SPEECHDPR: END-TO-END SPOKEN PASSAGE RETRIEVAL FOR OPEN-DOMAIN SPOKEN QUESTION ANSWERING

Chyi-Jiunn Lin[1], *Guan-Ting Lin*[1], *Yung-Sung Chuang*[2], *Wei-Lun Wu*[1], *Shang-Wen Li*[3], *Abdelrahman Mohamed*[4], *Hung-yi Lee*[1], *Lin-shan Lee*[1]

[1]National Taiwan University [2]Massachusetts Institute of Technology
[3]Meta AI [4]Rembrand

ABSTRACT

Spoken Question Answering (SQA) is essential for machines to reply to user's question by finding the answer span within a given spoken passage. SQA has been previously achieved without ASR to avoid recognition errors and Out-of-Vocabulary (OOV) problems. However, the real-world problem of Open-domain SQA (openSQA), in which the machine needs to first retrieve passages that possibly contain the answer from a spoken archive in addition, was never considered. This paper proposes the first known end-to-end framework, Speech Dense Passage Retriever (SpeechDPR), for the retrieval component of the openSQA problem. SpeechDPR learns a sentence-level semantic representation by distilling knowledge from the cascading model of unsupervised ASR (UASR) and text dense retriever (TDR). No manually transcribed speech data is needed. Initial experiments showed performance comparable to the cascading model of UASR and TDR, and significantly better when UASR was poor, verifying this approach is more robust to speech recognition errors.

***Index Terms*—** Spoken Question Answering, Spoken Language Understanding, Spoken Content Retrieval

1. INTRODUCTION

Spoken Question Answering (SQA) is to find the answer span out of a passage for a given question by a machine, where the question, passage and answer span are all in the form of audio waveform (the question may be in text form in some cases though) [1, 2, 3, 4]. Open-domain SQA (openSQA) is one step further, in which the passage is not given, and the machine has to find in addition one or more passages containing the answer from a large spoken dataset or the spoken content over the Internet, referred to as ”a spoken archive” below, before performing SQA. These technologies are certainly important when machines have to reply to the user's questions.

Substantial effort has been made and reasonable results have been obtained for the text version of the above two tasks, Text Question Answering (TQA) and Open-domain TQA (openTQA), in which everything is in form of text instead. The latter is usually achieved by cascading a text retriever in front of a text reader (or TQA) [5, 6, 7]. So a natural approach for SQA and openSQA discussed here is simply cascading an ASR module in front of TQA or openTQA, or performing TQA or openTQA on top of the ASR transcriptions.

However, building reliable ASR modules requires extensive training with large quantities of labeled data, which is prohibitively difficult for many low-resourced languages over the world. To render openSQA affordable for the low-resourced languages, it is imperative to develop models that require no paired speech-text data for both training and inference. Therefore, this paper specifically emphasizes on the scenario in which *no paired speech-text data is available*.

A simple solution is to replace the supervised ASR with an unsupervised ASR (UASR) [8, 9] trained on unpaired speech and text data. However, current unsupervised ASR models are still less performant and stable than the supervised ones. Besides, the recognition errors produced by the inaccurate UASR module can not be recovered by the following TQA or openTQA, but instead cause more errors. Moreover, the correct answers to the questions often include named entities or out-of-vocabulary (OOV) words which can never be recognized. These explain why end-to-end approaches directly performing retrieval and identifying the answer span over the audio signals rather than over the UASR transcriptions are highly desired.

SpeechBERT [10] performed end-to-end SQA by properly aligning the phonetic embeddings of spoken words and semantic embeddings of text words in a hidden space, so the semantics can be extracted directly from the audio signals. SPLAT [11] achieved similar goals later on with improved performance. DUAL [12] was then the first successfully performing SQA without any paired speech-text data by transforming the audio signals into sequences of quantized discrete units such that textless NLP [13, 14] can be performed bypassing ASR errors. However, all of these above mentioned approaches achieved SQA only, not openSQA. In these approaches, the answer spans were identified from given passages, while the problem of how passages containing the answer to the user's question can be obtained from a large spoken archive was totally left out.

