Review
Towards Multi-Lingual Audio Question Answering

Kerner Tobias

Ingolstadt, Germany

tok6578@thi.de

Authors of the reviewed paper: Swarup Ranjan Behera, Pailla Balakrishna Reddy, Achyut Mani Tripathi, Megavath Bharadwaj Rathod, Tejesh Karavadi

Abstract

The Paper "Towards Multi-Lingual Audio Question Answering" introduces the mClothoAQA Dataset, building upon the ClothoAQA Dataset to explore multi-lingual Audio Question Answering. With the assistance of Googles machine translation API, the team translated ClothoAQA to seven languages, resulting in a dataset featuring 1991 audio files each presenting four yes/no and two single-word questions. The models architecture consists of Bidirectional LSTM, pre-trained OpenL3 audio embeddings, and FastText text embeddings for feature extraction. Despite claiming "good performance across eight languages", results especially for the binary classifier suggest performance just slightly better than random chance. The studys optimistic conclusions contrast the models moderate accuracy, calling for cautious interpretation of the results.

Index Terms: Machine Translation, Multi-Lingual Audio Question Answering, Audio Dataset

1. How confident are you in your evaluation of this paper

As a third-semester student specializing in artificial intelligence, my exposure to the field, is still limited. While i am familiar with basic principles like supervised learning, training language models and popular tools like TensorFlow and PyTorch, my experience related to natural language processing is, while being of great interest to me, still lacking.

Expertise in the field of the paper one is reviewing is of great importance, since it directly affects the quality and reliability of the review. Expert reviewers may have more in-depth knowledge of the topic at hand and can point out mistakes in the paper that other would have missed, or notice even small missed opportunities. They can provide better feedback and suggestions to improve the paper. It can also be assumed that experienced reviewers are more familiar with the ethical standards in their field, being able to identify potential ethical issues, conflicts of interest or data manipulation in the paper and therefore making sure of the integrity of the provided research.

In my academical and personal studies so far, i worked with small-scale data augmentation and supervised learning for value prediction in financial markets given historical data; creating a token encoder-decoder from scratch and training a small language model in torch with scraped and preprocessed web data; as well as creating n-Gram models with KenLM. I also got a brief introduction to speech recognition in an optional subject at Technische Hochschule Ingolstadt, although missing prior knowledge made it hard to follow at some points.

In this area, which this paper is focusing on, i have yet to learn a lot, so there might be some mistakes in my review.

For this reason, i am not confident in my review and it should be read with a grain of salt. However, due to my perspective as a student, i hope to make a contribution on this paper by clarifying parts that may be harder to understand than necessary and could be elaborated better in the paper.

Rating: 1

2. Importance/Relevance

The paper presents an advancement in the realm of Audio Question Answering (AQA), particularly in multi-linguality which is indispensable considering the increasing demand for accessibility tools. Individuals with hearing impairments can greatly benefit from developments in AQA as it can potentially enhance their interaction with the environment. Improving communication in various languages through AQA can reduce communication gaps and enable greater global understanding and cooperation. Non-English speaking communities and those speaking indigenous or regional languages could connect with more people worldwide, enhancing cultural exchange and collaboration on a global scale.

Research by Gretchen Stevens et al.[1] in Figure 1-3 shows, how hearing impairment is a global problem and not limited to english speaking regions.

This papers exploration into solutions for multi-lingual AQA not progresses research in this field but also enables more inclusive user experiences. The significance of this work extends to regions other than English speaking ones, creating technological opportunities in countries with various local languages but less developed technological infrastructures.

By introducing the mClothoAQA dataset in eight different languages, the authors create the possibility for more inclusivity. This can prove exceptionally valuable in educational settings where AQA models could support language learning and comprehension. It allows students to engage with a larger set of linguistic resources, hearing and questioning in multiple tongues, which improve language education.

However, the paper appears to evade the question of how these models and datasets could be integrated into real-world applications. This integration is especially important for languages that are underrepresented in technological applications, often due to commercial unviability. Language inclusivity is a key factor in the enhancement of user experience and designing human-computer interfaces.

