**Al-Rafidain Journal of Computer Sciences and Mathematics (RJCM)**

**www.csmj.mosuljournals.com**

# A Comparative Study for Speech Summarization Based on Machine Learning: A Survey

**Hiba Adreese Younis [1], * Yusra Faisal Alarheam[2]**

*Department of computer science, College of Computer Science and Mathematics, University of Mosul, Mosul, IRAQ[1,2]*

**Corresponding author. Email: hibaadreese@uomosul.edu.iq[1]*

**Article information**

***Article history:***
Received : 30/6/2022
Accepted : 13/11/2022
Available online :

**Abstract**

The most important aspect of human communication is speech. Lengthy media such as speech takes a long time to read and understand. This difficulty is solved by providing a reduced summary with semantics. Speech summarization can either convert speech to text using automated speech recognition (ASR) and then build the summary, or it can process the speech signal directly and generate the summary. This survey will look at a various of recent studies that have used machine and deep learning algorithms to summarize speech. it discusses the speech summarizing literatures in terms of time restrictions, research methodology, and lack of interest in particular databases for literature searches. As newer deep learning approaches were not included in earlier surveys, this is a new survey in this discipline where different approaches with various datasets were explored for speech summarization and evaluated using subjective or objective methods.

***Keywords***:
ASR, LSTM, speech summarization, deep learning, ROUGE

***Correspondence:***
Author : Hiba Adreese Younis
Email: hibaadreese@uomosul.edu.iq

## 1. INTRODUCTION

Humans use speech as the most natural and effective means of communication. In today's information technology era, spoken natural communication is of critical importance. The majority of IT researchers are working on developing highly effective spoken language understanding (SLU) systems using machine learning and artificial intelligence approaches applied to speech processing technology and natural language processing algorithms. Speech summarization is a hot issue in the SLU domain right now. Speech summarizing aids in the extraction of relevant information from a variety of spoken material sources, including meetings, broadcast news, talk shows, lectures, voice mails, and YouTube videos. Speech summarizing is an important method for dealing with the large quantity of data available in low-information-density audio recordings. By pointing to a short summary that will be listened to by an analyst at the end, speech summarizing is also particularly beneficial for commercial marketing analysts and government intelligence [1]. Abstractive and extractive summarization are the two categories into which text summarization is typically divided. For parsing, word reduction text summary generation, and abstractive summarization natural language processing (NLP) techniques are applied [2]. When compared to abstractive summarization, extractive summarization is stated to be flexible and time-efficient. In extractive summarization, the phrase is taken into account in matrix form, and the key sentences are then identified using some feature vectors. A feature vector is referred to as an n-dimensional vector with numerical features that represent the object. A text summary is produced by speech summarizing using inputted speech data. In order to build a concise representation of the content, it must process extensive speech data (a series of

utterances) and extract crucial information. Additionally, speech involves fillers, disfluencies, redundancies (such as using the same sentences

## 2. MOTIVATION FOR SPEECH SUMMARIZATION

Speech summarizing has sparked a lot of interest in research due to the growing number of audio recordings because people are generally quite busy and locating adequate and appropriate knowledge base information has become extremely challenging [5]. Regarding broadcast news people find it difficult to identify news from vast news broadcast archives, and as a result, television program space is wasted. They may be preoccupied with other vital tasks and miss the news program that is carried out every day at a fixed time and they are rarely watch the entire news presentation; instead, they are interested in watching a specific news piece [6].

## 3. AUDIO SIGNAL PROCESSING USING DEEP LEARNING

To extract audio features, audio machine learning systems previously relied on classic digital signal processing techniques. However, since deep learning has become more widely used in recent years, it has had amazing success in handling audio. Traditional audio processing approaches are no longer required, and we can rely on standard data preparation rather than a lot of manual and customized feature development. Using deep learning, we don't work with audio data in its raw form. Instead, audio data are transformed to images and then process those images using deep learning methods such as conventional neural network(CNN) architecture. This is done by generating spectrograms from the audio signal [6].

## 4. CONTRIBUTION

Previous surveys till 2018 focused on researches, some of them include traditional methods for summarization and others, not includes recent methods for deep learning. So in this survey we explore the most techniques of deep learning were used recently. It also presents a speech summarization categorization based on a new different criterion. To the best of our knowledge this is the first survey that includes recent work on speech summarization using machine learning (beyond 2016) as well as recent studies for deep learning. It focuses on speech summarization models, Source of input corpus, number of stages in summarization process and evaluation metrics.

