Computer Speech and Language 72 (2022) 101305

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Symbolic and Statistical Learning Approaches to Speech Summarization: A Scoping Review

Dana Rezazadegan [1,2,*], Shlomo Berkovsky [2], Juan C. Quiroz [3,2], A. Baki Kocaballi [4,2], Ying Wang [2], Liliana Laranjo [5,2], Enrico Coiera [2]

[1] *Department of Computer Science and Software Eng, Swinburne University of Technology, VIC, Australia*
[2] *Australian Institute of Health Innovation, Macquarie University, NSW, Australia*
[3] *Centre for Big Data Research in Health, University of New South Wales, NSW, Australia*
[4] *School of Computer Science, University of Technology Sydney, NSW, Australia*
[5] *Westmead Applied Research Centre, University of Sydney, NSW, Australia*

ARTICLE INFO

KEYWORDS:
Speech summarization
Spontaneous speech
Automatic speech recognition
Extractive summarization
Abstractive summarization

ABSTRACT

Speech summarization techniques take human speech as input and then output an abridged version as text or speech. Speech summarization has applications in many domains from information technology to health care, for example improving speech archives or reducing clinical documentation burden. This scoping review maps close to 2 decades of speech summarization literature, spanning from the early machine learning works up to ensemble models, with no restrictions on the language summarized, research method, or paper type. We reviewed a total of 110 papers out of a set of 188 found through a literature search and extracted speech features used, methods, scope, and training corpora. Most studies employ one of four speech summarization architectures: (1) Sentence extraction and compaction; (2) Feature extraction and classification or rank-based sentence selection; (3) Sentence compression and compression summarization; and (4) Language modelling. We also discuss the strengths and weaknesses of these different methods and speech features. Overall, supervised methods (e.g. Hidden Markov support vector machines, Ranking support vector machines, Conditional random fields) performed better than unsupervised methods. As supervised methods require manually annotated training data which can be costly, there was more interest in unsupervised methods. Recent research into unsupervised methods focusses on extending language modelling, for example by combining Uni-gram modelling with deep neural networks. This review does not include recent work in deep learning.

1. Introduction

Speech summarization seeks to identify the most important content within human speech and then to generate a condensed form, suitable for the needs of a given task. Summarized speech should also be more understandable than a direct transcript of speech, as it excludes the breaks and irregularities, as well as the repairs or repetitions that are common in speech (Goldman et al., 2005). The steady improvement in automatic speech recognition accuracy, audio capture quality, and the increased popularity of natural language

[*] Corresponding author:
E-mail addresses: drezazadegan@swin.edu.au, dana.rezazadegan@gmail.com (D. Rezazadegan).

https://doi.org/10.1016/j.csl.2021.101305

Available online 29 September 2021

as a computer interface has underpinned the recent growth in interest for speech summarization methods.

Speech summarization has been applied in various settings such as broadcast news, meetings, lectures, TED talks, conversations, and interviews (Hori et al., 2002; Xie and Liu, 2011; Fung et al., 2008; Beke and Szaszák, 2016). The benefits of speech summarization range from improved efficiency and cost reduction in telephone contact centres (e.g. by identifying call topics, automatic user satisfaction evaluation, and efficiency monitoring of agents) (Riccardi et al., 2015) to more efficient progress tracking in project meetings (Murray et al., 2010; Murray and Renals, 2008) and facilitation of learning using online courses (Zhang and Yuan, 2013; Zhang and Yuan, 2016). In healthcare, speech summarization has the potential to create a new generation of digital scribes (systems which generate clinical records from spoken speech) and conversational agents which can interact with patients (Finley et al., 2018; Coiera et al., 2018; Laranjo et al., 2018).

A speech summarization system takes speech as its input and generates a summary as its output (Fig. 1). Speech summarization usually involves a series of technical components. An Automatic Speech Recognition (ASR) component first generates a direct transcription from audio into text. Next, summarization modules (which may include sentence segmentation, sentence extraction, and/or sentence compaction sub-modules) summarize key parts of the transcription. Some summarizers skip the transcription stage with speech signal going directly into the summarization modules (Maskey and Hirschberg, 2006; Sert et al., 2008; Yella et al., 2010).

The two key approaches to speech summarization are extractive and abstractive summarization (Fig. 2). *Extractive summarization* identifies the most relevant utterances or sentences from speech or a document that succinctly describe the main theme (Banerjee and Rudnicky, 2008; Lin and Chen, 2010; Chen and Lin, 2012). It concatenates these into a coherent summary with or without applying compression. *Abstractive summarization* attempts to generate a fluent and concise summary, paraphrasing the intent, but not necessarily the exact content, of the original (Pallotta et al., 2009; Banerjee et al., 2015; Liu and Liu, 2009). Abstractive summarization is more challenging than extractive summarization because of the need to infer semantic intent, as well as the need for natural language generation (Pallotta et al., 2009; Liu and Liu, 2009).

Existing literature reviews of speech summarization methods are now more than a decade old (Furui, 2007; Furui and Kawahara, 2008) and do not include recent technical advances in machine learning, natural language processing (NLP), and speech and text generation. This review aims to synthesize the existing literature on machine learning methods for speech summarization and focuses on application domains and the speech features and training corpora used. More recent works using deep learning for speech summarization were beyond the scope of this review, as this flourishing research area deserves a separate review.

2. Search Methods

This scoping review follows the methodology outlined by Peters et al (Peters et al., 2015). The databases searched included IEEE, Springer, Science Direct, ACM, and PubMed. We searched for papers with the keywords “speech summarization”, “conversation summarization” and “meeting summarization” in the title and abstract, using both British and American spelling.

Primary research studies focusing on summarizing individual or multi-party speech such as lectures, news, meetings, and dialogues, were included. Research papers could be published any time up to July 2018 and be written in any language. Those papers which focused solely on speech recognition, speech analysis without summarization, and summarization of emails or written social media conversations were excluded. After duplicates removal, papers were screened by title and abstract against the inclusion criteria. Where information in the title and abstract was not sufficient to reach a decision, full-text screening was conducted. Fig. 3 shows the PRISMA flow diagram, with an initial set of 188 papers reducing to 110 studies which met our criteria (Moher et al., 2009). Data was extracted from these included studies by four researchers using a standard data extraction form (Electronic Supplements, A.4). From each paper we extracted the first author’s name, year of publication, title, application domain and task, training datasets (corpora), methods of summarization, speech features, evaluation metrics, study results and key findings.

3. Results

There was significant heterogeneity across the summarized content, application domain, technical architecture, methods, and evaluation metrics. The articles were also heterogenous in the choice of methods and speech features used, but concentrated on

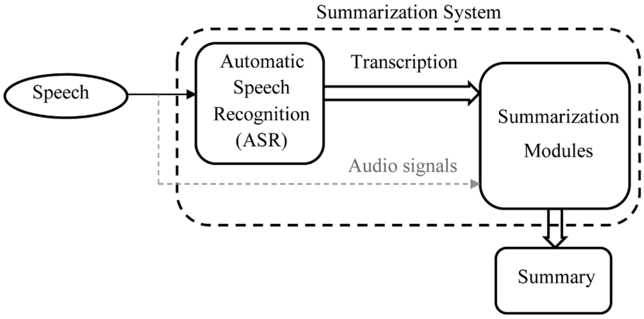
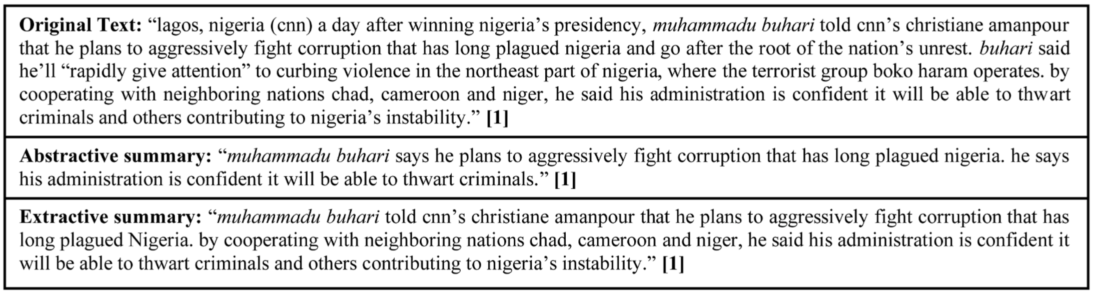


Fig. 1. General structure of speech summarization systems.



Original Text: "lagos, nigeria (cnn) a day after winning nigeria's presidency, *muhammadu buhari* told cnn's christiane amanpour that he plans to aggressively fight corruption that has long plagued nigeria and go after the root of the nation's unrest. *buhari* said he'll "rapidly give attention" to curbing violence in the northeast part of nigeria, where the terrorist group boko haram operates. by cooperating with neighboring nations chad, cameroon and niger, he said his administration is confident it will be able to thwart criminals and others contributing to nigeria's instability." **[1]**

Abstractive summary: "*muhammadu buhari* says he plans to aggressively fight corruption that has long plagued nigeria. he says his administration is confident it will be able to thwart criminals." **[1]**

Extractive summary: "*muhammadu buhari* told cnn's christiane amanpour that he plans to aggressively fight corruption that has long plagued Nigeria. by cooperating with neighboring nations chad, cameroon and niger, he said his administration is confident it will be able to thwart criminals and others contributing to nigeria's instability." **[1]**

Fig. 2. Text summaries may either be extractive (using key sentences verbatim) or abstractive (inferring the meaning of text). The original text and abstractive summary were selected from (See et al., 2017), while we highlighted the extractive summary to show the difference.

