International Journal of Computer Applications Technology and Research
Volume 11–Issue 03, 36-55, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1103.1001

Speech To Speech Translation: Challenges and Future

Sandeep Dhawan
Senior IT Director

Abstract: The different languages in the world and the various native mother tongues have distinct origins. Speech is the most widely recognized method of articulation, with an average individual talking more than 11000 words each day.

Speech can make each other understand actions and thoughts, be it a dialogue, conversation, or even a presentation. The passage of information would be ineffective if the other party did not understand the language of communication. As a result, a system capable of bridging the linguistic divide is required. Speech-to-speech translation is one such system that can be useful in facilitating communication between people who speak different languages.

Efforts are constantly taken all around the globe to accomplish this objective and set it up as a regular occurrence to assist everybody. This paper describes a significant international and inter-institutional effort in this direction, highlighting the current challenges being faced as well as the future of the technology of Speech-to speech translation. The developed language-specific technology, parallel corpora, and speech unit (segmental) database are all been described.

Keywords: Speech-to-speech translation, languages, technology.

1. INTRODUCTION

The present circumstances and situations in the world demand the need for communication among speakers of various languages. SPEECH-TO-SPEECH translation (S2ST) is a human pipe dream that allows communication between people speaking different languages. The importance of S2ST technology is growing by the day because the world is becoming more borderless by the day. Moreso, the need for the exchange of information is inevitable because of the global and borderless economy, and speech is one of the means of achieving such. The borderless economy in the world has made it critical for speakers of different languages to be able to communicate. Speech translation which has been named as one of the top ten technologies that will transform the world, has long been a human dream and the ultimate aim is to improve communication between people who speak different languages. Speech-to-Speech Translation (SST) research topic represents a relatively recent research area in the Human Language Technologies arena, which helps in producing a speech signal in the target language that conveys the linguistic information contained in the source language's speech signal [1].

All of the languages that exist in our world have distinct origins, as does the set of their native mother tongue. According to the study Karunesh et al., 2013, learning a new language as an "adult" is more complex and expensive than learning as a child. Due to geographical factors, almost all of the world's population finds learning foreign languages extremely difficult. As a result, speech-to-speech translation technology would be a massive boon to everyone. Manual Translation of Speech to Speech has only been used for important official documents, news items, and award-winning literary works. There is a massive backlog of materials that need to be translated for administration, education, commerce, tourism, and other purposes. It is critical to have technical assistance in the form of machine translation aids. The rapid increase in demand for translingual conversations, spurred by

Information Technologies and an increase in the number of borderless communities as evidenced by the increased number of EU countries, has increased the interest in S2ST research activities. The goal of speech-to-speech translation is to convert speech input from one language into speech in another. The technology facilitates communication between two or more people who speak different languages and can provide access to multimedia content in multiple languages [1].

Speech to speech translation is a technology that converts spoken language into another language's speech. S2ST is significant because it allows speakers of various languages from around the world to communicate with one another, eradicating the language divide in global business and cross-cultural exchange. It has been a colossal logical, social, and financial worth to humankind. One of the ten advances recorded in the article "10 Emerging Technologies That Will Change Your World" in the February 2004 issue of A MIT Enterprise Technology Review is "All-inclusive Translation." The piece features an assortment of interpretation innovations, with attention to speech translation technology.

Speech translation is a cycle that takes the conversational speech expressed in one language as info and deciphered speech phrases in one more language as a result. The three parts of speech-to-speech Translation are associated with a successive request. Automatic speech translation technology comprises of three separate advances: innovation to perceive (speech acknowledgment), innovation to decipher the perceived words (language interpretation), and innovation to integrate speech in the other individual's language (speech union). Late mechanical advances have made programmed interpretation of informal communication in Japanese, English, and Chinese for voyagers functional, and sequential interpretation of short, straightforward, conversational sentences spoken each, in turn, has become conceivable. Speech-to-Speech translation is a three-stage programming process that incorporates **Automatic Speech Recognition, Machine Interpretation, and voice synthesis**. ASR is liable for changing the communicated in expressions of source language over to the text in a similar language followed by machine interpretation which deciphers the source language close to the target language text. Lastly, the speech synthesizer is liable for the text-to-speech transformation of the target language [3]..

Speech-translation technology is essential because it allows speakers of many languages from all over the world to impart, connecting the language hole in worldwide business and cross-cultural communication. Speech translation would be highly beneficial to humanity in terms of science, culture, and economics. "Universal Translation" is one of the ten emerging technologies listed in the article "10 Emerging Technologies That Will Change Your World" in the issue of An MIT Enterprise Technology Review. The Universal Speech Translation Advanced Research Consortium (U-STAR) is an international research collaboration formed to develop a network-based speech to speech translation (S2ST) with the goal of breaking down language barriers and implementing vocal communication between different languages around the world. The ITU-T standardized and approved international communication protocols in 2010, thanks to a U-STAR initiative.; Recommendations F.745 [4], and H.625 [5], enabling speech-to-speech translation (S2ST) modules to be connected across the globe over networks.

Multilingual speech-to-speech translation technologies are critical for bridging the language divide, which is one of globalization's most vital issues. Machine translation is the leading technology used in S2ST to generate natural translations from raw data. As a result, S2ST's performance is significantly reliant on the machine translation system's performance.

The purpose of automatic voice-to-speech translation is to produce a speech signal in one (target) language that transmits the linguistic information contained in another (source) language's speech signal.

In the subject of human language technology, significant progress has been accomplished. Various tasks, such as continuous speech recognition with an extensive vocabulary, speaker and language identification, spoken information inquiry, information extraction, and cross-language retrieval in restricted domains, are now possible, and various prototypes and systems are in use. The difficulty of spoken Translation, on the other hand, remains a considerable challenge: "It was difficult enough to get a good text translation. It was more than traveling to the Moon for Speech to Speech MT - it was going to Mars..." [Steve Silbermann, Wired Magazine].

Speech translation (ST), the assignment of making an interpretation of acoustic speech signals into text in an unknown language, is a complex and diverse undertaking that expands upon work in automatic speech recognition (ASR) and machine translation (MT) [7].

1.1. Applications:

ST applications are diverse and include travel assistants, simultaneous lecture translation, movie dubbing/subtitling, language documentation, and crisis response, and developmental efforts [8].

Speech to speech translation is crucial for understanding cross-lingual spoken conversations and lectures and has been used in situations such as international travel and conferences. Existing speech to speech translation systems either use target text as a pivot (translating source speech into target text and then synthesizing target speech based on the translated text or directly translate source speech into target speech. The text matching to the target speech is used as pivots or supplemental training data in these translation systems [9]; otherwise, the Translation would not be possible, or the translation accuracy would drop dramatically [9]

1.2. Essential Features of the S2S Translation System

A speech recognition system should be able to recognize speech in loud situations and from people who talk in a variety of ways. Machine translation systems must be domain-agnostic and capable of translating a wide range of topics. Speech synthesis must achieve a more natural and expressive quality of speech. To develop the S2ST system, all of the researchers in this project—including speech processing and natural language researchers—are working together and closely. The speech recognition system must identify speaker-independent, ongoing, spontaneous conversational speech for the S2ST system to be successful [9] 10].

The centralized connections considered in this task are depicted in Figure 6. This will normalize the connection points and information organizations of the discourse interpretation engineering's modules, permitting them to convey over the Web. The data formats and interfaces of the modules of the speech translation architecture will be standardized as a result, allowing them to communicate over the Internet. It's also required to assemble standardized multilingual corpora and establish common speech recognition and translation dictionaries. Web-based HTTP 1.1 communication will be the primary communication interface, with a markup language called STML (speech translation markup language) being created as the data format for linking apps [2].

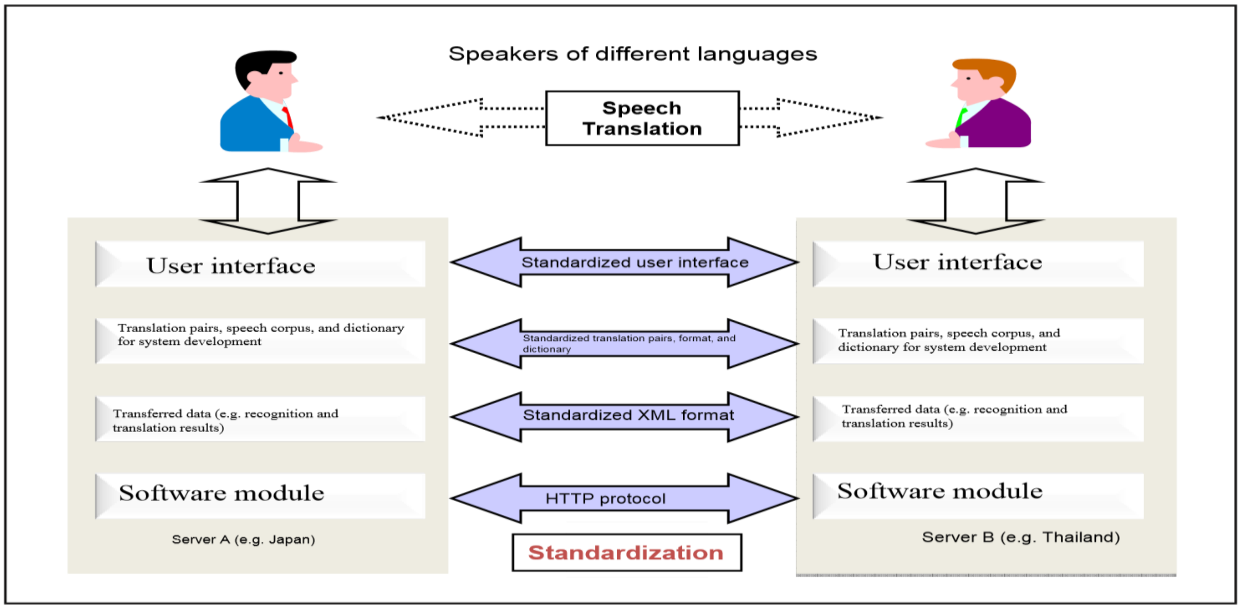

Speakers of different languages
Speech Translation
User interface
Standardized user interface
User interface
Translation pairs, speech corpus, and dictionary for system development
Standardized translation pairs, format, and dictionary
Translation pairs, speech corpus, and dictionary for system development
Transferred data (e.g. recognition and translation results)
Standardized XML format
Transferred data (e.g. recognition and translation results)
Software module
HTTP protocol
Software module
Server A (e.g. Japan)
Standardization
Server B (e.g. Thailand)

Figure 6 : Illustration of speech translation standardization

Source : Reference [12]

1.3. History and world scenario of S2S translation systems

Several previous studies have looked into the unsupervised conversion of speech to its corresponding phonetic categories (discrete tokens), which mimics how newborn children gain knowledge acoustic models in their mother tongue during their early years of life. (some of them only focus on a much easier task such as speech-to-text Translation. Among these works, vector quantized variational autoencoder (VQ-VAE) has been widely adopted and shown advantages over other methods. However, VQ-VAE is still purely unsupervised and cannot ensure the quality of the learned discrete representations. Therefore, although VQ-VAE performs very well on relatively more straightforward tasks like speech synthesis (Dunbar et al. 2019), it cannot achieve good accuracy on more complicated speech to speech translation where semantic representations of speech are necessary and more accurate phonetic words are required. Few works tackle speech to speech Translation for unwritten languages since it is incredibly challenging. [1,2, 3]

When NEC Corporation presented a display of speech translation as a conceptual demonstration at the 1983 ITU Telecom World (Telecom' 83), it drew a lot of interest. Realizing that speech translation will demand several decades of basic study, the Advanced Telecommunications Explore Institute International (ATR) was created in 1986 and launched a mission to study the subject. This initiative drew researchers from a wide spectrum of Japanese and foreign academic institutes [1]. In 1993, the ATR, Carnegie Melon University (CMU), and Siemens collaborated on a voice translation project that involved three locations around the world: the ATR, CMU, and Siemens. Speech translation initiatives were established all around the world after ATR's effort began. The Verbmobil plan was initiated in Germany, the Nespole and TC-Star programs have been established in the European Union, and the Transpac and GALE projects were launched in the United States. The GALE project began in 2006 with the goal of mechanically translating Arabic and Chinese into English. The aim of this assignment is to automate the extraction of critical multilingual information, which was previously done by humans; the project design is a batch text-output system. The ATR and NEC, on the other hand, aim to provide real-time voice translation for face-to-face and non-face-to-face cross-language communication. Online speech-to-speech Translation is thus an integral component of this Research, and immediacy of processing is a key factor [2, 3].

The first time speech translation was observed was during the 1983 ITU Telecom World (Telecom'83), when NEC Corporation demonstrated it as a proof of concept. In 1993, the ATR, Carnegie Melon University (CMU), and Siemens collaborated on a speech translation experiment that involved three locations around the world: the ATR, CMU, and Siemens. The Verbmobil project was launched in Germany, the Nespole! and TC-Star projects were launched in the European Union, and the Transpac and GALE projects were launched in the United States. The Translation of Research and development has developed from relatively simple to more complicated Translation. There have been a number of attempts to construct S2S Translation systems, with some success stories. The significant works in S2S Translation are shown in Table 1 in the annexure. [1, 3]

The history of speech-translation technology is shown in Table 1. From meeting scheduling to hotel bookings to travel discussion, Research and development have steadily developed from relatively simple to more complicated Translation. Nevertheless, in the future, the supported fields will need to be expanded to allow a wider variety of basic and complex commercial interactions

.

Table 1: Trends in the Research and Development of Speech Translation

Research Phase	1980s Confirmation of Feasibility	1990s Extension of Technology	2000s Attempts at Practical Systems
Fields	Simple reservations (ATR-phase 1)	Reservations and scheduling (ATR-phase 2, Verbmobil)	*Everyday travel conversation (ATR-phase 3) *Translation of keynote speeches (TC-Star) *Conversation for military use (TranTac) *Intelligence collection (Gale)
Linguistic features	Expressions that are grammatically accurate	Everyday idioms that may or may not be grammatical or context-dependent	Topics and proper nouns are included in the expressions.
Phonological features	Pronounced correctly	Pronunciation is ambiguous.	Background noise is included in the audio.
Translation method	Translation based on rules. Artificial intermediate language translation	Translation based on examples English is used as an intermediary language in the translation.	Translation based on statistics Multiple languages are directly translated.

Recent improvements in speech translation technology have contributed significantly to the achievement of autonomous speech translation technology, which is among the three aspects of speech translation. Speech translation technology has a lengthy history dating over half a century. In 1946, immediately after the development of the very first computer. Warren Weaver of the Rockefeller Foundation urged for studies into automatic translation technology. In the first place, the Rockefeller Foundation had a substantial impact on US scientific and technology policy. Then, in 1953, Georgetown University and IBM collaborated on automatic translation research using the 701 computers (the first commercial computer developed by IBM). This computer was used to create the world's first automatic translation system in 1954, proving the feasibility of translating from Russian to English. The translation capabilities of this system, which consisted of a lexicon of 250 phrases and six rules, were extremely restricted, but the presentation had a major influence on society. The linguistic barrier, it was thought at the time, would be broken relatively shortly. Following that, as a reaction to the shock of Sputnik's launch, the US government-funded an astounding $20 million on automatic translation research. [1, 2]

ALPAC delivered a somber report to the National Academy of Sciences in the United States in 1965. Because automatic Translation will not be practicable for the years ahead, the paper recommends that extensive Research be focused on language theory and comprehension to operate as the technology's foundations. Budgets for machine translation were subsequently curtailed in the United States, and the focus shifted to fundamental Research, with interpretation and comprehension as the key ideas. Winograd's language comprehension utilizing global knowledge in 1970 is a well-known result from this period. However, the information basis for this type of study was insufficient and therefore cannot be considered to have positively linked with increased automatic translation performance in a broad or realistic sense.

In the early 1980s, Japan saw three major technology waves: rule-based Translation, example-based Translation, and statistically-based Translation. The Mu project, which intended to interpret abstracts from the Scientific and Technology Agency's science and technology literature, was a breakthrough in Japan. As a result, Research & innovation into lexicon and guidelines (analytic grammar rules, conversion rules, and generative grammar rules) for rule-based machine translation managed to receive traction. Bravis, a company, has begun selling a commercial translation tool. This prompted big-name IT businesses, including Fujitsu, Toshiba, NEC, and Oki Electric Industry, to commercialize automatic-translation software. This rule-based technology is used in practically all commercial software packages today, as well as nearly all Web-based applications. Although improved and more thorough specialized lexicon was an efficient strategy to enhance translation quality, gradual but constant attempts to increase dictionary sizes beyond a few tens of thousands to millions of entries have been made.

Meanwhile, in 1981, Kyoto University professor Makoto Nagao drew inspiration from human Translation to design an example-based translation approach that uses phrases identical to the input text and their translations (together known as "example-based translations"). This example-based Translation joined with other Research at Kyoto University and ATR around 1990, spawned a second wave that swept the globe. Some commercialized rule-based systems have included this strategy; It is already being utilized as the foundation for a Japanese-to-Chinese technical and scientific publications translation initiative conducted by the National Institute of Information and Communications Technology (NICT). [5].

Later, in 1988, IBM presented an approach known as statistical machine translation, which combines a bilingual corpus with pure statistical processing that removes grammatical and other expertise. For a long time, nevertheless, this strategy received relatively little attention for a variety of reasons: the paper was tough to comprehend, the operating system was lacking, translation corpora were too minimal, the execution method was only published in patent specifications, and it was ineffective for languages other than related languages such as English and French. However, in the year 2000, a new approach known as phrase-based statistical machine translation was devised, and this ushered in the third significant wave, which was aided by larger bilingual corpora and more powerful computers.

ATR created its speech-translation system [1,2] by assembling a corpus of commonly spoken travel discussions is designed to enable travel conversation speech translation. The Research has so far produced a Basic Travel Expression Corpus (BTEC) with 1,000,000 matched pairs of Japanese and English sentences, as well as 500,000 matched pairs of Japanese-Chinese and Japanese-Korean utterances. This is the world's biggest multilingual travel conversation translation corpus. The average length of the English sentences in the corpus is seven words, and they include subjects including greetings, difficulties, purchasing, transit, hotel, tourism, meals, communication, airports, and business. Computers with increased processing power and memory, as well as more pervasive networks, are allowing portable speech translation systems to be implemented. Developments are taking place in the establishment of standalone systems on small hardware as well as distributed implementations that connect mobile phones and other devices to high-performance servers across the network. Due to concerns such as size, weight, and battery life, the standalone technique is not possible to apply on a portable computer. In instances when wifi and other facilities are unavailable, there is also anticipated to be a demand. As a result of these concerns, attempts are made to commercialize specialized mobile devices with built-in speech-translation capabilities. NEC produced the earliest commercial mobile device with onboard Japanese-to-English voice translation in 2006 (with hardware requirements of a 400-MHz MPU and 64 MB of RAM). Meanwhile, in November 2007, ATR created a distributed version of a speech translation system for the DoCoMo 905i family of mobile phones utilizing mobile phones and network servers. The technique, termed "shabette honyaku" (see Figure 4), was developed by ATR-Trek and is the world's first mobile phone-based speech translation service. Then, in May 2008, the DoCoMo 906i series launched a Japanese-to-Chinese speech-translation service. The design of the speech recognition module utilized in distributed speech translation is shown in Figure 5.[2].

Background noise reduction, acoustic analysis, and ETSIES 202 050-compliant encoding are all performed on the mobile device (front end), and only the bitstream data is sent to the speech recognition server. The back end of the speech recognition server extends the incoming bitstream, conducts speech recognition, and determines word reliability. One of the advantages of this design process is that it is not constrained by the mobile phone's information-processing limits, allowing for the deployment of large-scale, extremely exact phonological and linguistic models. Because these models are stored on a server rather than a mobile phone, they are simple to update, allowing them to be kept up to current at all times. The technology is already in widespread use: as of June 2008, it has been used by a total of 5,000,000 people. [1, 2].