## 5. CATEGORIZATION OF SPEECH SUMMARIZATION PROCESS

This survey comprises fundamental studies based on speech summarizing processes. Speech summarization process can be classified based on a set of criterions as describes below.

### 5.1 Speech Summarization Models

#### A. Conventional Summarization Models

Techniques for converting speech audio files to text files, as well as text summarization on the text files are presented which are utilized for text summarization in the latter situation. words are allocated weights based on the number of times they appear in the text file. The recorded speech can be converted to text with the help of Google API[7]. To distinguish sentences, python string tokenization employs periods. The sentence is summarized once the index is identified and ranked according to the sentence's weight. We can deduce from the recognition time that sentences with a period and a question mark are recognized faster than those without [8]. To avoid having to read the entire text each time, a more efficient method of summarizing books into important keywords was developed. The suggested model summarizes the content using a weighted TF-IDF (Term Frequency Inverse Document Frequency), and then converts it to speech [9]. These models are considered out of our study.

#### B. Machine Learning and Deep Learning based Summarization Models

The basic studies which implemented machine and deep learning models were as follows:

Chen, B., et al. presented an empirical study of the merits of two schools of training criteria for reducing the negative effects of imbalanced-data problem and improving summarization performance. One method is to train a summarizer's classification capabilities using pair-wise ordering information of phrases in a training document according to importance. The other method is to train the summarizer by directly optimizing the associated evaluation score or an objective that is tied to the final evaluation. Experiments on the summary of broadcast news show that these training criteria can provide significant gains over a few existing summarizing methods[10]. Liu, Z., Ng, A., Lee, S., Aw, A. T., & Chen, N. F. presented a neural architecture that works well and efficiently. It was shown that the hierarchical structure of dialogues is used to integrate topic-level attention mechanisms in pointer-generator networks. The suggested model outperforms competitive bust when it comes to long dialogue samples, and it also works well with limited training data[11]. Zhao, Z., et al. presented a hierarchical neural encoder based on adaptive recurrent networks to learn the semantic representation of meeting discussion. Then, for abstractive meeting summarizing, a reinforced decoder network was created to generate high-quality summaries constructing the HAS Hierarchical encoder-decoder network learning framework with adaptive dialogue segmentation[12]. Atsunori Ogawa., et al. suggested a compressive speech summarizing approach(performs both extraction and compression at the same time)based on integer linear programming (ILP), which maximizes the coverage of content words in the resulting summary. It was considered the first trial of compressive speech summarization at that time. A single step summary using the confidence, TF IDF

and bigram LM scores, was conducted for each produced ASR hypothesis sequence that corresponds to a full lecture speech (12 min on average). The results showed that the compressive approach outperformed the extractive method, and compressive CN over compressive 1-best**[**13**]**. M Menacer, et al. designed a two-stage ASR and text summarization pipeline to provides an end-to-end dialog summarizing system. Users can first see a high-level summary of the information, then drill down into longer and more thorough summaries or listen to the raw audio itself using hierarchical summarization. PEGASUS the current state-of-the-art abstractive summarization model was used, which is capable of producing summaries of significantly higher human quality [14]. Kumar, B. D. (n.d.) used an Essence vector (EV) modeling which is an unsupervised paragraph embedding method aims to derive the most important information from a paragraph while also including general background information. First speech is transformed to text and fed into the Essence Vector (EV) model as a high-dimensional vector input, which summarizes the information in a low-dimensional space. The EV model's performance was improved by adding an Attention mechanism and employing LSTM to handle the speech to text module's faulty and erroneous speech conversions [15]. González-Gallardo, et al. offered a hybrid technique during the training phase, whereas summary creation was text-independent. It is based on employing textual information to learn an in formativeness representation based on probability distribution divergences, which is not taken into account by normal audio summarization with audio features. The length of the summary was set to equal 35% of the original audio length[16]. Weng, S.-Y., Lo, T.-H., Chen, B. extended the BERT-based method for supervised extractive speech summarization that is capable of performing robust summary on spoken documents containing erroneous ASR transcripts. To improve the summarization performance of a spoken material, many supplementary structural and linguistic features were applied to enrich the embeddings of the sentences. The results showed that in both the text document(TD) and spoken document(SD)cases, neural network-based methods (skip-gram(SG) and continuous bag of word (CBOW)) always outperform classic vector-based methods (vector space model(VSM) and latent semantic analysis (LSA)) when used unsupervised. In the TD scenario, supervised summarizers such as deep neural network (DNN), convolution neural network (CNN), and Refresh outperform SG and CBOW, as do most of them in the SD case. 10% was chosen as the summary ratio [17]. Sharma, R., Palaskar, S., Black, A. W., Metze, F. proposed three transformer-based modules: speech segmentation, speech recognition, and extractive text summarization. An extractive summarization module employs BERTSUM using the self-supervised learning model BERT for text summarization to select the most essential sentences from a text that contains recognized sentences. Each sentence is binary classified by the module, which then creates a summary that meets the desired repeatedly), and colloquial terminology in contrast to text input [3]. The automatic speech recognition (ASR) engine, which transcribes spoken documents into textual representation, is one of the most important parts of the speech summarizing technique. The ASR component is extremely dependent on spoken language, and academics all across the world have created language-specific ASR engines. Transcribing speech to text and then summarizing it, is the common two-stage process for speech summarization. Producing a summary directly from speech without having to transcribe it, is a different approach that could help to avoid some fundamental summarization issues [4].