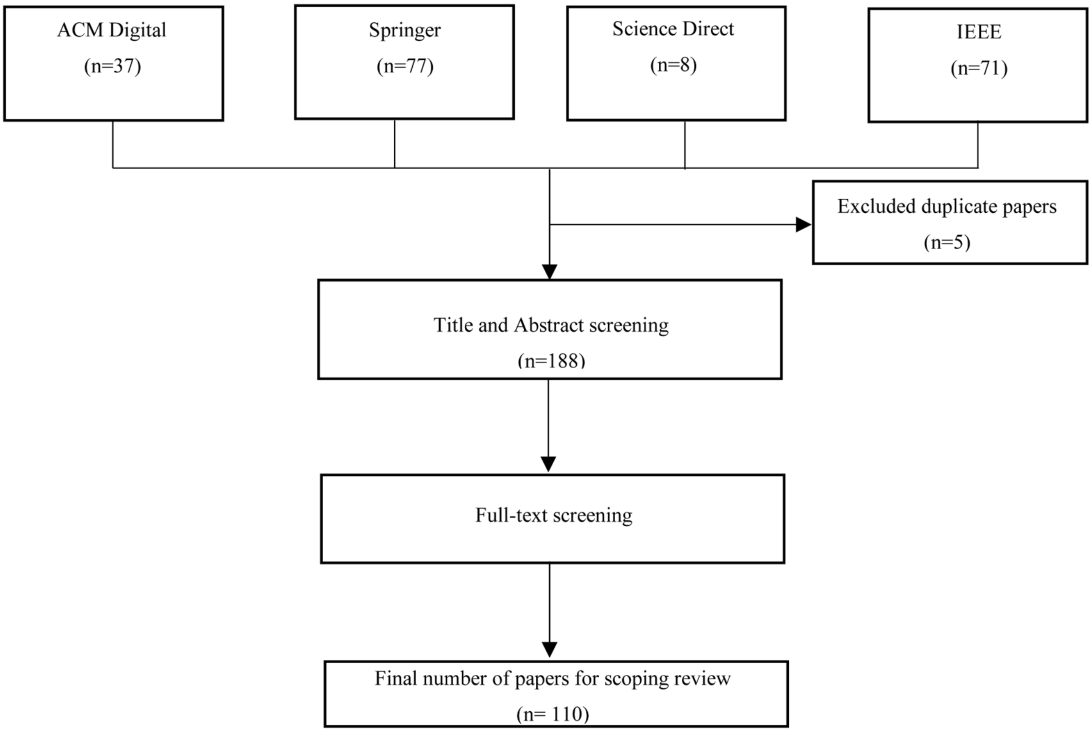


Fig. 3. Literature search results in PRISMA format.

broadcast news, lectures and talks, meetings, interviews, and spoken conversations. This important limitation makes quantitative meta-analysis of performance results impossible and only permits characterization of algorithms and architectures.

3.1 Application Domain

Most studies focused on summarizing broadcast news (43 studies) and meetings (41 studies) (Table 1). This may be due to the wide availability of public, labelled datasets for broadcast news and meetings compared to the other domains where dataset availability is

Table 1
Frequency of study domains for speech summarization studies

Scope/Year interval	2002-2006	2007-2010	2011-2014	2015-2018	Total
Broadcast news	12	7	12	12	43
Lectures/TED talks	4	6	6	1	17
Meetings	2	23	8	8	41
Conversation/ interview	1	1	3	4	9
Total	19	37	29	25	110

lower (Table 2, see Electronic supplements, A.1 for descriptions). Seventeen papers focused on summarizing lectures and TED talks. Although only nine studies applied speech summarization to conversations and interviews, the interest in this application class seems to have increased in recent years. Table 2 shows the corpora used in the reviewed papers, which are publicly available or can be provided upon request. Characteristics of available corpora in the speech domain include: Corpora mostly (1) are in English, (2) have one or two speakers in each session, (3) are of a small/moderate size and only three go over 500 hours (See Electronic Supplements, A.1 for more details).

3.2 Speech Features

Speech analysis can take advantage of a number of different features within an audio signal, and studies varied widely in the features used. Eight feature classes were identified (Table 3) with lexical, acoustic, and structural features most commonly used (Fig. 4).

Most studies did not provide a clear comparison of different feature types, and there was little consensus on which features were most useful for summarization. However, studies that used more than two feature types typically concluded that the best results could be obtained by combining different features (Fung et al., 2008; Murray and Carenini, 2008; Hasan et al., 2016; Zhang and Fung, 2007; Zhang and Yuan; 2014; Chen et al., 2013). Some studies claimed that the use of only one or two features was sufficient and performed on par with the combined use of lexical, acoustic, structural, and relevance features (Zhang and Yuan, 2013; Maskey and Hirschberg, 2006; Zhang and Fung, 2007; Zhang et al., 2010; Koto et al., 2014).

The importance of features differed by task and domain. For broadcast news summarization, lexical, acoustic, and structural features achieved the best performance (Chen and Lin, 2012; Hasan et al., 2016; Zhang and Fung, 2007; Zhang and Yuan, 2014) . Although, Chen et al. found these features to be complementary and reported the best results in combination (Chen and Lin, 2012), later on, they showed that using relevance features in isolation outperformed the combination of structural, lexical, and acoustic features (Chen et al., 2013). Lin et al. also found relevance features effective in summarization performance, whilst using the relevance features combined with lexical and acoustic features achieved the highest performance (Liu et al., 2014).

For lecture summarization, the best indicative features were relevance and discourse features that appeared as rhetorical units (underlying message in lecture speech and corresponding slides) (Fung et al., 2008; Zhang and Yuan, 2016; Lee et al., 2012; Zhang et al., 2008). These papers argued that the strong performance of relevance and discourse features is due to their resilience to problems like synonyms and recognition errors. Speaker-normalized acoustic features were found to be important when summarizing lectures because of the different speaking styles of speakers (Zhang et al., 2010; Zhang and Fung, 2009; Xie et al., 2009).

For meeting summarization, lexical features were identified as the most useful feature set in six papers (Banerjee and Rudnicky, 2008; Murray and Carenini, 2008; Galley, 2006; Murray and Renals, 2008; Liu and Xie, 2008; Tokunaga and Shimada, 2014). Five papers found Term Frequency - Inverse Document Frequency (TF-IDF) to be a highly effective lexical feature (Beke and Szaszák, 2016; Banerjee and Rudnicky, 2008; Banerjee and Rudnicky, 2009; Metze et al., 2013; Basu et al., 2008). Two other papers introduced the Speaker use - Inverse Document Frequency (SU-IDF) as a feature, where term weighting considers how variable is term usage across speakers in multi-speaker dialogue, and claimed that it works competitively with or better than TF-IDF (Murray and Renals, 2008; Murray and Renals, 2007). The combination of lexical, structural, and acoustic features was also found to achieve the highest score of Recall-Oriented Understudy for Gisting (ROUGE) in (Murray and Renals, 2008; Liu and Xie, 2008) (See Electronic Supplements, A.3. for definition of ROUGE as the most popular evaluation metric for summarization).

3.3 Speech Summarization Types – Extractive vs Abstractive Summarization

Extractive summarization methods were more popular than abstractive methods (99 out of 110 papers, Fig. 5). However, the rate of abstractive summarization usage increased 5.7% for 2007-2010 and 17.7% for 2011-2018.

Extractive summarization has been criticized for propagating word errors to the summary (Pallotta et al., 2009; Wang and Cardie,

Table 2
Publicly Available Corpora for Speech Summarization Studies

Corpus	Summarized Content	Language	No. of speakers	Size	Source
AMI	Meeting	English	More than 2	100 hours	(XX)
ICSI	Meeting	English	More than 2	70 hours	(YY)
MATRICS	Multimodal meeting	English	More than 2	10 hours	(Nihei et al., 2014)
TEDe	Lecture	English	1	50 hours+75 hours	(ZZ)+ (XXX)
CSJ	Lecture; Task-oriented dialogue	Japanese	1 or 2	658 hours	(YYY)
TDT2	Broadcast news	English	1 or 2	518-1036 hours	(ZZZ)
RT-03 MDE	Broadcast news+ Telephone Speech (a portion of switchboard)	English	1 or 2	20 hours+40 hours	(XXXX)+ (YYYY)
MATBN	Broadcast news	Mandarin	1or 2	198 hours	(ZZZZ)
ALERT	Broadcast news	Portuguese	1or 2	300 hours	(XXXXX)
Switchboard-1	Telephone Speech	English	2	260 hours	(YYYY)
Fisher	Telephone Speech	English	2	2000 hours	(YYYYY)
MAMI	Spoken word	English	1	-	(ZZZZZ)
BEA	Interview	Hungarian	2	250 hours	(XXXXXX)

Table 3
Summary of used features in speech summarization systems

Features	Definition	Examples	Frequency
Lexical (also called textual or linguistic in the literature)	Features extracted from the speech text, based on linguistic, lexical, syntax, and grammatical analysis, such as the words' size, type and semantic relationships.	• Number of words in current, previous, next utterance (Zhang and Yuan, 2014) • Number of stop-words • Number of NE, e.g. person, location, and organization names (Zhang and Yuan, 2014) • Number of NE which appear in the utterance at the first time in a story (Zhang and Yuan, 2014) • Ratio of the number of unique NE to the number of all NE (Zhang and Yuan, 2014) • TF • TF-IDF • POS • Sentiment polarity • Number of bi-gram • Cosine similarity between the current utterance and the entire speech • SU-IDF	85
Acoustic (also called prosodic or spectral in the literature)	Features extracted from the analysis of the audio, that can cover both features of the speaker (e.g. speaker's intention, emotion, mood, etc) and the utterance (e.g. the mode of the utterance: statement, question, command, duration, frequency, etc).	• Pauses between turns • Pitch • Intonation • Accent • Intensity • Log energy • Duration • Spectral characteristics • MFCC • Frequency of utterance (F0) • RMS slope • Speaking rate • The peak normalized cross-correlation of pitch	40
Structural	Features extracted from the structure of the utterances' transcriptions, including position of each utterance respect to t.	• Centroid scores • Length of current, previous, next utterance • Position of the turn in the overall speech	50
Relevance	Features extracted from the semantic similarity between all the utterances/document and each one of its utterances/sentences.	• Average similarity • Relevance score obtained by VSM, Relevance score obtained by WTM, Relevance score obtained by LSA, Relevance score obtained by MRW (Lin et al., 2010)	39
Discourse	Features that typically target the presence or absence of critical words showing a planned course of action, such as decide, discuss, result, conclude, and/or phrases such as "we should".	• Number of new discourse cue/clause • Discourse cue/clause position (first, second, other) (Wang and Cardie, 2012) • Position to the first discourse cue/clause (Wang and Cardie, 2012)	10
Visual	Features extracted from videos that capture body language.	• Visual semantic concepts • Objects • Actions • Visual attention (head-gaze target) • Body behaviour • Head motion • Hand gestures	4
Phonetic	Features extracted from the audio which is used for computing the similarity between sentences and sentence-like units (SUs), to overcome speech recognition errors and disfluencies (Ribeiro and de Matos, 2013).	• Type (vowel/constant) • Vowel length, height and front-ness • Lip rounding • Consonant type • Place of articulation • Consonant voicing • (All were taken from (Ribeiro and de Matos, 2013))	1

