Is Synthetic Data Sufficient for Extractive Spoken Question Answering?

Anonymous ACL submission

Abstract

Spoken language understanding is essential for extracting meaning from spoken language, particularly in low- or zero-resource language settings relying on speech in the absence of text data. This work investigates the effectiveness of using synthetic speech data in Spoken Question Answering (SQA). By manipulating prosody in human-read test sets, as well as proposing a new SQA dataset for fine-tuning, we demonstrate that models trained solely on synthetic speech can utilise prosodic cues. Moreover, synthetic speech fine-tuned models outperform those fine-tuned on natural speech, even with the same or restricted lexical information. Our findings suggest that current text-to-speech systems can simulate sufficient prosody for SQA models, and that the contribution from natural prosody is limited within the current textless SQA framework.

1 Introduction

Spoken language understanding (SLU) aims to extract meaningful information from spoken language input. Unlike natural language understanding (NLU), which relies on text, SLU is particularly valuable for so-called low-resource languages with limited text data; speech serves as the primary linguistic signal (Bloomfield, 1933) and there is generally an abundance of speech data to harvest for spoken languages. Traditional SLU systems, however, consist of two separate components: an automatic speech recognition (ASR) model and an NLU model, with only the NLU model fine-tuned for the downstream task. This cascaded approach is easy to implement, as the models can be trained separately with external datasets. However, errors from ASR propagate to the NLU model, significantly impacting performance due to ill-formed inputs to the NLU model. Recently, there have been efforts to bypass the explicit transcription step by using end-to-end models (Chuang et al., 2020) or discrete units as pseudo-text (Lin et al., 2022). Training such models requires task-oriented speech datasets, that typically would only have been created for text previously. Since most SLU tasks build on corresponding NLU datasets and collecting annotated audio recordings is labor-intensive, applying Text-to-Speech (TTS) techniques to generate large training datasets is common (Lee et al., 2018; Lin et al., 2022; Ünlü Menevşe et al., 2022).

However, research has shown people perceive natural speech differently from synthetic speech: listeners often have more difficulty understanding synthetic speech due to limited acoustic-phonetic cues and the lack of natural variability (Winters and Pisoni, 2004; Wester et al., 2016). Additionally, modeling prosody has been shown to benefit various tasks, from segmentation-related tasks to meta-information and paralinguistics tasks (Tran et al., 2018; Cho et al., 2022). Therefore, we are interested in exploring the role of prosody in SLU.

In this paper, we focus on spoken question answering (SQA) as our main task, which predicts the start and end points of an answer span from its input. We first modify the prosodies of the audios in the human-read test set and demonstrate that a model trained solely on synthetic speech can still leverage prosodic cues to answer questions. We then explore whether natural speech is necessary for training a SQA system by comparing systems fine-tuned on natural and synthetic speech with the same lexical information. Since there are no existing natural SQA datasets [1] for English, we propose a novel data extension approach. Our findings reveal that synthetic speech fine-tuned systems not only perform competitively with natural speech fine-tuned systems, but can also maintain competitive performance even when lexical information within the speech data is severely restricted.

[1]We are only interested in factoid SQA datasets with the context from human speech. Other related SQA datasets are discussed in Section 2

2 Related work

Linguistics research has long confirmed that prosody can aid across a variety of tasks, from disambiguating homographs to conveying a speaker's sentiments (Tran, 2020). While the role of prosody has been studied for decades in human speech perception and production, its use in spoken language technology has been limited due to the challenges in computational modeling (Cutler et al., 1997; Tran, 2020). Since recovering prosody from text is difficult (Talman et al., 2019), recent work has focused on incorporating coarse acoustic features into ASR outputs for downstream SLU tasks (Tran et al., 2018; Tran, 2020; Tran and Ostendorf, 2021; Cho et al., 2022). Additionally, perturbing the input audio to omit certain sources of prosodic information has been explored to investigate whether models can learn to pick up prosodic cues (Ekstedt and Skantze, 2022). In this work, we apply a similar idea to investigate if the SQA models trained on synthetic speech can pick up any prosodic features.

SQA was used to refer to work on spoken documents (manual or ASR transcripts) rather than audio files (Umbert, 2012). Recently, it has evolved to resemble extractive textual QA tasks, involving *a spoken context, a question, and an answer within the context*, which is also our focus here.