While the problem of retrieving passages from spoken archives has been extensively investigated and reported [15, 16, 17, 18, 19], prior works have not specifically addressed the end-to-end resolution of the openSQA retrieval task. The openSQA retrieval task involves semantic retrieval from sentence to sentence, where the question is a spoken sentence that may not necessarily share overlapping content with the gold passage. Instead, previous end-to-end approaches were limited to either query-by-example spoken-term detection [20, 21], where the question (or query) typically consists of a spoken short phrase and the gold passages must contain this phrase, or semantic search limited to visually-grounded speech [22, 23].

This paper addresses the untranscribed spoken-question-to-spoken-passage semantic retrieval problem for openSQA by proposing an end-to-end model, SpeechDPR (Speech Dense Passage Retriever), to retrieve the passages from a spoken archive without supervised ASR or manually transcribed speech data. SpeechDPR adopts the bi-encoder retriever framework and learns a sentence-level semantic representation space by distilling knowledge from

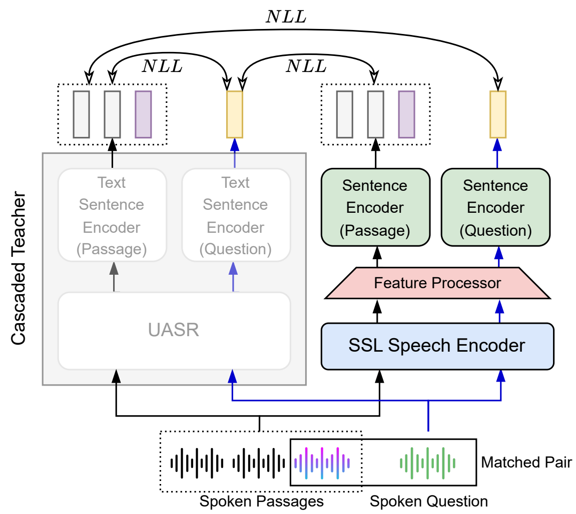

Fig. 1: The framework of SpeechDPR.

the cascading model of UASR and text dense retriever (TDR). SpeechDPR assesses the similarity between the question and each passage in the spoken archive by calculating the inner product of their sentence representations, and thus can find the most semantically relevant passage.

The main contributions are summarized as follows:

- This research proposes SpeechDPR, the first end-to-end model to tackle the challenge of untranscribed spoken passage semantic retrieval for openSQA without requiring any supervised ASR transcriptions or speech-text paired data for training and inference.
- SpeechDPR achieves a competitive retrieval accuracy, comparable to the cascading baselines, which involves performing the corresponding modules trained on text on top of UASR transcriptions, and outperforms significantly in scenarios with relatively poor ASR accuracies.

2. BACKGROUND

We propose the never-attempted retrieval-SQA framework to address the openSQA problem, and this paper focuses on improving the retrieval component of openSQA, where all questions and passages are in the form of raw spoken sentence waveforms. Given a spoken question and a spoken archive containing numerous long spoken passages, such as broadcast news or podcast database, the retrieval model calculates the similarity between each passage and the question. It then returns a small number of passages with the highest similarity so that the downstream SQA model can identify the answer span from these passages. A retrieval is considered successful when the returned passages includes the gold passage containing the answer to the question. We assume neither paired speech-text data nor off-the-shelf supervised ASR is available.