summary rate [18]. Kano, T., Ogawa, A., Delcroix, M., & Watanabe, S. introduced a single model optimized end to end for speech summarization. The suggested model learns to directly summarize speech and it outperforms the earlier proposed cascaded model by three points. On ROUGE.ASR was used to pre train the sequence model, because training summarization models from start is difficult. Then, for speech summarization, the encoder-decoder model is fine-tuned.The effects of different window sizes and dilations on summarization were investigated, with the conclusion that larger window sizes are required for better models [19]. Dammak, N., &BenAyed, Y. presented the cascade connection of applying automatic speech recognition and text summarization that allows for the use of state-of-the-art modules which are optimized for each task separately, without the need for a large amount of paired data made up of speech data and associated summaries. Posterior probability fusion and Attention-based multi-hypothesis fusion approaches were used for speech summarization [20].Li, D., Chen, T., Tung, A., & Chilton, L. B. suggested a deep encoder-decoder model based on the attention mechanism (DEDA) for ASR transcripts. it takes advantage of the deep structure of RNNs based on a Long Short-Term Memory (LSTM) network. The key difference is to incorporate a powerful attention mechanism into the encoder-decoder structure to address the sequence problem in the summarization area. Experiments on the AMI Dataset show that the proposed strategy outperformed the state-of-the-art on both extractive and abstractive models. The performance of summarized utterances and the reduction of occurrence repetition in summaries, were also highlighted in the experimental analyses [21]. Because these studies applied their methods on different corpus, it is not possible to compare them accurately and identify which machine or deep learning algorithm was the best.

## 5.2 Source of Corpus

A corpus is a group of authentic text or audio that has been arranged into datasets. 'Authentic' in this context refers to text produced by a native speaker of the language or dialect or audio produced by that speaker. Newspapers, books, recipes, radio broadcasts, television shows, motion pictures, and tweets can all be included in a corpus.

A corpus of text and speech data used for natural language

processing can be utilized to train AI and machine learning systems[22].

The characteristics of the speech domain corpora include:

Corpus are (1) of different languages, (2) comprise one or two speakers, (3) range in size from small to moderate, and big [4].

Table 1 shows both Audio Speech and Multimodal Corpus for various studies which are publicly available.

**A. Audio Speech Corpus**

A Corpus of spontaneous Japanese (CSJ) was used which includes a training set with 3,212 spontaneous speeches from lectures and conferences, as well as three evaluation sets (eval1, eval2,and evel3). Every speech was recorded at a sampling rate of 16 kHz and a bit depth of 16[3]. In[7] the pyaudio module is used to record and process audio. Continuously adding frames of audio is used to record audio. This phase's ultimate output is an audio file in the wave format. The MATBN corpus, (which contains roughly 200 hours of Mandarin Chinese TV broadcast news) collected by Academia Sinica and the Taiwan Public Television Service Foundation between November 2001 and April 2003, was used[10]. A training set of 100k dialogues was used, followed by a validation set of 1k dialogues. The test set was produced from 490 multi-turn talks between nurses and patients in a healthcare setting[11]. In[12] the AMI meeting corpus consists of 142 meeting records(which are 100 dialogs in the training set, 20 dialogs in the validation set, and 22 dialogs in the testing set) and associated abstractive summaries written by humans .