Incorporating these technological advancements into daily use does not get elaborated in the paper. While the paper does

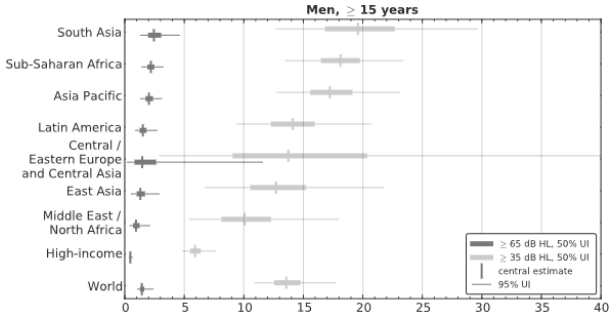

Figure 1: *Occurence of hearing impairment for men*

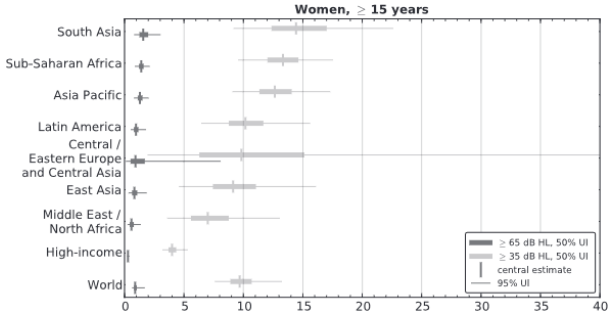

Figure 2: *Occurence of hearing impairment for women*

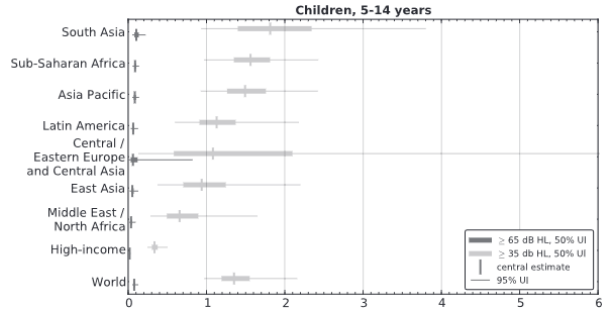

Figure 3: *Occurence of hearing impairment for children*

a good job at introducing multi-linguality to AQA, it does not address scalability beyond the eight languages in the dataset. Exploration into further widening the linguistic scope and methods for seamless integration into systems could be investigated in future research.

In the current world, the importance of speech recognition and AQA is high. With the growing world-wide digital connection, the ability to interact seamlessly with technologies through natural language is becoming increasingly important. With acceleration of adopting advanced technology into daily life, advancements in AQA could be an important step in improving human-machine interfaces.

Rating: 4

3. Novelty/Originality

The authors of the paper introduce a new concept to the field of Audio Question Answering (AQA) with their multilingual dataset, mClothoAQA. While the use of machine translation is not a novel approach, as machine translations have been utilized in numerous applications for converting text across languages, the authors application of this approach uniquely expands the ClothoAQA [2] dataset into multiple languages. The creation of a multilingual dataset for AQA using machine translation is an innovation considering the previous focus had mostly been limited to english datasets.

This expansion addresses a notable gap and shows commitment to diversifying linguistic representation. The mClothoAQA datasets coverage of eight diverse languages advances AQA research into a more inclusive direction that is no longer limited to english speakers. This enhancement potentially opens up new research possibilities and increases the applicability of AQA systems across different language communities, which is especially critical for a global audience.

This is an important step towards making AQA systems more accessible globally. However, one must not overlook the technological milestone represented by the baseline model shown in the paper. This offers a reference for the development of future multi-lingual AQA systems. There is still concern about the impact the machine translation has on the quality of the translations, even though it is claimed that all translations are validated by humans. The reliance on machine translation necessitates further methodological refinements to enhance the models reliability and to mitigate any biases that language models may introduce.

It is noted that the quality of translations from the google machine translation API was compared to open-source alternatives, but it does not fully address how the use of machine translation might introduce errors or reduce the quality of the data. Future research might explore the use of alternative translation services into the translations. There is also a lack of detailed discussion about the human verification process that was used to verify the machine translated question-answer pairs.