1.4. The basic model for Speech-to-Speech Translation

The high-tech Speech-to-Speech Translation system that empowers such multilingual communications supports a pipelined architecture of automatic speech recognition, machine translation, and speech synthesis or text-to-speech that mainly depends on linguistic information while disregarding some other valuable insights existing in speech and spoken discourse such as noise and human utterances. There is minimal connection between the fundamental components prior to the pipeline, and no people are involved in the loop for autonomous learning, adaptation, or collectively controlling the interaction. To transcend these fundamental problems, pipeline resilience must be built at all levels.

The system's foundation is that it should take advantage of the rich context that exists outside of the dictated words while also being aware of and cooperating with various cultures to improve information transfer, communication efficiency, and social co-presence in order to enable successful multilingual interactions. Various projects in this study capture that model and transmit highly linked information for voice prosody, discourse, and user state behavior in order to allow robust translation and considerable synthesis of the target language [3].

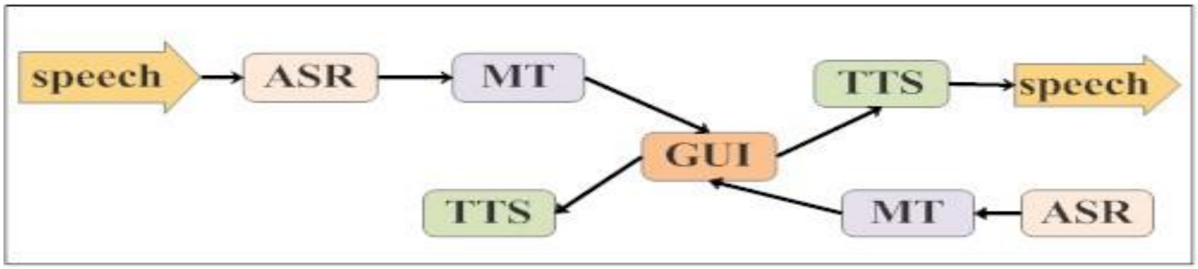


Figure 1: Overall Speech-to-Speech translation system

2. CHALLENGES

Computer-assisted cross-lingual conversation by automatic speech-to-speech Translation has been one of the most challenging problems in spoken language technologies in decades. Recent remarkable advances in speech and language processing led by deep learning techniques benefit this challenge by real-time and accurate speech translation. One crucial problem in automatic speech-to-speech Translation is its delay. Spoken language processing tasks are usually handled at the utterance or sentence level. Their application to speech-to-speech Translation suffers from a long delay that is proportional to the input length because the process starts after the observation of the end of an utterance. That is similar to consecutive interpretation and is not useful for long monologues such as lecture talks. On the other hand, in such situations, simultaneous interpretation is often used for an audience not proficient in the language of a talk. Simultaneous interpretation is a challenging task to listen to the talk and speak its interpretation in a different language.

Attributed to the reason that state-of-the-art automatic speech recognition and machine translation systems are far from flawless, there are substantial outstanding research concerns that impede the implementation of normal and unrestrained voice-to-speech translation systems, even for relatively limited application domains. Furthermore, unlike interpreting written information, colloquial spoken words are frequently communicated with sloppy grammar and informal, spontaneous speech [Ref 1.2].

In fact, while creating demonstration systems, substantial limits on the application domain and the sort and format of allowable utterances are often imposed, i. e., both in terms of the range and breadth of user input that can be provided at any time during the encounter. As a result, the system's flexibility and naturalness of use are jeopardized.

S2ST between Western languages and non-Western languages, such as English-from/to-Japanese or English-from/to-Chinese, necessitates the use of technologies to overcome the linguistic differences. A translation from Japanese to English, for example, necessitates 1) a word separation process for Japanese due to the lack of explicit spacing information, and 2) transforming the source sentence into a target sentence with a drastically different style due to their word order and word coverage, among other factors. Another factor for S2ST is that the technology must be portable across all domains because S2ST systems are frequently utilized for applications in specialized situations, such as facilitating nonnative language discussions by tourists. As a result, the S2ST technique in voice recognition, machine translation, and speech synthesis must incorporate (semi-)automated mechanisms for adapting to unique situations/domains and language pairs [3].

Some other challenges in exploiting rich contextual features in machine-mediated speech-to-speech Translation include:

1. Designing appropriate user interfaces that can augment the translation hypotheses with contextual information.
2. Detecting and exploiting source language prosody and dialog information in target text-to-speech synthesis.
3. Actively learning and adapting the system from a user feedback or user in the loop.

However, there are indeed numerous obstacles to overcome before this technology becomes viable. The requirement to encourage new languages, as well as the automatic adoption of town names, given names, and other common words, are only a few examples. Simultaneous interpretation, in which a continuous stream of speech is interpreted, should also be developed using technology. Speech translation systems have a wide range of applications, including speech information retrieval, interactive navigation, dictation, summarization, and archiving, and more applications are predicted to arise.

S2S technology's three aspects (speech recognition, language translation, and speech synthesis) each have their own set of challenges. This innovation requires the recognition and Translation of spoken language, which is far more challenging than interpreting text since spoken language comprises ungrammatical, conversational idioms and lacks punctuation such as question marks, exclamation marks, and quote marks. Speech recognition errors also lead to significant mistranslations. As a result, instead of enabling all types of discussion from the start, researchers have selected a development approach that focuses on boosting accuracy to a useable level by firstly constraining the system to comparatively straightforward speech [Ref 1.2].

An instance of spoken English translations of a Japanese statement can be found underneath. "Mado o akete mo ii desu ka" means "mado o akete mo ii desu ka" in Japanese. Here are the English sentences that correspond:

1. Is it okay if I leave the lid open? 2. Is it alright if I open the curtains? 3. Is it possible for me to open the window? 4. could we breach the window 5. is it alright if I open the window 6. would you mind if I rolled down the window 7. is it alright if I open the window 8. do you mind if I open the window 9. would it be alright if I open the window 10.

As all these examples demonstrate, speech translation utterances are not entire statements — they frequently lack subjects, and no capitalization is utilized in subjects and proper nouns – and perhaps even questions lack question marks. Highly conversational terms must also be dealt with. A dataset called Field Experiment Data was also analyzed, and so was information obtained from a corpus of roughly 10,000 utterances of dialog captured under real-life situations and mediated by a speech translation system called Machine Aided Data (MAD) (FED). This information was obtained over the course of five days at Kansai International Airport, with the assistance of the Osaka Prefecture, throughout December 2004 and January 2005. The information comprises approximately 2,000 utterances of discussion handled by a speech-translation system involving foreign-language speakers (39 English speakers and 36 Chinese speakers) and facilitators at a tourist resort [4].

However, there are still numerous research hurdles to overcome; specifically, there is a high level of speaker reliance and expression variability; also, new terms and ideas are continuously being developed in response to societal changes. Speech translation technology is now limited to short utterances of roughly seven words, such as a trip chat. As a result, there are many other unanswered questions before speech translation can handle long, complex speeches like those seen in newspapers or lectures. Some of the immediate technical challenges include:

1. Standardization for using the technology to connect speech translators all around the world.
2. Monitoring and assessing usability in practical applications.
3. Relaxation of copyright to allow web usage of example translations
4. Using the most recent proper nouns based on the user's present location
5. Multiple language support is available.

There are thousands of unwritten languages in the world, which are purely spoken and have no written text. Just like in Jia et al. (2019), prolonged speech (which generally includes information, context, communicating style, and other factors) is far more adaptable than discrete symbols in representing semantic meanings (text), this renders speech translation more difficult than text translation. As a result, reducing the continuous fluid space of speech into a more constrained discrete area is the best way of making speech translation easier for unwritten languages.

The creation of S2S systems necessitates the collection of diverse types and quantities of spoken language data: several hours of acoustic speech data, adequately reflecting dialectal variants, for language modeling, there are several hundred thousand active words of domain data, as well as millions of words of parallel text in target language pairings for machine translation.

3. LISTS OF CHALLENGES

3.1. Existing and collected data

The initial barrier in both Pashto and Persian was indeed the lack of speech/language input in any form — lexicon, translations, or acoustic data. The First Discourse data set, which contains 20 read sentences from 300 individuals going in age, sex, schooling level, and tongue, for a sum of 6000 expressions, was the sole information accessible on the Persian side (available from ELDA). Because the transcripts were unsuitable for voice recognition, they had to be reconstructed. To supplement these transcripts, we enlisted the help of Persian speakers from all across Los Angeles, who provided read and semi-spontaneous speech data. In a Wizard of Oz-style scenario, the semi-spontaneous speech was elicited, while read speech was gathered and confirmed by the speakers themselves using an interactive data collection tool.

These data couldn't be used to simulate language: Parallel to creating the data mining approaches mentioned below, 300 Standardized Patient (medical student-actor patient) sessions were conducted at the USC campus in partnership with the Medical School to gather data for language models [1]. Following that, the SP data was transcribed, yielding almost 300,000 words of in-domain data in both English and Persian. The only pre-existing corpus of recorded Pashto we could uncover was a series of untranscribed Voice of America (VOA) Pashto service broadcasts, which the Linguistic Data Consortium had captured from the broadcasts. The data, on the other hand, were not well suited for the task: the broadcasts were controlled by less than twelve speakers, exhibited just a proportion of Pashto dialect variety, were of poor audio quality, and did not represent dialogue speech pattern. As a result, we gathered about 80 Pashtuns from a nearby emigre population and asked each of them to record 100-200 randomly created utterances, including responses to inquiries. The entire amount of speech recorded was around 7 hours. The VOA data was also translated for 5 hours. These 12 hours of speech, or nearly 100,000 running words, were used to develop the acoustic model.

3.2. Transcriptions & Lexicons

The transcription schemes of the languages, particularly those that can enable machine spoken language processing, have been the singular biggest impediment to all parts of the Pashto and Persian languages in this work. Because Pashto lacks a uniform, standard system of writing or spelling rules, a single word can be written in a variety of ways, and distinct words can be spelled in the same way. We attempted to translate acoustic data straight into a phonemic description to avoid the non-standardization of the original orthography. However, we discovered that translators struggled to recognize the phonemes, which was made even more complicated by the fact that our phonology assessment was designed to encompass a wide variety of Pashto dialects. As a result, we switched to indigenous (Arabic-based) writing, which, despite its challenges, proved to be more trustworthy. The set of words was, therefore, phonemically translated in indigenous script orthography. Each script format was often connected with several phonemic representations. These representations could indicate significantly different pronunciations, totally different words written in just the same style, or syntactic patterns that are distinct but equivalent. Similarly, any given phonemic representation can be connected with one or more script forms, each with one or more intended meanings. Such texts might be used for language modeling speech recognition processing using an isomorphism of the indigenous script. This isomorphism was creating ambiguity both at the input and at the output of the recognizer. Each "isomorphic class," which served as a "word" for all practical purposes of recognizer training and testing, had many pronunciations, sometimes quite a few, affecting acoustic model training and search accuracy. The recognizer's output, on the other hand, matched to a series of word classes and did not distinguish between word meanings within the same class; therefore, it was handed to the translation engine to handle. Although Persian has a standardized writing system, it is not well-suited for use in a speech-to-speech translation system. The Modern Persian alphabet is based on the Arabic alphabet, with the addition of four letters and changes to two-character designs. Despite the capability to indicate vowel sounds in the written script, this process is rarely used in Persian transcription, resulting in a lossy encoding of the required orthography (this problem incidentally is common, at varying degrees, to all languages that use the Arabic script). Three alternative encoding techniques were devised as a result of the solution: A one-to-one mapping from the Arabic script to the Latin alphabet (USCPers), an augmented version (USCPers+) that encodes the extra vowel information found in spoken Persian, and a phonetic transcription scheme (USCPron) that allows the ASR and TTS components to be created [3]. Given the structure of the foregoing transcription strategies, the very next step was to gather enough data to create a lexicon with mappings all along three dimensions. Clearly, generating pronunciation from the Arabic script is an ill-posed problem, necessitating the need of humans to create dictionaries. Transcribers and transliterators turned clean English utterances into these numerous formats during the first data collection, building pronunciation dictionaries at the same time. Because some early data was available, statistical learning methods might be used to automate the procedure.

3.3. Acoustic modeling

We employed front-ends with 16 kHz sampling rate, 10 ms frame advance rate, 12 mel frequency cepstral coefficients plus normalized energy, and first- and second-order differences for both Persian and Pashto systems (39 features). Persian contains 34 phonetic units (29 phonemes, silence, br, ls, ga, and LG), whereas Pashto has 43. (41 phones, silence, reject). State clustering was used to train 3-state triphone hidden Markov models (HMMS). 4207 clustered states were utilized in the Persian, with an average of 14 Gaussians per state. To match the limits of the predicted tiny footprint platform, the Pashto system employed a substantially lower model size (129 phone-state Gaussian clusters with 32 Gaussians each, trained via discriminative maximum mutual information estimation) (MMIE, [5]).

The Persian system employed the SONIC [6] speech recognition engine, whereas the Pashto system used DYNASPEAK [7]. Both techniques employed an English phoneme mapping into the target language to initialize the models, which were then altered or reassigned using the limited amount of information available. For Pashto, we only employed a knowledge-based (linguistic) phone mapping approach; however, for Persian, we looked at three distinct methods: knowledge-based, data-driven phoneme mapping, and data-driven state mapping. The Earth Movers Distance (EMD) method was used in the data-driven strategies to try to reduce the amount of effort required to turn one GMM into another. We improved phoneme level recognition by 2.7 percent utilizing EMD at the sub-phoneme level, compared to using simply Persian speech data. This benefit, while appealing for little amounts of data, becomes unimportant as the quantity of the data grows. The models produced from cross-lingual phonetic alignment were exclusively employed for alignment reasons in that scenario. As we expand our translation system into languages with fewer resources, we want to take advantage of cross-lingual expertise even more. We currently have speech recognition engines in multiple languages (English, Persian, Arabic, Greek, and so on), so the pool of potential GMM mixtures is growing, and we anticipate that the potential benefits will be greater and that a procedure for rapid language portability into new languages and dialects will be developed.

3.4. Language modeling

To offer enough surface form coverage for these 2-way S2S systems, language modeling requires orders of magnitude more data than acoustic modeling. Written material can, however, be utilized as a rough approximation to spoken language transcripts, as is common practice (although easy access to text data may be difficult, e.g., Persian, or even not possible for some of the target languages, e.g., Pashto). We faced two key challenges in developing adequate language models for the Persian-English and Pashto-English translation systems. The first was a lack of (medical) domain data, and the second was a lack of generic background data for bootstrap in Persian and Pashto. Take the Persian-English system, for example, where there were some currently accessible textual materials (medical domain data in English and some Persian text). Multiple simultaneous procedures were used to generate relevant domain data: The initial stage was to find any existing medical domain text, which we accomplished through the use of medical phrasebooks, paraphrase, and Wizard of Oz data collectors, among other methods. This content is clearly restricted, which was used as a starting point for mining web data. The Web has a lot of text and even some transcribed information, but it's also tough to find and automatically filter the relevant in-domain content. Our first attempts were centered on a bag-of-words technique [8], but we eventually created far more powerful algorithms. The current approach [9] is centered on an iterative Web crawling strategy that employs a competitive set of adaptive models, including a generic topic, a noise framework indicating spurious text commonly faced in web-based data (Web data), and a topic-specific model to yield query strings for WWW search engines using a relative entropy-based approach and to weigh the available to download Web data appropriately for building topic-specific language models. When compared with the results of a generic model built with only 5K words of in-domain data as a seed corpus, this method yielded a 14 percent improvement.

In addition to supplying in-domain English data, we utilized the simplified bag-of-words web data technique to mine Persian text, allowing us to develop a Persian background language model. Caused by a lack of relevant Persian web resources, this initiative did not achieve the same level of success. In addition to establishing data mining tools, 300 Standardized Patient (medical student-actor patient) sessions were held on the USC campus in partnership with the Medical School. The SP data was then translated, yielding more than 300,000 words of in-domain data. The language models that resulted were built on layers of data, most notably medical phrasebooks and paraphrases (English and Persian) in domains that were personally gathered (generic web-data, domain web-data, and generic models).

Approximately 21,000 English words and more than 8000 Persian words are represented in the models. Moreover, both the English and Persian models are class-based, allowing us to supplement crucial classes such as prescription names, pleasantries, connections, and various sorts of lexical features with human knowledge. Language modeling was made considerably more difficult due to Pashto's morphological complexity and the limited amount of accessible training data. We solved the problem by adopting the technique described in and creating a language model with finer backoff layers than a standard word language model. To do so, we first created a vocabulary clustering tree, with the root representing the whole vocabulary and each node representing a node with all terms in its descendent nodes. The tree is created using a measure of similarity depending on the left and right contexts of a word and the least discriminative information clustering technique. When assessing the transition probabilities of a word based on its n-gram prefix, we first back off to its context, replacing the most distant word with its class, from the most specific to the most general, and then back off to the normal lower-order (n-1)-gram prefix if none of these backoffs could guarantee a minimum number of occurrences. The language model that results has a relative perplexity reduction of over 10% and a sig.

3.5. Speech recognition

Several speech recognition tasks have been explored and assessed during the last 15 years. Each task proffered its own set of difficulties. The following characteristics distinguish such tasks: kind of speech (pre-rehearsed vs. spontaneous), communication goal (computer, audience, person), and bandwidth (FWB, full bandwidth TWB, telephone bandwidth, FF, far-field). Dictation (WSJ), broadcast news, switchboard, voicemail, and meetings are just a few of the responsibilities.

The following is a list of them in order of word error rate (wer) 7 percent dictation, well-formed, computer, FBW 12 percent, varied, viewer, broadcast media FBW Voicemail: 30 percent spontaneous, individual, TWB Switchboard: 20-30% spontaneous, person, TWB Meetings are 50-60% spontaneous, with person FF. At the moment, the trait having the greatest impact on word mistake rate is spontaneous speech, preceded by environmental influence and domain dependency.

3.6. Speech synthesis

In a voice-to-speech translation system, speech synthesis is essential. One of the most difficult goals for speech synthesis is to mimic the human voice. Gender, age, and cultural adaptation are all new challenges in the multilingual human-to-human communication framework. Emotion and prosody are also crucial considerations. [7] [8]. Concatenation of different acoustic units is currently the most effective way to generate synthetic speech. This method differs from previous rule-based synthesis, which requires explicit knowledge and competence to construct the deterministic units. Because the unit selection process in a corpus-based approach encompasses a combinational search across the whole speech corpus, fast search algorithms have been proposed as an integral part of the current synthesis process. The primary components of corpus-based approaches for specifying the speech segments necessary for concatenative synthesis are a unit selection algorithm, some objective measurements utilized in the selection criteria, and lastly, the development of the needed speech corpus. From an application standpoint, the massive amount of storage required to leverage the concatenation of speech units severely restricts the type of application. The other two major challenges in speech synthesis are prosody and speaker characteristics, as well as speech segment design. It is vital to provide sufficient tone and emphasis, rhythm, pace, and accent in order to manage prosody. There is a need for segmental duration control and fundamental frequency control. The global spectral features reflecting vocal tract characteristics, as well as the glottal waveform of voice excitation, contain not only language information but also speaker voice characteristics. Furthermore, as indicated by variations in voice quality and prosody, paralinguistic elements impact speaking styles. The majority of improvements are anticipated to come from prosodic modeling. Investigating this field and attempting to master the language and extralinguistic phenomena will almost certainly address multicultural issues, which are crucial in a multilingual communication process.