2.1. Bi-encoder dense retriever framework

SpeechDPR and all TDRs in this paper adopt the bi-encoder dense retriever framework commonly used in openTQA [7, 24]. The bi-encoder dense retriever encodes the question and the passage with two separate encoders. Given a bi-encoder retriever model (Q, P) with a question encoder $Q(.)$ and a passage encoder $P(.)$, before performing retrieval, the retrieval model encodes each passage $\boldsymbol{p}$ in the archive into a low-dimensional continuous vector $P(\boldsymbol{p})$. During retrieval, the model similarly encodes the input question $\boldsymbol{q}$ into a vector $Q(\boldsymbol{q})$ so that the similarity between $\boldsymbol{q}$ and each $\boldsymbol{p}$ can be expressed by calculating the dot product of $Q(\boldsymbol{q})$ and $P(\boldsymbol{p})$:

$$sim(\boldsymbol{q}, \boldsymbol{p}) = Q(\boldsymbol{q}) \cdot P(\boldsymbol{p}) \tag{1}$$

The bi-encoder dense retriever (Q, P) learns a sentence representation that minimize the negative log likelihood (NLL) between the representation of the question and its paired passage. For each pair of question and its relevant (positive) passage $(\boldsymbol{q}, \boldsymbol{p}^+)$ in the dataset, n irrelevant (negative) passages $\{\boldsymbol{p}_i^-\}_{i=1}^n$ are sampled to calculate the NLL loss:

$$NLL_{Q,P} = -log \frac{e^{Q(\boldsymbol{q})\cdot P(\boldsymbol{p}^+)}}{e^{Q(\boldsymbol{q})\cdot P(\boldsymbol{p}^+)} + \sum_{i=1}^n e^{Q(\boldsymbol{q})\cdot P(\boldsymbol{p}_i^-)}} \tag{2}$$

The negative passages we use to calculate the NLL loss are the gold passages of the other questions in the same batch.

3. PROPOSED APPROACH

3.1. SpeechDPR model

The end-to-end SpeechDPR model is composed of SSL speech encoder, feature processor, question sentence encoder, and passage sentence encoder. In forward propagation, the input waveform is passed through the SSL speech encoder, and downsampled by feature processor. After that, the question or passage encoder is used for extracting semantic embedding according to the type of the speech.

SSL speech encoder is pre-trained on raw speech waveforms via self-supervised learning (SSL) and turns the waveform into contextualized frame-level representations, or a sequence of vectors. The SSL encoder is only used for feature extraction, the model parameters are frozen in our experiments. Feature processor instance-normalizes the extracted representations and passes them through a two-layer CNN to shorten the very long sequence length for reducing memory consumption. Both sentence encoders use the Roberta-base [25] encoder to encode the processed speech representations. This design choice is based on the finding of BERT-like encoders' cross-disciplinary transferability by Kao et al. [26]. The representation sequence is first concatenated with Roberta word embedding of the [CLS] token at the beginning and fed into the sentence encoder to derive the sentence-level representation, a 768-dim vector, from the representation at the [CLS] token.

3.2. Knowledge distillation from Cascading Teacher model

Based on our preliminary finding that SpeechDPR performs incompetently if it is trained by just minimizing the NLL loss in Eq. 2, we additionally distill knowledge from Cascading Teacher model. The Cascading Teacher consists of a UASR and a TDR. In our case, neither UASR nor TDR requires paired speech-text data for training. Specifically, UASR transcribes speech waveforms to text sentences and is trained on unpaired speech-text data via adversarial learning. TDR is a bi-encoder text retriever that converts the UASR-transcribed text sentence to the sentence-level representation. The TDR is trained on UASR transcripts but not the ground truth transcripts.

3.3. Training Objective

We propose to distill the sentence representation encoded by TDR to improve SpeechDPR training. Apart from learning to minimize the NLL loss in Eq. 2, SpeechDPR additionally distills knowledge from Cascading Teacher model by minimizing the NLL between the question sentence representation encoded by SpeechDPR and the passage sentence representation encoded by the teacher. It similarly minimize the NLL between the passage representation encoded by SpeechDPR and the question representation encoded by the teacher. Therefore, we set the total loss to the weighted sum of the above-mentioned three NLL losses. Let Q_T, P_T respectively denote the question and passage encoder of the teacher model, and Q_S, P_S denote those of SpeechDPR. The total loss function L is:

$$L = NLL_{Q_S,P_S} + \alpha NLL_{Q_S,P_T} + \beta NLL_{Q_T,P_S}, \quad (3)$$

where α and β are tuned hyperparameters.