Applying other AI-based methods could further improve translation quality. To add more robustness, verification strategies, such as expert linguistic reviews or crowdsourced validation methods, could be used to enhance dataset fidelity. Human verification can greatly improve the quality of machine translated text, but details regarding the exact methods of verification and its effect on preserving data quality in the translated dataset are missing.

While the authors have made efforts in diversifying the AQA datasets with mClothoAQA and introducing a baseline multi-lingual AQA model, future research would benefit from a more detailed description of the quality control measures to ensure the datasets reliability.

Rating: 4

4. Technical Correctness, Theoretical Development

The technical correctness of the paper appears to be generally good. The methods and results are detailed and clear. Particularly, the usage of a Bidirectional Long Short-Term Memory (BiLSTM) network is a theoretically good approach, given that audio data is sequential. This network architecture uses both past and future context by processing data in two directions, advantageous for understanding the temporal aspects of audio signals. This aligns well with the interactive use of a question-answering system, where the context before and after a certain point in the audio clip could both be important for generating accurate responses.

The dataset creation process is described in detail and the use of pre-trained embeddings and the BiLSTM network for feature extraction and temporal understanding is elaborated

well. The preprocessing steps, including the extraction of Mel spectrograms and leveraging OpenL3 [3] embeddings for audio and FastText [4] embeddings for Text before inputting the data into the BiLSTM, are well suited for environmental audio data. These steps aim to condense the audio information into an abstract representation that the model can efficiently process, improving its ability to learn relevant features that are essential for the AQA task.

The paper outlines the architecture of the proposed multilingual model, however a more detailed discussion on the choice of model structure and its processes would improve the readers overall understanding of the choices made. Regarding the dataset, it is important to consider the potential implications of using machine translated content for training the AQA system. Theoretically, machine translation could introduce noise and semantic inaccuracies, which could misalign the models understanding of questions and answers. This challenge is further increased in a multi-lingual setting, where nuances could be lost in translation, leading to poorer model performances for certain languages. However, the paper attempts to mitigate this by evaluating the translations using metric such as BLEU [5], ROUGE [6], and METEOR [7], and applying human verification and adjustment, which is a practical approach to enhance the quality of machine translated data.

Rating: 3

5. Experimental Validation

The dataset mClothoAQA is extensively detailed, which is a strong aspect of this paper providing clarity on the data used for training the model and contributes to the reproducibility of the research. The performance metrics selected - accuracy, top-1, top-5, top-10 accuracies -, aim to help at evaluating the models performance to a certain extent. However, these metrics are not fully comprehensive, and other metrics like precision, recall, or F1-score could give a better insight into the models performance, especially for multi-class classification.

A possible problem is that that some classes are more represented than others. Confusion matrices or ROC curves might help better understand if the model is actually learning or is biased towards certain classes or question patterns. Using various of these evaluation metric together would possibly provide a better understanding of the models performance.

The authors could have included other existing AQA datasets like CLEAR[8] and DAQA[9] in their multi-lingual conversion to provide a more exhaustive analysis. By applying the multilingual training model to these additional datasets, the models effectiveness could be measured across different data sources and the results could be interpreted more generally.

The model architecture and the training setup are explained in detail with a focus on the multi-lingual aspect. It is also important to mention the computational resources that have been utilized, such as the number of GPUs/CPUs, memory requirements, which are missing in the paper. This could have been useful for replicating or extending the study.

During the walkthrough through the experiment steps, the paper suddenly mentions the training of "all the models", which introduces ambiguity or confusion in the reader since there is no previous mention of using multiple models, nor is there an explanation of the differences between these models.

Testing the robustness of the presented model on noisy or incomplete data, which is quite common in real-world scenarios, might have added more value to the results of the proposed solution, which seems to be overlooked in the experimental validation.