3.7. Machine Translation

Besides speech recognition and synthesis, the translation component is the core of a speech-to-speech translation system. The standard machine translation (MT) issue, which involves translating a text from a source language, such as Italian, into a target language, such as Chinese, is not the same as the S2PT problem. To begin with, no individual is engaged in the standard MT dilemma. The procedure is one-way only. The text is designed to be 'correct' linguistically. Two persons are participating in the S2ST process, the procedure is bi-directional, and the speech is colloquial, spontaneous, ungrammatical, and combined with non-verbal cues. Furthermore, the surroundings are a key concern in terms of acoustic noise and interaction mode. In S2ST, near-real-time Translation is required. Then, because people are effectively engaged in the process, the comprehending task is carried out collaboratively by individuals. Finally, because a machine is engaged in the interpretation in any case, there is a significant problem of human-machine communication to address. All of these elements must be considered in order to address the S2ST problem. Various designs were used, some of which use an intermediary language (interlingua, interchange format), and others use a direct translation approach. JANUS and NESPOLE architectures are two common examples of the first scenario. [9]. The analysis and synthesis chains are the two main processing chains in the Italian implementation of the NESPOLE [S2ST system architecture]. The analysis chain converts an Italian acoustic signal into a (sequence of) IF representation(s) by passing it through the recognizer, which also generates a succession of word hypotheses for the input signal; and the recognizing module, which delivers IF representations using a multi-layer argument extractor and a statistical-based classifier. The synthesis chain starts with an IF expression and ends with a synthesized audio message in the target language that expresses that content. It is made up of two modules. The generator transforms the IF representation into a more language-oriented model before combining it with the domain expertise to generate Italian phrases. These kinds of statements are sent into a voice synthesizer. The ATR-MATRIX architecture is an exemplification of the direct translation approach [10], as it employs a cascade of a speech recognizer with a direct translation algorithm, TDMT, whose produced text is then synthesized. Example-based algorithms are used to accomplish the direct translation technique. Starting with text translation, IBM[11] [12] pioneered the second example of direct Translation based on statistical modeling. Statistical Translation has also been created as part of the EU-TRANS project and the VERBMOBIL project in Germany. Currently, Research is underway to develop unified or integrated techniques. The quintessential aim of this model is to harmonize speech recognition, understanding, and Translation as whole statistical processing. "For a future study on spoken language translation, we consider this integrated methodology and its appropriate implementation to be an open question." The most important experience gained in the VERBMOBIL project, in particular a large-scale end-to-end evaluation, revealed that the statistical approach resulted in significantly lower error rates than three competing Translation approaches: the sentence error rate was 29%, compared to 52% to 62% for the other translation approaches. Furthermore, the end-to-end assessment process is a critical concern for S2ST systems. The focus is to create a strategy that is based on objective facts. VERBMOBIL, CSTAR, and a number of other organizations have suggested and created evaluation systems.

3.8. Improve the end-to-end performance significantly.

This is the main test to be tended to sooner rather than later. It appears to be that brought together techniques in light of factual displaying are exceptionally encouraging, given that a few central questions will be managed and appropriate arrangements worked out. This approach permits the incorporation of acoustics, phonetic setting, talking rate, speaker varieties, language highlights like sentence structure or semantics, and so forth into one bound together with the way. Then, at that point, this approach mutually streamlines acoustics, language, and speaker impacts. From the displaying point of you, it addresses, all in all, a shift from the source model. Significantly more work is required in proposing new computational instruments and developing them. This approach is likewise steady with the speech synthesis perspective: corpus-based and information-driven A test will likewise be the double-dealing of genuine applications in a restricted area, i.e., the travel industry, of frameworks in light of interlingua draws near. Central questions for this situation are movability and power.

3.9. Produce aligned multilingual corpora and lexica

Corpora and lexica are a vital problem in having to provide the difficulty of constructing new models in the hopes of greatly improving performance. There have been plans to gather and transcribe 5000 hours of spontaneous speech in order to solve the challenge of spontaneous speech recognition [14]. This is indeed a contentious issue, but it is what we have discovered from our previous speech recognition experience. The test data could come from both old and news sources. Multilingual text corpora that are aligned for Translation are also essential. A joint effort is underway with ATR and IRST, as well as the other members of the CSTAR III consortium, to create aligned text corpora composed of transcription and Translation of phrasebooks in the tourism domain. This phrasebook covers a broad range of situations: emergency, timetable, transport, sightseeing, directions, attractions, hotels, shopping. Aligned multilingual lexical are also important language resources for future S2ST systems development. Current activity is under development in LC-STAR [15], a newly funded project in the Vth framework by the EU.

3.9.1. Integrate speech to speech translation components in real applications

Real services and applications involving speech communication need to manage the "interface problem," i.e., the physical impact of the user with a device that involves multimodal multimedia in a ubiquitous environment. A wearable device, a PDA or 3G cellular, cannot be operated by a keyboard and requires sophisticated natural multimodal human interfaces. Speech, vision, and handwriting seem natural candidates for human-machine interaction. But how can a system provide seamless integration between human-machine services and human-human services? How can the system blend the two, provide assistance and guidance for a user to access and understand databases and information resources, but also serve as a go-between to facilitate the interaction with other humans or with a user's direct environment?

4. A NEW ACTION IN EUROPE

Given the challenges previously discussed and the experience carried on in the previous and ongoing projects, a new and innovative initiative is needed to tackle the problem. This initiative, in order to be successful, needs, first of all, a critical mass of researchers. Within Europe, few research groups have the capability to build up complete SST systems. Most research groups are small and work only on some research themes, i.e., prosody, acoustic modeling, language modeling, speech synthesis. Although these small groups may have excellent researchers, their work has less impact on the development of SST components. This new initiative should provide an appropriate infrastructure to use in an effective way the intellectual potential of European researchers. Given the big shift needed in order to set up this new action, a group of major European players in the spoken language technology, both research institutions, industrial entities, and ELDA, proposed a preparatory action, which acronym is TC-STAR_P (Technology and Corpora for speech translation).

4.1. Goals and Activities

The preparatory action, under negotiation, fits with the action line IST2002-III.5.2 c) "preparing for future research activities." It is scheduled to begin in July 2002. The duration will be one year with the purpose of preparing and getting ready an integrated project for the VI Framework. An integrated project is a large-scale action with the purpose of creating the European Research Area, ERA. The activity of the TC-STAR_P will be carried on by the cooperation of the four groups: an industrial group with proven experience in SST technology development, a research group with proven experience in Research in SST technologies, an infrastructure group with proven experience in producing language resources for SST components and with proven experience of evaluation of SST components and systems. Then a dissemination group will be in charge of using and spreading the project's results. Three are the main goals of this action: • developing research roadmaps and associated implementation models • identifying and bringing together all relevant actors in the Speech-to-Speech Translation (SST) area • investigating effective mechanisms for managing future activities 4.1.1 Preparing RTD roadmaps and associated implementation models The consortium is composed of different RTD communities: industrial, academics, and infrastructure entities. All these organizations will contribute to developing common visions and analyzing research requirements for SST systems. As a result of these tasks, industrial partners will prepare roadmaps for technical implementations and services; the scientific and academic groups will prepare roadmaps for technology improvements, and the infrastructure group will provide roadmaps for production and evaluation campaigns. The work will include a case study where industrial partners and research partners will provide application-oriented and Research input, respectively. The infrastructure group will focus on preparatory tasks for setting up production, evaluation, and validation centers for the needed LR. 4.1.2 Identifying and bringing together all relevant actors the consortium includes some of the most relevant actors in the SST field. One of the objectives during the lifetime of the project is to attract further key actors from the Industrial

Research and infrastructure groups, as well as SMEs working with SST applications and related fields. Within the infrastructure group, the key action is to attract and prepare contacts with national agencies for funding language-specific LR-production in the future FP6 and with entities working on the evaluation and validation of language resources. The development of language resources is a very expensive activity, which must be best tackled by coordinated funding actions at national and European levels. 4.1.3 Investigating a new management model According to the IST 2002 Work program, Action Line 3.5.2 should focus on building and strengthening RTD communities by encouraging Research, business, and user organizations to develop together common visions and analyze research requirements in order to identify common challenges and objectives; and on investigating effective mechanisms for managing future activities. Moreover, a cornerstone of the future work to be developed under the Integrated Project is the management structure. In accordance with Action Line 3.5.2., the work to be performed under TCSTAR_P includes exploring a new organizational model in order to allow partners to collaborate in pursuing the final goal smoothly. This important task will be investigated during the project. Issues such as distribution of work and resources, admission and withdrawal of participants, engagement of additional parties, scientific guidance, monitoring, etc., will be examined. The model has to be effective to reach the envisaged goal, to react to external new trends, needs, and demands coming from the market, society, and scientific community. Section 2

4.2. A Comparative Study with Human Speech-Translation Capability

Essentially, evaluating the efficiency of speech translation is exceedingly challenging. If the speech synthesis component is not included in the assessment, the efficiency of the result is assessed by entering a number of sample sentences into the system. The approach for assessing voice translation is largely the same as the model for assessing computerized text translation in this regard. Speech translation, on the other hand, evaluates utterances rather than strings of text. To assess translation quality, two approaches are used: one in which the translations are humanly rated on a five-point scale, as well as another in which the closeness between both the system's output and previously generated reference translations are compared. For the latter, a range of evaluation scales has been suggested, including BLEU, NIST, and word error rate (WER). These scales have become increasingly popular in recent years. These data can be used to compare two different systems because they are basic numerical values. What these scores can't tell you is how well the higher-scoring system will function in the actual world. A strategy for resolving the situation has been presented, which involves measuring system performance in human terms and calculating the system's Test of English for International Communication (TOEIC) score. To begin, native speakers having documented TOEIC scores ("TOEIC takers") are requested to pay heed to test Japanese statements and interpret them into spoken English.

Following that, Japanese-English bilingual assessors evaluate the TOEIC takers' interpretations of the results of the speech-translation system. The fraction of test phrases for which the humans' translations are better would then be measured as the human win rate. After calculating the human win rate for all TOEIC takers, regression analysis is utilized to generate the speech-translation system's TOEIC score.

Figure 2 depicts the system's performance in TOEIC scores. The speech-translation method is virtually always correct when employing relatively brief utterances like those in basic travel conversation (BTEC). However, the efficiency of the speech-translation system on conversational speech (MAD and FED) is comparable to a Japanese speaker's TOEIC score of 600. Additionally, when coping with prolonged, infrequent, or complicated utterances, performance suffers dramatically. As a result, there is still potential for development in terms of performance.

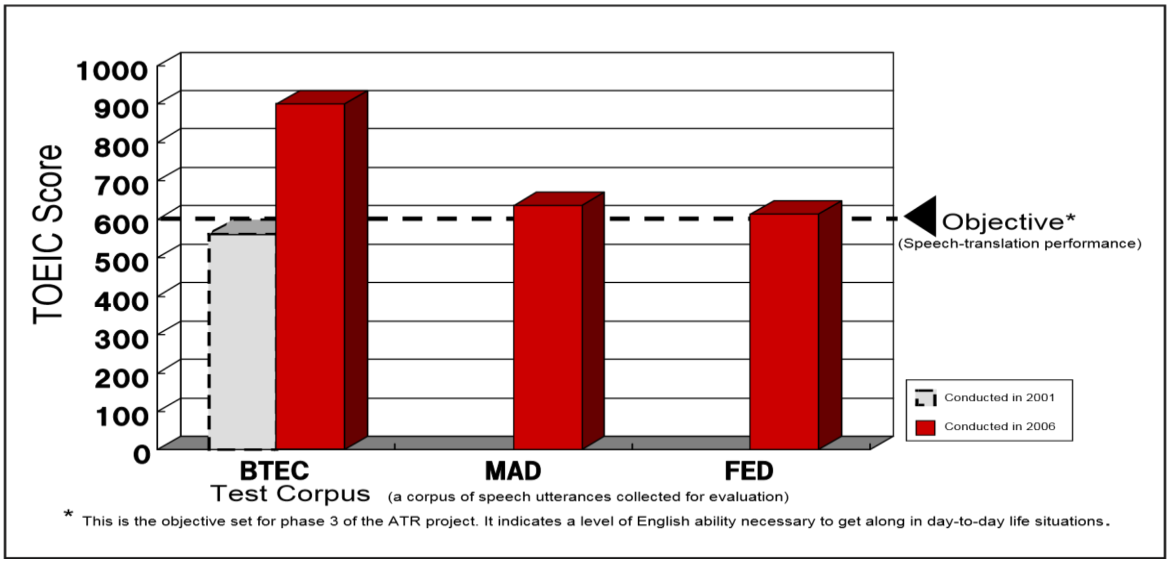


Figure 2: TOEIC scores are an instance of how to assess the reliability of speech translation.

Source (S.Nakamura et al., "ATR Multi-lingual Speech-To-Speech Translation System," IEEE Trans. ASLP, vol.14, no. 2 (2006)

5. THE FUTURE

Figure 7 depicts the history of speech translation to this point, as well as future research and development prospects. A worldwide research collaboration working on Asian languages hopes to prototype spoken translation via the Internet in 2010. By roughly 2015, the worldwide research collaboration is expected to release a prototype that includes Western European languages and has more standardized interfaces. After many field tests, Japan's Project to Accelerate Benefits to Society (detailed in the next section) expects to build technology for networked voice translation by 2012. Speech translation competence of continuous simultaneous interpretation of business and lectures is predicted to be ready by approximately 2015, and by 2025, it is expected to be available. Multilingual simultaneous interpretation with contextual information and summary is projected to become accessible, moving closer to the goal of simultaneous interpretation.

Speech translation has become more practicable as a result of technological advancement in speech and language research, especially for basic themes with a clear value of use. Speech translation, on the other hand, has only recently progressed to the stage of developing core technologies. Research and development should be accelerated in order to achieve more sophisticated speech translation. The following are some points that should be the focus of future attention. To begin with, corpus-based solutions have the distinguishing property of improving with use. As a result, securing opportunities for field and social testing, and making active use of technologies that have been developed, is important.

Multilingual speech-translation technology can be tested during activities like the Olympics and World Fairs when speakers of many various languages are anticipated to participate. As a result, it is vital to seize these opportunities to advance technology. NICT ran a monitor study at the Beijing Olympics, focusing on Japanese visitors. It created a speech translation system in Beijing that supported proper nouns, and it had monitors utilize speech translation devices to speak throughout the city, utilizing the devices for transit, tourism, and shopping. Users' satisfaction with the service was surveyed using a questionnaire, bringing voice translation technology one step closer to practical implementation.

If Japan wants to want to be a major tourist destination, providing continuous tourism information services to international tourists through spoken language translation might be an efficient method. Meanwhile, as the majority of international residents and employees in Japan grow, multilingual voice translation may become a necessary tool for local authorities, medical institutions, law enforcement, and schooling. It should assist cut interpretation expenses even if translators are present. However, if all of these methods are employed independently, only partial knowledge would be gathered, making feedback for Research and development wasteful. To enhance efficiency, national and local government, as well as the business sector, will almost certainly need to develop a cooperative framework. Distributing compact translation gadgets to public organizations where their use is anticipated, for example, maybe useful, as could lending these gadgets to foreign workers and visitors for free. Second, speech translation is a technique that converts spoken words into written words in many languages.

Although interpretation into English is important, it will be much more so if speech translation can function directly between Japanese and the local languages of a variety of nations. As a result, expanding the number of languages supported is critical. When it comes to gathering corpora, there are limits to how far this Research and development can go in Japan alone. A plan for collaboration across nations with a wide range of languages is required; in other words, a system is required to allow diverse countries to collaborate on speech translation, speech, and language research. As a system for collaborative R&D, establishing worldwide spoken language technology research centers and the likes should provide input from a wide variety of studies into the gathering of speech and dialect data language structure. Third, when several nations begin to explore and create speech translation, standardizing the interfaces that connect these multiple language processing modules will be important. The creation of connecting techniques, data formats, dictionaries, and other such tools must have a standardized mindset in mind. We must prevent a situation in which each nation creates its own system that is incompatible with the others. Japan has improved speech translation technology, and as a result, it may lead other nations in terms of standards. Finally, the copyright must be taken into consideration. Speech and language processing necessitate speech and text corpora, and the amount and quality of these corpora have a significant impact on speech translation performance. As a result, the utilization of news broadcasting companies, newspapers, and the Internet is incredibly successful. The current copyright legislation does not allow for secondary applications like these sorts of corpora. It will be required to amend and manage the legislation to make it more flexible in order to explore and create new technologies. The Cultural Council's Copyright Working Group is presently debating this issue, and a decision will be made soon. After the findings of this study are disclosed, it will be important to rearrange the themes for future full-scale speech translation services and review the response, including service models.

The key goal over the next several decades is to create speech recognition algorithms that are as accurate as human performance. This indicates that for both spontaneous and reading the speech, it is regardless of the surroundings, area, and job. The major priorities will be on enhancing spontaneous speech models (i.e., prosodic characteristics and articulatory models, multi-speaker speech, collecting an acceptable volume of conversational speech, etc.), as well as modeling and training strategies for multi-environment and multidomain scenarios. Then there'll be the issue of language modeling. Several dynamic language models are generally recognized to function better in particular domains. Establishing a language model that functions effectively across several domains will be a significant step toward emulating human performance. The project aims to achieve a very fast dynamic adaptability at the word/sentence level. Finally, the continual improvement of computer efficiency over time, the autonomy from vocabulary, and the engagement of all possible scholars in the area, not just a few universities, will be further elements pushing advancement. The two most important needs for increasing S2ST performance are to improve conversational speech performance and to introduce highly dynamic language models. This is perhaps the most crucial aspect because speaking at less than 10% in conversational discourse appears to be a difficult difficulty nowadays.

6. CONCLUSION

Propels in speech and language research have brought Translation of speech near the pragmatic level for basic points where there is a generally clear worth of utilization. At the current level, nonetheless, speech translation has just arrived at the phase of making the center innovations. To accomplish more modern speech translation, innovative work ought to be additionally sped up. The following are a few focuses that should be the subject of center moving advances.

First and foremost, one unmistakable component of corpus-based advances is that they improve with use. It is subsequently crucial to get open doors for field and social testing and to effectively utilize created innovations. Occasions like the Olympics and World Fairs, where speakers of a wide range of dialects can be anticipated to join in, are an optimal chance to handle test multilingual speech-translation technology. It is subsequently indispensable to use these potential chances to propel innovation. NICT directed a screen try at the Beijing Olympics, mostly focusing on explorers from Japan. It fostered a speech translation framework supporting formal people, places, or things in the city of Beijing and had screens use discourse interpretation gadgets to impart in the city, involving the gadgets for such purposes as transportation, touring, and shopping.

Besides, speech translation is an innovation that deciphers communicated words in various languages. Despite the fact that interpretation into English is obviously fundamental, it will likewise be profoundly huge assuming speech translation can work straightforwardly between various languages of the world. Therefore, it is indispensable to expand the number of languages upheld. A plan of a joint effort between nations with a wide range of languages is required; all in all, a component is expected to empower different nations to work in an organization to explore speech translation, speech, and language. Making global communication in language innovation research focuses and so forth as a plan for cooperative Research and development should deliver input from a wide scope of examination into the assortment of discourse and tongue information, language structure, and such.

Thirdly, when numerous nations start to really investigate and foster speech translation, it will be important to normalize the points of interaction to associate these different language handling modules. The improvement of association strategies, information arrangements, word references, and such should keep an eye toward normalization. We should keep away from a circumstance wherein every nation fosters its own framework, and the frameworks are not viable together. Speech translation technology is progressed in such nations like Japan, and these nations can, in this manner, lead different nations with connection to normalization.

At long last, consideration should be given to copyright. Speech and language handling require speech and text corpora, and the presentation of speech translation relies intensely upon the amount and nature of these corpora. Thusly, the utilization of corpora of information broadcasts, papers, and the Web is very successful. Current intellectual property regulation doesn't consider auxiliary uses like these sorts of corpora. To investigate and foster new advances, it will be important to reexamine and manage the law with the goal that it is more flexible.

7. REFERENCES

1. Arora, K., Arora, S., and Roy, M.K. 2013. Speech to speech translation: a communication boon. CSIT 1**,** 207–213. https://doi.org/10.1007/s40012-013-0014-4.

2. Satoshi, N. 2008. Overcoming the Language Barrier with Speech Translation Technology.

3. Mahak, D., and Sumanlata, G. 2015. Speech-to-Speech Translation: A Review. International Journal of Computer Applications, 129 – 13.