4. EXPERIMENTS

4.1. Data

We adopt a data setup similar to that used in openTQA research [5, 7]. These studies gathered questions from various TQA datasets and used a set of Wikipedia passages as the sole source for retrieval during both training and inference. In our paper, we use spoken questions from the SLUE-SQA-5 dataset [27] and spoken passages from the Spoken Wikipedia dataset [28].

SLUE-SQA-5 is a SQA dataset whose training, dev, test subsets all includes the human-spoken version of questions sourced from five TQA datasets: TriviaQA [29], SQuAD v1.1 [30], Natural Questions (NQ) [31], WebQuestions (WQ) [32] and CuratedTREC (TREC) [33]. Notably, the question speakers across different subsets are distinct. This characteristic makes SLUE-SQA-5 well-suited for openSQA experiments. The dataset contains about 46k questions in the training set, 1.9k in the dev set, and 2.4k in the test set. Each paired passage of SLUE-SQA-5 questions consists of a 40-second Spoken Wikipedia passage that includes the answer to the corresponding question.

Spoken Wikipedia contains the spoken version of 1.2k Wikipedia articles from about 400 human speakers. This paper follows the preprocessing procedure in SLUE-SQA-5 by splitting spoken articles into about 39k spoken passages with duration of 40 seconds and taking these passages as spoken archive for retrieval. The total duration of the spoken archive is 427 hours.

Note that currently due to a lack of suitable datasets for openSQA in other languages, we are limited to experimenting on English.

4.2. Evaluation

Top-K retrieval accuracy, or the percentage of questions for which the top-K passages returned by the retriever include any gold passage, is used to evaluate the retriever performance. We consider a passage as the gold passage to a question only when it is paired with that question in the original SLUE-SQA-5 setting. This paper reports the top-20 accuracy.

In addition, to investigate how different retrievers affect the whole openSQA accuracy, we passed the top-20 retrieved passages into a shared SQA module, whose implementation details are described in Section 4.3, to predict the answer span to the question. Following the works [10, 27] in SQA, we use the FF1 (Frame-level F1 Score, overlapping span length out of the reference span and predicted span respectively as precision and recall) as the metrics for evaluating the performance of openSQA. Note that sometimes the correct answer is in the answer span identified by SQA, but because the carrying passage is somehow not labeled as a gold passage for the question, and we assume the transcriptions are not available, this answer span can only be taken as incorrect and the FF1 is set to zero. This may underestimate the performance of the models.

4.3. Implementation details

Cascading Teacher model. We follow the training procedure of wav2vec-U 2.0 [9] to train our UASR module on the datasets mentioned in Section 4.1. We pick the unpaired speech data in SLUE-SQA-5 training set and the unpaired text data in TriviaQA training set excluding the ones included in SLUE-SQA-5 for the adversarial-training stage, while all spoken passages in Spoken Wikipedia plus the SLUE-SQA-5 training set are used for the self-training stage. The UASR model has word error rates of 24.1, 23.3, 39.8 respectively on SLUE-SQA-5 dev set, SLUE-SQA-5 test set, Spoken Wikipedia passages. For TDR, we use two Roberta-base models as the bi-encoder and train the models on the UASR transcriptions with 64 batch size, 4e-5 learning rate, and 100 warmup steps for 100 epochs.

SpeechDPR. We use the HuBERT-large model[1] with 24 transformer layers pre-trained on Libri-Light speech dataset [34] as SSL speech encoder and extract representations from its 22nd layer. The CNN in feature processor has a stride of 4 and 3 in its first and second layer, respectively. Feature processor and the two sentence encoders are jointly trained with 64 batch size, 1e-4 learning rate, and 500 warmup steps for 100 epochs. We set both α and β to 0.5 in the loss function (Eq. 3).

The above-mentioned hyperparameters and the best checkpoints of both TDR and speechDPR models are picked according to the top-20 retrieval accuracy on SLUE-SQA-5 dev set.

