An Overview of Speech-to-Speech Translation Framework and its Modules

Nasrin Ehassan
College of Computer Sciences and Information Technology
Sudan University of Science and Technology, Sudan

Cosimo Ieracitano
DICEAM, University Mediterranea of Reggio Calabria,
Via Graziella, Feo di Vito, Reggio Calabria 89060, Italy

Mandar Gogate
Edinburgh Napier University,
School of Computing,
Merchiston Campus, Edinburgh EH10 5DT, Scotland, U.K.

Kia Dashtipour
Edinburgh Napier University, School of Computing,
Merchiston Campus, Edinburgh EH10 5DT,
Scotland, U.K

Amir Hussain
Edinburgh Napier University, School of Computing,
Merchiston Campus, Edinburgh EH10 5DT,
Scotland, U.K

ABSTRACT
Speech is the most natural form of human communication and arguably the most efficient method of exchanging information. However, communication between people who only speak different languages is a very challenging task. Speech-to-Speech translation (S2ST) attempts to overcome this issue, making it one of the most promising research domains in speech and Natural Language Processing (NLP). This present article reviews the most recent S2ST systems employed for different languages in terms of their constituent modules, namely Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-To-Speech (TTS). Furthermore, the paper critically highlights the main advantages and disadvantages of state-of-the-art techniques in S2ST in order to provide researchers with an up-to-date picture of current systems and potential directions for future work.

General Terms
Speech Recognition.

Keywords
Automatic Speech Recognition, Machine Translation, Speech to Speech translation, Text to Speech.

1. INTRODUCTION
Speech-to-Speech translation (S2ST) is the process of translating speech phrases from a source to a target language to enable communication between people speaking different languages [1]. Research in S2ST was started in the 1990s with international joint studies such as CSTAR, NESPOLE, TCSTAR, and GALE, which were used in Iraq for military purposes by the U.S. Armed Forces. Multilingual S2ST can be useful in many applications since it allows real-time translation from one language to another. S2ST holds promise for several application sectors, including health (by facilitating patient-doctor contact), news broadcast translation, education (by lecture translation), social media, and online meetings. Three progressively connected modules make up the conventional S2ST system architecture: Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-To-Speech (TTS). A typical S2ST framework is shown in Figure

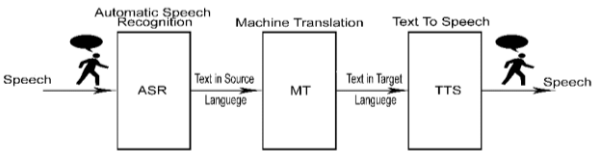

Fig. 1 Speech to Speech System is composed of the following modules: Automatic Speech Recognition, Machine Translation, and Text to Speech.

Automatic Speech Recognition (ASR) converts a speech signal into text [2]. Several techniques have been applied to the ASR problem, including Hidden Markov Models (HMM) [3], Dynamic Programming, Dynamic Bayesian Networks (DBN), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and hybrid models. Speech recognition technology is used in different devices such as robots, smart TVs, smartphones, domestic appliances, and digital cameras, and is also employed in health care to enhance the learning of students with disabilities [4], [5].

Machine Translation (MT) can be defined as "automatic translation from one language to another using computing devices and algorithms" [6]. This task of an automatic conversion from source to target language [7] belongs to the field of Natural Language Processing (NLP) [8] and combines statistics [9], computer science [10], [11], and linguistics [12]. One of the most popular translation systems is Google Translate (https://translate.google.com/), which is currently able to guarantee the translation of more than 100 languages and support search in 78 languages.

Approaches to MT include Rule-Based Machine Translation (RBMT), Corpus-based Machine Translation (CBMT), and hybrid approaches. RBMT consists of a set of rules that are used to identify words using prefix rules and to assign a translation to source words. Three steps are involved in the translation process: the input text is first analyzed, followed by the creation of an intermediary linguistic representation, and finally, the text is generated in the target language in accordance with morphological, syntactic, and semantic criteria. RBMT approaches include transfer, interlingua, and direct methods that require large lexicons and a set of rules created by professional linguists [13]. The transfer approach

generates the target text from an intermediate representation that captures the meaning of the original sentence. This approach proceeds in three different stages: i) analysis, ii) transfer, and iii) generation [14]. The interlingua approach is similar to the transfer strategy but its intermediate representation is intended to be language-independent. This makes the interlingua approach able to translate multiple target languages, whereas the transfer strategy depends directly on the original language representation [15]. The last type of RBMT is the direct or dictionary-based approach, which is able to translate texts without considering variations in word meaning. As such, it is typically used with hybrid techniques because it can be easily combined with other technologies [16], [17]. Corpus-based Machine Translation (CBMT), also referred to as data-driven machine translation, requires a large amount of raw data as source text (e.g., bilingual Arabic-to-English corpus. It should be noted that some languages still lack suitable data of this kind [18].

The two main data-driven MT approaches are Example-Based Machine Translations (EBMT) and Statistical Machine Translation (SMT). EBMT has a corpus that contains a set of sentences in the source language and the corresponding translation in the target language with point-to-point mapping. In contrast, SMT is generated on the basis of statistical models, whose model parameters are derived from the analysis of bilingual text corpora [18]. SMT can be defined as the process of finding and matching identical pairs from the source language and target language in parallel corpora. The goal of SMT is to find the optimal translation based on a statistical theory that uses a probability distribution function for two probabilistic models: language and translation. The main advantage of SMT is that it requires less human effort, but it is known to be computationally expensive [19], [20]. Finally, the hybrid approach is a combination of the strengths of the statistical method and one or more of the other MT strategies. This method can be implemented in different ways: for example, by using a rule-based approach and then statistical information for output correction, or by using rules to pre-process the input and to post-process the output from the statistical system [20].