Abbreviations: NE: Named Entities; TF: Term Frequency; TF-IDF: Term Frequency - Inverse Document Frequency; POS: Part-Of-Speech; SU-IDF: Speaker use - Inverse Document Frequency; MFCC: Mel-Frequency Cepstral Coefficients; VSM: Vector Space model; WTM: Word Topic Model; LSA; Latent Semantic Analysis; MRW: Markov Random Walks. Speech Summarization Types, Architectures, Methods, and Evaluation metrics

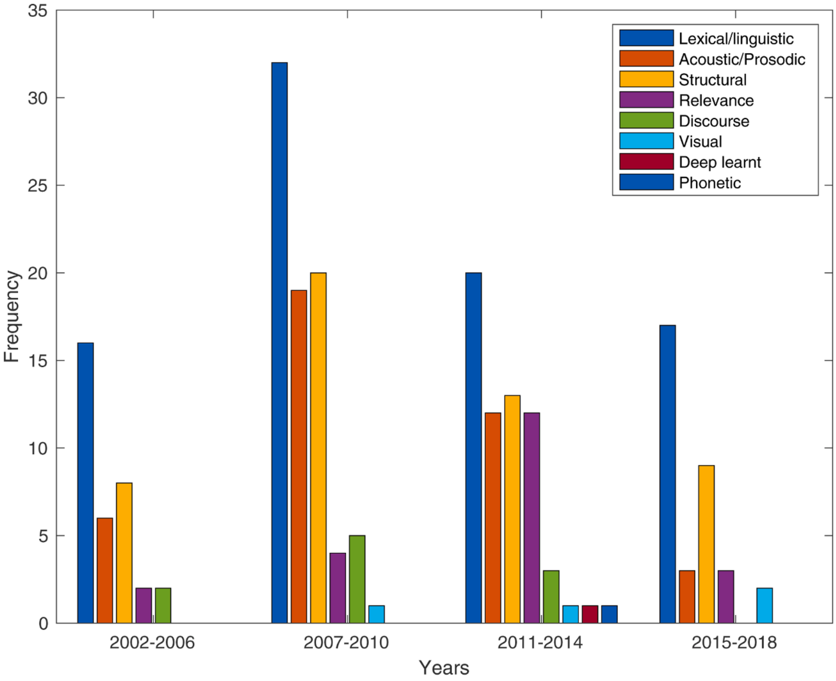


Fig. 4. Distribution of feature categories over time. There can be studies using more than one feature type.

2012), for summaries being repetitive (Liu and Liu, 2013) or containing incomplete sentences (Banerjee et al., 2015; Wang and Cardie, 2012; Liu and Liu, 2013), or for the creation of incoherent summaries (Wang and Cardie, 2012; Murray et al., 2010). Only one paper showed abstractive summarization performing better than extractive summarization, when evaluated using ROUGE (Murray et al., 2010). However, the results of this study are atypical, given that the ROUGE scores for both types of summarization were about 50% lower than reported by others. This may be because they used “human abstracts rather than human extracts” as reference summaries for the evaluation, and because the summarizers were configured to generate very short summaries that leaded to extremely low recall (Murray et al., 2010). Since other studies using abstractive summarization methods have not performed a direct comparison to the extractive methods or have not used standardized evaluation metrics, no clear conclusion regarding the performance of the two types of summarization can be drawn.

3.3.2 Format of output summary – Text vs Audio

Most papers (107 out of 110) generated a text summary, probably because of the convenience of using transcribed text as an intermediate representation prior to feature extraction and summarization (Furui, 2007). This is partly due to the wider selection of existing natural language processing techniques to pre-process the text and perform feature extraction (Banerjee and Rudnicky, 2008; Murray and Carenini, 2008; Galley, 2006; Murray and Renals, 2008; Liu and Xie, 2008; Tokunaga and Shimada, 2014).

Only three studies generated an audio summary, avoiding the need to have an intermediate step of transcribing speech and using text features (Maskey and Hirschberg, 2006; Sert et al., 2008; Flamary et al., 2011). These papers argued that there were advantages to preserving the original format of the data source, such as exploiting acoustic information, e.g. pauses, speaking style, and the emotion of speakers. Another reported benefit was avoiding possible speech recognition errors in the text. In general, there were two ways of

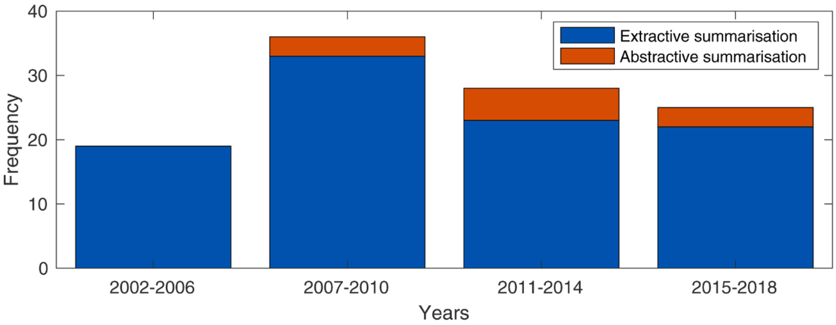


Fig. 5. Types of speech summarization-extractive vs abstractive.

audio summary generation: concatenating utterances that were extracted from the speech or generating audio from a text summary using a speech synthesizer (Furui, 2007). Although the former method is likely to generate unnatural sounds which is undesirable, it is more robust against ASR errors, compared to the latter.

3.3.3 Speech summarization architectures

Various summarization architectures have been applied to the speech summarization problem with four major approaches (Fig. 6): (1) sentence extraction and compaction, (2) binary classification of sentences or rank-based sentence selection, (3) sentence compression and summarization, and (4) language modelling.

We define several technical expressions used in the following sub-sections, in Table 4.

3.3.3.1 Sentence extraction and compaction. Sentence extraction and compaction (Fig. 6-a) was most popular between 2002 to 2006 and was used in six papers (Hori et al., 2002; Kikuchi et al., 2003; Hori and Furui, 2003; Hori et al., 2003; Chatain et al., 2006; Chatain et al., 2006). Using output generated by a speech recognizer, each sentence is scored using linguistic, significance, and confidence scores, with filler phrases removed. Then, sentence compaction is enforced to the most highly ranked sentences using dynamic programming. The summarization ratio is set experimentally by trading off the ratio of sentence extraction and sentence compaction to gain the highest summarization performance. Among the reviewed papers, there have been some observations on the performance and effect of different sentence scoring methods. Hori et al. argued that the linguistic score may reduce out-of-context word extraction from human disfluencies and recognition errors (Hori et al., 2002; Hori et al., 2003). However, only Huang et al. reported the linguistic score as the single most effective score for the summarization of the Mandarin language (Huang et al., 2005). Three papers found the significance score or its combination with the confidence score more important than the linguistic score, in achieving better results (Furui et al., 2004; Kikuchi et al., 2003; Hori et al., 2003), which was verified for the Japanese and English languages, but the effect of confidence score was higher for English. All the studies (Hori et al., 2002; Furui et al., 2004; Kikuchi et al., 2003; Hori et al., 2003; Huang et al., 2005) agreed that the summarization score could be improved by combining all three scores. Eight papers used dynamic programming for sentence compaction and finding the best summarization results (Hori et al., 2002; Furui et al., 2004; Hori et al., 2002; Kikuchi et al., 2003; Hori and Furui, 2003; Hori et al., 2003; Huang et al., 2005; Lee et al., 2017). Seven papers used only lexical features for sentence extraction and compaction (Hori et al., 2002; Kikuchi et al., 2003; Hori and Furui, 2003; Hori et al., 2003;

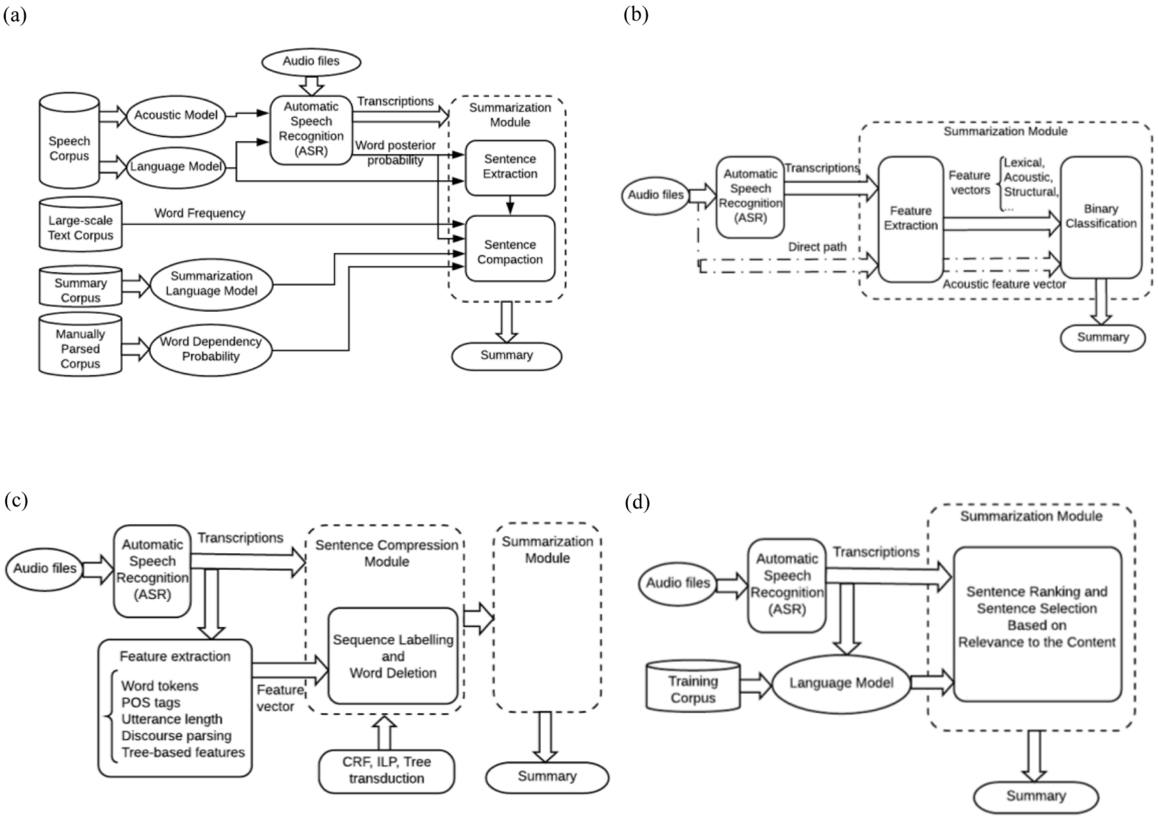


Fig. 6. Four different architecture for speech summarization: (a) Two-stage methodology based on sentence extraction and compaction (Furui et al., 2004); (b) Feature extraction and binary classification; (c) Two-stage methodology based on sentence compression and summarization; (d) The summarization task based on language modelling

Table 4
Definitions of technical terms used in different architectures

Technical term	Definition
Filler phrases (FPs)	Words that can be removed without losing any information, including "discourse markers (e.g., "I mean", "you know"), editing terms" (Liu and Liu, 2013), and commonly used terms (e.g. "for example", "of course," and "sort of") (Liu and Liu, 2013).
Summarization ratio	"The ratio of the number of words in the automatic summary to that in the original transcript of a spoken document" (Hori et al., 2002).
Summarization scores	Numerical scores, including "Linguistic score, significance score and confidence score", that indicate the appropriateness of a summarized sentence through the ranking process. Please see (Kikuchi et al., 2003) for details.
Sentence ranking	A ranking function which puts different weights on the summarization scores, and then combines them.