Several datasets have been developed for this task. Lee et al. (2018) introduced Spoken SQuAD, a dataset with spoken contexts and textual questions, using TTS on the SQuAD dataset (Rajpurkar et al., 2016). Lin et al. (2022) extended Spoken SQuAD by applying TTS to questions and providing a benchmark corpus read by humans: Natural Multi-speakers Spoken Question Answering dataset (NMSQA). Ünlü Menevşe et al. (2022) proposed a framework for generating SQA data by fine-tuning a language model to generate questions and answers, followed by TTS for audio. We focus only on ***factoid*** *QA* in this paper but there are other research directions including multi-turn conversational SQA datasets (You et al., 2022), QA from meeting transcripts and interviews (Archiki Prasad and Bansal, 2023; Shankar et al., 2024), and SLUE-SQA-5 retrieving only relevant but real natural speech from Spoken Wikipedia (Shon et al., 2023).

3 SQA Data and Model

Most English SQA datasets are generated by TTS systems, with limited human-read samples like NMSQA's testset, making detailed analysis or fine-tuning challenging. Therefore, we propose a novel dataset extension approach using natural speech as context passages. We select the Boston University Radio News Corpus (BURNC) as our natural speech source due to its rich prosody and mix of formal and communicative speech (Ostendorf et al., 1996) and employ DUAL (Lin et al., 2022) as our SQA model.

3.1 BURNC_QA dataset

We first segment transcripts into utterances and generate named entities (NE) as answers using the Flair toolkit (Akbik et al., 2019). This approach, compared to generating questions directly from the transcription using a language model as in (Ünlü Menevşe et al., 2022), offers better efficiency and accuracy. It ensures semantically relevant questions while providing greater control over the generation process.

The next step involves generating questions corresponding to the answer and the utterance. We fine-tune the FLAN T5-BASE model (Chung et al., 2022) on a question generation task with the SQuAD dataset, by concatenating the context and answers as input and use questions as output[2]. During the experiments, we observed that the model may generate questions containing the answers themselves or irrelevant information, and the examples are shown in Appendix A.1. We filter out self-answered questions and check if the NE extracted from the questions exists in the context.

Following (Lin et al., 2022), we apply Longformer[3] as our SOTA textual QA system on the generated pairs, achieving results of exact_match: 78.18, F1: 88.36. Assuming incorrect answers (under exact_match) may result from prior generation errors, we also filter them out, leaving in a total of 7,318 QA pairs. To align with the SQuAD dataset structure, we use *entire BURNC paragraphs*, as context instead of individual utterances.

After obtaining textual QA pairs, we utilise SPEECHT5_TTS (Ao et al., 2022) to generate synthetic speech for the questions. To ensure consistency in alignment across different datasets, we realign all datasets using the Montreal Forced Aligner framework (McAuliffe et al., 2017).

There are 7 speakers in the BURNC dataset. We assign all audios from speaker **m3b** to the test set

[2] eval_rouge1: 0.51, eval_rouge2: 0.28, eval_rougeL: 0.47, and eval_rougeLsum: 0.47 on the SQuAD testset

[3] https://huggingface.co/valhalla/longformer-base-4096-finetuned-squadv1

test set	original		shiftpitch		flatpitch		flatintensity		avg		lowpass	
	FF1	AOS	FF1	AOS	FF1	AOS	FF1	AOS	FF1	AOS	FF1	AOS
NMSQA	61.08	54.44	61.55	53.82	55.54	48.86	52.31	46.31	29.60	24.35	27.83	23.85
BURNC	59.79	52.25	59.29	51.84	59.67	52.16	58.44	51.12	58.72	51.29	32.96	27.12

Table 1: Performance of DUAL on modified NMSQA and BUNRC_QA testsets.

to also assess the ability of the model on unseen speakers. The remaining data is randomly [4] split so that eventually we obtain 7:2:1 for training, development and test.

3.2 DUAL framework and evaluation metrics

The DUAL framework comprises a Speech Content Encoder (SCE) and a Pre-trained Language Model (PLM). Unlike conventional cascade models, DUAL does not rely on ASR transcripts, thus avoiding ASR error propagation. Our SCE uses Hubert (Hsu et al., 2021), a self-supervised pre-trained model which also has demonstrated effectiveness for prosody-related tasks (Lin et al., 2023), to encode representations directly from raw waveforms. K-means clustering is then applied to transform representations into discrete units, which are then fed into the PLM. The PLM predicts the span of the answer in the context passage by identifying the start and end positions.