4. Satoshi, N., Konstantin, M., Hiromi, N., Genichiro, K., Hisashi, K., Takatoshi, J., Jin-Song, Z., Hirofumi, Y., Eiichiro, S., and Seiichi, Y. 2006. The ATR Multilingual Speech-to-Speech Translation System. IEEE Transactions on Audio, Speech, and Language Processing, 14-2.

5. Gianni, L. 2002. The VI framework program in Europe: Some thoughts about Speech to Speech Translation research.

6. Marija, O., and Martina, L. 2021. Assessing Speech-to-Speech Translation Quality: Case Study of the ILA S2S APP.

7. Lavie, A., Waibel, A., Levin, L., Finke, M., Gates, D., Gavalda, M., Zeppenfeld, T., and Zhan, P. 1997. JANUS

III: Speech-to-speech translation in multiple languages. In 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, 99–102. IEEE.

8. Nakamura, S., Markov, K., Nakaiwa, H., Kikui, G.-i., Kawai, H., Jitsuhiro, T., Zhang, J.-S., Yamamoto, H., Sumita, E., and Yamamoto, S., 2006. The ATR multilingual speech-to-speech translation system. IEEE Transactions on Audio, Speech, and Language Processing 14(2), 365–376.

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

10. Jia, Y. Weiss, R. J., Biadsy, F., Macherey, W., Johnson, M., Chen, Z., and Wu, Y. 2019. Direct speech-to-speech trans-lation with a sequence-to-sequence model.

11. Dunbar, E., Cao, X. N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X., and Dupoux, E. 2017. The zero resource speech challenge 2017. In 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 323–330. IEEE.

12. Jerneja, Ž. G., and Mario Ž. 2006. The VoiceTRAN Speech-to-Speech Translation Communicator. Proceedings of the 5th WSEAS International Conference on Applications of Electrical Engineering, Prague, Czech Republic, 79-83.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 03, 56-65, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1103.1002

A Systematic Literature Review of Meta-Learning Models for Classification Tasks

Jackson Kamiri
Murang'a University of Technology, Kenya

Geoffrey Mraiga
Murang'a University of Technology, Kenya

Aaron Oirere
Murang'a University of Technology, Kenya

Abstract: Meta-learning is a field of learning that aims at addressing the challenges of conventional machine learning approaches such as learning from scratch for every new task. The main aim of this study was to do a systematic literature review of the existing meta-learning models that have been developed, published, and can be used for classification tasks. Systematic literature review method was used, employing a search of journal articles and publications of conference proceedings. The process involved data collection, analysis, and reporting of the results. To achieve the objective, 30 primary papers published since 2016 and relevant to classification tasks in meta-learning were considered. Data was extracted from the papers, then the following was analyzed in each model as presented in the papers; techniques used, the contribution, and the research gap. Although a lot has been done so far in Meta-learning, the existing models are not yet optimal. They still have challenges in few-shot learning, computation time complexity, difficulty in continual learning, and generalizability across multiple related tasks during transfer learning.

Keywords: Machine Learning; Meta-Learning; Few-Shot Learning; Transfer-Learning.

1.0 INTRODUCTION

Meta-learning also known as learning-to learn is a field of learning that aims at addressing the challenges of conventional machine learning approaches such as generalization and learning from scratch for every new task [1]. Through meta-learning, deep learning models are able to learn from a variety of related tasks and then use transfer learning to enable them to solve a new but related task with just a few data samples also known as few-shot learning. This enables machine learning developers to develop models that are robust and with high-level performance even in areas where labeled training data is limited such as in the medical field [2]. Meta-learning can be used to perform a variety of tasks such as regression, classification, and reinforcement learning. However, this study mainly focuses on classification tasks using Meta-learning algorithms.

Finn et al. [1], further argues that meta-leaning enables machines to gain state-of the art learning capabilities almost similar to that of a natural human being. The main aim of this study is to do a systematic literature review of the existing meta-learning models that have been developed, published, and can be used for classification tasks. According to Kitchenham [3] a systematic literature review is means through which available research relevant to a particular research question, phenomenon or topic of interest is evaluated and interpreted. Unlike ordinary literature review, Systematic literature review is evidence-based in the sense that its approach is well documented, scientific and reproducible.

Previous studies such [4] and [5] have done a survey on Meta-learning. However, we fault the above studies since they did not do a systematic literature review. Our contribution therefore, is that we have used a scientific approach which is evidence-based and guided by the guidelines of Kitchenham [3].

2.0 METHODOLOGY

We employed systematic literature review approach in reviewing meta-learning models for classification tasks.

2.1 Research Questions

This study was guided by three research questions (RQ) which are:

1. Which techniques have the existing meta-learning models used?
2. What problems have the models addressed?
3. What are the gaps that the existing models have left that can be addressed by future studies?

2.2Inclusion criteria

For a paper to be included in the review it needs to meet the following criteria; first, published between 2016 to 2021. Second, published in English language. Third, must be a peer-reviewed journal paper or published conference proceedings. Fourth, the paper is available in google scholar. Fifth, must be using meta-learning to perform classification tasks. Sixth, has a well-documented methodology, contribution and results. Seventh, the paper must be cited in google scholar citations.

2.3 Exclusion Criteria

The papers that did not meet the criteria discussed in section 3.2 were excluded from the study. On top of the above-discussed criteria, in cases where more than one version of the same paper were available, the most comprehensive version was included while the others were excluded.

2.4 Identification of papers

To identify potential papers for this study, we combined three techniques which are; searching for papers in google scholar search engine, identification of papers using references from included studies, and manual search in IEEE Explore repository. The key words combination that we used in the search include: “Meta-learning” “Meta-learning for classification”, “classification using meta-learning “, “Transfer learning”, and “Learning-to-learn”. The first keyword provided the largest output of papers in google scholar. We used several search key words to overcome the threat of omission where the initial key word was not very conspicuous in a study. The total number of papers obtained from the search was 58 papers.

2.5 Quality Assessment.

According to Kitchenham [3] quality is sometimes subjective. Therefore, we set our quality threshold to be the capability of a paper to answer all our research questions. To determine how well each paper answered our research questions, we used the following criteria:

First, determining if the study is a meta-learning classification task: in this we read the abstract and introduction sections of the paper to know if the paper was performing classification using meta-learning techniques.

Second, establishing the contribution of the paper: we scrutinized the papers thoroughly to determine the problem they have solved and how they have contributed to the field of meta-learning. In this we considered any improvement to the studies that previously existed as noted in the literature survey of the paper.

Third, determining if the paper has documented its methodology properly: we ensured that all papers included have a clear methodology that documents the following critical elements; the techniques used, the dataset used, the results obtained, and discussion of the results.

After subjecting all the papers obtained to the inclusion and quality assessment criteria, only 30 of the 58 papers met the needed threshold. Therefore, the results that we present in this paper are based on the 30 analyzed papers. Table 1 shows a summary of the number of papers sourced and how the numbers faired in each stage. The table follows the structure proposed by [6]

Table 1: Papers Included after Applying Quality Criteria

Total Sourced	Numbers of papers that failed Stage 1	Numbers of papers that failed Stage 2	Numbers of papers that failed Stage 3	Number of papers that qualified all stages
58	20	5	3	30

2.6 Data extraction and analysis

Data extraction was guided by the research questions. We developed data extraction forms which we used to extract data from each primary study. Attached in appendix 2 is a sample data extraction form. The researchers concentrated with what the papers reported rather than their personal interpretation of what the papers documented. The key elements that were extracted from each paper include:

RQ1, the techniques used in the paper: we focused on the specifics of the algorithms, techniques, datasets, and the general research approach used in the paper. For instance, [7] combined gradient-based meta-learning with model-based meta-learning. We also considered the application area in which the paper was applied, this was mainly informed by the dataset used.

RQ2, Contribution of the paper: we considered the value that each paper brings on-board. In this we captured what the authors of each paper have documented as their contribution. In cases where comparison was available, we considered contribution in comparison with previous studies in the same context such as in the case of [8] which was compared to [1]. All the papers reviewed had a contribution which is discussed in the results sections. We also extracted quantitative data in the form of performance metrics of each paper with respect to the experiment conducted and the datasets used.

RQ3, limitations of each paper: here we considered the gap that each paper has left out. Then we checked if any other paper has filled the gap by solving the limitations identified. We then documented in our results the gap that has not yet be covered by either of the papers analyzed.

2.7 Deviations from protocol

A review protocol defines the methods that will be used to undertake a specific systematic literature review [6]. The protocol is composed of all the elements of the review and some additional planning information. This study followed the protocol as defined in the methodology section of the paper. The researchers committed to follow the protocol in order to ensure that the results of the study are not influenced by researchers' bias.

3.0 RESULTS

Appendix 1: shows the papers that met the inclusion, exclusion, and quality assessment criteria. This section demonstrates how the researchers answered each of the research questions that formed the basis of our study.

3.1 RQ1: Which techniques have the existing meta-learning models used

The results of this study demonstrate that each of the 30 studies reviewed used techniques which incorporated a baseline and a technique unique to the specific problem that the paper intended to solve. For instance Goldblum et al. (2020) used adversarial querying together with the baseline techniques to achieve the objective of the model while [10] used hierarchical structuring on top of the base-structure. The baseline across all the papers reviewed involved the following techniques Neural network, meta-learner, multi-task learning, an optimizer for adaptation, and few-shot learning.

Another notable theme is that the model Agnostic Meta-Learning proposed by [1] has formed a standard that 86% of the analyzed papers extended. Of these 86% of the papers, 1 paper by [2] extended an extension of MAML known as

Reptile proposed by [11]. Therefore, these models that extended MAML used the MAML architecture as their baseline and then added new techniques to it. For instance [12] introduced noise gradient-based learning as way of extending MAML while [7] extended MAML by combining model-based learners with gradient-based learning.

All the papers considered in this research used secondary datasets to develop the models. 73.4% used more than one dataset while 26.6% used only one dataset.

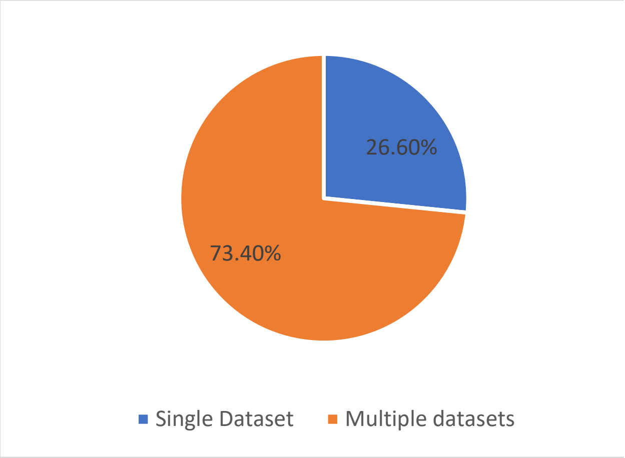


Figure 1: Data set Usage analysis

The most used datasets by the reviewed papers are Omniglot (36.67%), MiniImagent (56.67%), MINST (16.67%), and Cifar-10+Cifar-100(16.67%).

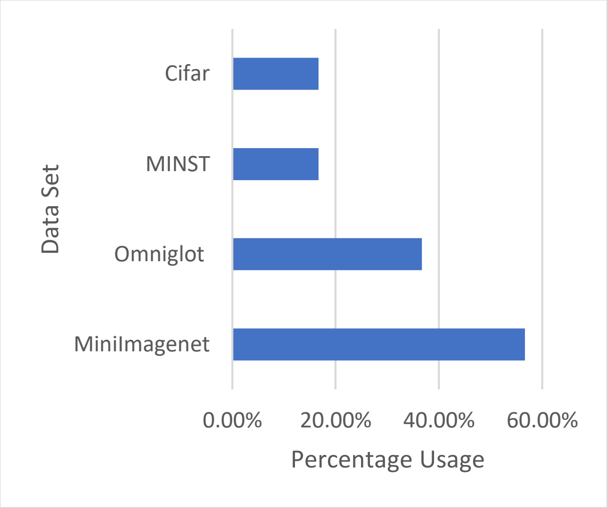


Figure 2: Most used datasets

The other datasets used are unique to the papers that used them. Papers such as [2] used the MiniImagenet dataset to pretrain the model then adapted the model through transfer learning to fit the Diabetic Retinopathy dataset. Also [13] used a pretrained ResNet50 which was trained on ImageNet then finetuned through transfer learning to fit labeled Retina Images.

3.2 RQ2: Which Problem have the papers solved.

Each of the papers analyzed in this section made some considerable contribution to the field of meta-learning in classification tasks. Some papers improved on classification accuracy while others introduced new and better approaches towards solving classification problems in meta-learning. In order to present the contributions of each paper clearly and in a summarized way, we have used paper codes to refer to the specific papers in appendix 1.

S01: Developed an optimization by gradient descent through learning rather than hand-crafted optimizers. S02: combined gradient-based learning with external memory modules. S03: proposed an LSTM based meta-learner model to learn the exact optimization algorithm used to train another learner neural network classifier in the few-shot regime. S04: proposed a model agnostic meta-learning model. S05: proposed a model that is able to acquire meta-learning prior for new tasks for multimodal task distribution. S06: proposed a model that solved catastrophic forgetting in meta-learning. S07: proposed a model that efficiently obtains a task posterior of a novel task. S08: proposed a model that improved MAML by introducing task robustness.

S09: solved the problem of task ambiguity in few-shot learning. S10: proposed an online hyperparameter adaptation scheme that eliminates the need to tune learning rates and meta-learning hyperparameters of the MAML. S11: proposed a model that dynamically updates the learning rate. S12: introduced a model that decouples meta-gradient computation from the choice of inner loop optimizer. S13: incorporated latent embedding optimization in gradient-based learning. S14: proposed a model in which they designed meta-regularization objective using information theory. S15: proposed an algorithm that hierarchically structures the transferrable knowledge into different clusters of tasks. S16: proposed an algorithm that is capable of maintaining an equilibrium between all the encountered tasks.

S17: proposed a model that scales without overfitting and is robust to task specific learning rate. S18: Combined hierarchical Bayesian models and gradient-based models. S19: improved Reptile by training it on transfer learning. S20: introduced unsupervised meta-learning to MAML.S21: combined meta-learning with the traditional fine-grained classification algorithms to improve on classification. S22: identified hyperparameters that generate optimal performance of MAML in image classification. S23: Expanded on MAML by demonstrating that first-order meta-learning algorithms perform well on some well-established benchmarks for few shot image classifications and provided theoretical analysis aimed at understanding why those algorithms worked.

S24: Proposed a model that uses trained linear classifiers as base leaners to learn representations for few-shot learning. S25: advanced few-shot classification paradigm towards a scenario where unlabeled examples are also available within each episode of learning. S26: proposed adversarial querying algorithm that is robust to adversarial examples and perform well in few-shot learning. S27: proposed a model to guide the learning of network parameters so that they are optimal for adapting to the target fine-grained classification task. S28: proposed a model that stores local states that extract the experience information from the seen task in transfer learning. S29: Proposed a meta-learning model that uses soft noisy labels. The model is trained using conventional approach and transfer learning. S30: the model fully integrates neural architecture search with gradient-based meta-learning.

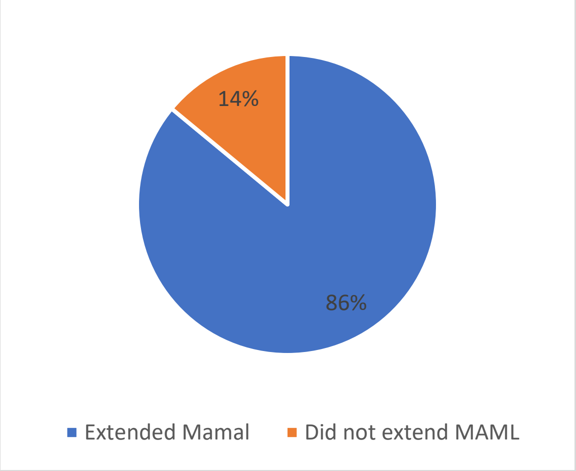


Figure 3: Analysis of papers that solved a problem that involved extending MAML

3.3 RQ3 What are the gaps that the existing models have left that can be addressed by future studies?

To answer this question, we considered the gaps or limitations of each of the 30 papers reviewed. Apart from our detailed analysis, we as well considered future works proposed by the authors of the papers to be part of the gap. We also considered whether or not either of the papers in the pool solved gaps left by other papers in the pool. For instance, the proposed future work by S04 was addressed by S05. This section therefore, reports outstanding gaps that have not yet been solved by either of the papers considered in the study.

First, S07 uses Stein variational gradient descent (SVGD) and hierarchical Bayesian and thus it does not have the full coverage of the task-posterior, a neural network would deliver better in this. Second, still in S07 the model suffers the shortcomings of ensemble approaches such as space/time complexity proportional to the number of particles. Third, SVGD in S07 is limited in that its performance is sensitive to the parameters of the kernel function[14]. Fourth the model developed by S09 provides impoverished estimator of posterior variance, thus, its effectiveness in gauging where task have different degrees of uncertainty is low[12]. Fifth, in S11 the method has not yet achieved convergence results of non-transitioning variant of hyper gradient descent, and also has not established the convergence rate [15].

Sixth, S13 the researchers proposed a future work that replaces the pre-trained feature extractor with one learned jointly through meta-learning, or using LEO for tasks in reinforcement learning or with sequential data [16]. Seven, in S18 the researcher records that the Laplace approximation is inaccurate in cases where the integral is highly skewed, or is not unimodal and thus is not amenable to approximation by a single Gaussian mode [17].

Eight, S19 the researcher proposes that in future an advanced neural network architecture can be used for transfer learning[2]. Also, the researcher uses default learning rates rather than dynamic learning rates thus limiting the capability of the model to achieve optimal performance quickly. Nine S20 the model traded off accuracy and reduction in the number of labeled images required. Therefore, it cannot be useful where classification accuracy is essential [18].

Ten, S24 proposed that future work can explore other convex base-learners such as kernel Support Vector Machines [19]. Eleven, S23 the researchers record that further research should: pay close attention to transduction's use of batch normalization during testing, seek to understand to what extent SGD automatically optimizes for generalization, explore if regularization can improve few-shot learning[11]. Twelve, S25 the researchers proposed that future work can extend their work by incorporating fast weights[20].

Thirteen S29 the model noisy labels impairs the convergence of SGD. Also, the model only does binary classification yet Diabetic retinopathy is multi-class classification task[13]. Fourteen, all studies that have used gradient-based learners have faced a generalization challenge as a result of poor convergence of the loss function in both the local and global minima[1], [21]. Thirteen models such as S18 consumed a lot of execution time [17].

3.1 Discussion

This research demonstrates that meta-learning has become a better alternative to deep learning in performing classification tasks. This is mainly due to the capabilities of meta-learning to leverage on previous knowledge during transfer learning, perform automatic hyperparameter optimization and learn from few data samples. According to Fin et al [1], the capability to leverage on previous knowledge is the hallmark of human intelligence and thus meta-learning is a clear path towards artificial general intelligence.

The study has also demonstrated that the analyzed models use baseline which mainly constitutes of a neural network, meta-leaner, multi-task learning, and an optimizer. Then, researchers build their models on the basis of this baseline. Meta-learning will continue to attract a lot of research interest among researchers since there are many grey areas as enumerated in section 3.3 that are yet to be addressed.

3.2 Limitations of the study

This study only covered Meta-learning models used for classification tasks. Therefore, it has not covered other areas in which meta-learning has been applied such as in reinforcement learning, regression, speech recognition, and text analytics.

4.0 CONCLUSION

In this study we have conducted a thorough systematic literature review of meta-learning models used for classification task. This work is part of continuing research. This study forms a strong foundation for researchers interested in the field of Meta-Learning since it provides sufficient information on the current status of meta-learning research in classification tasks. The study has demonstrated clearly how the analyzed models fair in the research questions of this study.

In future researchers can consider solving the gaps that have been identified in section 3.3 of this paper.

References

[1] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” *34th Int. Conf. Mach. Learn. ICML 2017*, vol. 3, pp. 1856–1868, 2017.