Chatain et al., 2006; Chatain et al., 2006; Mrozinski et al., 2006), whereas two others used both lexical and acoustic features (Hori et al., 2002; Furui et al., 2004).

3.3.3.2 Binary classification of sentences or rank-based sentence selection. Fig. 6-b illustrates the process of summarization by feature extraction and binary classification. This casts the problem of summarization as a binary classification problem, where a classifier determines whether each sentence, phrase, or speaker turn should be extracted for the summary. This approach relies on supervised learning. Simple machine learning methods, binary support vector machines (SVM) in particular, were popular until 2009, with recent papers using them for baseline comparisons (Fung et al., 2008; Chen and Lin, 2012; Hasan et al., 2016; Chen et al., 2013; Zhang et al., 2010; Liu et al., 2014; Tokunaga and Shimada, 2014; Lin et al., 2010; Wang and Cardie, 2012; Lin et al., 2011; Lo et al., 2012; Parthasarathy and Hasan, 2015; Liu et al., 2015; Liu et al., 2015; Liu et al., 2017; Xie et al., 2008; Xie and Liu, 2010; Juneja et al., 2010; Kim and Rudin, 2014). Among the reviewed papers, the following methods were used: simple SVM (Fung et al., 2008; Murray and Carenini, 2008; Zhang and Fung, 2007; Zhang and Yuan, 2014; Zhang et al., 2008; Xie et al., 2009; Metze et al., 2013), Ranking SVM (Lin et al., 2010), hidden Markov SVM (HMSVM) (Zhang and Yuan, 2016; Zhang and Yuan, 2014; Zhang and Fung, 2010; Zhang and Fung, 2010), Bayesian classifier (McKeown et al., 2005), multi-layer perceptron (Christensen et al., 2003; Christensen et al., 2004), conditional random field (CRF) (Zhang and Yuan, 2016; Zhang and Yuan, 2014; Galley, 2006), logistic regression (Zhu and Penn, 2006), random forests (Nihei et al., 2016), and decision trees (Banerjee and Rudnicky, 2008). Zhang and Yuan compared three classifiers and found CRF to outperform HMSVM and binary SVM (Zhang and Yuan, 2014).

Imbalanced datasets were considered an issue due to insufficient number of positive class samples (summary sentences) and data resampling was used in three papers to deal with imbalanced datasets (Xie et al., 2008; Xie and Liu, 2010; Juneja et al., 2010). Two studies handled the imbalanced data problem with Ranking SVM and AdaRank summarizers using training criteria based on evaluation metric, rather than sets of features (Chen et al., 2013; Lin et al., 2010). These improved the summarization performance by maximizing the correlated evaluation score or by optimizing an objective function linked to the final evaluation. Eight papers studied the ranking score method (rather than binary classification) after the feature selection process. This method chooses the highest ranked sentences according to their features' scores (Beke and Szaszák, 2016; Murray and Renals, 2008; Murray and Renals, 2007; Furui et al., 2004; Chatain et al., 2006; Chatain et al., 2006; Mrozinski et al., 2006; Cheng et al., 2014). Key phrase extraction was applied as the initial step for summarizing broadcast news (Huang et al., 2005; Marujo et al., 2012), lectures (Lee et al., 2014), spoken conversations (Basu et al., 2008), and meetings (Lee et al., 2014; Gillick et al., 2009; Riedhammer et al., 2008; Riedhammer et al., 2010).

3.3.3.3 Sentence compression and summarization. Fig. 6-c illustrates the process of summarization starting with sentence compression followed by the final summary generation (Liu and Liu, 2010). Sentence compression has been formulated as a word deletion and sequence labelling task, with the aim of retaining the most important information in the form of grammatical sentences (Liu and Liu, 2010; Liu and Liu, 2013). Both supervised and unsupervised methods have been used for this purpose. Two papers exploited CRFs to compress the spoken utterance through a sequence labelling task while effectively integrating acoustic features (Liu and Liu, 2010; Liu and Liu, 2013). A filler phrase detection module and Integer Linear Programming (ILP) have been also deployed, without the need for human annotations (Liu and Liu, 2013). The sentences were first compressed and then summarized using maximum marginal relevance (Liu and Liu, 2010; Liu and Liu, 2013). One paper applied compression methods, such as Integer Programming (IP) and Markovization of Synchronous Context Free Grammar, to extractive summaries to generate abstractive summaries (Liu and Liu, 2009).

3.3.3.4 Language modelling. Language modelling (LM) has attracted more attention since 2014. LM is an approach ranking each sentence or utterance of a document, based on the individual word frequency, semantic relationship, or sentence position in a document. It can be deployed either independently or in conjunction with classification methods and autoencoders (Fig. 6-d). With unsupervised extractive summarization, LM was used to select sentences, mostly from the transcriptions generated by ASR engines (Lin et al., 2011; Liu et al., 2015; Chen et al., 2013). This approach used a probabilistic generative paradigm to rank every sentence of an utterance and build a unigram language model (ULM) in accordance to the individual word frequency in the sentence. However, it has not leveraged long-span context dependence cues that could render another proof for existing the semantic relationships among words or between a given sentence and the whole document. This is taken into account in recurrent neural network LM (RNNLM) as a promising modelling framework for speech recognition (Mikolov and Zweig, 2012). An RNNLM framework, consisting of the input, hidden and output layers, has been explored for sentence modelling formulation in LM-based summarization approaches (Chen et al., 2015; Chen et al., 2014).

Position-aware LM for extractive summarization was introduced in (Liu et al., 2017). The motivation behind this method was the assumption that a sentence at a particular position of the transcription might contain more important content or might be resided closer to the parts of the transcription that are richer in content than the rest, and consequently these sentences are better candidates for the summary. The authors tested four types of position-based LMs and the passage-based model was found to outperform others (Liu et al., 2017). Translation-based LM was applied to score a document based on a translation model (Chen et al., 2016). This LM approach calculates matching degree between a word in a sentence and semantically similar words in the spoken document and scores the sentence, accordingly (Chen et al., 2016).

3.3.4 Methods of summarization

Machine learning methods applicable to the summarization task can divided into three categories: supervised (sample labels required for model training), unsupervised (no labels required), and semi-supervised (small set of sample labels required)[1]. We refer readers to Electronic Supplements, A.2 to see examples of supervised, unsupervised, and semi supervised methods, supported by methods' descriptions. The distribution of the utilized methods over time (2002-now) is shown in Fig. 7.

It shows supervised and unsupervised methods were used about the same. Supervised methods with balanced labelling of the summary and non-summary classes achieved a higher ROUGE score than unsupervised methods (Chen et al., 2013, Parthasarathy and Hasan, 2015). Only two studies used semi-supervised learning (Zhang and Fung, 2012; Zhang and Yuan, 2014).

Directly comparing the performance of the deployed methods and frameworks was challenging, because the proposed methods were tested using different types and volumes of data with varying assumptions, restrictions, and ASR engines. Here, we only report on the main methods applied in the reviewed papers, and group them based on the domain of summarization.

Broadcast news

Supervised methods such as AdaRank, Ranking SVM, and traditional SVM with balanced labels achieved a higher performance than unsupervised methods including MRW, VSM, WTM, LSA (Chen et al., 2013). For instance, Kullback-Leibler divergence Measure (KLM) and SVM outperformed graph-based and vector space-based methods (Liu et al., 2015). Among the supervised methods, CRF, HMSVM, and Ranking SVM outperformed SVM (Zhang and Yuan, 2014; Parthasarathy and Hasan, 2015). Among the unsupervised methods, Uni-gram and/or Bi-gram language modelling did not beat neither the graph-based methods e.g. LexRank and MRW, nor the combinational optimization methods such as Submodularity (SM) and ILP (Chen et al., 2015; Chen et al., 2014). Nevertheless, combining Recurrent Neural Network and Uni-gram language modelling techniques for extractive summarization outperformed all the previously-mentioned methods (Chen et al., 2015; Chen et al., 2014). Density peaks clustering algorithm demonstrated better results than graph-based, vector-based, ULM, and word embedding-based methods (Chen et al., 2015).

Lectures

Lecture summarization initially exploited methods based on sentence ranking (Furui et al., 2004; Chatain et al., 2006; Chatain et al., 2006; Mrozinski et al., 2006). However, recently there has been a large number of papers that emphasized the usefulness of rhetorical information in a shallow or deep structure (Fung et al., 2008; Zhang and Yuan, 2016; Zhang et al., 2010; Zhang et al., 2008; Zhang and Fung, 2010; Zhang and Fung, 2010). HMM states and SVM classifiers have been a common trend in these papers, compared to using PLSA (Lee et al., 2012; Lee et al., 2014) and leveraging active learning combined with SVMs (Zhang and Fung, 2009; Zhang and Fung, 2012).