The suggested approach is trained and tested using the CNN/Daily Mail summarization dataset. The documents and summaries from CNN news articles are included in this dataset[15]. A set of empirical experiments are done on a mandarin benchmark broadcast new (MATBN) corpus, while the training and validation sets were constructed using simulated data, the summarizing studies used a subset of 205 broadcast news documents collected between November 2001 and August 2002[17]. Three corpora were used to train and evaluate speech summarization systems. CNN-Daily Mail (CNNDM), How2 and TED corpus[19].The AMI meeting corpus was used(100 hours of meetings recorded using multiple synchronized recording devices).The AMI corpus contains ASR transcripts for 137 meetings that were spoken by four participants who were assigned certain roles. Each meeting lasts an average of 35 minutes and includes about 800 unprocessed utterances, or nearly 6700 words[20]

**B. Multimodal Corpus**

The total video corpus consists of roughly 300 hours of video, with approximately 100 hours in each of the languages (French, English and Arabic) in [14]. How-2 dataset includes 2000 hours of instructional videos, as well as text transcripts, speech, video, translations, and summaries was employed[18]. A cascaded multimodal abstractive speech summarization model was described, which learned semantic concepts as an intermediary step. The benefits of using multimodal inputs (How2 data set ) includes speech, video, human annotated transcription and a summary rather than unimodal inputs were tested for intermediate concepts, and consistent advantages were discovered[24].

The anchor person-based story identification and lexical chain algorithm were used to implement multimedia summarization of news broadcasts. This system allows users to create multimedia summaries of one or more input news broadcasts and search for and retrieve selected news stories using a keyword-based search and retrieval system[25].

**Table 1.** Corpora for Speech Summarization Studies that are publicly available.

| Corpus | Size & Material |
|---|---|
| **CSJ** | 3,212 of spontaneous speeches(lectures and conferences) |
| **CNN-DailyMail(CNNDM)** | news documents |
| **How** | Videos(YouTube) 2000 hours(instructional videos) |
| **TED** | TED Talks |
| **LinguisticData Consortium(LDC)** | 20 hours(broadcast news) |
| **MATBN** | 205 broadcast news, 200 hours of Mandarin Chinese TV broadcast news |
| **AMI** | 100 hours(meetings) |

## 5.3 ONE VERSUS TWO STAGE SUMMARIZATION PROCESS:

Most researches focus on a two-stage process for extracting summary output from an audio file input. Speech to Text conversion is the first step, followed by text summarization. However, the output's efficiency is determined by the audio file's clarity [7].In[8] the proposed method uses automated speech recognition (ASR) to transcribe the audio, then summarize the transcript, and lastly returns the audio associated with thetext summary. Speech summarization is accomplished by integrating two primary sub-modules: an automated speech recognition (ASR) module and a text summarization system (TS)[13].

In [23] two stage summarization process was implemented. First, the speech is converted into text using Pocket Sphinx engine for transcribing spontaneous speech using three models. An acoustic model which contains phones acoustic properties, a phonetic model which involves word–to-phone mapping and a language model which limits the matching process by defining which words can come after previously recognized terms, then natural language processing preprocessing approaches were performed.

Speech transcripts, on the other hand, may be costly, unavailable, or of poor quality, which has an impact on summarizing performance. This leads to one stage summarization process. The complexity of the model structure with cascade topologies, and faults in the ASR which reduces summarization performance offers motivation for end-to-end(E2E) modeling for speech summarization[18].

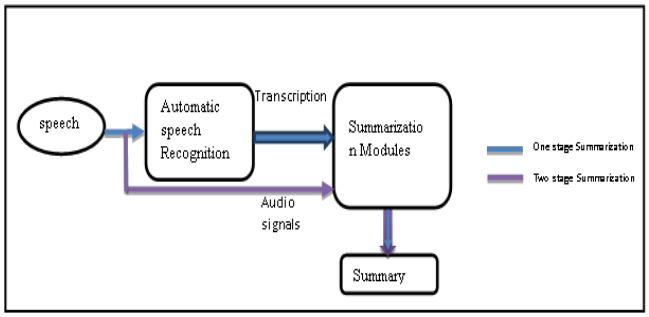


Fig1. ONE AND TWO STAGE SUMMARIZATION PROCESS[2]

## 6. DEEP LEARNING APPROACHES FOR SPEECH SUMMARIZATION

Deep learning approaches are now the most popular term in machine learning for extracting complicated data representations at a high level of abstraction, which is especially useful for exceedingly complex issues. It is a data-intensive method that produces better results than classic approaches (Naïve Bayes, HMM, SVM, and so on).