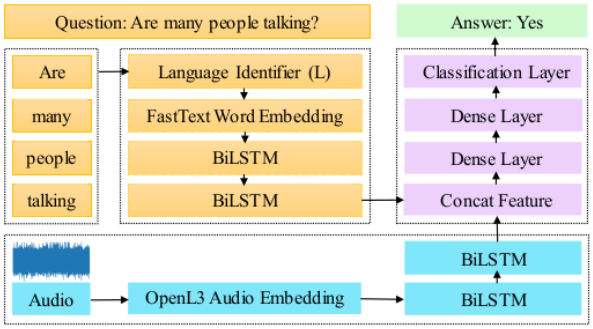

Figure 4: *Illustration of the model architecture*

The rest of the experiment appears to be replicable, given that the base ClothoAQA dataset is publicly available and the procedure for its modification is elaborated in the paper, although the human verification of the translation is hard to replicate exactly. Considering the diversity in acoustic environments, the performance of the model might be different when used on languages and acoustic environments not included in the training data. Given that the dataset size is relatively small, this could limit the models performance in generalized applications. The team published their code at `https://github.com/swarupbehera/mAQA`.

Table 1: *The table showing results for binary classification*

Languages	Unfiltered	Unanimous	Majority
English (en)	62.05	**71.50**	63.00
French (fr)	60.28	67.87	62.94
Hindi (hi)	**62.61**	70.64	**65.09**
German (de)	61.20	68.95	62.31
Spanish (es)	61.36	69.43	61.36
Italian (it)	60.97	67.14	62.47
Dutch (nl)	62.35	69.67	64.21
Portuguese (pt)	62.57	69.43	64.21

Table 2: *The table showing results for multi-class classification*

Languages	Top-1 Acc	Top-5 Acc	Top-10 Acc
English (en)	**53.73**	**91.68**	**98.57**
French (fr)	51.96	90.24	95.93
Hindi (hi)	53.31	91.54	97.41
German (de)	51.29	89.50	96.52
Spanish (es)	52.07	90.19	97.09
Italian (it)	51.62	89.69	95.70
Dutch (nl)	53.09	91.05	95.38
Portuguese (pt)	52.48	90.70	96.96

An explanation of the columns in Table 1 and Table 2 taken from the paper:

- Unfiltered: All the question-answer pairs are considered, even if they have contradicting answers.
- Unanimous: Only those question-answer pairs are considered where all three annotators have responded unanimously.
- Majority: For each question, the answer provided by at least

two of the three annotators is considered

In the experiments results, it is stated that the proposed multi-lingual AQA model shows "good performance across eight languages". However, this might not reflect the actual capability of the model. More robust or tailor-made evaluation methods could perhaps reveal the strength or weaknesses of the model more clearly.

While the datasets detailed description and the clear explanation of the model architecture are strengths of the research paper, the lack of details on the hardware and time, the opportunity for dataset comparison that was not taken, the sudden mention of multiple models without clarification, and a possibly optimistic characterisation of the results are areas that require improvement.

Rating: 3

6. Clarity of Presentation

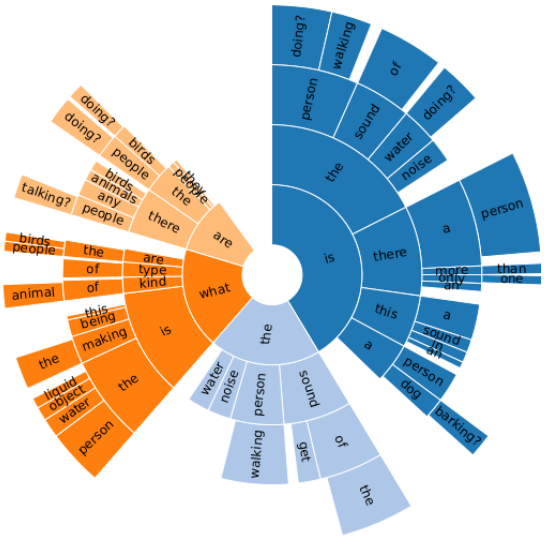

Figure 5: *The figure illustrating the first four words for all questions in mClothAQA-en*

The diagrams provided in the paper are hard to read. The types of diagrams used, resulting in impractical text angles and sizes, could have been chosen better, which would significantly enhance readability. There is a non-negligible amount of overlapping text elements in Figure 5, which implies a lack of thoroughness in verifying the papers overall clarity.