Meetings

In the reviewed papers that summarized meetings, the focus has mostly been on the finished actions and items/decisions as the key parts of the output summary (Murray and Renals, 2008; Murray et al., 2006). Some researchers leveraged personal notes of meeting participants as potential predictors of important meeting parts, through abstractive (Bothin and Clough, 2012) or extractive summarization process (Banerjee and Rudnicky, 2008; Banerjee and Rudnicky, 2009), using a binary (summary, non-summary) or ternary (definitely show in the summary, maybe show, do not show) labelling. Graph-based methods achieved a better performance than vector-based methods and combinational optimization methods (Bokaei et al., 2016), while the positive effect of word stemming and noise filtering for utterances has been shown to achieve a better performance on ASR transcripts than manual ones (Chen and Metze, 2012). Resampling techniques have been explored and demonstrated good results, using SVM and learning-based sampling method (Juneja et al., 2010).

There have also been attempts to conduct abstractive meeting summarization. One way was to apply sentence compression to extractive summaries, using integer programming and Markovization of Synchronous Context Free Grammar (Liu and Liu, 2009). Others focused on the combination of compression and abstractive synthesis (Murray, 2015), manually creating abstractive summaries (Bothin and Clough, 2012) and using a two-stage method on ILP/CRF based sentence compression and MMR-based summarization (Liu and Liu, 2013). Another approach relied on discourse relations, in order to learn a general semantic structure and link the turns in a conversation (Pallotta et al., 2009).

[1] We remind that deep learning methods were left beyond the scope of this review.

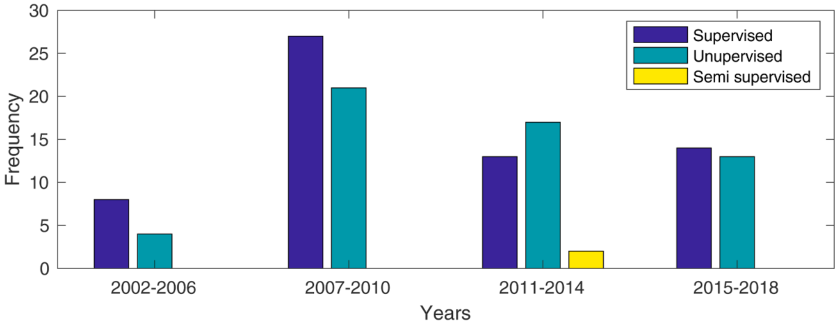


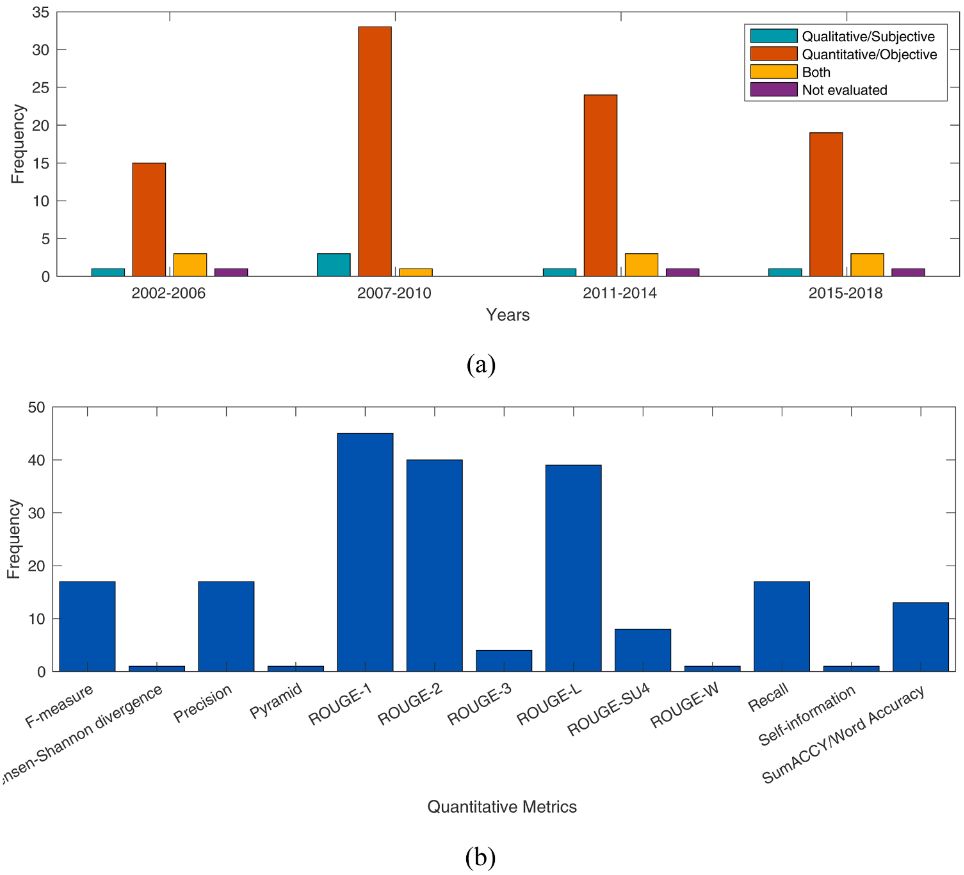
Fig. 7. Utilized methods of speech summarization, over time

3.3.5 Evaluation metrics

In this section, we report on the evaluation metrics used in studies and their strengths or limitations. A description of evaluation metrics including quantitative and qualitative metrics and the reported results are given in Electronic Supplements, A.3.

Automatically generated summaries have been evaluated using both subjective/qualitative metrics such as readability, coherence, usefulness, completeness, and objective/quantitative metrics such as ROUGE, Precision, Recall, F-measure, word accuracy, and Pyramid. Fig. 8-a shows the distribution of the qualitative and quantitative evaluation metrics over time, while the more prevalent quantitative metrics are further broken down in Fig. 8-b.

Among the quantitative metrics, ROUGE and its variants were dominant. ROUGE has been criticized for only performing string matching between the summaries without incorporating the meaning of words or words sequences (Lloret et al., 2018; Sjöbergh, 2007). Therefore, it is possible to get a high ROUGE score for a poor summary (Sjöbergh, 2007). Furthermore, obtaining perfect ROUGE scores may be impossible even for humans (Schluter, 2017). The correlation between a ROUGE score and human evaluations was lower than expected for meeting summaries, although ROUGE-SU4 showed a better correlation than ROUGE-1 (Liu and Liu,



(a)

(b)

Fig. 8. (a) Distribution of the utilized evaluation metrics; (b) Distribution of quantitative evaluation metrics

2008). These scores could be improved by removing disfluencies and exploiting speaker information (Liu and Liu, 2008). One paper reported that weighted precision, recall, and F-measure are more stringent and stable evaluation mechanisms than ROUGE for meeting summarization (Murray et al., 2006).

Two papers recommended the use of extrinsic metrics over intrinsic ones. Intrinsic metrics evaluate summaries based on how well the information content can match the reference summaries' content (Murray et al., 2009). Extrinsic evaluation not only aims to find the informativeness of the summary, but also to understand its usefulness in including an actual information which is required (Murray et al., 2009). In other words, intrinsic evaluation focuses on informativeness and coherence of output summaries, whereas extrinsic evaluation targets their utility in a given application (Lloret et al., 2018). Decision audit is an example of applying this to summarize useful information related to meeting decision making (Murray et al., 2009).

One common observation amongst the reviewed studies was that human evaluation can be subjective, either as a reference summary to be used in a quantitative metric or as a direct qualitative metric. Since this affects the evaluation of automatically generated summaries, using multiple reference summaries written by different human subjects was proposed as a way to overcome this problem, although precision and recall can fail for evaluating the automatically generated summary (Jing et al., 1998). However, this is still subjective and it requires extra analysis of the agreement among human subjects (Liu and Liu, 2008). A good example of examining the consistency of human evaluation and the critical factors for human agreement was presented in (Liu and Liu, 2008). Through several measurements, e.g. Kappa coefficient, ROUGE and their proposed sentence distance score and divergence, it has been shown that the number of speakers could be related to the human subjects' agreement, while the speech length was found non-critical. Another observation was the lack of agreement between subjective and objective evaluations on the performance of lexical and acoustic features, for lecture summarization, possibly due to the relatively large number of fillers included in a lecture (Furui et al., 2004).

4. Discussion

4.1 Main findings

This scoping review identified several important gaps in the speech summarization research literature. Although various techniques for speech summarization have been proposed, there is still a considerable gap between the quality of automatic speech summarization and manual summarization by humans. Despite their potential usefulness, there has been little research on abstractive summarization. This is partially due to the lack of suitable resources, corpora, and reference summaries in the speech domain. Another gap is the scarcity of extrinsic or task-based evaluations, which indicates that most studies focussed on traditional summarization without paying attention to the usefulness for a specific task. The use of different corpora or different batches of the same corpus makes replication and comparison across studies difficult.

Factors such as audio quality, structured speech, and number of speakers, affect the quality of the speech-to-text conversion, selection of methods and/or features, and the overall quality of summarization (Furui and Kawahara, 2008).

In broadcast news, the audio is recorded under ideal acoustic conditions and the speech is produced by professional following a structured script. These factors facilitate low word error rate in ASR, also making it easier to summarize (McCallum et al., 2012). Lectures are less structured, speakers are usually not trained, and speaking styles and/or accents can vary widely. Conversations and meetings tend to be even less structured, include disfluencies, interruptions, multiple speakers, and grammatically wrong utterances.

In terms of the speech features used, there was substantial variation, suggesting that the choice of feature types depends on the task, dataset, method applied, and language characteristics.

In broadcast news, language modelling is increasingly popular, but not in other domains where the data is not structured. In lecture summarization, recent studies shifted from sentence ranking-based method (Furui et al., 2004; Chatain et al., 2006; Chatain et al., 2006; Mrozinski et al., 2006) to rhetorical information-based methods in a shallow or deep structure (Fung et al., 2008; Zhang and Yuan, 2016; Zhang et al., 2010; Zhang et al., 2008; Zhang and Fung, 2010; Zhang and Fung, 2010), due to their higher performance. However, there was one common observation about the superiority of relevance features over structural, lexical, and acoustic features used in isolation (Fung et al., 2008; Zhang and Yuan, 2016; Chen et al., 2013; Liu et al., 2014; Lee et al., 2012; Zhang et al., 2008; Marujo et al., 2012). This can be due to the nature of the "relevance features that capture the relevance of a sentence to the whole document and the relevance between sentences" (Chen et al., 2013), or can be associated with their lower vulnerability to problems like synonyms and speech recognition errors. In meeting summarization, graph-based methods showed a better performance than other methods. Although capturing short important utterances is still challenging for these methods, DNNs showed a strong potential to address this problem (Nihei et al., 2017).