Speech synthesis or Text-To-Speech (TTS) is a process that automatically converts text into spoken words or speech [21]. TTS has two phases: text analysis and speech signal generation. Techniques can be divided into three categories: i) articulatory synthesis, ii) formant synthesis and iii) concatenative synthesis. TTS systems are useful for a wide range of speech-enabled applications such as personal assistants [22], navigation systems, and assistive technologies for visually impaired people [23]– [25].

The rest of this paper is organized as follows: Section 2 presents a brief overview of the recent works about S2ST systems. Section 3 discusses the state-of-the-art of S2ST system and its modules. Section 4 concludes the paper.

2. RELATED WORKS

This section reviews the main studies related to S2ST technology. However, it should be noted that although various surveys related to the specific modules of an S2ST system exist in the literature (Fig. 1), comprehensive review papers on the complete S2ST framework are lacking. For example, Dureja et al. [26] considered two well-known S2ST systems: IBM Masters and Verbmobil. The authors examined various issues related to the English and French languages, the methodology for ASR, MT, and TTS, and finally the error rates for English language performance.

Other work has focused on the ASR module and its components [26]–[29]. Saini et al. [29] studied different approaches to ASR and compared their performance, whereas [27], [28] authors covered ASR approaches in more detail. Jinyu et al. [30] presented a more specific review discussing the concept of ASR from a pattern recognition perspective. The authors reviewed and compared the prominent feature extraction and acoustic modeling techniques, addressed the challenges and the advanced topics related to the ASR field, and made an exclusive overview of modern noise-robust techniques for ASR. Arora et al. [31] reviewed the evolution of ASR systems in several languages by comparing the different techniques used but did not take into account the mathematical formulations of these systems. Benzeghiba et al. [32] outlined advanced topics related to this field such as speech variability resources, speakers, and foreign and regional accents.

Focusing on MT modules, Alqudsi et al. [33] reviewed MT techniques used to translate the Arabic Language into English. Costa-Jussa et al. [34] focused on the rule-based structure (hybridization) and its applications, whereas Gaspari et al. [35] considered MT quality assessment and post-editing aspects. Khan et al. [36] focused on the performance of phrase-based Statistical Machine translation (SMT) in multiple Indian languages. Chakrawarti et al. [37] reviewed the most effective Hindi-to-English MT techniques and discussed the structure of the Indian and English languages, as well as methods, resources, and tools related to MT in full detail. Another study conducted by Chakrawarti et al. [38] examined different MT approaches, systems, benefits, and limitations as well as idiomatic problems.

With regard to Text to Speech (TTS) modules, considerable work has been carried out for a number of different languages. Research has so far addressed many issues, including accents and dialects. Rashed et al. [39] and Mattheyses et al. [40] surveyed the various text-to-speech techniques, such as rule-based synthesis techniques (formant synthesis and articulatory synthesis), concatenative, and unit selection synthesis.

3. STATE OF THE ART OF SPEECH-TO-SPEECH TRANSLATION

In this section, a critical and comprehensive review of different S2ST systems is carried out by discussing different techniques used for the three constituent modules, datasets used, languages supported, challenges, and system performance. S2ST systems allow people who speak different languages to communicate easily. S2ST is a service that recognizes the speech in one language, translates the obtained text to a target language, and finally synthesizes the translation into speech [41]. There are several projects that implement such a system, e.g., STAR, TC-STAR, GALE, Verbmobil, and NESPOLE. Matsuda et al. [41] proposed the VoiceTra system, which was the first network-based multilingual system available for smartphones. For ASR, VoiceTra used a large vocabulary continuous speech recognition decoder (ATRASR) and for MT, Matsuda et al. employed both a corpus-based and a phrase-based statistical translation framework.

Yun et al. [42] established Genietalk, a multi-language S2ST-based network system for Korean-English, Korean-Japanese, and Korean-Chinese translation that works on Android and iOS mobile devices. Genietalk is intended for the specific domain of live conversation, and the massive corpus used in its construction distinguishes it from other systems. This study designed a user-friendly S2ST UI, reduced errors in translation, and improved the performance of their system in different environments. The ASR module in Genietalk consists of a

speech recognition engine that was trained as an HMM-based acoustic model, a trigram language model, and a WFST (Weighted Finite State Transducer) decoder. The MT module comprises i) an MT engine using a statistics-based method for the translation of Korean-Japanese and Japanese-Korean, and ii) a regulation-based method for the translation of Korean-English, English, Korean-Chinese, and Chinese-Korean. Finally, Genietalk contains a T2S module and also log features.

Chen et al. [43] introduced a novel multilingual web conferencing system using S2ST to facilitate real-time communication between conference participants located in different places. The proposed system translates English to Spanish in both directions by using a session initiation protocol (SIP). The study investigated the optimal segmentation strategy that maximizes translation accuracy whilst also minimizing latency in order to allow real-time translation by using incremental speech recognition and segmentation.

Abdelali et al. [44] developed QAT2, an S2ST system able to convert Arabic speech into English using Aljazeera News as a speech corpus. Three off-the-shelf systems were pipelined together: i) the KALDI toolkit used for speech recognition, ii) Moses for machine translation based on Statistical Machine translation (SMT) with a 5-gram language model, and iii) MaryTTS for speech synthesis. The S2ST systems just reviewed are summarized in Table 1.