The shared SQA module for openSQA evaluation. The SQA module is a cascading model of UASR mentioned above and a Deberta-base [35] TQA model fine-tuned on SQuAD v1.1 dataset questions excluding the ones included in the SLUE-SQA-5. This module obtains 36.82 FF1 on SLUE-SQA-5 test set when given the gold passage to each question. During openSQA inference, we pick the answer span with highest answer score from the retrieved 20 passages as the finally predicted answer span. Following [6], the answer score is defined as the linear interpolation of the passage similarity predicted by the retriever and the span score predicted by the SQA model. The weights for linear interpolation are tuned on the SLUE-SQA-5 dev set. We use Montreal Forced Aligner [36] to obtain the time intervals of the TQA-predicted answers in seconds for evaluating FF1.

4.4. Baseline: the cascading approach (UASR + TDR)

Because there did not exist any known prior work for the openSQA retrieval task considered here, we take the cascading approach (UASR + TDR) as the baseline to compare with. Apart from Cascading Teacher described above, for a fairer comparison, we additionally train another cascading model, referred to as "Cascading Student", with Cascading Teacher by applying the same objective function with SpeechDPR. Cascading Student shares the same UASR module and hyperparameters of TDR training with the teacher.

[1] https://huggingface.co/facebook/hubert-large-ll60k

Table 1: Top-20 retrieval accuracy (%) (**Top-20**) and openSQA accuracy (**FF1**) on SLUE-SQA-5 test set questions. "Knowledge distillation from Cascading Teacher" is abbreviated as "KD" in row (d). The openSQA accuracy is derived by passing the top-20 retrieved passages to a shared SQA model to evaluate FF1.

Retrieval Model	Top-20	FF1
Single model:		
(a) Baseline: Cascading Teacher	**20.59**	5.46
(b) Baseline: Cascading Student	20.46	**5.81**
(c) Proposed: SpeechDPR	20.11	5.57
(d) Ablation: SpeechDPR w/o KD	00.04	0.00
Ensemble model:		
Baseline: Ensemble of (a) and (b)	25.60	7.14
Proposed: Ensemble of (a) and (c)	**28.88**	**8.00**

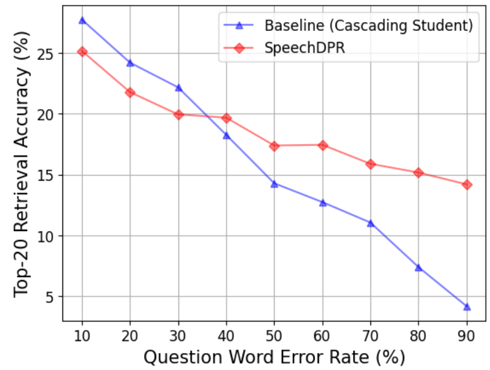

Fig. 2: Top-20 retrieval accuracy evaluated on subsets of SLUE-SQA-5 test set questions with different levels of UASR WER.

5. RESULTS

5.1. Retrieval results

The top-20 retrieval accuracies are listed in the second column of Table 1. By comparing Cascading Student in row (b) to Cascading Teacher in row (a), we see that Cascading Student gets a similar top-20 accuracy (20.46%) to Cascading Teacher does (20.59%), indicating that knowledge distillation from text retrievers to text retrievers does not lead to significant improvement. By comparing SpeechDPR in row (c) to baselines of row (a) and row (b), we observe that SpeechDPR achieves a comparable accuracy (20.11%) with the two cascading baseline models. The difference between their retrieval accuracies is less than 1%. This result shows that SpeechDPR can extract useful information for retrieval directly from continuous speech waveforms instead of the intermediate textual transcripts while maintaining a reasonable performance.

Besides, we also conduct an ablation study by training another SpeechDPR model without knowledge distillation (KD) from Cascading Teacher, or without NLL_{Q_S,P_T} and NLL_{Q_T,P_S} terms in the loss function (Eq. 3). This trained model, denoted by "SpeechDPR w/o KD" in row (d), gets a significantly lower accuracy (0.04%) compared to the original SpeechDPR in row (c). This indicates that knowledge distillation from a cascading model is crucial for SpeechDPR training.