4.2 Study Quality

There was significant heterogeneity across the reviewed articles in terms of language (English, Mandarin, Japanese, Portuguese, Persian, Hungarian, Turkish, etc.), content and size of the utilized datasets (some used their own collected data which is not publicly available, e.g. (Banerjee and Rudnicky, 2008; Cheng et al., 2014), some used different parts of the same public dataset (Chatain et al., 2006; Mrozinski et al., 2006)). Evaluation metrics used for assessments varied widely. This makes any quantitative meta-analysis of results impossible, and only allows the reporting of qualitative patterns across studies. This also means that there was little to no attempt by researchers to replicate earlier results.

Some researchers did not make their own collected data available, e.g. (Chen et al., 2016; Chen and Metze, 2012), or did not mention how they chose a portion of data from a public corpus (Chatain et al., 2006; Mrozinski et al., 2006). Therefore, reproducibility was a big limitation of several papers. For example, Chatain et al. used nine talks from the TED corpus and five news stories from the CNN corpus, without specifying the talks or stories. Chen et al. used 14000 text news documents (in addition to MATBN) for estimating the models' parameters in the paper, which was not released as a new corpus to enable replication of their work.

4.3 Strengths and Limitations

Our review has analysed speech summarization methods and features based on the domain, content, and type of data. We surveyed more than 100 papers deploying a range of methods and architectures, which is useful for selecting the right methodology for a specific domain. A major strength of our review is that we used a systematic approach based on published guidelines (Moher et al., 2009; Arksey and O'Malley, 2005) to review five large repositories of research papers. We provide the detailed information in Electronic Supplement, including the definitions of methods and evaluation metrics, a brief description of publicly available corpora, the table of quantitative evaluation results and the table of information extracted from the reviewed papers. However, due to the search strategy, this review may have omitted papers not indexed by the databases we searched.

This review does not include recent work published after our 2018 search and we do not include deep learning methods which are a rapidly emerging field that would require a separate and lengthy review (Dammak and BenAyed, 2021; Dammak and BenAyed, 2021; Goo and Chen, 2018; Koay et al., 2021; Liu et al., 2019; Manakul et al., 2020; Shang et al., 2018; Tardy et al., 2020; Weng et al., 2021; Zheng et al., 2020; Zheng et al., 2020). These deep learning methods are sufficiently mature to be included in commercial speech summarization systems (HATI; Reason8 is an AI assistant for managers and meetings; Mphasis).

Our analysis highlighted several findings around the prevalent domains, architectures, and methods for summarization. These are indicative of research trends in NLP and machine learning at large, which may guide the selection of summarization methods. However, we also observed a substantial variability across the experimental setting of the analysed papers, in particular in terms of the corpora, training and test datasets used, features used summarization, and importantly the evaluation metrics. Hence, the direct comparison of the obtained results and the performance analysis of the proposed methods in these papers is difficult.

4.4 Comparison to Prior Work

To the best of our knowledge, this is the first scoping review of automatic speech summarization. There has been only one survey paper on automatic speech summarization exploring different types of output summaries, features, methods, and evaluation metrics (Furui, 2007). That work considered a smaller number of papers published before 2006 and only investigated a two-stage summarization method containing key sentence extraction and sentence compaction. A follow-up review of spontaneous speech transcriptions was published by the same authors in 2008 and discussed speech corpora issues, speech recognition, acoustic models, speech structure extraction, and speech summarization (Furui and Kawahara, 2008). However, only a small part of the paper focused on speech summarization, explaining different types of speech summarization and reviewing methods for key sentence extraction and summary generation. More than a decade passed after the publication of these reviews, and since then new speech corpora in various domains and new methods of speech summarization have been proposed, which are captured in our scoping review. Also, we analyse the surveyed papers according to a wider range of criteria, thus, painting a more encompassing picture of the speech summarization research.

4.5 Open problems and challenges

Despite the progress to date in speech summarization, several open problems still require significant effort. Whilst some of these problems are shared with text summarization, many challenges are unique to the world of speech.

4.5.1 Speech Recognition Errors

Despite rapid advances in automatic speech recognition, these technologies still suffer problems that have a flow on effect when summarizing speech content. Most summarization studies assume that sentences or spoken units are correctly rendered following transcription (Liu et al., 2015), but this may not be the case for transcripts produced by ASR. ASR engines can have error rates of up to 40%. (Liu and Liu, 2013). ASR error rates are lower for well-defined single-speaker tasks like interpreting TED talks but remain problematic, when tasks are less structured or involve multi-speakers e.g. audio from meetings. Machine learning, deep learning and language models can be used to correct ASR output errors. In general, language model-based methods such as ULM, RNNLM and their combinations are better able to handle of the effect of imperfect speech recognition on speech summarization compared to graph-based methods. RNNLMs with syllable level units for example, can convert ASR generated words into overlapping syllables to form a vocabulary of syllable pairs for indexing (Chen et al., 2015). In contrast, graph-based methods are particularly error-prone because of the poor performance of similarity measures like SM and ILP - used to compare pairs of sentences (Chen et al., 2015).

4.5.2 Speaker turn identification

Summarizing speech involving multiple speakers remains challenging. Diarization is the task of segmenting an audio stream into homogeneous units associated with different speakers. Deep learning techniques including those used by Google and Microsoft can alleviate the speaker diarization problem. However, none of the ASR engines, including commercial ones, has solved the diarization problem, perfectly (Sell et al., 2018; Ryant et al., 2019). Accurately detecting sentences or spoken unit boundaries is a particular challenge, as incorrect unit boundary detection may lead to incorrect speaker identification across a sequence of sentences. Summarization is also challenged when information is distributed across a sequence of sentences e.g. a question-answering sequence from multiple speakers, where answers may be short ("yes" or "no") and directly refer to the utterances of other speakers. As a result, the summarization process might exclude either the question or the answer in a question-answer pair because of apparent low value. There have been exploratory efforts to determine the relevance of spoken units, speaker turn identification, and creating continuous links between cross-speaker information and question-answer pairs, for example based on speaker activity or interaction and dialogue acts (Murray and Renals, 2008; Yella et al., 2010; Murray and Carenini, 2008; Murray et al., 2006).

4.5.3 Speech disfluency

In any conversation, it is common to have different kinds of disfluency, including interruptions, overlapped speech, interleaved wrong starts (e.g. "I'll, let's talk about it"), filler phrases (e.g. "of course", "ok", "you know"), non-lexical filled pauses (e.g. "umm", "uh") (Quiroz et al., 2019) and redundancies. These disfluencies complicate the identification of the semantics content of speech and consequently can impede summarization. Speech in broadcast news is the closest to structured text in having the lowest number of disfluencies, due to the professionally training of the presenters (McCallum et al., 2012). In meetings and interviews, disfluencies, filler phrases, redundancies and a lack of structure make them more challenging for summarization. Summarizing lectures with untrained speakers can also be problematic. Researchers have tried filler phrase detection (Liu and Liu, 2013) and MMR to remove redundant information as an extra criterion (Liu et al., 2015; Bokaei et al., 2016), but with mixed success.

4.4.4 Direct speech summarization

Speech summarization typically follows a two-stage process of transcribing speech to text and then summarizing the transcription. Generating a summary directly from speech without transcription is an alternate pathway that may side-step some fundamental problems in summarization. A speech summary in this case would be an audio file that contains an extractive or abstractive concatenation of spoken words drawn from the original speech. Very few studies have explored the possibility of direct summarization from speech (Maskey and Hirschberg, 2006; Sert et al., 2008; Flamary et al., 2011), partly because speech to text is such a popular approach. Direct speech summarization in the WordCloud study clustered recurrent patterns in speech (Flamary et al., 2011). Hidden Markov Models using acoustic and prosodic features (Maskey and Hirschberg, 2006) have also been applied to identifying repetitive speech patterns, as have a combination of computer vision techniques and a similarity matrix of spectral features (Sert et al., 2008). Although performance of these systems is promising, there is space for new approaches. For example, once could allocate higher weights to the most frequent sound patterns using deep learning, computer vision techniques using spectrogram images or indeed, other audio signal representations.

4.4.5 Abstractive speech summarization

Because of the complexity of natural language, abstractive summarization is an open problem, even in the text domain (See et al., 2017; Nallapati et al., 2016; Rush et al., 2015). However, the task is harder in the speech domain because of the higher possibility of error propagation in transcriptions. Abstractive approaches to summarization have attracted more attention than extractive approaches. This in part is because of studies showing abstractive methods achieve superior ROUGE and readability scores compared to extractive approaches (Liu and Liu, 2009; Murray et al., 2010), as well as research that shows users prefer abstract summaries over extracts (Murray et al., 2010). Additionally, the need for real-time summarization is critical for some applications e.g. a healthcare digital scribe that aims to listen to and summarize conversations on the fly. In such cases a summarizer does not have access to the full final content of the conversation but may need to draw intermediate conclusions that require repair as more information becomes available as a conversation or speech progresses. It may be that RNNs and LSTM architectures along with language models will be of value in tackling these abstractive challenges, given their potential to discover semantic relationships between sequences of utterances needed to generate meaningful summaries (Wu et al., 2019).

5. Conclusion

We framed our scoping review of speech summarization methods across four main domains (broadcast news, lectures, meetings and spontaneous conversation/interview) and also reviewed publicly available training corpora. As sentence classification using feature vectors is so prominent in the literature, we identified widely used features and discussed their performances. Methods in the review were also distinguished by output (text/speech), use of extractive or abstractive methods, and architecture.

Speech summarization is a fast-growing field of research that has the potential to contribute to many application domains and tasks. At present however, the evidence for their effectiveness remains limited. The wide variety of approaches, tasks and study designs limits our ability to reliably compare the effectiveness of much of the published research. For this reason, future research should report in a more standardized way, and use standard public corpora to assist with performance comparisons.

Supplemenary Material

Supplemenary Material_V2.docx

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests

ACKNOWLEDGMENTS

This paper was funded by the National Health and Medical Research Council (NHMRC) grant APP1134919 (Centre for Research Excellence in Digital Health) and Program Grant APP1054146.

REFERENCES

Arksey, H., O'Malley, L., 2005. Scoping studies: towards a methodological framework. International journal of social research methodology 8 (1), 19–32.

Banerjee, S., Mitra, P., Sugiyama, K., 2015. Abstractive meeting summarization using dependency graph fusion. In: Proceedings of the 24th International Conference on World Wide Web. ACM.