A Long Short Term Memory (LSTM) network architecture is a special type of RNN network that can learn long term dependencies. In more than 1000 steps, an LSTM can learn to fill the gap between time intervals. To handle information in both directions, bidirectional LSTM has two hidden layers. The input sequences are processed forward by the first hidden layer, and backward by the second hidden layer. Both are then linked to the same output layer, which gives them access to the sequence's future and past context. As a result, BLSTM outperforms both traditional LSTMs and RNNs, as well as providing a substantially faster and more accurate model.

The gated recurrent unit (GRU) is a recurrent neural network extension that tries to handle memories of sequences of data by storing the network's prior input state and planning to target vectors depending on the prior input.

A recurrent neural network (RNN), on the other hand, is a type of artificial neural network model in which the connections between the processing units create cyclic paths. It is recurrent because they take inputs, update the hidden layers based on previous calculations, and provide predictions for all sequence members.

In a traditional Recurrent Neural Network (RNN), data passes through only one layer on its way to the output layer before being processed. However, Deep Neural Networks (DNN) is a hybrid of deep neural networks (DNN) and RNNs[26].

## 7. EVALUATION METRICS

Since the early 2000s, there has been a set of measures for automatically evaluating summaries. The most extensively used metric for automatic evaluation is ROUGE.ROUGE comes in a variety of flavors, and the most popular ones are: ROUGE-n, ROUGE-L, ROUGE-SU+. Objective and subjective evaluation criteria were used in speech summarization. Objective evaluations includes ROUGE metric(ROUGE1, ROUGE2, ROUGE l) which are based on word and phrase overlaps in summary documents prepared automatically and manually, whereas subjective evaluations were used to assess the quality of the generated summaries based on interest, informativeness, abruptness, attractiveness, and overall quality. Users were particularly sensitive to the audio stream's linguistic coherence and continuity. The ROUGE measure, which includes ROUGE-1, ROUGE-2, and ROUGE-Lwas utilized as an evaluation metric in[10][11][17][19]. The suggested system's ROUGE value was 0.34343, which is likely comparable to other unsupervised summarization techniques like Lexrank and latent semantic analysis [15]. 1A subjective scaled opinion metric of 1-5 was used to assess the quality of the generated summaries and their components. Two objective metrics were also used: full score and average score metrics [16]. In order to obtain an overall score suitable for DEDA method, ROUGE scores for every meeting transcript in the test set were computed and then the macro-averaging method was used in [20]. ROUGE scores (ROUGE-1,ROUGE-2, ROUGE-SU) use n-gram overlap and skip-gram overlap to compare machine summaries to human gold-standard summaries were used in [12][13]. A BERT Score is used to assess coherence, while the cosine similarity between sentence transformer embedding of a reference ASR segment and a model generated output summary is employed to determine what information is kept in [21].

Fig2. Shows number of researches which were evaluated using subjective and objective measures, whereas Fig3. Shows Measures types used for Objective Evaluation.

TABLE 2. THE SURVEYED PAPERS CLASSIFIED BASED ON DIFFERENT CRITERIONS.

| **Source and Year of publication** | **Methodology Used** | **Corpus** | **One versus Two Stage Summarization** | **Evaluation Metric** |
|---|---|---|---|---|
| [10], 2012 | Summarization using single, multiple features and evaluation metric-related training criteria | MATBN | two stage summarization | ROUGE-1, ROUGE-2, ROUGE-L |
| [11], 2019 | PGN with attention mechanism | training set: 100k dialogues validation set: 1k dialogues | two stage summarization | ROUGE-1, ROUGE-2, ROUGE-L |
| [12], 2019 | a hierarchical neural encoder based on adaptive recurrent networks and a reinforced decoder network | AMI | two stage summarization | ROUGE scores(ROUGE-1, ROUGE-2, ROUGE-SU) |
| [13], 2019 | integer linear programming (ILP) | CSJ | two stage summarization | ROUGE scores(ROUGE-1, ROUGE-2, ROUGE-SU) |
| [14] 2019 | Artex algorithm | 300 hours of video, with approximately 100 hours in each of the languages (French, English and Arabic) | two stage summarization | Word Error Rate , Subjective Evaluation(score:1:5) |
| [15], 2020 | Essence Vector (EV) model Attention mechanism, LSTM | CNN/Daily Mail | two stage summarization | ROUGE |
| [16], 2020 | Probability Distribution Divergences | 5,989 audio broadcasts which corresponds to more than 310 hours of audio in French, English and Arabic | One stage summarization | A subjective scaled opinion metric of 1-5, full score and average score metrics |
| [17], 2020 | BERT-based method for supervised extractive speech summarization | ___ | two stage summarization | ROUGE-1, ROUGE-2, ROUGE-L |
| [3], 2021 | transformer based modules | CSJ | two stage summarization | ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-L |
| [18], 2021 | three transformer based modules | How-2 | One stage summarization | ROUGE, BERTS METEOR score |
| [19], 2021 | Encoder decoder model | CNN-DailyMail (CNNDM), How2 and TED corpus | one stage summarization | ROUGE-1, ROUGE-2, ROUGE-L |
| [20], 2021 | Posterior probability fusion and Attention-based multi-hypothesis fusion | AMI | two stage summarization | ROUGE, Macro-Averaging |
| [21], 2021 | deep encoder-decoder model based on the attention mechanism (DEDA) | AMI | two stage summarization | BERT, Cosine Similarity |