Frame-level F1 (FF1) score (Chuang et al., 2020) and Audio Overlapping Score (AOS) (Lee et al., 2018) are used as the evaluation metrics. FF1 score is similar to F1 score in textual QA, but are calculated on frames instead of tokens. AOS measures the overlap between predicted and ground-truth spans with the intersection-over-union ration on frames. The detailed illustration is Appendix A.2.

4 Prosodic variation

Method. To investigate if the model has learned prosodic cues, we modify the audio prosodies in both the NMSQA human-read subset and proposed BURNC_QA testset. Inspired by Ekstedt and Skantze (2022), we explore the following prosodic details using Parselmouth (Jadoul et al., 2018; Boersma and Weenink, 2021). These modifications are also illustrated in Appendix A.3.

Pitch flatten: Flattens F0 to the average value of each utterance.

Pitch shift: Shifts the pitch by 90% of its original value for each utterance.

Intensity flatten: Flattens intensity to the average value of each utterance.

Low pass filter: Removes high-frequency phonetic information using a cutoff frequency of 800Hz.

Average phone duration: Scales each phone to its average duration obtained from the corpus, check Appendix A.4 for their values.

Experiments and results. We evaluate the DUAL checkpoint released by (Lin et al., 2022), trained exclusively on synthetic speech, on both NMSQA and BURN_QA. From the results in Table 1, we observe the performance drops when prosodic features are modified, except for the FF1 score with shiftpitch on NMSQA. That indicates the utilisation of prosodic cues despite the model's lack of exposure to natural speech.

Similar to Ekstedt and Skantze (2022), we find DUAL is most sensitive to the low-pass transform, which preserves intensity and F0 contour while removing most high-frequency phonetic information. This underscores the significant impact of phonetic details in SQA tasks. The average phone duration transform impacts differ between the two datasets, possibly because the average duration is calculated from 11 hours of BURNC data compared to just 2 hours of NMSQA data. Additionally, BURNC audios are read by professional news announcers, resulting in less disfluencies and prosodic errors (Ostendorf et al., 1996)

Between shiftpitch and flatpitch, we observe that DUAL performs better with shiftpitch on NMSQA but better with flatpitch on BURNC, illustrating the complex nature of prosody. Variations and patterns learned from synthetic data may enhance the model by introducing contextual cues, yet they may also introduce noise and ambiguity compared to natural prosody patterns. Interestingly, our results indicate that flattening intensity has a greater impact than pitch variation, although pitch typically is considered more crucial for language understanding, as it conveys nuances such as intonation and stress.

Therefore, our study shows that synthetic speech can effectively simulate reasonable prosodies, and training models on such data enables effective utilisation of prosodic cues for SQA tasks.

[4] In BURNC, some news stories are read by multiple speakers. We ensure there is no overlap of identical stories between the test and other sets.

5 Natural-BURNC vs Synthetic-BURNC

Method. To investigate the necessity of natural speech in training a SQA system, we synthetically generate two variations of the BURNC_QA training dataset with SpeechBrain (Ravanelli et al., 2021): **UTT_TTS** where each audio is generated using the speaker embedding extracted from the corresponding natural audio, and **SPK_TTS** where utterance embeddings from all utterances by the same speaker are mean-normalised to obtain a single embedding per speaker. This latter approach allows us to generate 6 audios (still excluding speaker *m3b* from 7 speakers in BURNC) for every audio.

Experiments and Results. We finetune [5] DUAL on all three datasets. FF1 results are illustrated in the first three bars on each modified testset in Figure 1, with detailed numbers in Appendix A.5. The model trained on SPK_TTS data performs the best, and the worst when trained on UTT_TTS, even though both contain identical lexical information. This suggests that the quantity of speech data plays a more important role than the prosody embedded within it, especially under the assumption that TTS may not fully capture natural prosody.

To further investigate the influence of the lexical information, e.g. number of unique QA texts, we randomly incorporate 6 additional speaker embeddings extracted from the CMU ARCTIC dataset (Kominek and Black, 2004) into our experiments. Following the same generation process, we now have 12 audios (6 generated using speaker embeddings from BURNC as described before and 6 from CMU ARCTIC) for each utterance. We first sample from 50% to 100% of unique textual QAs from BURNC_QA, and then randomly select synthetic speech accordingly, ensuring we ultimately have the same amount of speech data as SPK_TTS. The FF1 results are also shown in the bars to the right side on each modified testset Figure 1, and the detailed numbers are also presented in Appendix A.5. We observe that even with only 50% of the unique QA pairs, when the speech data is increased six-fold compared to the quantity of natural speech, their scores are already on par. Generally, within the same amount of speech data, exposure to more lexical information during training improves the final results. This suggests that as the PLM component in DUAL has already learned how to answer textual QA effectively, it requires only a limited amount of data to adapt to similar tasks on discrete units. Furthermore, SPK_TTS, which uses speaker embeddings exclusively from BURNC, performs significantly better than using external speaker embeddings that do not match the testset domain, even when they contain identical lexical information.