Banerjee, S., Rudnicky, A.I., 2008. An extractive-summarization baseline for the automatic detection of noteworthy utterances in multi-party human-human dialog. In: Spoken Language Technology Workshop, 2008. SLT 2008. IEEE. IEEE.

Banerjee, S., Rudnicky, A.I., 2009. Detecting the noteworthiness of utterances in human meetings. In: Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Association for Computational Linguistics.

Basu, S., et al., 2008. Scalable summaries of spoken conversations. In: Proceedings of the 13th international conference on Intelligent user interfaces. ACM.

Beke, A., Szaszák, G., 2016. Automatic summarization of highly spontaneous speech. In: International Conference on Speech and Computer. Springer.

Bokaei, M.H., et al., 2016. Summarizing Meeting Transcripts Based on Functional Segmentation. In: IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 24, pp. 1831–1841.

Bothin, A., Clough, P., 2012. Participants' personal note-taking in meetings and its value for automatic meeting summarisation. Information Technology and Management 13 (1), 39–57.

Chatain, P., et al., 2006. Class model adaptation for speech summarisation. In: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. Association for Computational Linguistics.

Chatain, P., et al., 2006. Topic and stylistic adaptation for speech summarisation. In: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on. IEEE.

Chen, B., et al., 2013. Extractive speech summarization using evaluation metric-related training criteria. Information Processing & Management 49 (1), 1–12.

Chen, B., Chang, H.-C., Chen, K.-Y., 2013. Sentence modeling for extractive speech summarization. In: Multimedia and Expo (ICME), 2013 IEEE International Conference on. IEEE.

Chen, B., Lin, S.-H., 2012. A risk-aware modeling framework for speech summarization. IEEE Transactions on Audio, Speech, and Language Processing 20 (1), 211–222.

Chen, K.-Y., et al., 2014. A recurrent neural network language modeling framework for extractive speech summarization. In: Multimedia and Expo (ICME), 2014 IEEE International Conference on. IEEE.

Chen, K.-Y., et al., 2015. Incorporating paragraph embeddings and density peaks clustering for spoken document summarization. In: Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on. IEEE.

Chen, K.-Y., et al., 2015. Extractive broadcast news summarization leveraging recurrent neural network language modeling techniques. IEEE Transactions on Audio, Speech, and Language Processing 23 (8), 1322–1334.

Chen, K.-Y., et al., 2016. Novel Word Embedding and Translation-based Language Modeling for Extractive Speech Summarization. In: Proceedings of the 2016 ACM on Multimedia Conference. ACM.

Chen, Y.-N., Metze, F., 2012. Intra-speaker topic modeling for improved multi-party meeting summarization with integrated random walk. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics.

Cheng, D.-Y., et al., 2014. Designing and Implementing a Real-Time Speech Summarizer System. In: Computer, Consumer and Control (IS3C), 2014 International Symposium on. IEEE.

Christensen, H., et al., 2003. Are extractive text summarisation techniques portable to broadcast news?. In: Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on IEEE.

Christensen, H., et al., 2004. From text summarisation to style-specific summarisation for broadcast news. In: European Conference on Information Retrieval. Springer.

Coiera, E., et al., 2018. The digital scribe. npj Digital Medicine 1 (1), 58.

Dammak, N., BenAyed, Y., 2021. Abstractive meeting summarization based on an attentional neural model. In: Thirteenth International Conference on Machine Vision. International Society for Optics and Photonics.

Finley, G., et al., 2018. An automated medical scribe for documenting clinical encounters. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations.

Flamary, R., Anguera, X., Oliver, N., 2011. Spoken wordcloud: Clustering recurrent patterns in speech. In: Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on. IEEE.

Fung, P., Chan, R.H.Y., Zhang, J.J., 2008. Rhetorical-state hidden Markov models for extractive speech summarization. In: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. IEEE.

Furui, S, 2007. *Recent advances in automatic speech summarization.* in *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*. In: LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

Furui, S., et al., 2004. Speech-to-text and speech-to-speech summarization of spontaneous speech. IEEE Transactions on Speech and Audio Processing 12 (4), 401–408.

Furui, S., Kawahara, T., 2008. Transcription and distillation of spontaneous speech. Springer Handbook of Speech Processing. Springer, pp. 627–652.

Galley, M., 2006. A skip-chain conditional random field for ranking meeting utterances by importance. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Gillick, D., et al., 2009. A global optimization framework for meeting summarization. In: Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE.

Goldman, J., et al., 2005. Accessing the spoken word. International Journal on Digital Libraries 5 (4), 287–298.

Goo, C.-W., Chen, Y.-N., 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In: 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE.

Hasan, T., et al., 2016. Automatic composition of broadcast news summaries using rank classifiers trained with acoustic and lexical features. In: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE.

HATI, Y., *Lando: Deep Learning used to summarize conversations.*
Hori, C., et al., 2002. Automatic speech summarization applied to English broadcast news speech. In: Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on. IEEE.
Hori, C., et al., 2002. Automatic summarization of english broadcast news speech. In: Proceedings of the second international conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc.
Hori, C., et al., 2003. A statistical approach to automatic speech summarization. EURASIP Journal on Applied Signal Processing 128–139, 2003.
Hori, C., Furui, S., 2003. A new approach to automatic speech summarization. IEEE Transactions on Multimedia 5 (3), 368–378.
Huang, C.-L., Hsieh, C.-H., Wu, C.-H., 2005. Spoken document summarization using acoustic, prosodic and semantic information. In: Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on. IEEE.
Jing, H., et al., 1998. *Summarization evaluation methods: Experiments and analysis*. in *AAAI symposium on intelligent summarization*. , Palo Alto, CA.
Juneja, V., Germesin, S., Kleinbauer, T., 2010. A learning-based sampling approach to extractive summarization. In: Proceedings of the NAACL HLT 2010 Student Research Workshop. Association for Computational Linguistics.
Kikuchi, T., Furui, S., Hori, C., 2003. Automatic speech summarization based on sentence extraction and compaction. In: Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on. IEEE.
Kim, B., Rudin, C., 2014. Learning about meetings. Data mining and knowledge discovery 28 (5-6), 1134–1157.
Koay, J.J., et al., 2021. A Sliding-Window Approach to Automatic Creation of Meeting Minutes. arXiv preprint arXiv:2104.12324.
Koto, F., et al., 2014. The use of semantic and acoustic features for open-domain TED talk summarization. In: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific. IEEE.
Laranjo, L., et al., 2018. Conversational agents in healthcare: a systematic review. Journal of the American Medical Informatics Association 25 (9), 1248–1258.
Lee, C.-S., et al., 2017. FML-based robotic summarization agent and its application. In: Fuzzy Systems (FUZZ-IEEE), 2017 IEEE International Conference on. IEEE.
Lee, H.-y., et al., 2014. Spoken knowledge organization by semantic structuring and a prototype course lecture system for personalized learning. In: IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 22, pp. 883–898.
Lee, H.-y., Chen, Y.-n., Lee, L.-s., 2012. Utterance-level latent topic transition modeling for spoken documents and its application in automatic summarization. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE.
Lin, S.-H., et al., 2010. Leveraging evaluation metric-related training criteria for speech summarization. In: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE.
Lin, S.-H., Chen, B., 2010. A risk minimization framework for extractive speech summarization. In: Proceedings of the 48th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics.
Lin, S.-H., Yeh, Y.-M., Chen, B., 2011. Leveraging Kullback–Leibler Divergence Measures and Information-Rich Cues for Speech Summarization. IEEE Transactions on Audio, Speech, and Language Processing 19 (4), 871–882.
Liu, F., Liu, Y., 2008. What are meeting summaries?: an analysis of human extractive summaries in meeting corpus. In: Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue. Association for Computational Linguistics.
Liu, F., Liu, Y., 2008. Correlation between rouge and human evaluation of extractive meeting summaries. In: Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers. Association for Computational Linguistics.
Liu, F., Liu, Y., 2009. From extractive to abstractive meeting summaries: Can it be done by sentence compression?. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers Association for Computational Linguistics.
Liu, F., Liu, Y., 2010. Using spoken utterance compression for meeting summarization: A pilot study. In: Spoken Language Technology Workshop (SLT), 2010 IEEE. Citeseer.
Liu, F., Liu, Y., 2013. Towards abstractive speech summarization: Exploring unsupervised and supervised approaches for spoken utterance compression. IEEE Transactions on Audio, Speech, and Language Processing 21 (7), 1469–1480.
Liu, S.-H., et al., 2014. A margin-based discriminative modeling approach for extractive speech summarization. In: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific. IEEE.
Liu, S.-H., et al., 2015. Positional language modeling for extractive broadcast news speech summarization. In: Sixteenth Annual Conference of the International Speech Communication Association.
Liu, S.-H., et al., 2015. Incorporating proximity information in relevance language modeling for extractive speech summarization. In: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific. IEEE.
Liu, S.-H., et al., 2015. Combining relevance language modeling and clarity measure for extractive speech summarization. IEEE Transactions on Audio, Speech, and Language Processing 23 (6), 957–969.
Liu, S.-H., et al., 2017. A position-aware language modeling framework for extractive broadcast news speech summarization. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 16 (4), 27.
Liu, Y., Xie, S., 2008. Impact of automatic sentence segmentation on meeting summarization. In: Acoustics, Speech and Signal Processing. *ICASSP 2008. IEEE International Conference on*. 2008. IEEE.
Liu, Z., et al., 2019. Topic-aware pointer-generator networks for summarizing spoken conversations. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE.
Lloret, E., Plaza, L., Aker, A., 2018. The challenging task of summary evaluation: an overview. Language Resources and Evaluation 52 (1), 101–148.
Lo, Y.-T., Lin, S.-H., Chen, B., 2012. Constructing effective ranking models for speech summarization. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE.
Manakul, P., Gales, M.J., Wang, L., 2020. Abstractive Spoken Document Summarization Using Hierarchical Model with Multi-Stage Attention Diversity Optimization. In: INTERSPEECH.
Marujo, L., et al., 2012. Key phrase extraction of lightly filtered broadcast news. In: International Conference on Text, Speech and Dialogue. Springer.
Maskey, S., Hirschberg, J., 2006. Summarizing speech without text using hidden markov models. In: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. Association for Computational Linguistics.
McCallum, A., et al., 2012. Ecological validity and the evaluation of speech summarization quality. In: Spoken Language Technology Workshop (SLT), 2012 IEEE. IEEE.
McKeown, K., et al., 2005. From text to speech summarization. In: Acoustics, Speech, and Signal Processing. *Proceedings.(ICASSP'05). IEEE International Conference on*. 2005. IEEE.
Metze, F., et al., 2013. Beyond audio and video retrieval: topic-oriented multimedia summarization. International Journal of Multimedia Information Retrieval 2 (2), 131–144.
Mikolov, T., Zweig, G., 2012. Context dependent recurrent neural network language model. In: 2012 IEEE Spoken Language Technology Workshop (SLT). IEEE.
Moher, D., et al., 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Annals of internal medicine 151 (4), 264–269.
Mphasis, *Mphasis DeepInsights Text Summarizer.*
Mrozinski, J., et al., 2006. Automatic sentence segmentation of speech for automatic summarization. In: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on. IEEE.
Murray, G., et al., 2006. Incorporating speaker and discourse features into speech summarization. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Association for Computational Linguistics.
Murray, G., et al., 2009. Extrinsic summarization evaluation: A decision audit task. ACM Transactions on Speech and Language Processing (TSLP) 6 (2), 2.
Murray, G., 2015. Abstractive meeting summarization as a Markov decision process. In: Canadian Conference on Artificial Intelligence. Springer.
Murray, G., Carenini, G., 2008. Summarizing spoken and written conversations. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Murray, G., Carenini, G., Ng, R., 2010. Interpretation and transformation for abstracting conversations. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics.