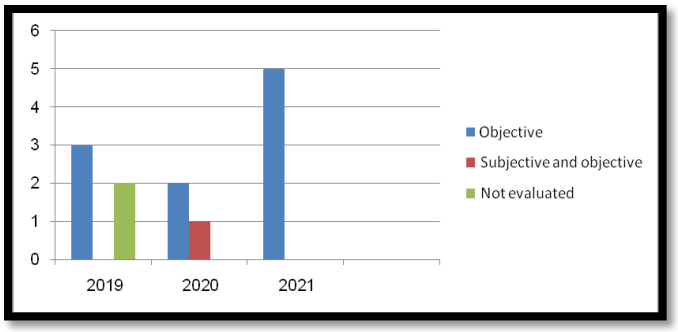


Fig2. NUMBER OF RESEARCHES EVALUATED OBJECTIVELY VERSUS SUBJECTIVELY SINCE 2019

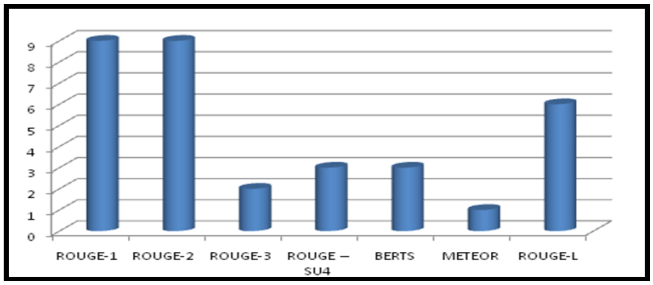


Fig3. MEASURES TYPES USED FOR OBJECTIVE EVALUATION

## 8. CHALLENGES AND BASIC LIMITATIONS FOR SPEECH SUMMARIZATION PROCESS

Regarding ASR, the basic difficulties and limitation are speech recognition errors, when tasks are well-defined single-speaker tasks, such as analyzing TED lectures, ASR error rates are reduced, but they remain difficult when tasks are less organized or involve multiple speakers, such as audio from meetings. To solve ASR output faults, machine learning, deep learning, and language models can be employed. Speaker turn identification is another problem for speech summarization. In accurate unit boundary detection can result in incorrect speaker identification across series of utterances. Although the number of speakers in BN is typically higher, speaker turns occur less frequently than in conference meeting data, resulting in a longer average speaker turn length in BN. Interruptions, overlapped speech, interleaved false starts, filler phrases (e.g. "of course", "ok", "you know"), non-lexical filled gaps (e.g. "umm", "uh"), and redundancies are all examples of speech disfluency. These disfluencies make identifying the semantic substance of speech more difficult, which can make summarization difficult. speech in broadcast news is the closest to organized text in terms of having the fewest disfluencies due to the presenters' professional training [4][27].

## 9. CONCLUSION AND FUTURE WORK

This survey paper emphasized various extractive and abstractive approaches for speech summarization. Most of the studies followed a two-step summarization process due to their commonly used and simplicity, while a few generates summary directly from speech. Deep learning and computer vision approaches can help in generating summary from speech without transcript it. In this survey, different studies which applied machine and deep learning models were analyzed. It is noticed an advance in the area of abstractive summarization with the presence of recent deep learning approaches. The source of materials in this survey were ranged from meeting, lectures, conferences and broadcast news with the focus on broadcast news because of their structural form in addition to ideal recording environment. No deterministic dataset was used and for the same dataset, different portions were employed for different studies, this leads to the difficulties in comparing these studies and evaluating them. It is recommended to employ existing datasets for latter studies and researches for accurate analysis and comparison.

**Acknowledgement**

The authors would express they're thanks to college of Computer Science and Mathematics, University of Mosul to support this report.