Although SpeechDPR distills from the Cascading Teacher, it is also trained end-to-end with speech directly. Therefore, SpeechDPR can potentially achieve superior performance on the one performs poorly with Cascading Teacher. The model ensemble technique is a common technique to combine models with different specialty, resulting in better performance. We tried to ensemble the SpeechDPR model in row (c) and with Cascading Teacher in row (a) by linearly interpolating the similarities scored by the two models. The weights for linear interpolation are tuned according to top-20 retrieval accuracy on the dev set. This ensemble model gets a retrieval accuracy (28.88%), which significantly surpasses the accuracy (25.60%) of the ensemble of Cascading Student in row (b) and Cascading Teacher in row (a). The result verifies that end-to-end SpeechDPR model can learn extra knowledge complementary to that learned by the cascading model, so the ensemble of the two models achieves superior performance to the ensemble of two cascading models.

5.2. OpenSQA results

The openSQA accuracies (FF1) are listed in the third column of Table 1. We can observe a trend similar to that in the top-20 retrieval accuracy. The two baseline models in row (a), (b) and SpeechDPR in row (c) achieve similar FF1 scores (5.46, 5.81, 5.57, respectively), and the ensemble of SpeechDPR and Cascading Teacher outperforms the ensemble of the two cascading baselines (8.00 vs. 7.14). This result shows the efficacy of SpeechDPR in the whole openSQA task. Besides, the ablated SpeechDPR in row (d) fails to identify any correct answers obviously because it has a poor retrieval accuracy.

It is worth noting that the openSQA accuracy is constrained by the performance of the downstream SQA module. This is due to the fact that questions from the SLUE-SQA-5 dataset pose a significant challenge for current SQA models. According to the SLUE-SQA-5 paper [27], even the highest-performing model without utilizing paired speech data, achieves only a 21.8 FF1 score on SQA.

5.3. Retrieval performance at different WERs

In real-world scenarios, ASR accuracy can be very poor due to noisy acoustic conditions, out-of-vocabulary (OOV) and other problems, referred to as ASR conditions here. We thus investigate the effect of these ASR conditions on the retrieval performance by splitting the questions in SLUE-SQA-5 test set into several subsets according to UASR WER in the question. The result of top-20 retrieval accuracy within each of these subsets are plotted respectively in Figure 2. We can see the retrieval accuracy of the Cascading Student baseline degrades seriously when UASR WER gets higher, while that of the proposed SpeechDPR is more robust to ASR conditions, obviously because SpeechDPR does not rely on UASR for inference. SpeechDPR actually significantly outperforms the cascading baseline when WER exceeds 40%. Because many of the questions and passages in openSQA task include name entities and OOV words, this makes the proposed SpeechDPR highly attractive.

6. CONCLUSION

Here we propose SpeechDPR, an end-to-end model for the semantic retrieval task for openSQA using knowledge distillation from the cascading model of UASR and TDR, in which no supervised ASR or any manual transcript is needed. Experimental results indicate that SpeechDPR achieves competitive performance compared to the cascading baseline (UASR-TDR) and outperforms significantly in scenarios of poor UASR performance. This is attractive because many spoken questions and passages include name entities and OOV words which cannot be recognized by UASR. The initial experiments are only done on English, but our framework can be extended to other low-resourced languages in future research.

7. REFERENCES

[1] Jordi Turmo et al., "Overview of QAST 2008," in *Workshop of the Cross-Language Evaluation Forum for European Languages*, 2008.

[2] Pere R Comas et al., "Sibyl, a factoid question-answering system for spoken documents," *ACM Transactions on Information Systems (TOIS)*, 2012.

[3] Chia-Hsuan Lee et al., "Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension," *Proc. Interspeech 2018.*

[4] Chia-Hsuan Lee et al., "ODSQA: Open-Domain Spoken Question Answering Dataset," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018.