[2] M. Welling, “Meta-Learning for Medical Image Classification,” no. Midl, pp. 7–9, 2018.

[3] B. Kitchenham, “Kitchenham , B .: Guidelines for performing Systematic Literature Reviews in software engineering . EBSE Technical Report EBSE-2007-01 Guidelines for performing Systematic Literature Reviews in Software Engineering,” no. January 2007, 2021.

[4] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey, “Meta-Learning in Neural Networks: A Survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–20, 2021, doi: 10.1109/TPAMI.2021.3079209.

[5] I. Khan, X. Zhang, M. Rehman, and R. Ali, “A Literature Survey and Empirical Study of Meta-Learning for Classifier Selection,” *IEEE Access*, vol. 8, pp. 10262–10281, 2020, doi: 10.1109/ACCESS.2020.2964726.

[6] T. Hall, S. Beecham, D. Bowes, D. Gray, and S. Counsell, “A systematic literature review on fault prediction performance in software engineering,” *IEEE Trans. Softw. Eng.*, vol. 38, no. 6, pp. 1276–1304, 2012, doi: 10.1109/TSE.2011.103.

[7] R. Vuorio, S. H. Sun, H. Hu, and J. J. Lim, “Multimodal model-agnostic meta-learning via task-aware modulation,” *Adv. Neural Inf. Process. Syst.*, vol. 32, no. NeurIPS, pp. 1–22, 2019.

[8] L. Collins, A. Mokhtari, and S. Shakkottai, “Task-Robust Model-Agnostic Meta-Learning,” pp. 1–30, 2020.

[9] M. Goldblum, L. Fowl, and T. Goldstein, “Adversarially robust few-shot learning: A meta-learning approach,” in *Advances in Neural Information Processing Systems*, 2020, vol. 2020-Decem, no. 1, pp. 1–15.

[10] A. A. Liu, Y. T. Su, W. Z. Nie, and M. Kankanhalli, “Hierarchical Clustering Multi-Task Learning for Joint Human Action Grouping and Recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 102–114, 2017, doi: 10.1109/TPAMI.2016.2537337.

[11] A. Nichol, J. Achiam, and J. Schulman, “On First-Order Meta-Learning Algorithms,” pp. 1–15.

[12] C. Finn, K. Xu, and S. Levine, “Probabilistic model-agnostic meta-learning,” *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. NeurIPS, pp. 9516–9527, 2018.

[13] G. Algan, I. Ulusoy, Ş. Gönül, B. Turgut, and B. Bakbak, “Deep Learning from Small Amount of Medical Data with Noisy Labels: A Meta-Learning Approach,” 2020.

[14] J. Yoon, T. Kim, O. Dia, S. Kim, Y. Bengio, and S. Ahn, “Bayesian Model-Agnostic Meta-Learning,” no. NeurIPS, pp. 1–11, 2018.

[15] D. Mart, F. Wood, R. Cornish, and M. Schmidt, “O NLINE L EARNING R ATE A DAPTATION WITH,” no. 2015, pp. 1–11, 2018.

[16] A. A. Rusu *et al.*, “M ETA -L EARNING WITH L ATENT E MBEDDING O PTIMIZATION,” pp. 1–17, 2019.

[17] E. Grant, C. Finn, S. Levine, T. Darrell, T. Griffiths, and C. Sciences, “R g -b m -l h b,” pp. 1–13, 2017.

[18] S. Khodadadeh, L. Bölöni, and M. Shah, “Unsupervised meta-learning for few-shot image classification,” in *Advances in Neural Information Processing Systems*, 2019, vol. 32, no. NeurIPS.

[19] K. Lee and S. Maji, "Meta-Learning with Differentiable Convex Optimization."

[20] M. Ren *et al.*, "Meta-learning for semi-supervised few-shot classification," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018, pp. 1–15.

[21] L. Zintgraf, K. Shiarlis, V. Kurin, K. Hofmann, and S. Whiteson, "Fast Context Adaptation via Meta-Learning," no. 2018, 2019.

[22] M. Andrychowicz *et al.*, "Learning to learn by gradient descent by gradient descent," *Adv. Neural Inf. Process. Syst.*, no. Nips, pp. 3988–3996, 2016.

[23] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-Learning with Memory-Augmented Neural Networks," *33rd Int. Conf. Mach. Learn. ICML 2016*, vol. 4, pp. 2740–2751, 2016.

[24] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," *5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc.*, pp. 1–11, 2017.

[25] R. Vuorio, D.-Y. Cho, D. Kim, and J. Kim, "Meta Continual Learning," 2018.

[26] H. S. Behl, A. G. Baydin, and P. H. S. Torr, "Alpha MAML: Adaptive Model-Agnostic Meta-Learning," 2019.

[27] A. Rajeswaran, S. M. Kakade, C. Finn, and S. Levine, "Meta-learning with implicit gradients," *Adv. Neural Inf. Process. Syst.*, vol. 32, no. NeurIPS, pp. 1–12, 2019.

[28] M. Yin, G. Tucker, M. Zhou, S. Levine, and C. Finn, "M -l m," pp. 1–21, 2020.

[29] H. Yao, Y. Wei, J. Huang, and Z. Li, "Hierarchically Structured Meta-learning," 2019.

[30] J. Rajasegaran, S. Khan, M. Hayat, F. S. Khan, and M. Shah, "iTAML : An Incremental Task-Agnostic Meta-learning Approach," pp. 13588–13597.

[31] X. Ruan, H. Liu, W. Pang, and S. Lu, "Fine-grained Classification Algorithm based on," pp. 2019–2022, 2019.

[32] Y. Zhang, H. Tang, and K. Jia, "Fine-Grained Visual Categorization using Meta-Learning Optimization with Sample Selection of Auxiliary Data."

[33] D. Alp, E. Acar, R. Zhu, and V. Saligrama, "Memory Efficient Online Meta Learning," 2021.

[34] T. Elsken, B. Staffler, J. H. Metzen, and F. Hutter, "Meta-Learning of Neural Architectures for Few-Shot Learning," pp. 12365–12375.

Appendix 1

Table 2: List of papers Reviewed

Paper Code	Citation
S01	[22]
S02	[23]
S03	[24]
S04	[1]
S05	[7]
S06	[25]
S07	[14]
S08	[8]
S09	[12]
S10	[26]
S11	[15]
S12	[27]
S13	[16]
S14	[28]
S15	[29]
S16	[30]
S17	[21]
S18	[17]
S19	[2]
S20	[18]
S21	[31]
S22	[31]
S23	[11]
S24	[19]
S25	[20]
S26	[9]
S27	[32]
S28	[33]
S29	[13]
S30	[34]

Appendix 2:

Table 3: Sample Data Extraction Form

Paper Code	**S06**
Title	Meta Continual Learning
Authors	Risto Vuorio, Dong-Yeon Cho, Daejoong Kim, and Jiwon Kim
Techniques Used	ANN (predictor) which updated a loss function on current and previous tasks.
Dataset Used	MINST
Main Contribution	proposed a meta-learning model that aimed at solving catastrophic forgetting problem in deep learning
Gaps/ Limitations	Although this model achieved continual learning, the researchers record that it was faced with the challenge of learning to optimize. Poor optimization affected the generalizability of the model.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 03, 66-71, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1103.1003

An Email Spam Filtering Model Using Ensemble of Machine Learning Techniques

Aju Omojokun Gabriel
Department of Computer Science
Adekunle Ajasin University
Akungba-Akoko, Nigeria

Adedeji Ayomiposi Joy
Department of Computer Science
Adekunle Ajasin University
Akungba-Akoko, Nigeria

Abstract: The growth of spam emails is on the increase responsible for larger portions of the global email traffics. Aside the annoyance and the time wasted sifting through the unwanted messages; spam emails can also cause immeasurable harms through malicious software capable of damaging systems and compromising confidential information. The risks of filtering spam emails is that sometimes, legitimate mails are marked as spam, yet the results of not filtering spam are the constant flood of spam clogs on networks that adversely impacts users inboxes while draining valuable resources on the networks such as bandwidth and storage capacity, productivity loss and interfere with the expedient delivery of legitimate emails. Several researchers had worked on the design of models for spam email filtering using different techniques, however the detection accuracy of these models have also become subject of discussions. This study developed spam email filtering model using Ensemble of Decision Tree, Support Vector Machine and Multilayer Perceptron (DT-SVM-MLP) technique as a solution approach to solving issues of low spam emails detection accuracy. The ensemble model was trained using forward propagation training technique and the performance was evaluated using five performance metrics of Accuracy, False Positive (FP) Rate, Precision, Recall and F-Measure.

Keywords: Spam Email, Email Filtering, Ensemble Machine Learning, Forward Propagation Training, Performance Metrics.

1. INTRODUCTION

The internet has become an integral part of everyday life and electronic mail (email) has become a powerful and indispensable tool for information exchange. It is one of the most commonly used features over communication networks that may contain texts, files, images, or other attachments. Email messages are sent through email servers and uses multiple protocols within the Transmission Control Protocol/Internet Protocol (TCP/IP) suite which allows users to send and receive messages anywhere in the world because the access mobility to email system is independent of physical locations.

The email is significant for many kinds of group connection and is being widely used by many people; individuals and organizations for both official and personal correspondence (Naem et al, 2018). In the 1990s, there was an increase use of email facilities as more companies and institutions joined the Internet system, as the significant advances made in telecommunication technologies, couple with the reduced costs of computers and telecommunication devices made the internet system more accessible. Email allows users to send and receive messages anywhere with an email address, the system can also be accessed from anywhere in the world and can deliver messages instantaneously. Because the mobile access to email is neither attached to a physical location nor restricted to a fixed place, rather the mobility of email allows people to work and communicate from anywhere. Due to these factors, email communication is used over other modes of communication because it is economical, flexible and reasonable (Palival et al., 2018).

Today, e-mail has become an efficient, rapid and cheap means of communication. Likewise, the dramatic growth in the spread of unwanted email messages, otherwise known as Spams cannot be overemphasised. One of the fast rising and costly problems linked with the internet today is the spam email which are predominantly mercantile and mostly have attractive links to famous websites that lead to meddlesome sites (Naem et al, (2018).

In recent times, unwanted commercial bulk emails have become a huge problem on the email systems. In April 2021, it was estimated that 89.35% of all emails were accounted as spam mails and 482.65billion daily spam mails were sent globally (Palmote et al., 2021). The huge volume of spam mails flowing through the internet networks have destructive effects on the memory space of email servers, communication bandwidth, central processing unit, power consumption and user time (Dada et al., 2019). Aside the cost of spam mails on the internet networks infrastructures, it has also been reported that the spread in spam mails has resulted to untold financial loss for many internet users who have fallen victim of internet scams and other fraudulent practices of spammers who send emails, pretending to be from a reputable source with the intention to persuade individuals to disclose sensitive personal information like passwords, Bank Verification Number (BVN) and credit card numbers. The cost of spam mails to companies worldwide in 2019 was estimated to be US$260 billion (Palmote et al., 2021).

The risk in filtering spam is that sometimes, legitimate mails may be rejected or marked as spam, however, the risks of not filtering spam are the constant flood of spam clogs on networks which adversely impacts the users inboxes, drain valuable network resources such as bandwidth and storage capacity, productivity loss and interfere with the expedient delivery of legitimate mails (Mallampati, 2019). Different machine learning algorithms have been used in the development of email filtering techniques to solve the problem of spam emails wreaking havoc on email users. These machine learning algorithms have been successfully applied to classify emails into either spam or non-spam. These algorithms include Logistic Model Tree Induction, Decision

Tree, Artificial Immune System, Support Vector Machine, and Artificial Neural Networks (Dada et al., (2019).
These algorithms have been giving varying accuracy in the filtering process and accuracy rates has become a point of research. This paper went further in increasing the performance of these machine learning algorithms by developing an ensemble algorithm that combines the three Support Vector Machine (SVM), Decision Tree (DT) and Multilayer Perceptron (MLP) algorithms to form an optimal model.

2. LITERATURE REVIEW

The number of spam email has increased for several reasons such as advertisements, multi-level marketing, chain letters, political emails, stock market advice, among others. Email Filtering have usually relied on keyword patterns, to be more efficient and prevent the danger of accidental removal of ham messages which are called Ham or allowed messages. These patterns need to be checked with each user's received emails. However, detailed setting of such patterns needs time and proficiency which are unfortunately not always available (Takhmiri and Haroonabadi, 2016).

In restricting spam email, several methods and spam filtering algorithms have been developed using machine learning techniques such as Naive Bayes, Support Vector Machine, K-Nearest Neighbor, Bayes Additive Regression, KNN Tree, Decision Tree and rules. Chan et al., (2010), the authors combined the Best Stepwise feature selection with a classifier of Euclidean nearest neighbor and created a Naïve Euclidean approach to develop email filtering system. Each email was represented in D-dimensional Euclidean space. Using SpamBase from the UCI repository, and a 10-fold cross validation, they achieved an accuracy of 82.31% compared to 60.6% for the Zero rule.

Rathi et al., (2013) proposed a data mining technique approach for finding the best classifier for email classification. They analyzed various data mining technique for measuring the performance of several classifiers through “with feature selection algorithm” and “without feature selection algorithm”. After selecting the Best feature selection algorithm, they considered the selected algorithm for their feature selection purpose. They experimented their data using Naïve Bayes, Support vector machine, J48, Random Forest and Random Tree algorithms. The dataset used consists of 58 attributes and 4601 instances. Bhat et al., (2014) proposed some community-based topological features to learn improved classification models for identifying spammers in online social networks. However, the results only spanned over single classifiers. Mahmoud et al., (2014) The proposed a combined Naïve Bayes, Clonal selection and Negative selection algorithms filtering technique that consists of four phases of Training phase, Classification phase, Optimization phase and Testing phase to classify the email messages. The worked used 2,500 spam messages and 2,500 non-spam messages to train the system.

Rusland et al., (2017) performed email spam filtering analysis using Naïve Bayes algorithm on two datasets which are evaluated based on the accuracy, recall, precision and F-measure metrics. The Naïve Bayes algorithm as a probability-based classifier counts the frequency and combination of values in a dataset. The work performed through three phases such as pre-processing, Feature Selection, and implementation. Abdulhamid et al., (2018) studied the analysis based on the classification of algorithms and their efficiencies. For this study various methodologies considered and their efficiencies were measured in terms of basic metrics. Any function collection or efficiency improve approach was used to provide a holistic view of the efficiency of classification techniques. Study shows that there are a variety of classification techniques that are more reliable if better investigated by way of selecting features. Of all the various methodologies utilized, Rotation Forest is the most reliable classifier of 94.2 percent.

Agarwal and Kumar (2018) proposed a combined methodology of machine learning techniques such as the NB algorithm and optimization algorithm namely, the PSO algorithm for identification of spam emails. NB algorithm is mainly utilized for classification of the obtained emails into two categories such as spam or non-spam. PSO algorithm is utilized for the optimization parameters that are of the NB algorithm. The implementation of this algorithm was made with the aid of the popular dataset of Ling spam evaluated the efficiency based on the popular metrics. PSO outperforms relative to individual NB approaches based on the validated findings. Palival et al., (2018) presented an email spam filtering model using ID3 Decision Tree based Algorithm. ID3 is a non-incremental algorithm used to build a decision tree from a fixed set of observations. The resulting tree is then used to classify test observations and each observation is represented by features or attributes and a class to which it belongs. ID3 uses information gain measure to select decision node. Enron dataset was used for training as well as testing the filter system. The Enron dataset contains emails of both types stored in plain text format with 3672 legitimate (ham) emails and 1500 spam emails.

Dada et al., (2019) analyzed the core principles, attempts, performance, and spam filtering study patterns. The latest study investigates the implementations of machine learning environments to the leading ISPs, including Gmail, Yahoo, and Outlook spam filters, to the spam processing e-mail process. There has been debate about the general approach of spam filtering and the efforts of different researchers to tackle spam using machine learning techniques. The study contrasts the advantages and disadvantages of the existing methodologies of machine learning and brings new problems with spam filter growth. The study suggested broad and strong opposing education as the strategies for managing spam e-mail risks to cope successfully with the potential.

Mallampati et al., (2019) presented a kernel function SVM approach to build a spam detection. System, when the support vector machine algorithm analyzes a single mail then it returns a 0 else it returns a 1. The authors considered the dataset from the UC Irvine Machine Learning Repository for spam emails. Olatunji (2019) proposed a model based on support vector machines that are suggested for spam identification when carefully searching for optimized parameters for better results. Experimental findings indicate that all earlier models on the same common dataset used in this work succeeded the model suggested. 95.87 and 94.06% accuracy for preparation is reached and collections of testing respectively

3. METHODOLOGY

The email spam filtering method is designed to separate the spam (unwanted emails) from the non-spam (wanted emails). The recent spam mail classification is mostly handled by machine learning (ML) algorithms intending to differentiate between spam and non-spam messages, the machine learning algorithms achieve this by using an automatic and adaptive technique, rather than depending on hand-coded rules that are susceptible to the continuously evolving features and varying characteristics of spam messages. Machine learning techniques have the capacity to obtain information from a set of messages provided, and then use the acquired information to classify new messages that it just received.

This research study used machine learning ensemble technique for the email classification. The machine learning ensemble technique combines the Decision Tree (DT), Multilayer Perceptron (MLP) and Support Vector Machine (SVM) base models in order to produce one optimal predictive model. The study used Python programming language (version-3.9) and the Jupyter notebook editor. The activities involved in the chosen methodology to ensure the success of the study include the: Data Collection, Data Pre-processing, Ensemble Model Design, Model Training and Testing, and Model Performance Evaluation.

3.1 Data Collection

The Enron dataset is used for training as well as testing the model. The Enron dataset contains emails of both types stored in plain text format. The Enron directory contains 3672 legitimate (ham) emails and 1500 spam emails, the first attribute contains the subject and body of the email, and the last attribute of the dataset is the nominal attribute, which consists of the value 0’s and 1’s to represent whether a mail is spam or not. The dataset is divided into a ratio of 80:20 wherein the 80% data is used for training the model and the remaining 20% is used for testing the accuracy of the developed model.

3.2 Data Preprocessing

Pre-processing is a very crucial step in spam email filtering techniques. There are three steps involved in the pre-processing: tokenization, stop word removal and stemming. The initial step consists of the process called tokenization. In the process, all of the unnecessary word, the punctuations and the symbols are removed from the sentences. The strings that are left is split up into various tokens. The next step is stop word removal. Stop-words are basically nothing but unnecessary and non-informative words, e.g. ‘a’, ‘an’, ‘the’, and ‘is’, among others that doesn’t add any sense and information to the message. In the second step all such words which carry no information are removed. English language has around 300-400 stop words.

The last step is the stemming which is the reduction of inflection in words and bringing it to their root form is known as stemming. The root word can just be a canonical form of the original word. Word-stemming is a term used to describe a process of converting words to their morphological base forms, mainly eliminating plurals, tenses, prefixes and suffixes. Stemming is closely related to lemmatization which while reducing a word considers the part of speech and the context of the word.

3.3 Ensemble Model Design

The model design for this study consists of three machine learning techniques algorithms ensembled to form a more optical model (DT – MLP – SVM model).The Decision Tree generates the output as a binary tree like structure called a decision tree, in which each branch node represents a choice between a number of alternatives, and each leaf node represents a classification or decision. MLP networks are general- purpose, flexible, nonlinear models consisting of a number of units organized into multiple layers. The complexity of the MLP network can be changed by varying the number of layers and the number of units in each layer. Given enough hidden units and enough data, it has been shown that MLPs can approximate virtually any function to any desired accuracy. In other words, MLPs are universal approximators. SVM considers data as points in space mapped in a way such that the difference between the closest data points is maximum.