**References**

[1] Basanta Kumar Swaina, SanghamitraMohantyb, and DillipRanjanNayakc, "Extractive summarization of recorded Odia spoken feedback",Proceedings of 2ndInternational Conference on Artificial Intelligence and Speech Technology, (AIST2020), November 19–20, 2020, Delhi, India.

[2] Mythreagi. R1, Dr. N. Yuvaraj," Automatic Document Summarization Using Deep Learning Mechanismwith Competent Analysis", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 7 (2019) pp. 1709-1714 © Research India Publications. http://www.ripublication.com.

[3] Tomoki Hayashi, et. al. ,"Spontaneous Speech Summarization: Transformers All The Way Through", European Association for Signal Processing. (n.d.). 29th European Signal Processing Conference (EUSIPCO 2021) : proceedings : 23-27 August 2021, Dublin, Ireland.

[4] Rezazadegan, D., Berkovsky, S., Quiroz, J. C., Kocaballi, A. B., Wang, Y., Laranjo, L., &Coiera, E., "Symbolic and Statistical Learning Approaches to Speech Summarization: A Scoping Review", Computer Speech and Language 72 (2022) 101305 https://doi.org/10.1016/j.csl.2021.101305.

[5] Khan, S., Madane, A., Sayyed, N., Halallimath, K., &Deshmukh, S, "Review on Multimedia Summarization System using Machine Learning", International Research Journal of Engineering and Technology, 2021, www.irjet.net.

[6] https://towardsdatascience.com/audio-deep-learning-made-simple-automatic-speech-recognition-asr-how-it-works-716cfce4c706?gi=709f7fbc3a67.

[7] Khandare, P., Gaikwad, S., Kukade, A., Panicker, R., Thamke, S., "audio data summarization system using natural language processing". International Research Journal of Engineering and Technology, 2019.

[8] Vinnarasu, A.,Jose, D. v., "Speech to text conversion and summarization for effective understanding and documentation", 2019, International Journal of Electrical and Computer Engineering, 9(5), 3642–3648. https://doi.org/10.11591/ijece.v9i5.pp3642-3648.

[9] Basheer, S., Anbarasi, M., Sakshi, D. G., Vinoth Kumar, V. , "Efficient text summarization method for blind people using text mining techniques", International Journal of Speech Technology, 23(4), 713–725, 2020. https://doi.org/10.1007/s10772-020-09712-z.

[10] Chen, B., et al. , "Extractive speech summarization using evaluation metric-related training criteria", Information Processing and Management (2012), https://doi.org/10.1016/j.ipm.2011.12.002.

[11] Liu, Z., Ng, A., Lee, S., Aw, A. T., & Chen, N. F. (2019), "Topic-aware Pointer-Generator Networks for Summarizing Spoken Conversations", arXiv: 1910.01335v1 [cs.CL] 3 Oct 2019. http://arxiv.org/abs/1910.01335.

[12] Zhao, Z., Liu, Y., Pan, H., Li, L., Cai, D., Fan, C., & Yang, M. , "Abstractive meeting summarization via hierarchical adaptive segmental network learning". The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019, 3455–3461. https://doi.org/10.1145/3308558.3313619.

[13] Atsunori Ogawa, Tsutomu Hirao, Tomohiro Nakatani, and Masaaki Nagata, "Ilp-based compressive speech summarization with content word coverage maximization and its oracle performance analysis", IEEE international conference on acoustics, speech and signal processing : proceedings : april 15-20, 2019, calgarytelus convention center, calgary, Alberta, Canada.

[14] M Menacer, C González-Gallardo, K Abidi, Dominique Fohr, Denis Jouvet, et al. , "Extractive Text-Based Summarization of Arabic videos: Issues, Approaches and Evaluations". ICALP: International Conference on Arabic Language Processing, Oct 2019, Nancy, France. pp.65-78, 10.1007/978-3-030-32959-4_5, hal-02314238.

[15] Kumar, B. D. (n.d.), "Speech Summarization using Essence Vector Modeling", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 9, Issue 05, May-2020, https://www.ijert.org.

[16] González-Gallardo, C.-E., Deveaud, R., SanJuan, E., Torres-Moreno, J.-M., "Audio Summarization with Audio Features and Probability Distribution Divergence", arXiv:2001.07098v1 [cs.CL] 20 Jan 2020, http://arxiv.org/abs/2001.07098.

[17] Weng, S.-Y., Lo, T.-H., Chen, B., "An Effective Contextual Language Modeling Framework for Speech Summarization with Augmented Features", arXiv:2006.01189v1 [cs.CL] 1 Jun 2020,http://arxiv.org/abs/2006.011899.