Table 1 List of S2ST systems in different languages

Ref	Languages	ASR/MT/TTS Technique	Comment
[41]	English to Spanish	AT&T WATSONSM speech recognize Moses toolkit AT&T Natural Voices TM text-to-speech	Flexible
[42]	Korean - English, Korean-Japanese, Korean-Chinese	HMM-based acoustic model Trigram language model	Massive corpus Specific domain
[45]	English, Hindi	MDL-SSS and N-gram phrase-based SMT Example-based machine translation	Bluetooth is limited Size
[46]	English, Brazilian, Portuguese	Google Translate	Noise and Speech overlapping not covered
[47]	Hindi, English, and Malayalam	Sphinx toolkit for speech recognition AnglaMT/ Google translate Festival synthesizer Google TTS	Small corpus size
[48]	Hindi, Malayalam, Tamil	No Translation MBROLA based TTS engine	Not covering the noise in ASR module
	and English.		
[49]	English-Japan Japan-English French – English	Corpus-based	-

3.1 Automatic Speech Recognition

Automatic speech recognition (ASR), known also as speech recognition, is a process that automatically recognizes speech and converts it into text by using the information included in the speech signal [50]. However, the performance of speech recognition remains a technical challenge despite the extensive research that has been carried out in the previous decades. The problem is inherently difficult due to signal noise and large vocabulary size, and also due to significant variations between speakers, domains, and languages [50]. This section will critically review the speech recognition techniques for different languages and show how machine learning algorithms can improve performance. Nahar et al. [51] introduced a hybrid algorithm to recognize continuous open-vocabulary speech in Arabic. The hybrid algorithm consisted of Learning Vector Quantization (LVQ) and a Hidden Markov Model (HMM). Experimental results demonstrated a recognition rate of 72% using LVQ alone but this increased to 89% with the hybrid LVQ-based HMM algorithm. Alshutayri et al. [52] proposed an Arabic dialect identification system using the Waikato environment (WEKA). The DSL corpus collection dataset was employed for training, and the best accuracy achieved was 50% using the Sequential Minimal Optimization (SMO) algorithm.

Emotion recognition is one of the most challenging sub-problems in speech recognition [53] due to the difficulty of extracting effective emotional features. To overcome this limitation, Han et al. [54] employed Deep Neural Networks (DNNs) to extract features from raw data. The proposed DNN was able to identify relevant emotion states which were then input to an Extreme Learning Machine (ELM) in order to identify utterance-level emotions. The Interactive Emotional Dyadic Motion Capture (IEMO-CAP) benchmark database was employed to evaluate the developed DNN. The achieved results were compared with other emotion approaches, such as an HMM-based method and OpenEAR (which uses global statistical features and an SVM for emotion recognition). The experimental results showed that this approach was able to learn emotional information from low-level features effectively and performed about 20% better than the compared approaches. Bahreini et al. [53] developed a real-time voice emotion recognition system intended to recognize the following emotions: happiness, surprise, anger, disgust, sadness, fear, and neutrality. This software is an implementation of the authors' FILTWAM framework (Framework for Improving Learning through Webcams and Microphones) and has potential applications in e-learning systems. For development and testing, the study used two existing tools: the Praat toolkit (for speech analysis) and the OpenSMILe tool (for audio feature extraction). Twelve participants were recruited for testing purposes and two experts annotated and rated the participants' recorded behavior. By comparing the voice emotion recognition software output with the experts' ground truth, an overall accuracy of 67% was achieved. Sarma and Else [55] presented a bilingual speech identification system for the Assamese and English languages. For speech recognition, the authors used MFCC as a feature

extraction technique and for classification purposes, they developed an ANN (multi-layer perceptron using the backpropagation algorithm). 20 mixed sentences were recorded with ten speakers (male and female) uttering each word ten times, for a total of 2000 recordings (split 75% / 25% for training and testing). The open-source Praat tool was employed for preprocessing and MFCC was used for feature extraction. Testing showed that the system was able to successfully identify the words in both languages (Assamese and English) for both seen and unseen speakers.

Sailor and other scholars [56] developed a Convolutional Restricted Boltzmann Machine (ConvRBM) based on unsupervised learning and Noisy Rectified Linear Units (NReLUs) for speech signal recognition. The proposed model was used as a front-end for feature learning, which was then applied to address the speech recognition task. For speech recognition experiments, this work used two databases: TIMIT [57] and WSJ0 [58]. ConvRBM features perform better than MFCC when using GMM-HMM systems, with relative improvements of 5% on the TIMIT test set and 7% on the WSJ0 database for both Nov'92 test sets. The study's DNNHMM systems outperformed MFCC and Mel filterbank (FBANK) by 3% on the TIMIT test set. On WSJ0 Nov'92 test sets, using ConvRBM features over MFCC features and ConvRBM filterbank over FBANK features respectively, the results showed a relative improvement of 4–14% and 3.6–5.6%, respectively.

Inspired by multimodel learning and the effectiveness of convolutional neural networks (CNN), the authors [59] proposed an Audio-Visual Deep CNN model (AVDCNN). Video recordings of 320 utterances of Mandarin sentences were used for training and testing and simulation results proved that combining visual and audio information was effective for Speech Enhancement (SE). The results of this experiment demonstrate that, in terms of five instrumental assessment measures, its performance for the SE task is superior to that of three audio-only baseline models, demonstrating the value of incorporating visual information with audio information into the SE process. By contrasting the model with other audio-visual SE models, was able to further establish its effectiveness.

The ASR systems just reviewed are summarized in Table 2.

Table 2 List of ASR systems in different languages

Ref	Feature Extraction technique / ML/ASR	Dataset	Evaluation	Comment
[51]	LVQ HMM	Arabic TV news	Performance recognition rate LVQ only LVQ HMM	LVQ - 72% Hybrid - 89% Dependent - 90% Do not use any phoneme bigrams model
[53]	openSMILe Praat toolkit	Record Voice	Accuracy	Accuracy - 67% Not able to recognize uncovered emotions
[54]	DNN/ELM/ ANN	IEMOCAP	Accuracy	Accuracy - 81%
[55]	MLP & BP	Speech signals	Identification and recognition rate	High identification rate
[56]	ConvRBM	TIMIT and WSJ0	PER WER	Better results than standard GMM-HMM and hybrid DNN-HMM
[59]	Audio-visual deep CNN	video recordings	PESQ, STOI, HASQI, HASPI	It used the Mouth area instead of the whole face
[60]	Kaldi Toolkit	Hausa Global Phone Speech	WER CER	WER - 0.9% WER - 0.1 %
[61]	Kaldi Toolkit	MSA	WER	With Hamza - 12.2% Baseline - 14.42%
[62]	FF & DNN	Telephonic conversation	Recognition rate & Time complexity	Outperform other approaches
[64]	MEL, VQ	TIMIT speech	PESQ, BER Accuracy	Accuracy - 97.85%
[65]	Novel algorithm	NOISEX-92 Database	PESQ, LLR	Not work to enhance observed speech frames in real time.

