hande Celikkanat, Sofoklis Kakouros, Jörg Tiedemann, and Martti Vainio. 2019. Predicting prosodic prominence from text with pre-trained contextualized word representations. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 281–290, Turku, Finland. Linköping University Electronic Press.

Trang Tran, Shubham Toshniwal, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Mari Ostendorf. 2018. Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 69–81, New Orleans, Louisiana. Association for Computational Linguistics.

Trang Tran, Jiahong Yuan, Yang Liu, and Mari Ostendorf. 2019. On the role of style in parsing speech with neural models. In *Interspeech*.

Merve Ünlü Menevşe, Yusufcan Manav, Ebru Arisoy, and Arzucan Özgür. 2022. A framework for automatic generation of spoken question-answering data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4659–4666, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kai Wei, Dillon Knox, Martin Radfar, Thanh Tran, Markus Muller, Grant P. Strimel, Nathan Susanj, Athanasios Mouchtaris, and Maurizio Omologo. 2022. A neural prosody encoder for end-ro-end dialogue act classification. *Preprint*, arXiv:2205.05590.

Mirjam Wester, Oliver Watts, and Gustav Eje Henter. 2016. Evaluating comprehension of natural and synthetic conversational speech. In *Proc. Speech Prosody 2016*, pages 766–770.

Stephen J Winters and David B Pisoni. 2004. Perception and comprehension of synthetic speech. *Research on spoken language processing report*, 26:95–138.

Yijing Wu, SaiKrishna Rallabandi, Ravisutha Srinivasamurthy, Parag Pravin Dakle, Alolika Gon, and Preethi Raghavan. 2024. Heysquad: A spoken question answering dataset. *Preprint*, arXiv:2304.13689.