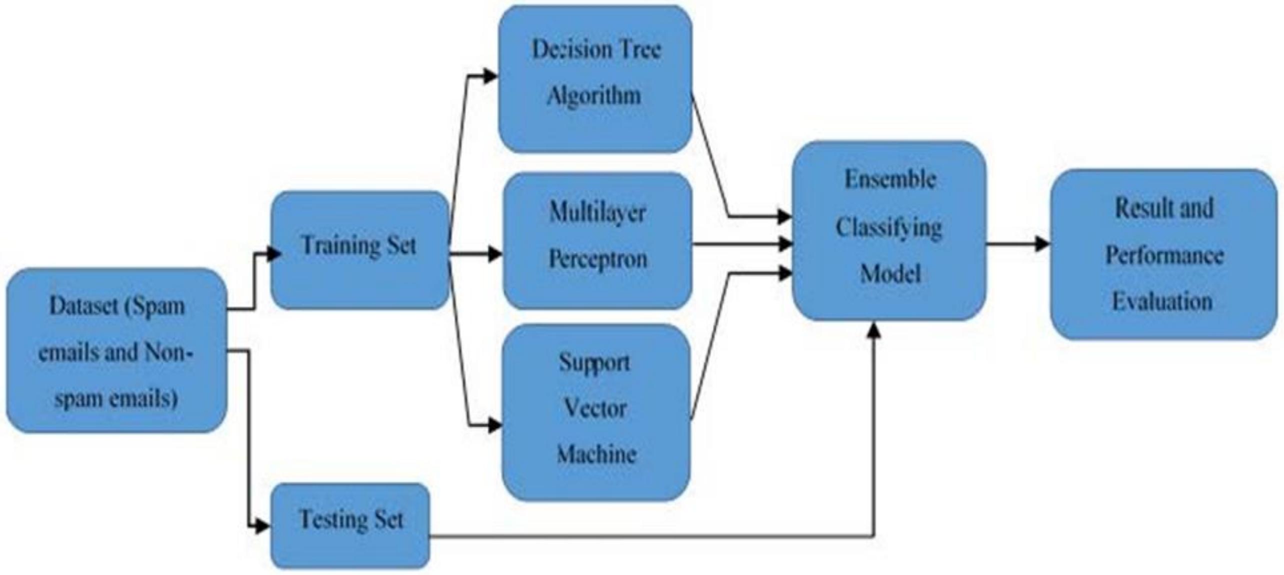
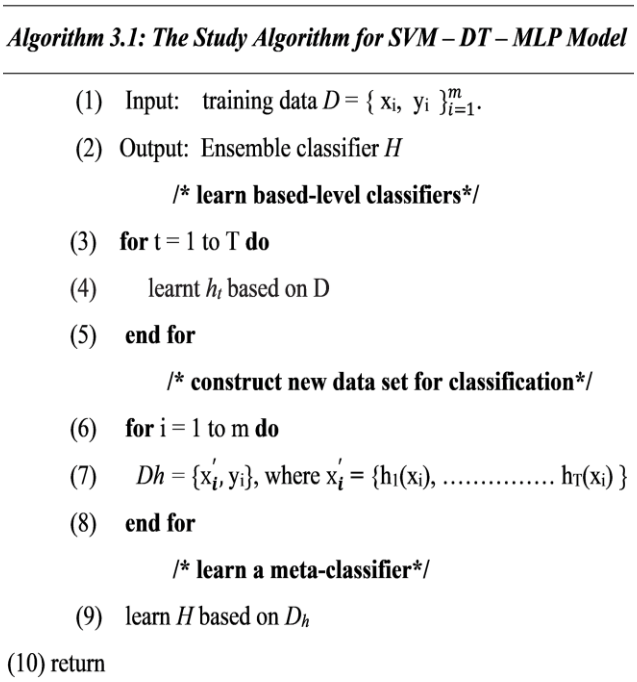

Decision Tree Algorithm
Training Set
Multilayer Perceptron
Ensemble Classifying Model
Result and Performance Evaluation
Dataset (Spam emails and Non-spam emails)
Support Vector Machine
Testing Set

Figure 3.1: The Ensemble Model Architecture

In this study, an ensemble Decision Tree – Multilayer Perceptron – Support Vector Machine (DT- MLP -SVM) model is developed to form a more optimal spam email filtering model by taking the advantages of each base models into consideration. The DT- MLP -SVM) model algorithm is as shown in algorithm 1. The outputs of the ensemble model (DT-SVM-MLP) were compared with that of the base models.



Algorithm 3.1: The Study Algorithm for SVM – DT – MLP Model

```
Input:   training data $D = \{ x_i,\ y_i \}_{i=1}^{m}$.
Output:  Ensemble classifier $H$
          /* learn based-level classifiers*/
for t = 1 to T do
    learnt $h_t$ based on D
end for
          /* construct new data set for classification*/
for i = 1 to m do
    $Dh = \{x_i^{'}, y_i\}$, where $x_i^{'} = \{h_1(x_i), \ldots\ldots\ldots\ldots\ldots h_T(x_i)\}$
end for
          /* learn a meta-classifier*/
learn $H$ based on $D_h$
return
```

3.4 Model Training and Testing

The forward propagation technique is used in this study for the training of the network. In the forward propagation, the input data is fed in the forward direction through the network. Each layer accepts the input data, processes it as per the activation function and passes to the successive layer. The dataset is divided into a ratio of 80:20 wherein the 80% data is used for training the model and the remaining 20% is used for testing the accuracy of the developed model. The model testing involves explicit checks for the behaviours that the model exhibits. Testing the model performance in terms of accuracy and other metrics on which the model is evaluated.

3.5 Model Performance Evaluation

The performance evaluation is done by measuring the percentage of spam detected and how many misclassifications are done by a particular technique and the ensemble model. The results obtained are then compared on the basis of the performance of each of the techniques (Sharaff, 2019). The ensemble model is evaluated using the five-performance metrics: Accuracy; FP rate; Precision; Recall and F-Measure.

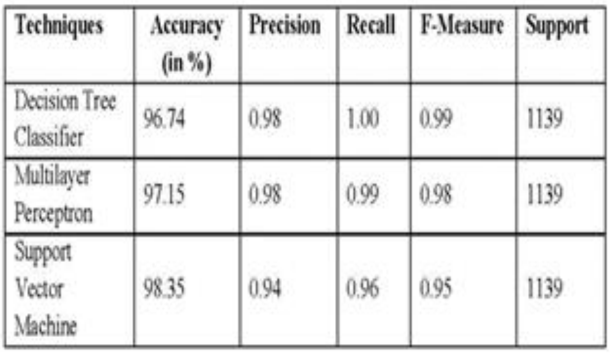
The model resulted into a confusion matrix which consists of four parts: True Positive (TP); True Negative (TN); False Positive (FP) and False Negative (FN). These values are used to determine model performances.

4. FINDINGS AND DISCUSSION

4.1 The Baseline Models

The training dataset (spam and legitimate message) was generated from the mails. The class labels are designated as spam to represent spam and ham to represent legitimate emails. The machine learning techniques (Algorithms): Decision Tree, Multilayer Perceptron and Support Vector Machine Algorithms were used for training the data on the Jupyter notebook environment. The performance of the trained models was evaluated using 10-fold cross validation for its predictive accuracy. Predictive accuracy is used as a performance measure for email spam classification. The prediction accuracy is measured as the ratio of number of correctly classified instances in the test dataset and the total number of test cases. The outputs process for the base models and the ensemble model are shown in Tables 4.1 and 4.2.

Table 4.1: Results of the Base Models Performance Evaluation



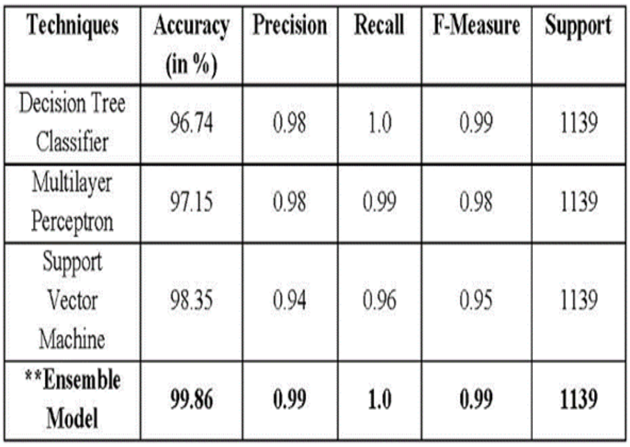
Techniques	Accuracy (in %)	Precision	Recall	F-Measure	Support
Decision Tree Classifier	96.74	0.98	1.00	0.99	1139
Multilayer Perceptron	97.15	0.98	0.99	0.98	1139
Support Vector Machine	98.35	0.94	0.96	0.95	1139

Accuracy, Precision, Recall and F-Measure Metrics: From the values obtained for the base models as shown in table 4.1. The SVM model has the best performance in terms of the models' accuracy, however, in term of precision, Recall and F-Measure, the SVM performed below the other two base models. Followed by the Multilayer Perceptron while the Decision Tree had the lowest value in term of the models' accuracy. In term of Precision, Recall and F-Measure; Decision Tree performed better than the Multilayer Perceptron. F-Measure is dependent on Precision and Recall..

4.2 The Developed Ensemble Model

The results of the ensemble model with the base models are shown in table 4.2. The main principle behind the technique is that the combined knowledge of multiple models, in this case, the Decision Tree, Multilayer Perceptron and Support Vector Machine Algorithms can performance better and give a more accurate results as compared to a single model considered for same task.

Table 4.2: Results of the Ensemble Model Performance Evaluation



Techniques	Accuracy (in %)	Precision	Recall	F-Measure	Support
Decision Tree Classifier	96.74	0.98	1.0	0.99	1139
Multilayer Perceptron	97.15	0.98	0.99	0.98	1139
Support Vector Machine	98.35	0.94	0.96	0.95	1139
****Ensemble Model**	**99.86**	**0.99**	**1.0**	**0.99**	**1139**

Figure 4.1 shows the performance of precision, recall and f-measure for the different models and Figure 4.2 shows the accuracy performance results of the models.

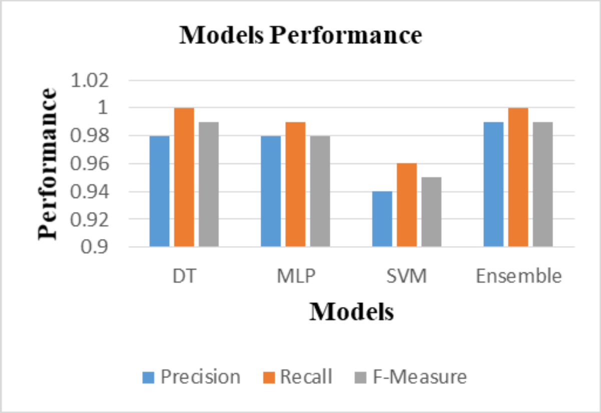


Figure 4.1: Models Precision, Recall & F-Measure Results

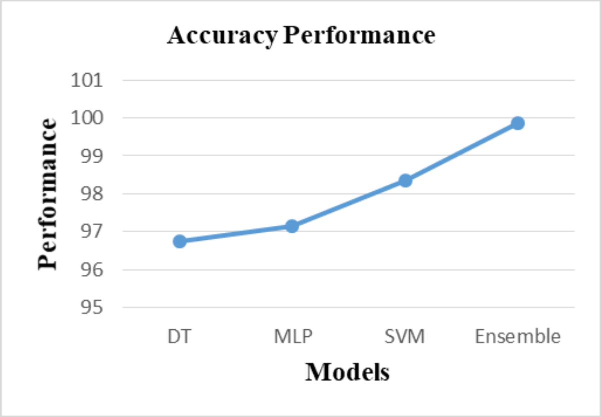


Figure 4.2: Models Accuracy Results

The developed ensemble model was evaluated using all the performance evaluation metrics used for the base models and the model gave an overall high accuracy of 99.86%. This resulted in a more promising approach of email spam filtering technique with more consistent and accurate results.

5. CONCLUSION

Email spam filtering is challenging but a highly desirable task. Different machine learning techniques have been used in different literatures for filtering genuine messages from spam messages. An ensemble model combining the three machine learning techniques of Decision Tree, Multilayer Perceptron and Support Vector Machine Algorithms is used and measured with the chosen performance evaluation metrics to observe the effectiveness and accuracy of each base techniques. Though a slight change was observed in the performance of the base models, this deviation indicates that the performance of a technique depends on the data used more than the algorithm. However, the developed ensemble model performed better than each of the base models. It resulted in a more promising approach of email spam filtering technique producing more consistent and accurate results.

6. REFERENCES

[1] Abdulhamid M. S, Shuaib M, Osho O, Ismaila I, and Alhassan J K, Comparative Analysis of Classification Algorithms for Email Spam Detection 2018, *Inter. J. Comp. Net. Inf. Sec.*, Vol. 10, pp. 60–67.

[2] Agarwal K and Kumar, T. (2018). Approach Of Naïve Bayes and Particle Swarm Optimization. *2018 Sec. Int. Conf. on Intel. Comp. and Cont. Sys.* pp. 685–90.

[3] Bhat, S. Y., Abulaish, M and Mirza, A. A. (2014). Spammer Classification Using Ensemble Methods over Structural Social Network Features. *Proceedings - 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2014*, *2*, 454–458.

[4] Chan, T. Y., Jie Ji, and Qiangfu Zhao. (2010). Learning to Detect Spam: Naive-Euclidean Approach. *International Journal of Signal Processing*. Issue 1 (2010), pp 31-38.

[5] Christina, V. (2010). Email Spam Filtering using Supervised Machine Learning Techniques. *International Journal on Computer Science and Engineering Vol. 02, No. 09. pp. 3126-3129*

[6] Cortez, P. (2010). Spam Email Filtering Using Network-Level Properties. *July*. https://doi.org/10.1007/978-3-642-14400-4

[7] Dada, E. G and Joseph, S. B. (2018). Random Forests Machine Learning Technique for Email Spam Filtering. Heliyon, Vol. *9, No.* 1. pp. 29–36.

[8] Dada E G, Bassi J S, Chiroma H, Abdulhamid S M, Adetunmbi A O, and Ajibuwa O. E. (2019). Machine Learning for Email Spam Filtering: Review, Approaches and Open Research Problems *Heliyon*, Vol. **5, No.** 6. https://doi.org/10.1016/j.heliyon.2019.e01802

[9] Delany, S. J., Cunningham, P., Tsymbal, A and Coyle, L. (2005). A Case-Based Technique for Tracking Concept Drift in Spam Filtering. *Knowledge-Based Systems,* Vol. *18 No.* 4–5. pp. 187–195. https://doi.org/10.1016/j.knosys.2004.10.002

[10] Enron mail dataset, “http://www2.aueb.gr/users/ion/data/enron-spam (Accessed: 24 March, 2021)

[11] Jantan, A., Ghanem, W. A. H. M and Ghaleb, S. A. A. (2017). Using Modified Bat Algorithm to Train Neural

Networks for Spam Detection. *Journal of Theoretical and Applied Information Technology*, Vol. *95, No.* 24. pp. 6788–6799.

[12] Kaur, H and Sharma, A. (2016). Improved Email Spam Classification Method using Integrated Particle Swarm Optimization and Decision Tree. *2016 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India* 14-16 October 2016.

[13] Mahmoud, T. M., El-hafeez, T. A and Khairy, M. (2014). *An Efficient Three-phase Email Spam Filtering Technique An Efficient Three-phase Email Spam Filtering Technique*. *Journal of Theoretical and Applied Information Technology*, Vol. *62, No.* 8. pp. 1742–1751.

[14] Mallampati, D., Shekar, K. C and Ravikanth, K. (2019). Supervised Machine Learning Classifier for Email Spam Filtering. (Issue January). Springer, Singapore. https://doi.org/10.1007/978-981-13-7082-3

[15] Naem, A. A., Ghali, N. I and Saleh, A. A. (2018). Antlion Optimization and Boosting Classifier for Spam Email Detection. *Future Computing and Informatics Journal*, Vol. *3, No.* 2. pp.436–442. https://doi.org/10.1016/j.fcij.2018.11.006.

[16] Olatunji S O, Improved email spam detection model based on support vector machines 2019, *Neu. Comp.and App.*, **31**, pp. 691–99.

[17] Palimote, J., Anireh, V.I.E and Nwiabu, N. D. (2021). International Journal of Advanced Research in Computer and Communication Engineering. Vol. 10, Issue 4. pp. 17-24

[18] Palival, D., Printer, K., Devre, R and Lemos, N. (2018). Email Spam Filtering Using Decision Tree Algorithm. International Journal of Scientific and Engineering Research, Vol. 9, Issue 3. pp. 40-42.

[19] Park, I., Sharman, R., Rao, H and Upadhyaya, S. (2007). The Effect Of Spam And Privacy Concerns on E-Mail Users’ Behaviour. *J. Info. Syst. Security*, Vol. *3, No.* 1. pp. 40–62.

[20] Takhmiri, H and Haroonabadi, A. (2016). Identifying Valid Email Spam Emails Using Decision Tree. International Journal of Computer Applications Technology and Research. Volume 5, Issue 2, pp. 61 - 65.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 03, 72-77, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1103.1004

What Triggers Violation of Information Security Policies

Sandeep Dhawan
Senior IT Director

Abstract: This paper offers a framework for information security professionals to evaluate ISP and understand policy non-adherence. It explores the different types of non-adhering behaviors and the motivations behind them. It goes on to share information about the environmental/organizational climates in which non-compliant behaviors are more or less likely to occur and briefly touches on when users are more likely to commit them. Finally, it suggests a user review process as a critical part of information security policy design and implementation.

Keywords: ISP violation intention, information security policies, IS compliance, information security behavior, internal threat agent

1. INTRODUCTION

User non-adherence to existing information security policies accounted for 88% of all reported data breaches in 2021. [1] With each data breach costing the affected organization an average of $4.4 million [2], reducing instances of information security policy violation or non-adherence must be a priority for any professional involved in drafting information security policies.

Information security systems are constructed as a protective measure, guarding the information technology (IT) of an organization or company. The threats they protect against include natural disasters, physical damage, and human behavior.

Human behavioral threats to information security can be classified as internal or external. Hackers or attackers outside the organization may implement techniques such as malware, phishing, DDoS attacks, and ransomware in attempts to gain control over valuable resources. Inside the organization or network, each user and each access point represent a potential vulnerability.

In an attempt to address and prevent user-based internal threats while complying with regulatory requirements, the designers of information security systems craft policy guidelines for the access to and use of information technology.

These detailed instructions, known as information security policies (ISP), also define what actions constitute a violation of the policy and frequently include a description of the consequences for performing such a violation. ISP take into account the wider context within which the information security system is situated, such as the industry, region, and model or structure of the organization.

Once composed, edited, and approved, the distribution method of the ISP varies by organization. ISP may be distributed directly to users, or an overview may be delivered as part of the onboarding process. Users with higher levels of access may be required to participate in more formal training.

A sufficiently robust and considered information security policy, implemented perfectly will provide the appropriate level of protection for the digital assets of a given organization. Yet, implementation is far from perfect.
Professionals involved in information security policy development and implementation must therefore consider the reasons non-adherence occurs.

Investigating the origins of non-compliant behavior and the circumstances or environments that make non-adherence more or less likely allow the human element to enter the ISP at the earliest design phases. Building awareness and a culture of trust around information security improve the chances that ISP will be implemented consistently. [3] Costly data breaches could be mitigated, decreased, or avoided altogether.

2. CONTENT

2.1 ISP Creation

Not only must user behavior be a consideration from the very first draft, an ISP cannot be considered complete until it has been enthusiastically adopted by the workforce.

2.1 ISP Implementation

Information security policies must be implemented with a high degree of consistency in order to perform their stated function. A violation occurs when a user, for whatever reason, does not follow the procedures laid out in the ISP.

2.3 ISP Violation

Common ISP violations include password and log-in sharing, using old or weak passwords, and clicking on phishing links. Less common violations include the access to or theft of privileged information.

2.4 Human Element

Human behavior is complex, and therefore a certain degree of inconsistency in implementation must be expected. Cybersecurity behavior is influenced by individual decision-making styles. [4] Organizing and understanding the triggers at the root of non-compliant information security behavior is a necessary step to creating novel ways of increasing cyber-resiliance.