3.2 Machine Translation

This section critically reviews the MT techniques for different languages and also shows how various machine-learning algorithms can lead to improvements in accuracy and performance. Truong et al. [66] proposed a new model based on neural networks for emphasis transfer in speech translation. This model transfers emphasis information across languages

instead of transferring linguistic content only and was able to solve the limitation of Conditional Random Fields (CRFs) in handling long-distance dependencies between words in a sentence. A Long Short-Term Memory (LSTM) neural network was used in this new model. The proposed system consisted of two stages: an LSTM encoder (able to encode features from the source language) and an LSTM decoder (able to use the encoded features to generate an emphasis sequence in the target language). The study carried out emphasis translation experiments from English to Japanese using an English-Japanese emphasized speech corpus. The obtained results showed an improvement of 4% over state-of-the-art models for the objective evaluation and 2% for the subjective evaluation. Nair et al. [67] proposed a hybrid Machine Translation (HMT) architecture for the conversion of English to Hindi which was based on declension rules-based MT and SMT According to the study's findings, Google Translate had an accuracy rate of just 80% whereas RBMT had a rate of 96%. It achieved 94% accuracy when tested with AnglaBharathi. Ma et al. [68] developed a system able to take images shared between different languages using both the encoding and decoding stages. The Flickr30K and MSCOCO datasets were used to test the system, with the best performance (52.3%) achieved on the MSCOCO dataset. Firat et al. [69] proposed a multiway and multilingual attention-based neural MT system, which is based on a single neural machine translation. This machine was able to translate between multiple languages by using several parameters that grow linearly with the number of languages. For the Arabic language, Shquier and Else [70] implemented the AE-TBMT system that allowed translation into English. This system achieved an accuracy of about 96.6%, outperforming several popular machine translation systems, namely: Alkafi, Google, and Tarjim. Almahairi et al. [71] used neural machine translation to translate between English and Arabic in both directions. The proposed model was compared with a standard phrase-based translation system. Experimental results showed that the neural network outperformed the phrase-based system. Moreover, the study found that using better preprocessing results improved translation quality. Phrase-based and neural systems each gain as much as +4.46 and +4.98 BLEU above the baselines, on MT05, by utilizing normalization and morphology-aware tokenization (Tok+Norm+ATB).

One of the most important challenges faced by the sequence-to-sequence emphasis translation system is that emphasis translation is currently sub-optimal and cannot handle the long-term dependencies of words and emphasis levels. Several studies have tried to address this problem such as [72]. In this work, authors introduced an approach able to handle emphasis translation levels for continuous sentences on sequence-to-sequence models by combining machine and emphasis translation into a single model. The proposed method used an LSTM-based encoder-decoder system to capture long-term dependencies and handle continuous emphasis. Experiments showed that the model could perform joint translation of words with one-word delays instead of full-sentence delays. In another work, Su et al. [73] used sequence-to-sequence attentional Neural Machine Translation to obtain the optimal model parameters for long parallel sentences. They proposed a hierarchy-to-sequence attentional NMT framework based on a two-layer RNN encoder able to convert input sentences into a hierarchical structure of short sentences. For the evaluation, the BLUE and NIST datasets were employed, reporting 36.65% and 19.17% for Chinese–English and English–German translations, respectively. The MT systems just reviewed are summarized in Table 3.

3.3 Text to speech

Text-to-speech (TTS), or speech synthesis, is an important technology that has found widespread consumer use in recent years. Considerable research has been conducted in various languages although English and French have been the main focus. TTS in some languages, such as Arabic, is still undeveloped in terms of quality [74] . For this reason, we will review TTS systems for Arabic in this section. AlRouqi et al. [74] evaluated five Arabic TTS synthesizers (VoiceOver, uSpeech, Acapella, Adel, and SVOX) when applied in a mobile-device navigation

Table 3 List of MT systems in different languages

Ref	Technique	Language	Dataset	Evaluation	Comment
[66]	LSTM	English Japanese	English-Japanese speech	Accuracy	Good result over CRF
[67]	Rule-based	English Hindi	-	Accuracy	Require more rules
[68]	NMT	English French German	Flickr30K MSCOCO	BLEU	Simple and effective
[69]	NMT	Different Language	WMT'15 dataset	BLEU	Only cover ten languages
[70]	Transfer Based Technique	Arabic English	Ten suit	Accuracy	Limit with spelling
[71]	Neutral Network	Arabic English	LDC2004T	BLEU	Easy to develop

[72]	LSTM	English Japanese	English Japanese dataset	F-measure	Reduce the complexity Prevent the overfitting Create delay
[73]	RNN	Chinese English German	NIST	BLEU	Work with sentence
[75]	NMT	English German	WMT14	BLEU	Handling Long sentences
[76]	RNN	English French	Selection dataset	BLEU	Cannot handle long sentences

system for visually-impaired people. The authors carried out two subjective evaluations, namely MOS (Mean Opinion Score/scale) and SUS (Semantically Unpredictable Sentences) tests for intelligibility and naturalness. 16 visually-impaired undergraduate students evaluated the systems and the experimental results showed that Adel was rated highest and Acapella the worst. Rebai and Else [81] described a TTS synthesis system for modern standard Arabic (MSA) based on a statistical parametric approach and Mel-Cepstral coefficients. The proposed approach used deep neural networks (i.e., stacked generalization techniques) and contained a diacritization stage to solve the problem of missing Arabic diacritic marks. In addition, Rebai and Else used various methods to solve the problem of acoustic model accuracy. The authors concluded that the diacritization system can generate a diacritized text with high accuracy and can generate intelligible and natural speech. The subjective evaluation was based on the MOS, with ratings ranging from 1 (Bad) to 5 (Excellent). The average score for all listeners was 3.9, with a naturalness evaluation of 3.65. The average recognition rate was 93.5%.