[5] Preliminary experiments indicated that training the DUAL framework from scratch requires at least 150 hours. Finetuning is chosen due to the limited 11 hours of data available.

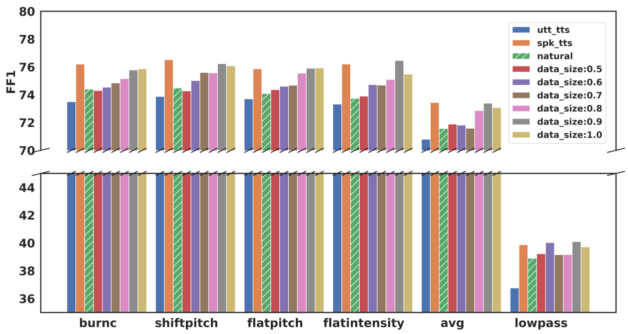

Figure 1: FF1 of DUAL finetuned (with 10 epoches, lr 1e-6, batch of 4 on 4GPUs) on natural and different synthetic data on modified BURNC. The y-axis is broken for better viewing of the difference between models.

6 Conclusion

Synthetic datasets are commonly used to train SQA systems, yet their effectiveness compared to natural speech remains unclear. In this work, we first demonstrated that models trained solely on synthetic data can still capture prosodic features by showing performance changes when modifying these features on human-read test sets. We then compare models fine-tuned on natural and synthetic datasets and find that the quantity of speech data is more crucial than the embedded prosody. Synthetic speech fine-tuned systems achieve similar results using only half the lexical information of natural speech by augmenting the same text.

Our findings indicate that current TTS systems can simulate sufficient prosody for SQA models to utilise prosodic cues and the use of discrete units carry enough lexical information, enabling language models to adapt efficiently to new domains with limited data. Thus, while building realistic spoken QA datasets is important, simply collecting speech data without explicit instructions may not significantly benefit model training, at least for factoid SQA tasks. Therefore, the answer to our title is affirmative, given that the PLM-like component in the current textless SQA framework can effectively answer factoid questions, limiting the contribution of natural prosody.

7 Limitation

This study primarily focuses on prosodic differences in context passages, overlooking their presence in questions, which also directly influences question intent. Moreover, only named entities are considered as answers, ensuring specificity and relevance of QA pairs but limiting question scope and depth. We employ SOTA textual QA for filtering potentially incorrect QA pairs, which might exclude those answerable by SQA systems instead of textual QA systems. Additionally, our study is limited to English datasets, a rich-resource language with more advanced TTS systems compared to other languages. Finally, the factoid question types examined may diminish the relevance of prosody compared to communicative QA types.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738.

Seunghyun Yoon Hanieh Deilamsalehy Franck Dernoncourt Archiki Prasad, Trung Bui and Mohit Bansal. 2023. Meetingqa: Extractive question-answering on meeting transcripts. *ACL*.

Leonard Bloomfield. 1933. *Language*. Holt, Rinehart and Winstons, New York.

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/`.

Yeonjin Cho, Sara Ng, Trang Tran, and Mari Ostendorf. 2022. Leveraging Prosody for Punctuation Prediction of Spontaneous Speech. In *Proc. Interspeech 2022*, pages 555–559.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

Anne Cutler, Delphine Dahan, and Wilma van Donselaar. 1997. Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, 40(2):141–201.

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.

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. *CoRR*, abs/2106.07447.

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

John Kominek and Alan W. Black. 2004. The cmu arctic speech databases. In *Speech Synthesis Workshop*.

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.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Mari Ostendorf, Patti Price, and Stefanie Shattuck-Hufnagel. 1996. Boston University Radio Speech Corpus.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624. ArXiv:2106.04624.

Natarajan Balaji Shankar, Alexander Johnson, Christina Chance, Hariram Veeramani, and Abeer Alwan. 2024. Coraal qa: A dataset and framework for open domain spontaneous speech question answering from long audio files. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13371–13375.

Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan S 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.

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. 2020. *Neural models for integrating prosody in spoken language understanding*. University of Washington.

Trang Tran and Mari Ostendorf. 2021. Assessing the use of prosody in constituency parsing of imperfect transcripts. *ArXiv*, abs/2106.07794.