2.5 Awareness

As it relates to information security, the human element can be organized broadly into two categories: compliant, and non-compliant. While remaining compliant with ISP may seem obvious to professionals rooted in cybersecurity, not all users have the same level of awareness of ISP and security in general.

2.6 Compliance

Compliance has costs and benefits and carries the risk of sanctions or the chance of praise. Individual beliefs about these costs, benefits, risks, and choices further complicate our understanding of the factors that go into cybersecurity behaviors. [5] (Fig. 2)

2.7 User Intention

Non-compliant behavior can be further segmented in several different ways. One such segmentation is based on the user's intention. Actions that violate ISP unintentionally and without malice can be considered 'misbehavior'. 'Non-malicious deviant behavior' occurs when the ISP is violated intentionally, but without malice. 'Deviant behavior' involves a violation of the ISP with malicious intent. [6]

2.8 Corrective Action

2.8.1 Misbehavior

Misbehavior occurs when security awareness is low. Training and education aimed at raising security awareness should therefore be a part of ISP implementation and may need to be repeated at regular intervals to prevent the awareness from fading.

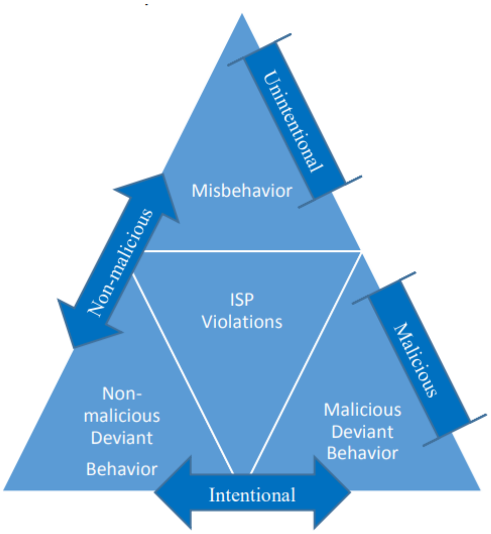
2.8.2 Malicious Deviance

Users with malicious intent cannot be relied upon to implement ISPs consistently no matter how much training or education they receive. Malicious deviation from the ISP in an attempt to damage it or profit off it is beyond the scope of the ISP to prevent.

2.8.3 Non-Malicious Deviance

Between these two extremes is a user group caught in a contradiction. They intentionally, knowingly work against their organizations' information security goals, which can have profound ramifications for the organization, but without malicious intent. At times, their intent is altruistic; when forced to choose between implementing ISP consistently and achieving their job-related tasks, these users will consistently choose the latter. [6]

Figure 1. The Triad of ITA Behaviors [6]



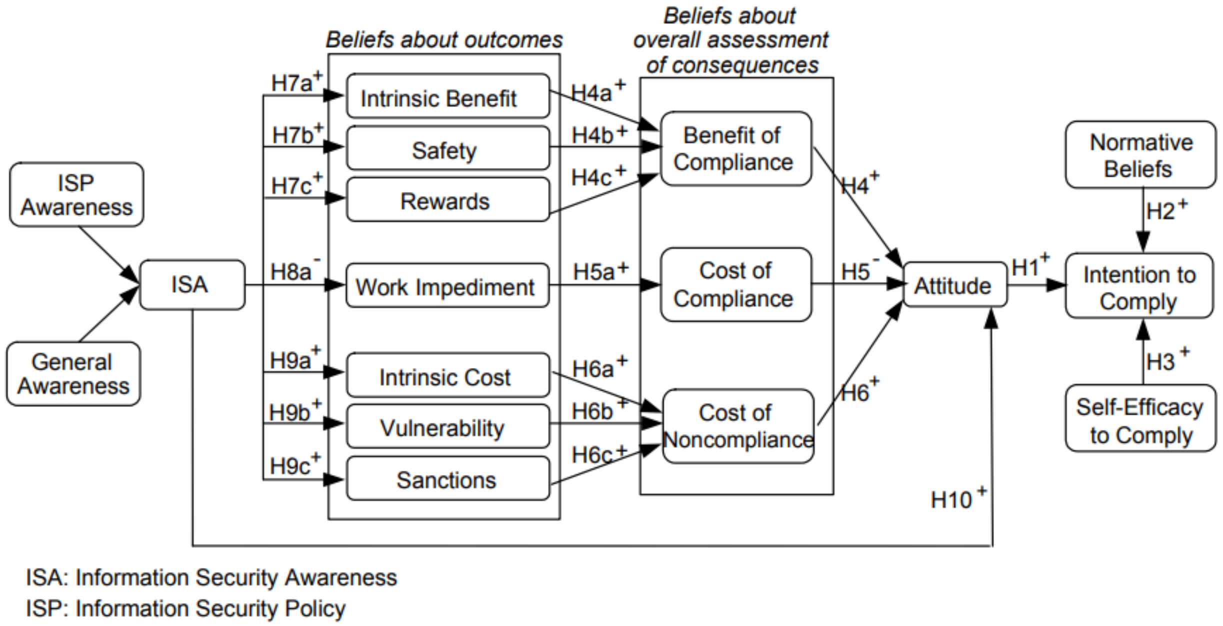
While it is possible to categorize individual instances of ISP non-compliance, it is not possible to characterize individual employees or users with the same labels. ISP compliance is fluid, rather than stable [7], and interacts with various stimuli

Fig.2 Model of Compliance Antecedents [5]

such as job demands, organizational structure, company culture, and personal affect.

2.9 Security-Related Stress

Security-related stress (SRS) has also been linked with security non-compliance, which is closely associated with fatigue and frustration. [8]



2.10 Situational Moral Beliefs

Situational moral beliefs affect security behaviors like sharing passwords or selling confidential data.

This effect has been shown to be mitigated somewhat by the severity and certainty of being sanctioned for violations. [9]

2.11 Consequences and Rewards

Monetary rewards for remaining compliant are more effective than harsh formal consequences. Informal consequences and perceived benefits more effectively encourage compliance. [10]

2.13 Environment

During the development and training period, information security policies are functioning in a controlled environment under expected circumstances. Once training is complete, employees return to their work, where they must balance their individual motivations, values, and temperament with competing priorities in an uncontrolled environment while also sticking to security protocols. Perfect implementation of the ISP during training exercises does not guarantee the same level of adherence on the job.

2.13.1 Organizational Ethical Climate

An instrumentalist-based organizational ethical climate (OEC) can positively modify the relationship between moral disengagement and ISP violation, while a negative modification is associated with rule-and-law based OEC. [11]

2.13.2 Organizational Structure

Organizational structure also has an effect on compliance. When compared to organizations, bureaucratic institutions have the best chance of keeping information safe. Organizations with a bureaucratic organizational structure must recognize the benefit provided by this structure and weigh it carefully when confronted with opportunities for more customer-centric modes of operation. [12]

2.13.3 Professional Subculture

In terms of the intention to violate an ISP, professional subculture plays a role. [13] Individuals who share similar training experiences tend to also share similar values and attitudes about security procedures.

2.14 Individuality

Personally held values, beliefs, and attitudes toward data protection also influence user behavior. [5] A consciously developed culture of data protection can increase cyber-resiliency. [3]

In environments that function on trust and collective responsibility, peer monitoring reduces ISP violation intention. [14]

2.15 Negative Emotions

Employees with negative emotions are more likely to intentionally violate ISP. While some negative emotions are inevitable and enter the workplace from outside, others are generated by experiences in the workplace. These negative emotions can be mediated by measures designed to increase organizational support, psychological ownership, and work engagement. [15]

3. DISCUSSION

Understanding what triggers a violation of ISP presents opportunities to decrease the likelihood of non-compliance. IS professionals responsible for the development and implementation of ISP must be aware of these possibilities in order to develop novel approaches and interventions that increase compliance.

These opportunities are distributed throughout the design and implementation process. IS professionals could begin the drafting of policies by defining the structure of the organization within which the policies will be implemented.

Another important factor to establish before the first draft begins is the cultural context within which security-affecting behaviors occur. A toxic work culture should be considered a valid security threat. While changing the culture of the organization may be beyond the power of the information security team, ISP training and education should take place in a trust-based environment.

It may also be possible to leverage non-IS-based users with an interest in cybersecurity to encourage the kind of peer-monitoring that has been shown to increase compliance. [16]

An ISP that has not been implemented in the real-world environment can only be considered a first draft. Many organizations choose to wait until after a security event has occurred to review the interaction between employees, job task completion, and security protocols. This is a costly mistake.

A user review period has the potential to identify areas of disagreement between job requirements and security expectations. Done thoughtfully, it may also encourage the kind of trust-based environment that reduces the risk of non-compliant behavior.

Further research could include investigation of novel user review processes to establish the efficacy of the proposed tool. The relationship between user review efficacy and organizational ethical climate is also ripe for exploration. Studies that track user-related variables such as security awareness, personal motivations, individual values and attitudes, professional subculture, stress, and situation moral beliefs could investigate if and how these phenomena are influenced by a user review process.

Longitudinal studies could be used to track the long-term effects of adding a user review process to the work system, including whether they increase or decrease compliance with ISP.

Since environment, culture, and peer-monitoring are crucial pieces of the ISP violation puzzle, continued study is needed on the complex social interactions that interact with and have an effect upon cybersecurity.

Taking a holistic view of the various triggers behind ISP violations allows information security professionals to identify potential threats and implement appropriate countermeasures throughout the ISP development and implementation process.

4. CONCLUSION

Users violate information security policies for a number of different reasons.

At a high level, organizational structure, ethical climate, peer relations and professional subculture all play a role.

Zooming in to look at the individual, many factors can affect both the intention to violate or comply with ISP and whether that violation actually occurs. They include security awareness, personally held moral and ethical beliefs, the agreement (or lack thereof) between job-related tasks and security policies, and the user’s conception of benefits and rewards for remaining compliant or violating the ISP.

During training and implementation, empowered users can function as a review board, pointing out inconsistencies between workflow and policy. Policies that force employees to choose between task completion and security protocols are the most likely to be violated. The solutions are often low-cost and easily implemented. [17]

Empowered users can be trusted to give valuable feedback on security protocols and policies. Reconciling user intention and user execution through formal and informal review processes has the potential to close a critical gap in information security systems.

The cost of a data breach includes not only the loss of information assets but also the loss of trust from critical stakeholders. With the extraordinarily high costs and risks associated with data breaches, information security professionals must use every tool at their disposal to protect information security systems.

A formal user review process during ISP development and implementation is one such valuable tool.

5. REFERENCES

[1] IBM. (2022). *Cyber Security Intelligence Index Report.* https://www.ibm.com/security/data-breach/threat-intelligence/

[2] IBM. (2021). *Cost of a Data Breach.* https://www.ibm.com/security/data-breach
D'Arcy, J, Lowry, PB. Cognitive-affective drivers of employees' daily compliance with information security policies: A multilevel, longitudinal study. *Info Systems J.* 2019; 29: 43– 69. https://doi.org/10.1111/isj.12173

[3] Huang, K., & Pearlson, K. (2019). For What Technology Can't Fix: Building a Model of Organizational Cybersecurity Culture. *Proceedings of the 52nd Hawaii International Conference on System Sciences*. URI: https://hdl.handle.net/10125/60074

[4] Charlette Donalds, Kweku-Muata Osei-Bryson, Cybersecurity compliance behavior: Exploring the influences of individual decision style and other antecedents, International Journal of Information Management, Volume 51, 2020, 102056, ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2019.102056.

[5]. S. Dhawan, "Information and Data Security Concepts, Integrations, Limitations and Future," IJAIST, vol.30, no.30, pp.09-13, Sep.2014

[6]van den Bergh, M., & Njenga, K. (2016). *Information Security Policy Violation: The Triad of Internal Threat Agent Behaviors*. https://www.researchgate.net/profile/Maureen-Van-Den-Bergh/publication/303408537_Information_Security_Policy_Violation_The_Triad_of_Internal_Threat_Agent_Behaviors/links/5742abfc08aea45ee84a4ef9/Information-Security-Policy-Violation-The-Triad-of-Internal-T

[7] D'Arcy, J., & Lowry, P. B. (2019). Cognitive-affective drivers of employees' daily compliance with information security policies: A multilevel, longitudinal study. *Information Systems Journal*, *29*(1), 43–69. https://doi.org/10.1111/isj.1217

[8] John D'Arcy, Pei-Lee Teh, Predicting employee information security policy compliance on a daily basis: The interplay of security-related stress, emotions, and neutralization, Information & Management,
Volume 56, Issue 7, 2019, 103151, ISSN 0378-7206, https://doi.org/10.1016/j.im.2019.02.006.

[9] Li, Han; Luo, Xin (Robert); and Chen, Yan (2021) "Understanding Information Security Policy Violation from a Situational Action Perspective," *Journal of the Association for Information Systems*, 22(3), .
DOI: 10.17705/1jais.00678

[10] Li, Yuanxiang John and Hoffman, Elizabeth, Behavioral compliance theory: an experimental and behavioral economics approach to information security policy compliance (November 15, 2021). Available at SSRN: https://ssrn.com/abstract=3252742 or http://dx.doi.org/10.2139/ssrn.3252742

[11] Chen, H., Chau, P.Y.K. and Li, W. (2019), "The effects of moral disengagement and organizational ethical climate on insiders' information security policy violation behavior", *Information Technology & People*, Vol. 32 No. 4, pp. 973-992. https://doi.org/10.1108/ITP-12-2017-0421

[12] Karlsson, M., Karlsson, F., Åström, J., & Denk, T. (2021). The effect of perceived organizational culture on employees' information security compliance. *Information & Computer Security*.

[13] Sarkar, S., Vance, A., Ramesh, B., Demestihas, M., & Wu, D. T. (2020). The Influence of Professional Subculture on Information Security Policy Violations: A Field Study in a Healthcare Context. *Information Systems Research*. Informs Pubs Online. https://doi.org/10.1287/isre.2020.0941

[14] Adel Yazdanmehr, Jingguo Wang, Can peers help reduce violations of information security policies? The role of peer monitoring, European Journal of Information Systems, 10.1080/0960085X.2021.1980444, (1-21), (2021).

[15] Jie Zhen, Zongxiao Xie, Kunxiang Dong & Lin Chen (2021) Impact of negative emotions on violations of information security policy and possible mitigations, Behaviour & Information Technology, DOI: 10.1080/0144929X.2021.1921029

[16] Kam, H.-J., Ormond, D. K., Menard, P., & Crossler, R. E. (2021). That's interesting: An examination of interest theory and self-determination in organisational cybersecurity training. *Information Systems Journal*, 1– 39. https://doi.org/10.1111/isj.12374

[17] Martha Nanette Harrell. 2014. Factors impacting information security noncompliance when completing job tasks. Doctoral dissertation. Nova Southeastern University. Retrieved from NSUWorks, Graduate School of Computer

and Information Sciences. (21) https://nsuworks.nova.edu/gscis_etd/21.

[18] Bulgurcu, B., Cavusoglu, H., & Benbasat, I. (2010). Information security policy compliance: an empirical study of rationality-based beliefs and information security awareness1. *MIS Quarterly*, *34*(3), 523-548.

[19] S. Dhawan, “Information and Data Security Concepts, Integrations, Limitations and Future,” IJAIST, vol.30, no.30, pp.09-13, Sep.2014

International Journal of Computer Applications Technology and Research
Volume 11–Issue 03, 78-81, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1103.1005

An Android Based E-Commerce Application for Farmers

Ayush Kumar
Department of Electronics and Telecommunication,
College of Engineering, Bharati Vidyapeeth (Deemed to Be University) Pune, Maharashtra, India

Anchal Thakre
Department of Electronics and Telecommunication,
College of Engineering, Bharati Vidyapeeth (Deemed to Be University) Pune, Maharashtra, India

Sudhir Kadam
Department of Electronics and Telecommunication,
College of Engineering, Bharati Vidyapeeth (Deemed to Be University) Pune, Maharashtra, India

Abstract: We have developed a mobile application for farmers, which will help them to sell their products directly to the consumers through this application. This mobile application helps the farmers sell as well as buy products through its easy and convenient interface. Farmers face a lot of challenges when it comes to selling their products and therefore the primary goal of our app is to provide a means for farmers to market their commodities at fair prices. Updated market prices and filters are also provided in the application to make it more convenient and efficient for farmers. This system has a simple interface that also provides certain filters that lets the consumer choose from a large variety of products from which they can select and purchase products according to their requirements. The main objective of this application is to take into concern the needs of the farmers as well as the buyers and fulfil their requirements accordingly. The fast and updated delivery system is one of the priorities of our android application.

Keywords: Intermediaries, Economy, Market Pricing, Retailers, Agriculture Industry.

1. INTRODUCTION

India is a country whose economy largely depends upon agriculture in other words we can say that farmers are the spine of India. The economic welfare of India heavily depends upon the development of agriculture. There have been many advancements in agricultural technologies worldwide but in India, most farmers are completely unaware of the market rates of their commodities. The farmers struggle in getting the correct price for their products and goods due to the lack of information about the actual rates of the products and a lot of time they fail to sell their products in time due to which they undergo a heavy loss.

Farmers take the help of intermediaries to sell their products in the market. The farmers do not get enough wages for their products due to the involvement of the intermediaries. The involvement of the intermediaries makes farmers lose a huge share of their income. Due to this, the farmers undergo heavy loss while the middleman makes the most of it.

As Android is the fastest growing mobile operating system in today's world, Android smartphones are becoming more accessible due to being cost effective. That is why we are making our application natively in android. Our application is solely aimed at the development of farmers and agricultural businesses through smartphones, and it commits to make selling and buying agriculturally based products, simple and appropriate.

2. METHODOLOGY

In this chapter, we present some basic ideas behind the making of our app and how they are utilized in our research work. The used methods are shortly presented and discussed. We propose to use a native application that will be built in Kotlin language specifically for the Android operating system since 95.7% of people in India use Android devices according to stat-counter. Agricultural commodities are traded in the market at the district level, the government sets support prices to stabilize the prices, but the market prices are dynamic [1].

The design of our mobile application is established on a client-server model where the user operating the gadget is designated as the client and the server used is Firebase. Firebase is a Backend-as-a-Service (Baas) and is a NoSQL database program.

Figure 1 shows the functioning of the application. This application is aimed to have a user-friendly interface so that the end user can have the best experience of the app. The application provides the user with a basic Firebase authentication using an Email link.

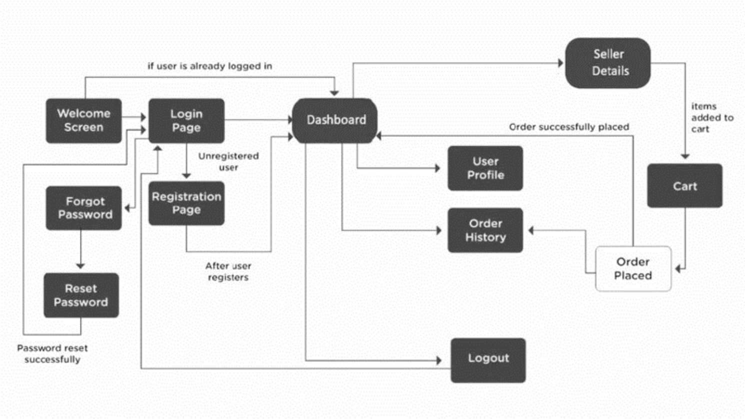


Figure.1. System Flowchart

3. RESULTS

3.1 Authentication

It will verify the identity of a user and will be used by the server to know who is accessing the information. The application will provide basic authentication services such as Google login and Email-password login.

3.1.1 *User Login*

This feature will grant the users to log into the system by entering their usernames and passwords in the respective fields after the completion of the registration part. If the user enters invalid credentials, they will not be allowed to gain access to the system.