Murray, G., Carenini, G., Ng, R., 2010. Generating and validating abstracts of meeting conversations: a user study. In: Proceedings of the 6th International Natural Language Generation Conference. Association for Computational Linguistics.

Murray, G., Renals, S., 2007. Term-weighting for summarization of multi-party spoken dialogues. In: International Workshop on Machine Learning for Multimodal Interaction. Springer.

Murray, G., Renals, S., 2008. Meta comments for summarizing meeting speech. In: International Workshop on Machine Learning for Multimodal Interaction. Springer.

Murray, G., Renals, S., 2008. Detecting action items in meetings. In: International Workshop on Machine Learning for Multimodal Interaction. Springer.

Nallapati, R., et al., 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. arXiv preprint arXiv:1602.06023.

Nihei, F., et al., 2014. Predicting influential statements in group discussions using speech and head motion information. In: Proceedings of the 16th International Conference on Multimodal Interaction. ACM.

Nihei, F., Nakano, Y.I., Takase, Y., 2016. Meeting extracts for discussion summarization based on multimodal nonverbal information. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM.

Nihei, F., Nakano, Y.I., Takase, Y., 2017. Predicting meeting extracts in group discussions using multimodal convolutional neural networks. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. ACM.

Pallotta, V., Delmonte, R., Bristot, A., 2009. Abstractive summarization of voice communications. In: Language and Technology Conference. Springer.

Parthasarathy, S., Hasan, T., 2015. Automatic broadcast news summarization via rank classifiers and crowdsourced annotation. In: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE.

Peters, M.D., et al., 2015. Guidance for conducting systematic scoping reviews. International journal of evidence-based healthcare 13 (3), 141–146.

Quiroz, J.C., et al., 2019. Challenges of developing a digital scribe to reduce clinical documentation burden. npj Digital Medicine 2 (1), 1–6.

Reason8 is an AI assistant for managers and meetings.

Ribeiro, R., de Matos, D.M., 2013. *Improving Speech-to-Text Summarization by Using Additional Information Sources*, in *Multi-source, Multilingual Information Extraction and Summarization*. Springer, pp. 277–297.

Riccardi, G., et al., 2015. The sensei project: Making sense of human conversations. In: International Workshop on Future and Emergent Trends in Language Technology. Springer.

Riedhammer, K., Favre, B., Hakkani-Tur, D., 2008. A keyphrase based approach to interactive meeting summarization. In: Spoken Language Technology Workshop, 2008. SLT 2008. IEEE. IEEE.

Riedhammer, K., Favre, B., Hakkani-Tür, D., 2010. Long story short–global unsupervised models for keyphrase based meeting summarization. Speech Communication 52 (10), 801–815.

Rush, A.M., Chopra, S., Weston, J., 2015. A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685.

Ryant, N., et al., 2019. The second DIHARD diarization challenge: Dataset, task, and baselines. arXiv preprint arXiv:1906.07839.

Schluter, N., 2017. The limits of automatic summarisation according to ROUGE. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers.

See, A., Liu, P.J., Manning, C.D., 2017. Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.

Sell, G., et al., 2018. Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge. In: Interspeech.

Sert, M., Baykal, B., Yazici, A., 2008. *Combining Structural Analysis and Computer Vision Techniques for Automatic Speech Summarization.* in *Multimedia*. In: ISM 2008. Tenth IEEE International Symposium on. IEEE. 2008.

Shang, G., et al., 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. arXiv preprint arXiv:1805.05271.

Sjöbergh, J., 2007. Older versions of the ROUGEeval summarization evaluation system were easier to fool. Information Processing & Management 43 (6), 1500–1505.

Tardy, P., et al., 2020. Leverage Unlabeled Data for Abstractive Speech Summarization with Self-supervised Learning and Back-Summarization. In: International Conference on Speech and Computer. Springer.

Tokunaga, Y., Shimada, K., 2014. Multi-party conversation summarization based on sentence selection using verbal and nonverbal information. In: Advanced Applied Informatics (IIAIAAI), 2014 IIAI 3rd International Conference on. IEEE.

Wang, L., Cardie, C., 2012. Focused meeting summarization via unsupervised relation extraction. In: Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Association for Computational Linguistics.

Weng, S.-Y., Lo, T.-H., Chen, B., 2021. An effective contextual language modeling framework for speech summarization with augmented features. In: 2020 28th European Signal Processing Conference (EUSIPCO). IEEE.

Wu, Y., et al., 2019. A sequential matching framework for multi-turn response selection in retrieval-based chatbots. Computational Linguistics 45 (1), 163–197.

Xie, S., et al., 2009. Integrating prosodic features in extractive meeting summarization. In: Automatic Speech Recognition & Understanding. *ASRU 2009. IEEE Workshop on.* 2009. IEEE.

Xie, S., Liu, Y., 2010. Improving supervised learning for meeting summarization using sampling and regression. Computer Speech & Language 24 (3), 495–514.

Xie, S., Liu, Y., 2011. Using N-best lists and confusion networks for meeting summarization. IEEE Transactions on Audio, Speech, and Language Processing 19 (5), 1160–1169.

Xie, S., Liu, Y., Lin, H., 2008. Evaluating the effectiveness of features and sampling in extractive meeting summarization. In: Spoken Language Technology Workshop. *SLT 2008. IEEE.* 2008. IEEE.

XX http://groups.inf.ed.ac.uk/ami/download/, *AMI Meeting Corpus.*

XXX http://catalog.elra.info/en-us/repository/browse/ELRA-S0031/.

XXXX https://catalog.ldc.upenn.edu/LDC2004S08.

XXXXX https://www.l2f.inesc-id.pt/w/ALERT_Corpus.

XXXXXX http://metashare.nytud.hu/repository/browse/bea-hungarian-spontaneous-speech-database/808c4c306ba911e2aa7c68b599c26a062458e40404d44e4087901b5b720d2765/.

Yella, S.H., Varma, V., Prahallad, K., 2010. Significance of anchor speaker segments for constructing extractive audio summaries of broadcast news. In: Spoken Language Technology Workshop (SLT), 2010 IEEE. IEEE.

YY http://groups.inf.ed.ac.uk/ami/icsi/download/.

YYY https://pj.ninjal.ac.jp/corpus_center/csj/en/.

YYYY https://catalog.ldc.upenn.edu/LDC97S62.

YYYYY https://catalog.ldc.upenn.edu/LDC2004T19.

Zhang, J., Fung, P., 2007. Speech summarization without lexical features for Mandarin broadcast news. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers. Association for Computational Linguistics.

Zhang, J., Fung, P., 2010. A rhetorical syntax-driven model for speech summarization. In: Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics.

Zhang, J., Yuan, H., 2013. Speech Summarization without Lexical Features for Mandarin Presentation Speech. In: Asian Language Processing (IALP), 2013 International Conference on. IEEE.

Zhang, J., Yuan, H., 2014. A Certainty-based active learning framework of meeting speech summarization. In: Computer Engineering and Networking, pp. 235–242. Springer.

Zhang, J., Yuan, H., 2014. A comparative study on extractive speech summarization of broadcast news and parliamentary meeting speech. In: Asian Language Processing (IALP), 2014 International Conference on. IEEE.

Zhang, J., Yuan, H., 2016. A novel decoding framework for extractive speech summarization with Rhetorical Structure modeling. In: Asian Language Processing (IALP), 2016 International Conference on. IEEE.

Zhang, J.J., Chan, R.H.Y., Fung, P., 2010. Extractive Speech Summarization Using Shallow Rhetorical Structure Modeling. IEEE Trans. Audio, Speech & Language Processing 18 (6), 1147–1157.

Zhang, J.J., Fung, P., 2009. Active learning of extractive reference summaries for lecture speech summarization. In: Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora. Association for Computational Linguistics.

Zhang, J.J., Fung, P., 2010. Learning deep rhetorical structure for extractive speech summarization. In: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE.

Zhang, J.J., Fung, P., 2012. Active learning with semi-automatic annotation for extractive speech summarization. ACM Transactions on Speech and Language Processing (TSLP) 8 (4), 6.

Zhang, J.J., Huang, S., Fung, P., 2008. RSHMM++ for extractive lecture speech summarization. In: Spoken Language Technology Workshop, 2008. SLT 2008. IEEE. IEEE.

Zheng, C., et al., 2020. A Two-Phase Approach for Abstractive Podcast Summarization. arXiv preprint arXiv:2011.08291.

Zheng, C., et al., 2020. A Baseline Analysis for Podcast Abstractive Summarization. arXiv preprint arXiv:2008.10648.

Zhu, X., Penn, G., 2006. Comparing the roles of textual, acoustic and spoken-language features on spontaneous-conversation summarization. In: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. Association for Computational Linguistics.

ZZ https://catalog.ldc.upenn.edu/LDC2002S04.

ZZZ https://catalog.ldc.upenn.edu/LDC99S84.

ZZZZ http://universal.elra.info/product_info.php?cPath=37_46&products_id=1673.

ZZZZZ http://mm2.tid.es/mamidb/mamidb.tar.gz.