[18] Sharma, R., Palaskar, S., Black, A. W., Metze, F. "End-To-End Speech Summarization Using Restricted Self-Attention", arXiv:2110.06263v1 [cs.CL] 12 Oct 2021, http://arxiv.org/abs/2110.06263.

[19] Kano, T., Ogawa, A., Delcroix, M., & Watanabe, S., "Attention-based Multi-hypothesis Fusion for Speech Summarization",arXiv:2111.08201v1 [eess.AS] 16 Nov 2021. http://arxiv.org/abs/2111.08201.

[20] Dammak, N., &BenAyed, Y. , "Abstractive meeting summarization based on an attentional neural model", Proc. of SPIE Vol. 11605, 1160504 · © 2021 SPIE · CCC code: 0277-786X/21/$21. https://doi.org/10.1117/12.2587172.

[21] Li, D., Chen, T., Tung, A., & Chilton, L. B. , " Hierarchical Summarization for Longform Spoken Dialog", UIST 2021 - Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology, 582–597. https://doi.org/10.1145/3472749.3474771.

[22] https://www.defined.ai/blog/the-challenge-of-building-corpus-for-nlp-libraries.

[23] HarikrushnaVanpariya , "Video summarization using a machine learning approach", International Journal of Scientific & Engineering Research Volume 10, Issue 2, February-2019 40, ISSN 2229-5518, http://www.ijser.org.

[24] Palaskar, S., Salakhutdinov, R., Black, A. W., Metze, F. , "Multimodal speech summarization through semantic concept learning", Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 5, 3826–3830, 2021. https://doi.org/10.21437/Interspeech.2021-1923.

[25] Mr. N.V. Bhalerao1, Dr. Mrs. S.S. Apte2, Prof. Mrs. A.R. Kulkarni, "Multimedia Summarization and Retrieval of News Broadcast", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, Volume: 05 Issue: 11, Nov. 2018.

[26] Chiche, A., Yitagesu, B., "Part of speech tagging: A systematic review of deep learning and machine learning approaches". In Journal of Big Data (Vol. 9, Issue 1), Springer Science and Business Media Deutschland GmbH, 2022, https://doi.org/10.1186/s40537-022-00561-y.

[27] Verma, P., &Verma, A. , "A Review on Text Summarization Techniques", Journal of Scientific Research, 64(01), 251–257, 2020. https://doi.org/10.37398/jsr.2020.640148.

# دراسة مقارنة لتلخيص الكلام على أساس التعلم الآلي: دراسة استقصائية

**هبه أدريس يونس**
hibaadreese@uomosul.edu.iq

**يسرى فيصل الأرحيم**
Yusrafaisalcs@uomosul.edu.iq

**قسم علوم الحاسوب**
**كلية علوم الحاسوب والرياضيات**
**جامعة الموصل، موصل، العراق**

**تاريخ الأستلام:** 30/6/2022 **تاريخ القبول:** 13/11/2022

## الملخص

إن الجانب الأكثر أهمية في التواصل البشري هو الكلام. تستغرق وسائل الإعلام الطويلة مثل الكلام وقتا طويلا للقراءة والفهم. يتم حل هذه الصعوبة من خلال تقديم ملخص مختصر للكلام مع الحفاظ على الدلالات. يمكن لتلخيص الكلام إما تحويل الكلام إلى نص باستخدام التعرف التلقائي على الكلام (ASR) ثم إنشاء الملخص ، أو يمكنه معالجة إشارة الكلام مباشرة وإنشاء الملخص. تتناول هذه الدراسة الاستقصائية مجموعة متنوعة من الدراسات الحديثة التي استخدمت خوارزميات التعلم الآلي والعميق لتلخيص الكلام.حيث تناقش أدبيات تلخيص الكلام مع الأخذ بنظر الاعتبار القيود الزمنية ، منهجية البحث ، مع عدم الاهتمام بقواعد بيانات معينة للبحث في الأدب. نظرا لأن طرائق التعلم العميق الأحدث لم يتم تضمينها في الدراسات الاستقصائية السابقة ، فيعد هذا استطلاع جديد في هذا التخصص حيث تم استكشاف طرائق مختلفة مع مجموعات بيانات متنوعة لتلخيص الكلام وتقييمها باستخدام أساليب ذاتية أو موضوعية.

**الكلمات المفتاحية:** تلخيص الكلام ، التعلم بعمق ASR, LSTM, ROUGE