In [82], the authors presented an efficient Arabic TTS system based on a statistical parametric approach and non-uniform unit speech synthesis with a diacritization system. This work addressed the quality of speech synthesis. The authors proposed a simple approach based on deep neural networks, which are trained to predict the diacritic marks and the spectral and prosodic parameters. A new simple stacked neural network was also developed to improve the accuracy of the acoustic models. Experimental results showed that the diacritization system allowed the generation of full diacritized text with high precision and that the TTS system produced high-quality speech. In [83], the researchers developed ILATalk, a multilingual text-to-speech system based on a well-known inductive learning algorithm called ILA and a concatenative approach to speech production. The system was able to accept any language by adding a database of words from the selected language with their associated phonemes. The proposed system was comprehensively tested with a number of experiments using various parameters and training set sizes., and the obtained accuracy was compared with ID3 and ANN backpropagation approaches. When the dataset was split 70 / 30 for training and testing, ITALk was found to outperform both the ID3 and ANN backpropagation approaches with an accuracy of 84.93%, This is compared to accuracies of 81.63% and 83.88% for ID3 and ANN backpropagation, respectively.

Arau´jo et al. [84] developed a system based on a genetic algorithm (GA), which is able to automatically estimate input parameters for two formant synthesizers (Klatt and HLSyn). The GA-based system outperformed the baseline Winsnoori system in both objective and subjective tests. Objective evaluation was performed using four metrics: mean-squared log-spectral distance (DLE), signal-to-noise ratio (SNR), root-mean-square error (RMSE), and perceptual evaluation of speech quality (PESQ). The GA with a Klatt synthesizer outperformed the Winsnoori baseline in terms of

SNR with 98.7% and also performed slightly better than the baseline system in the subjective test. Birkholaz et al. [85] presented a study to confirm that articulatory speech synthesis can control secondary prosodic features by the application of rules. Vocal-TractLab 2.1 software was used to re-synthesize nine German words in the voice of a male German native speaker. This software was also used to manipulate the voice in order to increase precision at the articulatory level by using different values of vocal tract length, articulatory precision, and degree of nasality. Subjective word-level tests were performed using sixteen subjects (10 females and 6 males). These tests showed that most of the manipulated words achieved a recognition rate between 77% and 96%. This study demonstrated that rule-based articulatory manipulations could generate suitable features. The TTS systems just reviewed are summarized in Table 4.

Table 4 List of TTS systems in different languages

Ref	TTS Technique	Test and Evaluation	Comment
[81]	MEL DNN	MOS	Add Arabic Dictionary
[82]	statistical parametric and non-uniform units speech synthesis	RMSE MCD	Add Arabic Dictionary
[83]	concatenative approach inductive learning	Accuracy	Can use for different languages
[84]	Klatt and HLSyn	DLE, SNR	GA with Klatt synthesizer outperform the Winsnoori baseline
[85]	Articulatory speech synthesis	Different values for vocal tract length.	Rate recognition between 77% and 96%.
[88]	Google, and Nuance	Intelligibility Accuracy Naturalness	Acapella and Sakhr had the best pleasant voice to hear
[89]	HMM	MOS	Require a small dimensionality

4. CONCLUSIONS

This paper reviewed various S2ST systems, as well as their constituent ASR, MT, and TTS modules. It was found that most systems were developed for popular languages such as English, Hindi, and German and that S2ST systems for less-popular languages and dialects need significant improvement and require building appropriate corpora. In particular, we found that full S2ST systems for the Arabic language do not currently exist. However, there are several works-in-progress, which are in their infancy due to the complexity and rich morphology of the Arabic language. After a critical review of the S2ST state-of-the-art and their modules, we can conclude the following. First, one of the main problems facing S2ST systems is their monolithic architecture: low accuracy of the ASR module will affect the entire system. However, accuracy can be increased by using hybrid techniques based on statistical and learning models as noted by [51] and deep learning architectures as reported by [54]. Second, one of the most challenging issues affecting the performance of MT systems is that of translation delay, since the full sentence needs to be read before starting the translation. In [72] a method using an LSTM-based encoder-decoder was developed to handle this problem, and in [67] statistical and deep learning approaches outperformed other conventional and hybrid methods. Lastly, for Arabic speech synthesizers, naturalness is crucial and many studies have used a diacritization system with deep learning algorithms to improve naturalness and intelligibility [74].

5. ACKNOWLEDGEMENTS

Our thanks to the Sudan University of Science and Technology (SUST) and its researchers for supporting and helping us to finish this work.

6. REFERENCES

[1] W. Wahlster, Verbmobil: foundations of speech-to-speech translation. Springer Science & Business Media, 2013.

[2] R. Zbib et al., "Machine translation of Arabic dialects," in Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies, 2012, pp. 49–59.

[3] M. A. M. Abushariah, R. N. Ainon, R. Zainuddin, M. Elshafei, and O. O. Khalifa, "Natural speaker-independent Arabic speech recognition system based on Hidden Markov Models using Sphinx tools," in International Conference on Computer and Communication Engineering (ICCCE'10), 2010, pp. 1–6.