In terms of the figures, instead of a traditional cohesive image, each element in Figure 5 is represented as a separate shape object within the pdf file. This is not apparent to the reader but raises questions about the methods used for composing the contents of this paper. The figures in academic papers should be embedded as single objects when possible, as this practice prevents unforeseen rendering issues across different platforms and maintains the integrity of the visual information.

The choice of visualisation for word clouds in Figure 6 to present data is questionable. While it can sometimes offer a quick, intuitive understanding of text data, in this instance, it seems like a rather lacking choice. A different type of chart or a tabulated format would have likely offered a clearer, more

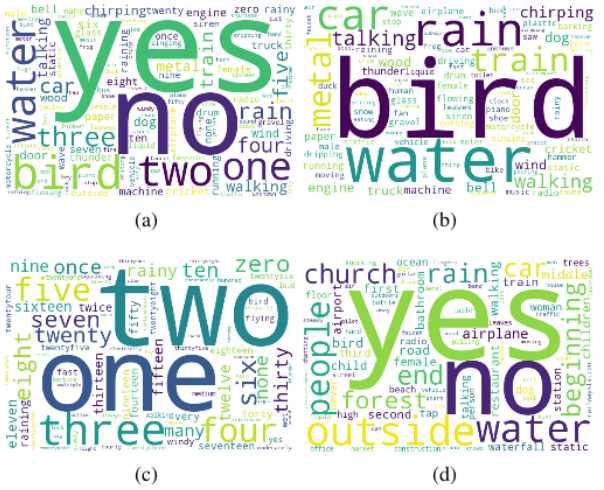

Figure 6: *The four wordclouds shown in the paper*

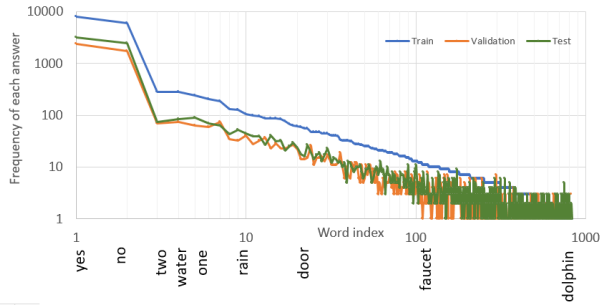

Figure 7: *Illustration of word frequencies in ClothoAQA*

precise representation of the data. Specificity regarding the relevance of the selected languages, emphasizing linguistic diversity in audio question answering research, might provide valuable context for the reader. A suggestion for alternative presentation is shown in Figure 7, taken from the ClothoAQA paper.

The resolution of the word cloud images is suboptimal. With a size of 9600x6400 pixels for each of the four images, the resolution is about three orders of magnitude higher than necessary, resulting in decreased performance when navigating the paper digitally, possibly not allowing readers using older devices to open the file at all. All other images in the paper have the similar problem of inadequate resolution, although not as extreme.

It is unclear why specifically French, Hindi, German, Spanish, Italian, Dutch, and Portuguese were selected for the dataset. The paper would benefit from an explanation of the criteria used when selecting the languages. Detailing the demographic or linguistic relevance could justify the language selection in terms of improving inclusivity and broader applicability in the AQA field.

The reason for choosing ClothoAQA as the base-dataset is well explained. Despite having only 1991 audio samples, which is quite small, the diversity it offers compared to the CLEAR and DAQA Datasets was the key factor for the choice. A structural change such as adding a methodology outline could increase the transparency of the research process.

The paper does not explicitly mention why embeddings of

size 300 were utilized for FastText. Those familiar with the subject might know that this size is the standard for FastText embeddings, but a note on this fact in the paper would provide clarity to readers new to the field.
Rating: 2

7. Reference to Prior Work

The limited number of works published in the field of Audio Question Answering is mentioned. Considering AQA is a relatively new field of research, there isn't a large amount of prior work to reference, emphasizing the novelty and the potential room for research in this area.