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.

Pere R Comas Umbert. 2012. *Factoid question answering for spoken documents*. Ph.D. thesis, Universitat Politècnica de Catalunya. Departament de Llenguatges i Sistemes ....

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.

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.

Chenyu You, Nuo Chen, Fenglin Liu, Shen Ge, Xian Wu, and Yuexian Zou. 2022. End-to-end spoken conversational question answering: Task, dataset and model. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1219–1232, Seattle, United States. Association for Computational Linguistics.

A Appendix

A.1 Examples of incorrect generated questions

Example 1 : *Self-contained question*

Context/utterance : *He will be in London at five*

Answer : *London*

Incorrect question : *Where will he be in **London** at five?*

Example 2 : *Irrelevant information*

Context/utterance : *He will be in London at five*

Answer : *at five*

Incorrect question : *At what time will **Schwarzenegger** be in London?*

A.2 Evaluation metrics

A.2.1 Frame-level F1

$$Precision = \frac{\textit{Overlapping Span}}{\textit{Predicted Span}}$$

$$Recall = \frac{\textit{Overlapping Span}}{\textit{Ground Truth Span}}$$

$$FF1 = \frac{2 \times \textit{Precision} \times \textit{Recall}}{\textit{Precision} + \textit{Recall}}$$

A.2.2 Audio Overlap Score

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

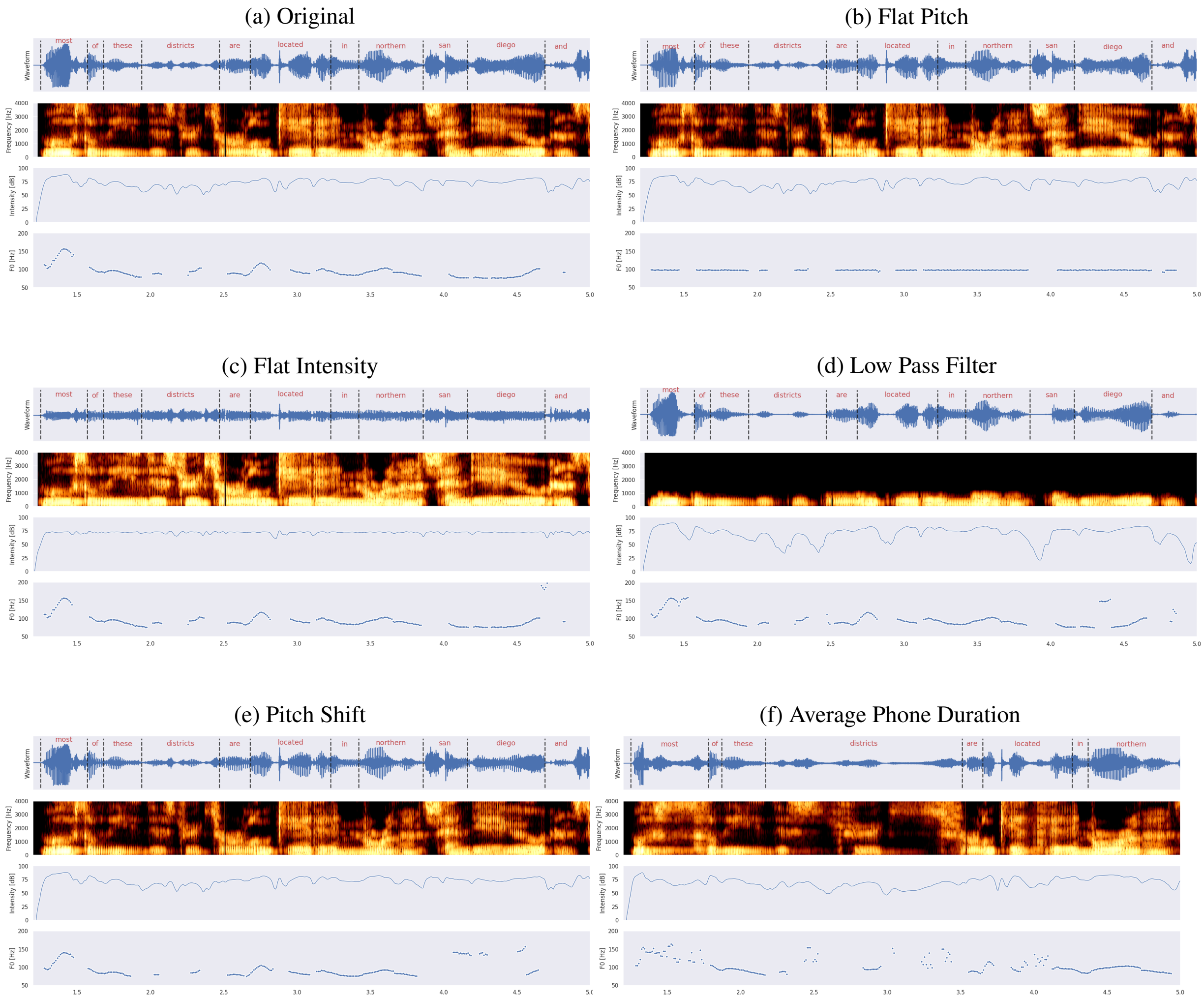

Figure 2: Waveforms, mel-spectrograms, intensity contours and F0 contours for an example audio with its modified versions.

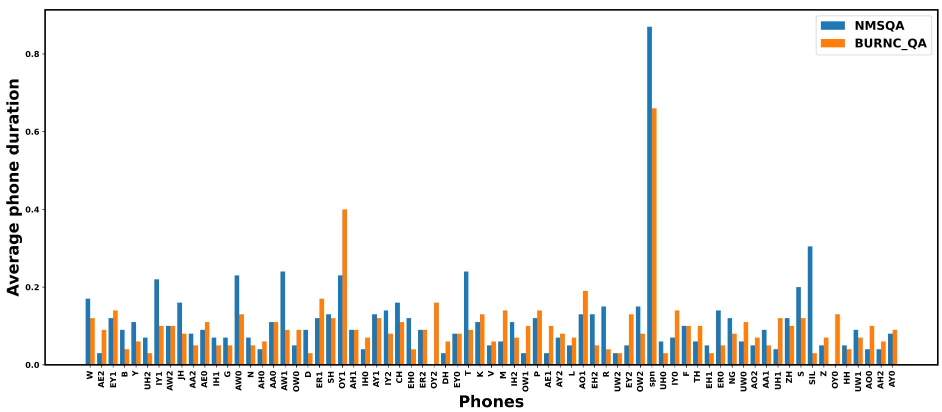