[5] Danqi Chen et al., "Reading Wikipedia to Answer Open-Domain Questions," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017.

[6] Wei Yang et al., "End-to-End Open-Domain Question Answering with BERTserini," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019.

[7] Vladimir Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[8] Alexei Baevski et al., "Unsupervised speech recognition," *Advances in Neural Information Processing Systems*, 2021.

[9] Alexander H Liu et al., "Towards end-to-end unsupervised speech recognition," in *2022 IEEE Spoken Language Technology Workshop (SLT).*

[10] Yung-Sung Chuang et al., "SpeechBERT: An Audio-and-Text Jointly Learned Language Model for End-to-End Spoken Question Answering," *Proc. Interspeech 2020*, 2020.

[11] Yu-An Chung et al., "SPLAT: Speech-Language Joint Pre-Training for Spoken Language Understanding," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

[12] Guan-Ting Lin et al., "DUAL: Textless Spoken Question Answering with Speech Discrete Unit Adaptive Learning," *arXiv e-prints*, 2022.

[13] Kushal Lakhotia et al., "On Generative Spoken Language Modeling from Raw Audio," *Transactions of the Association for Computational Linguistics*, 2021.

[14] Ann Lee et al., "Direct speech-to-speech translation with discrete units," *arXiv preprint arXiv:2107.05604*, 2021.

[15] Lin-shan Lee et al., "Spoken content retrieval—beyond cascading speech recognition with text retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015.

[16] John S Garofolo et al., "The TREC Spoken Document Retrieval Track: A Success Story.," *NIST SPECIAL PUBLICATION SP*, 2000.

[17] Martha Larson et al., "Spoken content retrieval: A survey of techniques and technologies," *Foundations and Trends® in Information Retrieval*, 2012.

[18] Ciprian Chelba et al., "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine*, 2008.

[19] Michael Witbrock and Alexander G Hauptmann, "Speech recognition and information retrieval: Experiments in retrieving spoken documents," in *Proceedings of the DARPA speech recognition workshop*, 1997.

[20] Dhananjay Ram et al., "Neural network based end-to-end query by example spoken term detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[21] Chia-Wei Ao and Hung-yi Lee, "Query-by-example spoken term detection using attention-based multi-hop networks," in *2018 IEEE ICASSP*, 2018.

[22] David Harwath et al., "Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech," in *2018 IEEE ICASSP*, 2018.

[23] Herman Kamper et al., "Semantic speech retrieval with a visually grounded model of untranscribed speech," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018.

[24] Lee Xiong et al., "Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval," in *International Conference on Learning Representations*, 2020.

[25] Yinhan Liu et al., "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[26] "Is BERT a Cross-Disciplinary Knowledge Learner? A Surprising Finding of Pre-trained Models' Transferability, author=Kao, Wei-Tsung and Lee, Hung-Yi," in *EMNLP 2021*, 2021.

[27] Suwon Shon et al., "SLUE Phase-2: A Benchmark Suite of Diverse Spoken Language Understanding Tasks," *arXiv preprint arXiv:2212.10525*, 2022.

[28] Timo Baumann et al., "The Spoken Wikipedia Corpus collection: Harvesting, alignment and an application to hyperlistening," *Language Resources and Evaluation*, 2019.

[29] Mandar Joshi et al., "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

[30] Pranav Rajpurkar et al., "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

[31] Tom Kwiatkowski et al., "Natural questions: a benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, 2019.

[32] Jonathan Berant et al., "Semantic parsing on freebase from question-answer pairs," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013.

[33] Petr Baudiš et al., "Modeling of the question answering task in the yodaqa system," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction, CLEF'15*, 2015.

[34] Jacob Kahn et al., "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020.*

[35] Pengcheng He et al., "DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION," in *International Conference on Learning Representations*, 2020.

[36] Michael McAuliffe et al., "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi.," 2017.