Table 3: *The table comparing the datasets, with the relatively high variety in column Ans being the reason for the choice of ClothAQA*

Dataset	#Audios	Dur(h)	MD(s)	AD(s)	QAS	L	#Ques	#Ans	Train	Val	Test
ClothoAQA [1]	1991	12.44	30	21	C	E	9153	830	1174	344	473
CLEAR [5]	50000	3.12	0.4	0.25	P	E	130957	47	35000	7500	7500
DAQA [4]	100000	2244.4	178.2	80.8	P	E	599294	36	80000	10000	10000

The paper references the foundational datasets in the field of AQA, namely ClothoAQA, CLEAR, and DAQA. The crowdsourced ClothoAQA dataset is particularly important as it is the base dataset which the authors chose to extend upon to create a multi-lingual version, mClothoAQA, for their work. By doing so, they are building on a robust foundation and expanding the accessibility of AQA to speakers of various languages, which is an important step towards inclusivity in the technological domain. The CLEAR and DAQA datasets are generated programmatically and offer a larger number of question-answer pairs, but they lack the in variety, which is a central component when attempting to design systems that can handle the unpredictability of real-world scenarios, making them less suitable for this experiment.

The choice to focus on multi-lingual aspects is relevant as it tries to address the linguistic bias often observed in AI research where english datasets dominate. The recognition of previous datasets proves the authors awareness of the groundwork laid already and creates the possibility for advancements in multilingual AQA.
Rating: 4

8. Overall evaluation of this paper

The paper makes an effort in the field of Audio Question Answering by trying to diversify and expand the existing ClothoAQA dataset into multiple languages, which is of value in an increasingly globalized world requiring technological inclusivity. This work creates possibilities for broader applications beyond the primarily English-speaking user base and further research. This is a positive contribution to the accessibility and utility of AQA systems for non-english speakers.

There are several aspects that could be improved. For instance, there is a disconnect between the claimed performance and the actual results presented in the document. The paper declares "good performance", yet the results, particularly for the binary classifier, suggest there is considerable room for enhancement. The results indicate a performance that is just slightly better than a random chance for the binary classifier. This marginal advantage raises questions regarding the models capabilities, and the applicability of this study in practical scenarios. This suggests that while the approach is valuable, the practical effectiveness of the model does not yet meet the expectations of the papers optimistic conclusion.

An improved presentation of figures and graphics is necessary. The clarity of the presentation could be improved with better figures.

It would be beneficial to include a more detailed analysis of the machine translations impact on the dataset quality and a comparison with other AQA datasets to test if the model is generally usable. An analysis of translation quality and its influence on the models training would clarify the robustness of the multi-lingual extension. A clarification on the selection of languages and reasoning behind methodological choices would also provide valuable information regarding the choices made by the team.

In summary, the paper represents a step forward in multilingual AQA research, exploring a new field by expanding an existing dataset. Despite improvable areas, the novelty of the study does not get overshadowed by its shortcomings. I would give this paper a "weak accept", as it holds value with several opportunities for improvement that could increase its quality. The potential impact for non-English speaking communities is significant, and with further improvements, the introduced methodology could become a valuable asset in AQA research.
Rating: 3

9. Additional comments to author(s)

Adjusting the image sizes would be an easy way to improve the overall digital usability of the paper.

10. Confidential Comments to technical program committee

11. References

[1] G. Stevens, S. Flaxman, E. Brunskill, M. Mascarenhas, C. D. Mathers, and M. Finucane, "Global and regional hearing impairment prevalence: an analysis of 42 studies in 29 countries," *European Journal of Public Health*, vol. 23, no. 1, pp. 146–152, 12 2011. [Online]. Available: https://doi.org/10.1093/eurpub/ckr176

[2] S. Lipping, P. Sudarsanam, K. Drossos, and T. Virtanen, "Clotho-aqa: A crowdsourced dataset for audio question answering," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 1140–1144.

[3] A. L. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3852–3856.

[4] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," 2017.

[5] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[6] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[7] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, C. Callison-Burch, P. Koehn, C. S. Fordyce, and C. Monz, Eds. Prague, Czech Republic: Association for

Computational Linguistics, Jun. 2007, pp. 228–231. [Online]. Available: https://aclanthology.org/W07-0734

[8] J. Abdelnour, G. Salvi, and J. Rouat, "Clear: A dataset for compositional language and elementary acoustic reasoning," 2018.

[9] H. M. Fayek and J. Johnson, "Temporal reasoning via audio question answering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2283–2294, 2020.