[4] M. Gogate, K. Dashtipour, P. Bell, and A. Hussain, "Deep neural network driven binaural audio visual speech separation," in 2020 international joint conference on neural networks (IJCNN), 2020, pp. 1–7.

[5] M. Gogate, K. Dashtipour, and A. Hussain, "Visual Speech In Real Noisy Environments (VISION): A Novel Benchmark Dataset and Deep Learning-Based Baseline System.," in Interspeech, 2020, pp. 4521–4525.

[6] Y. Bar-Hillel, "The present status of automatic translation of languages," Advances in computers, vol. 1, pp. 91–163, 1960.

[7] S. Rajpirathap, S. Sheeyam, K. Umasuthan, and A. Chelvarajah, "Real-time direct translation system for Sinhala and Tamil languages," in 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), 2015, pp. 1437–1443.

[8] S. Poria, O. Y. Soon, B. Liu, and L. Bing, "Affect recognition for multimodal natural language processing," Cognit Comput, vol. 13, no. 2, pp. 229–230, 2021.

[9] C. Ieracitano, A. Adeel, F. C. Morabito, and A. Hussain, "A novel statistical analysis and autoencoder driven intelligent intrusion detection approach," Neurocomputing, vol. 387, pp. 51–62, 2020.

[10] C. Ieracitano, A. Paviglianiti, M. Campolo, A. Hussain, E. Pasero, and F. C. Morabito, "A novel automatic classification system based on hybrid unsupervised and supervised machine learning for electrospun nanofibers," IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 1, pp. 64–76, 2020.

[11] Z. Cai and L. Shao, "Rgb-d scene classification via multi-modal feature learning," Cognit Comput, vol. 11, no. 6, pp. 825–840, 2019.

[12] K. Dashtipour, M. Gogate, J. Li, F. Jiang, B. Kong, and A. Hussain, "A hybrid Persian sentiment analysis framework: Integrating dependency grammar based rules and deep neural networks," Neurocomputing, vol. 380, pp. 1–10, 2020.

[13] H. U. Mullah, "A Comparative Study of Different Text-to-Speech Synthesis Techniques," Int J Sci Eng Res, vol. 6, no. 6, 2015.

[14] D. Moussallem, M. Wauer, and A.-C. N. Ngomo, "Machine translation using semantic web technologies: A survey," Journal of Web Semantics, vol. 51, pp. 1–19, 2018.

[15] D. W. Lonsdale, A. Franz, and J. R. R. Leavitt, "Large-Scale Machine Translation: An Interlingua Approach.," in Iea/aie, 1994, pp. 525–530.

[16] M. Madankar, M. B. Chandak, and N. Chavhan, "Information retrieval system and machine translation: a review," Procedia Comput Sci, vol. 78, pp. 845–850, 2016.

[17] M. D. Okpor, "Machine translation approaches: issues and challenges," International Journal of Computer Science Issues (IJCSI), vol. 11, no. 5, p. 159, 2014.

[18] M. Rushdi-Saleh, M. T. Mart\'\in-Valdivia, L. A. U. Lopez, and J. M. Perea-Ortega, "Bilingual experiments with an arabic-english corpus for opinion mining," in Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, 2011, pp. 740–745.

[19] B. A. Abdulsalami and B. J. Akinsanya, "Review of different approaches for machine translations," International Journal of Mathematics Trends and Technology (IJMTT), vol. 48, no. 3, pp. 197–202, 2017.

[20] S. Dubey, "Survey of Machine Translation Techniques," International Journal of Advance Research in Computer Science and Management Studies, Special Issue, vol. 5, no. 2, pp. 39–51, 2017.

[21] I. Isewon, J. Oyelade, and O. Oladipupo, "Design and implementation of text to speech conversion for visually impaired people," Int J Appl Inf Syst, vol. 7, no. 2, pp. 25–30, 2014.

[22] A. Zhang et al., "Clustering of remote sensing imagery using a social recognition-based multi-objective gravitational search algorithm," Cognit Comput, vol. 11, no. 6, pp. 789–798, 2019.

[23] M. Gogate, A. Adeel, R. Marxer, J. Barker, and A. Hussain, "DNN driven speaker independent audio-visual mask estimation for speech separation," arXiv preprint arXiv:1808.00060, 2018.

[24] M. Gogate, A. Adeel, and A. Hussain, "Deep learning driven multimodal fusion for automated deception detection," in 2017 IEEE symposium series on computational intelligence (SSCI), 2017, pp. 1–6.

[25] X. Yang, K. Huang, R. Zhang, and J. Y. Goulermas, "A novel deep density model for unsupervised learning," Cognit Comput, vol. 11, no. 6, pp. 778–788, 2019.

[26] M. Dureja and S. Gautam, "Speech-to-Speech Translation: A Review," Int J Comput Appl, vol. 129, no. 13, pp. 28–30, 2015.

[27] A. Katyal, A. Kaur, and J. Gill, "Automatic speech recognition: a review," International Journal of Engineering and Advanced Technology (IJEAT), vol. 3, no. 3, pp. 71–74, 2014.

[28] S. K. Gaikwad, B. W. Gawali, and P. Yannawar, "A review on speech recognition technique," Int J Comput Appl, vol. 10, no. 3, pp. 16–24, 2010.

[29] S. Preeti and K. Parneet, "Automatic speech recognition: A review," International Journal of Engineering Trends and Technology, vol. 4, no. 2, p. 2013, 2013.

[30] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," IEEE/ACM Trans Audio Speech Lang Process, vol. 22, no. 4, pp. 745–777, 2014.

[31] S. J. Arora and R. P. Singh, "Automatic speech recognition: a review," Int J Comput Appl, vol. 60, no. 9, 2012.