Figure 3: Average phone duration in NMSQA and BURNC

A.3 Example of prosodic modification

Figure 2 presents the change in waveform, mel-spectrograms, intensity contours and F0 contours when modifying different prosodic information on the utterance ... *most of these districts are located in northen san diego and ....*

Models	original		shiftpitch		flatpitch		flatintensity		avg		lowpass	
	FF1	AOS	FF1	AOS	FF1	AOS	FF1	AOS	FF1	AOS	FF1	AOS
Natural	74.40	67.63	74.49	67.90	74.08	67.34	73.74	66.84	71.57	64.63	38.90	33.49
UTT_TTS	73.49	67.01	73.87	67.49	73.70	67.17	73.33	66.85	70.79	64.12	36.76	31.41
SPK_TTS	76.20	69.80	76.52	70.34	75.86	69.55	76.20	69.76	73.44	66.85	39.87	34.45

Table 2: Performance of DUAL fine-tuned on natural, UTT_TTS and SPK_TTS on modified BUNRC_QA testset.

Data size	original		shiftpitch		flatpitch		flatintensity		avg		lowpass	
	FF1	AOS	FF1	AOS	FF1	AOS	FF1	AOS	FF1	AOS	FF1	AOS
0.5	74.30	67.83	74.27	68.07	74.36	68.03	73.90	67.53	71.89	65.46	39.22	33.62
0.6	74.54	68.24	75.02	68.77	74.60	68.22	74.72	68.34	71.81	65.26	40.03	34.47
0.7	74.85	68.46	75.60	69.35	74.68	68.36	74.69	68.25	71.59	65.08	39.15	33.76
0.8	75.16	68.80	75.57	69.35	75.55	69.20	75.10	68.71	72.86	66.31	39.17	33.84
0.9	75.78	69.49	76.23	70.12	75.90	69.50	76.46	70.13	73.39	66.92	40.10	34.59
1.0	75.87	69.54	76.08	69.87	75.93	69.55	75.48	69.03	73.08	66.53	39.72	34.21

Table 3: Performance of DUAL fine-tuned on TTS of different data sizes on modified BUNRC_QA testset.

A.4 Phone duration distribution

Figure 3 illustrates the final phone duration we used in the data perturbation.

A.5 Results

Tabel 2 and 3 presents the detailed results of DUAL fine-tuned on different datasets.