[32] M. Benzeghiba et al., "Automatic speech recognition and speech variability: A review," Speech Commun, vol. 49, no. 10–11, pp. 763–786, 2007.

[33] A. Alqudsi, N. Omar, and K. Shaker, "Arabic machine translation: a survey," Artif Intell Rev, vol. 42, no. 4, pp. 549–572, 2014.

[34] M. R. Costa-Jussa and J. A. R. Fonollosa, "Latest trends in hybrid machine translation and its applications," Comput Speech Lang, vol. 32, no. 1, pp. 3–10, 2015.

[35] F. Gaspari, H. Almaghout, and S. Doherty, "A survey of machine translation competences: Insights for translation technology educators and practitioners," Perspectives (Montclair), vol. 23, no. 3, pp. 333–358, 2015.

[36] N. J. Khan, W. Anwar, and N. Durrani, "Machine translation approaches and survey for Indian languages," arXiv preprint arXiv:1701.04290, 2017.

[37] R. K. Chakrawarti and P. Bansal, "Approaches for improving Hindi to English machine translation system," Indian J Sci Technol, vol. 10, no. 16, pp. 1–8, 2017.

[38] R. K. Chakrawarti, H. Mishra, and P. Bansal, "Review of machine translation techniques for idea of Hindi to English idiom translation," International journal of computational intelligence research, vol. 13, no. 5, pp. 1059–1071, 2017.

[39] M. Z. Rashad, H. M. El-Bakry, I. R. Isma'il, and N. Mastorakis, "An overview of text-to-speech synthesis techniques," Latest trends on communications and information technology, pp. 84–89, 2010.

[40] W. Mattheyses and W. Verhelst, "Audiovisual speech synthesis: An overview of the state-of-the-art," Speech Commun, vol. 66, pp. 182–217, 2015.

[41] S. Matsuda et al., "Multilingual speech-to-speech translation system: VoiceTra," in 2013 IEEE 14th International Conference on Mobile Data Management, 2013, pp. 229–233.

[42] S. Yun, Y.-J. Lee, and S.-H. Kim, "Multilingual speech-to-speech translation system for mobile consumer devices," IEEE Transactions on Consumer Electronics, vol. 60, no. 3, pp. 508–516, 2014.

[43] J. Chen, S. Wen, V. K. R. Sridhar, and S. Bangalore, "Multilingual web conferencing using speech-to-speech translation.," in INTERSPEECH, 2013, pp. 1861–1863.

[44] A. Abdelali, A. Ali, F. Guzmán, F. Stahlberg, S. Vogel, and Y. Zhang, "QAT2—The QCRI Advanced Transcription and Translation System," in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[45] M. D. F. Ansari, R. S. Shaji, T. J. SivaKarthick, S. Vivek, and A. Aravind, "Multilingual speech to speech translation system in bluetooth environment," in 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 1055–1058.

[46] F. Calefato, F. Lanubile, D. Romita, R. Prikladnicki, and J. H. S. Pinto, "Mobile speech translation for multilingual requirements meetings: A preliminary study," in 2014 IEEE 9th international conference on global software engineering, 2014, pp. 145–152.

[47] A. Gopi, T. Sajini, J. Stephen, V. K. Bhadhran, and others, "Multilingual Speech to Speech MT based chat system," in 2015 International Conference on Computing and Network Communications (CoCoNet), 2015, pp. 771–776.

[48] J. Stephen, M. Anjali, and V. K. Bhadran, "Voice enabled multilingual newspaper reading system," in 2013 IEEE Global Humanitarian Technology Conference: South Asia Satellite (GHTC-SAS), 2013, pp. 317–320.

[49] S. Nakamura, "Towards real-time multilingual multimodal speech-to-speech translation," in Spoken Language Technologies for Under-Resourced Languages, 2014.

[50] D. Kamińska, T. Sapiński, and G. Anbarjafari, "Efficiency of chosen speech descriptors in relation to emotion recognition," EURASIP J Audio Speech Music Process, vol. 2017, no. 1, pp. 1–9, 2017.

[51] K. M. O. Nahar, M. Abu Shquier, W. G. Al-Khatib, H. Al-Muhtaseb, and M. Elshafei, "Arabic phonemes recognition using hybrid LVQ/HMM model for continuous speech recognition," Int J Speech Technol, vol. 19, no. 3, pp. 495–508, 2016.

[52] A. Alshutayri, E. Atwell, A. Alosaimy, J. Dickins, M. Ingleby, and J. Watson, "Arabic language WEKA-based dialect classifier for Arabic automatic speech recognition transcripts," in Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), 2016, pp. 204–211.

[53] K. Bahreini, R. Nadolski, and W. Westera, "Towards real-time speech emotion recognition for affective e-learning," Educ Inf Technol (Dordr), vol. 21, no. 5, pp. 1367–1386, 2016.

[54] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in Interspeech 2014, 2014.

[55] S. Sarma and A. Barman, "MULTILINGUAL SPEECH IDENTIFICATION USING ARTIFICIAL NEURAL NETWORK," International Journal of Information Technology Convergence and Services (IJITCS) Vol, vol. 5, pp. 1–6.

[56] H. B. Sailor and H. A. Patil, "Filterbank learning using convolutional restricted Boltzmann machine for speech recognition," in 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2016, pp. 5895–5899.

[57] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-

1.1," NASA STI/Recon technical report n, vol. 93, p. 27403, 1993.

[58] D. B. Paul and J. Baker, "The design for the Wall Street Journal-based CSR corpus," in Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992, 1992.

[59] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," IEEE Trans Emerg Top Comput Intell, vol. 2, no. 2, pp. 117–128, 2018.

[60] E. Gauthier, L. Besacier, and S. Voisin, "Automatic speech recognition for African languages with vowel length contrast," Procedia Comput Sci, vol. 81, pp. 136–143, 2016.

[61] M. A. Menacer, O. Mella, D. Fohr, D. Jouvet, D. Langlois, and K. Smaili, "An enhanced automatic speech recognition system for Arabic," in The third Arabic Natural Language Processing Workshop-EACL 2017, 2017.

[62] S. Agarwalla and K. K. Sarma, "Machine learning based sample extraction for automatic speech recognition using dialectal Assamese speech," Neural Networks, vol. 78, pp. 97–111, 2016.

[63] H. B. Sailor and H. A. Patil, "Filterbank learning using convolutional restricted Boltzmann machine for speech recognition," in 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2016, pp. 5895–5899.

[64] A. Revathi, N. Sasikaladevi, and C. Jeyalakshmi, "Digital speech watermarking to enhance the security using speech as a biometric for person authentication," Int J Speech Technol, vol. 21, pp. 1021–1031, 2018.

[65] K. Huang, Y. Liu, and Y. Hong, "Reduction of residual noise based on eigencomponent filtering for speech enhancement," Int J Speech Technol, vol. 21, pp. 877–886, 2018.

[66] Q. T. Do, S. Sakti, G. Neubig, and S. Nakamura, "Transferring Emphasis in Speech Translation Using Hard-Attentional Neural Network Models.," in INTERSPEECH, 2016, pp. 2533–2537.

[67] J. Nair, K. A. Krishnan, and R. Deetha, "An efficient English to Hindi machine translation system using hybrid mechanism," in 2016 international conference on advances in computing, communications and informatics (ICACCI), 2016, pp. 2109–2113.

[68] M. Ma, D. Li, K. Zhao, and L. Huang, "Osu multimodal machine translation system report," arXiv preprint arXiv:1710.02718, 2017.

[69] O. Firat, K. Cho, and Y. Bengio, "Multi-way, multilingual neural machine translation with a shared attention mechanism," arXiv preprint arXiv:1601.01073, 2016.

[70] M. M. A. Shquier and K. M. Alhawiti, "Fully automated Arabic to English machine translation system: transfer-based approach of AE-TBMT.," Int. J. Inf. Commun. Technol., vol. 10, no. 4, pp. 376–391, 2017.

[71] A. Almahairi, K. Cho, N. Habash, and A. Courville, "First result on Arabic neural machine translation," arXiv preprint arXiv:1606.02680, 2016.

[72] Q. T. Do, S. Sakti, and S. Nakamura, "Sequence-to-sequence models for emphasis speech translation," IEEE/ACM Trans Audio Speech Lang Process, vol. 26, no. 10, pp. 1873–1883, 2018.

[73] J. Su, J. Zeng, D. Xiong, Y. Liu, M. Wang, and J. Xie, "A hierarchy-to-sequence attentional neural machine translation model," IEEE/ACM Trans Audio Speech Lang Process, vol. 26, no. 3, pp. 623–632, 2018.

[74] H. AlRouqi et al., "Evaluating Arabic Text-to-Speech synthesizers for mobile phones," in 2015 Tenth International Conference on Digital Information Management (ICDIM), 2015, pp. 89–94.

[75] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," arXiv preprint arXiv:1508.04025, 2015.

[76] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," arXiv preprint arXiv:1409.1259, 2014.

[77] O. Firat, K. Cho, and Y. Bengio, "Multi-way, multilingual neural machine translation with a shared attention mechanism," arXiv preprint arXiv:1601.01073, 2016.

[78] M. M. A. Shquier and K. M. Alhawiti, "Fully automated Arabic to English machine translation system: transfer-based approach of AE-TBMT," International Journal of Information and Communication Technology, vol. 10, no. 4, pp. 376–391, 2017.

[79] A. Almahairi, K. Cho, N. Habash, and A. Courville, "First result on Arabic neural machine translation," arXiv preprint arXiv:1606.02680, 2016.

[80] Q. T. Do, S. Sakti, and S. Nakamura, "Sequence-to-sequence models for emphasis speech translation," IEEE/ACM Trans Audio Speech Lang Process, vol. 26, no. 10, pp. 1873–1883, 2018.

[81] I. Rebai and Y. BenAyed, "Text-to-speech synthesis system with Arabic diacritic recognition system," Comput Speech Lang, vol. 34, no. 1, pp. 43–60, 2015.

[82] I. Rebai and Y. BenAyed, "Arabic speech synthesis and diacritic recognition," Int J Speech Technol, vol. 19, no. 3, pp. 485–494, 2016.

[83] S. M. Abu-Soud, "ILATalk: a new multilingual text-to-speech synthesizer with machine learning," Int J Speech Technol, vol. 19, no. 1, pp. 55–64, 2016.

[84] F. Araújo, A. Klautau, and others, "Genetic algorithm to estimate the input parameters of Klatt and HLSyn formant-based speech synthesizers," Biosystems, vol. 150, pp. 190–193, 2016.

[85] P. Birkholz, L. Martin, Y. Xu, S. Scherbaum, and C. Neuschaefer-Rube, "Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis," Comput Speech Lang, vol. 41, pp. 116–127, 2017.

[86] I. Rebai and Y. BenAyed, "Text-to-speech synthesis system with Arabic diacritic recognition system," Comput Speech Lang, vol. 34, no. 1, pp. 43–60, 2015.

[87] I. Rebai and Y. BenAyed, "Arabic speech synthesis and diacritic recognition," Int J Speech Technol, vol. 19, pp. 485–494, 2016.

[88] I. Abu Doush, F. Alkhatib, and A. A. R. Bsoul, "What we have and what is needed, how to evaluate Arabic Speech Synthesizer?," Int J Speech Technol, vol. 19, pp. 415–432, 2016.

[89] A. Jafri, I. Sobh, and A. Alkhairy, "Statistical formant speech synthesis for Arabic," Arab J Sci Eng, vol. 40, pp. 3151–3159, 2015.