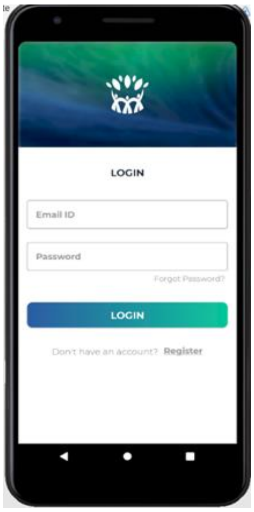


Figure.2. Login Screen

3.1.2 *New user registration*

A new user will have to click on the Register button to get signed up and take advantage of the application as shown in Figure 3 by providing the necessary details such as the first name, last name, email, etc.

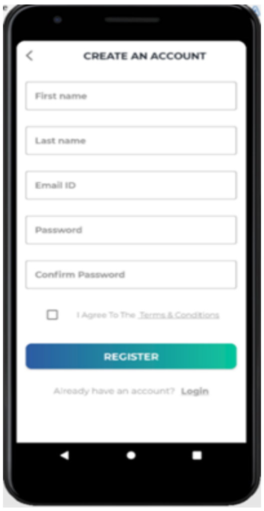


Figure.3. Login Screen

3.1.2.1 *Functional requirement*

The system should authenticate and confirm the details provided by the user during the signup process.

3.1.3 *Account recovery*

If the user needs help resetting the password, they can click the forgot password button and enter the email address with which they registered to get the reset password link as presented in Figure 4.

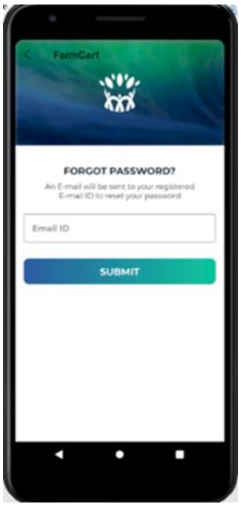

FarmCart
FORGOT PASSWORD?
An E-mail will be sent to your registered E-mail ID to reset your password
Email ID
SUBMIT

Figure.4. Forgot password screen

3.2 Product details and dashboard

Users will be able to upload the product details with real images which will be stored in the Firebase storage, and they will also be able to provide the fundamental details about their goods such as the title, cost, description, and stock availability. After uploading the image of the product, the user can see the edit option to change the image that was uploaded previously. Once the product is submitted, the user can see their product on the dashboard screen. The user will also have the option to remove an undesired product.

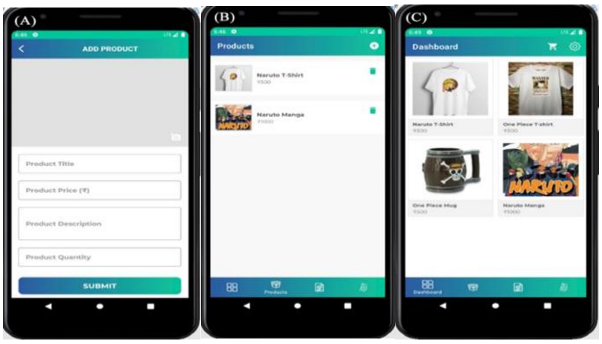
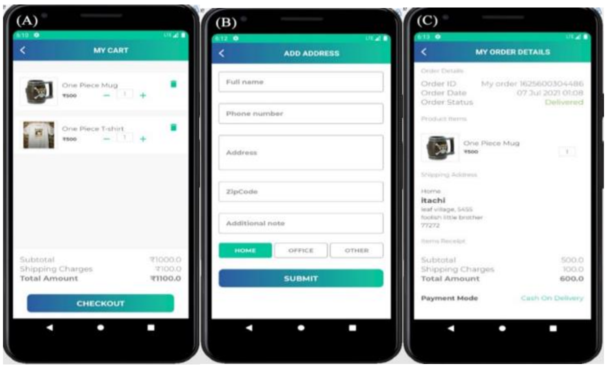


Figure. 5. (A) Add product screen, (B) Previously added products, (C) Dashboard screen.

3.3 Shopping cart and orders

After choosing the desirable products from the dashboard screen, the user can add them to the shopping cart by clicking the Add to Cart option where the cart will display the names of the added products with their respective prices and the total amount at the bottom of the screen and then the user can advance to the checkout screen. Before advancing to the checkout screen the user will also have the option to edit or delete the current items in the cart. After proceeding to the checkout screen, the user will have to add an address to fulfil the delivery requirements. After placing an order, the user can visit the orders screen to see the order ID, date, and status as well as the shipping address and the receipt. The seller can also check the products sold in the sold products screen.



MY CART
One Piece Mug ₹500 − 1 +
One Piece T-shirt ₹500 − 1 +
Subtotal ₹1000.0
Shipping Charges ₹100.0
Total Amount ₹1100.0
CHECKOUT

ADD ADDRESS
Full name
Phone number
Address
ZipCode
Additional note
HOME OFFICE OTHER
SUBMIT

MY ORDER DETAILS
Order Details
Order ID My order 1625600304486
Order Date 07 Jul 2021 01:08
Order Status Delivered
Product Items
One Piece Mug ₹500 1
Shipping Address
Home
itachi
leaf village, 5455
foolish little brother
77272
Items Receipt
Subtotal 500.0
Shipping Charges 100.0
Total Amount 600.0
Payment Mode Cash On Delivery

Figure. 6. (A) Shopping cart, (B) Add address screen, (C) My orders screen.

3.4 Database server

Database servers are used to store and manage databases that are stored on the server and to provide data access for authorized users [2]. It uses real-time processing techniques to handle the workloads. For our application, we will use Cloud Firestore which is a NoSQL database that our android application can access directly via native SDKs.

It is designed for very large, structured data. It can handle a larger workload. It stores your data in documents. Cloud Firestore also provides offline support for the app so one can read and write data without Internet connectivity.

In Cloud Firestore you can store data in form of collections which act as containers and are used to organize data. In our application, the collection will be named as products and users for which we will assign the documents with unique ID's and then the data will be stored following the ID.

3.5 Cloud functions and cloud storage

Firebase cloud functions is a serverless framework and helps us run the backend code automatically which saves a lot of time. Certain events get triggered by the features of Firebase and also due to the various HTTPS requests made by the client.

Cloud storage lets you store an incredible amount of data online rather than storing it on your computer. It securely saves the data and stores user-generated content such as pictures, videos, etc. This function will let the application store the pictures of the product uploaded or the display picture of the user to the cloud.

3.6 Analytics

Firebase analytics automatically captures key events such as active users, daily user engagement, top screens, revenue, retention reports, demographics, etc. Google Analytics for Firebase provides a vast array of charts, tables and graphs containing information covering not only the way in which users interact with the app, but also details about the users themselves [3]. The Android APIs will provide all the data to the user in a summarized manner so that the user can easily decipher through the graphs.

Firebase is a platform developed by Google and is free to use with certain limits.

It is a cloud-based application, and it can work individually as well as when implemented together which gives better performance. Sharing data and analytics between Firebase components make them integrated and ideal to be used in conjunction.

It will provide the admin with the top user property value, top location, best app version in form of a map, the conversion events, and will give more information on custom definitions.

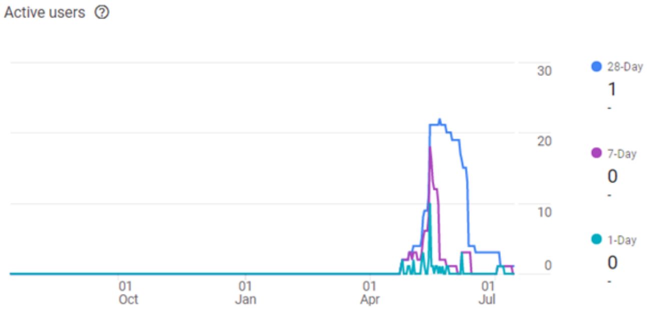
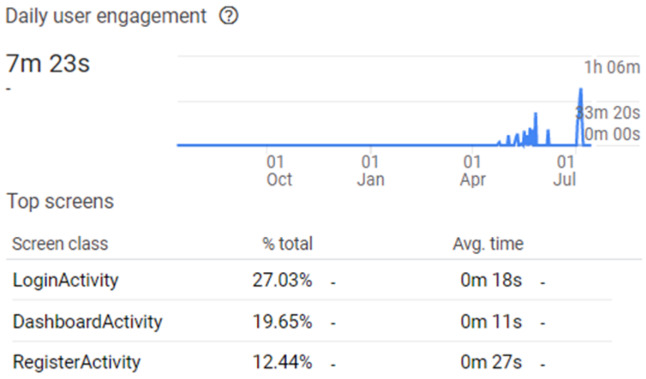


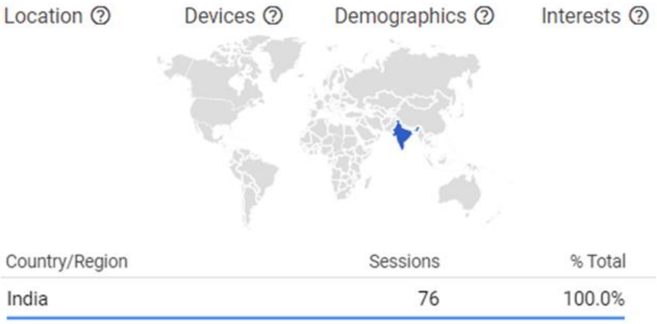
Figure. 7. Number of active users per day.



Top screens

Screen class	% total	Avg. time
LoginActivity	27.03% -	0m 18s -
DashboardActivity	19.65% -	0m 11s -
RegisterActivity	12.44% -	0m 27s -

Figure. 8. Shows average daily engagement with a graph displaying trends for the time period selected.



Country/Region	Sessions	% Total
India	76	100.0%

Figure. 9. Percentage of sessions from each of the top countries.

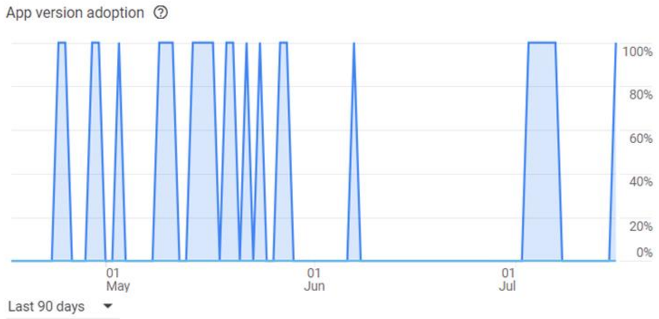


Figure. 10. Graph shows the % of active users for each app version in the past 90 days.

4. DISCUSSION

Currently, the deals are done physically as there exists no centralized platform for trading goods. There is very little connectivity between the farmers and the retailers as well as the agricultural department. Furthermore, the existing system prevents farmers from getting the most out of their products. Intermediaries are hired by the agriculture department to supply them with essential goods. Such operation leads to the use of a large amount of manpower and time, which includes human errors as well.

A variety of smartphone apps have been developed to make agricultural business easier for farmers, retailers, and buyers. There is a variety of mobile app that has been developed for easy agricultural business for farmers as well as retailers or buyers. Several mobile applications have been developed to provide farmers with information on agricultural enhancements.

Here, various research papers and Mobile applications have been discussed related to the agriculture sector.

4.1 Mahafarm

This paper talks about emerging technologies in the field of agriculture in India. It discusses the use of information and communication technology (ICT) in agriculture, which is a new field that focuses on improving agricultural development in rural India.

4.2 E-agro android application

E-Agro Android Program is a software application aimed at assisting farmers in their long-term development. Many times, farmers are perplexed when it comes to choosing a fertilizer, pesticides, and the best time to perform specific farming tasks.

4.3 Bigbasket

India’s first comprehensive online megastore, bigbasket.com, brings a whopping 20000+ products with more than 1000 brands, to over 4 million happy customers [4]. We were able to find practically everything we needed in a single store. Every category has a large choice of possibilities offered at reasonable rates. Products are purchased from vendors and stored in large-scale warehouses or tiny go-downs, from which customers' orders are fulfilled.

5. CONCLUSION

We proposed a stable system that will help the food processing industry which includes farmers as well as users to get the most out of their products. Considering the previous works, the focus of the paper is to define a new concept to allow the farmers to sell their products directly to the consumers and for the industry to reap the benefits. Through this application, farmers can maximize their profits since they will not earn from the distributors who quote their stock prices. This application is a medium that helps farmers to market their goods to customers directly, bypassing any intermediaries that might cause farmers to lower their prices.

The application is an excellent trading platform for the agricultural market. Farmers and retailers can use this platform to acquire and sell their products at competitive pricing. The application strikes a balance in the agricultural trading infrastructure by removing the intermediaries that are involved in the trading between the farmers and the retailers. Moreover, this new idea of online trading for agricultural products that the system represents will enhance the agricultural market.

6. REFERENCES

[1] P. Shriram and S. Mhamane. Android App to Connect Farmers to Retailers and Food Processing Industry. 2018 3rd International Conference on Inventive Computation Technologies (ICICT); 2018 Nov 15-16; Coimbatore, India. 2018; 284-287.

[2] Martin Grasdal, Laura E. Hunter, Michael Cross, Laura Hunter, Debra Littlejohn Shinder and Thomas W. Shinder, Chapter 2 - MCSE 70-293: Planning Server Roles and Server Security (Syngress, 2003), p. 53-146.

[3] Techotopia, A Guided Tour of the Firebase Analytics Dashboard (2021), https://techtopia.com/index.php/A_Guided_Tour_of_the_Firebase_Analytics_Dashboard.

[4] BigBasket, Online grocery store (2021), https://www.bigbasket.com/.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 03, 82-87, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1103.1006

Modified Algorithm Based on Quadratic Correlation Time Delay Estimation

Zheng Jing
College of Communication Engineering
Chengdu University of Information Technology
Chengdu 610225, China

Yu-Zheng Zheng
College of Communication Engineering
Chengdu University of Information Technology
Chengdu 610225, China

Jian-Yu Meng
College of Communication Engineering
Chengdu University of Information Technology
Chengdu 610225, China

Abstract: In the passive positioning system, the traditional generalized cross-correlation algorithm estimates the effect under the influence of noise and reverberation, while the delay estimation algorithm after homomorphic filtering reduces the anti-reverb ability of the cross-correlation algorithm, but also reduces the cross-correlation of the signal, resulting in the decrease in the noise resistance of the cross-correlation algorithm. Based on this, the quadratic cross-correlation operation of the full-pass component of the signal is performed after homomorphic filtering to improve its noise immunity. In addition, the Introduction of Hilbert Difference sharpens the secondary cross-correlation peaks to make peak detection that reflects the time delay more accurate. Experimental simulation shows that the proposed method can effectively suppress the influence of noise and reverberation in non-stationary speech signals, and improve the performance of delay estimation.

Keywords: Time delay estimation ；Reverberation; Homomorphic filter; Quadratic cross-correlation; Hilbert

1. INTRODUCTION

Time-delay estimation (TDE) is a fundamental method for identifying, locating, and tracking radiation sources, with the goal of measuring the relative arrival time difference (TDOA) between different channels. Recently, there have been more and more smart devices and applications using voice-based locators. In a room, the sensor receives not only signals from direct paths, but also signals after attenuation and delay of source signals absorbed through reflections from the walls of the room. This multipath propagation effect introduces echo and spectral distortion in the observation signal, called reverberation, which severely reduces the performance of the delay estimation algorithm.

Mainstream methods for delay estimation include adaptive estimation algorithms[1], methods based on higher-order statistics[2], and methods based on cross-correlation [3]. Adaptive algorithms for latency estimation are able to track real-time latency between signals, but estimation accuracy is often limited at low signal-to-noise ratios. In order to eliminate Gaussian noise and improve the estimation accuracy, an estimation method based on a high-order statistic was studied, but the computational complexity was too high. As the most commonly used method, cross-correlation has been deeply studied for its simplicity and good performance. In order to obtain a higher time resolution at a very low signal-to-noise ratio, many improved algorithms based on intercorrelation are proposed. The quadratic correlation algorithm is one of the effective methods to suppress quadratic cross-correlation noise. In this paper, the TDE problem is studied, focusing on the problems of homomorphic filtering against reverberation and Hilbert differential noise immunity.

2. SIGNAL MODEL

In the process of TDOA positioning, the signals emitted by the source are received by sensors with different positions for delay estimation. Take, for example, the continuous signals received by receiver 1 and receiver 2. The continuous signal is set to $x_1(\mathrm{t})$ and the discrete signal is set to $x_2(\mathrm{t})$ and the discrete signal is set to $x(k)$. τ represents the time delay between the signal reaching receiving station 1 and receiving station 2, D represents the number of sample points of the delay, a_1 and a_2 represent the amplitudes of $x_1(\mathrm{t})$ and $x_2(\mathrm{t})$, and $w_1(\mathrm{t})$ and $w_2(\mathrm{t})$ represent Gaussian white noise. The above signals are all smooth Gaussian processes and independent of each other, and the mathematical model of the continuous signal can be expressed as equation (2-1), and the mathematical model of the discretized signal can be expressed as formula (2-2).

$$\begin{cases} x_1(\mathrm{t}) = a_1 s(\mathrm{t}) + w_1(\mathrm{t}) \\ x_2(\mathrm{t}) = a_2 s(\mathrm{t}-\tau) + w_2(\mathrm{t}) \end{cases} \qquad (2-1)$$

$$\begin{cases} x_1(k) = a_1 s(k) + w_1(k) \\ x_2(k) = a_2 s(k-\mathrm{D}) + w_2(k) \end{cases} \qquad (2-2)$$

where $s(t)$ is the signal sent by the source, $w_1(t)$ is the noise signal during transmission, $x_1(\mathrm{t})$ is the signal received by the monitoring station 1, and $s(k)$ is the discrete sequence obtained after sampling $s(t)$.

There is a reverberation in the actual environment, where receiver 1 and receiver 2 receive signals, respectively

$$\begin{cases} x_1(k) = h_1(k) * s(k) + w_1(k) \\ x_2(k) = h_2(k) * s(k-\mathrm{D}) + w_2(k) \end{cases} \qquad (2-3)$$

where $h_1(g)$ represents the room impulse response and $*$ represents the convolution operation. The analysis in this paper is a reverberation-based model that is close to the real environment.

3. ALGORITHM

3.1 Generalized Cross Correlation

Generalized Cross Correlation [4](GCC) on the basis of the CC algorithm, using $x_1(\mathrm{t})$ and $x_2(\mathrm{t})$ as the receiving signal, $x_1(\mathrm{t})$ and $x_2(\mathrm{t})$ are first processed by the pre-filter $H_1(f)$ and $H_2(f)$ respectively to obtain the output signals $y_1(\mathrm{t})$ and $y_2(\mathrm{t})$, and then as a general stationary signal using the Fourier function to transform the time domain signal $y_1(n)$ and $y_2(n)$ into the frequency domain processing, and the weighted function processes the cross-correlation function, making the main peak more sharp and prominent. Then use the Fourier inverse transformation to the time domain. This allows for more accurate latency estimation and stronger tracking of signals. The GCC algorithm flow diagram is shown in Figure 1.

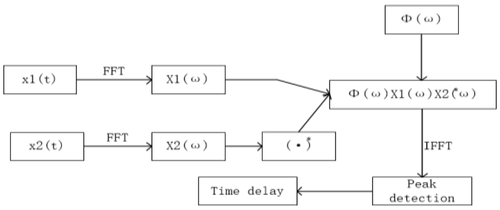


Figure 1 Schematic diagram of generalized cross-correlation algorithm flow

In the generalized cross-correlation weighting algorithm, in order to further improve the sharpness of the main peak and weaken other peaks, the GCC algorithm uses different weighted functions to process the cross-correlation power spectrum.

3.2 Homomorphic filtering

Homomorphic filtering[5], also known as reciprocal analysis, is a filtering technique used for anti-reverberation, where homomorphic filtering breaks down a signal into the smallest phase part and the full-pass component part. The minimum phase component of the signal has the characteristics of fast attenuation and small quadratic peak amplitude. The all-pass component of the signal provides position information useful for delay estimation. The reverberation intensity in the omni-pass component, especially early reverberation intensity, is greatly reduced. These features make the all-pass component of the signal not only retain the direct path delay information, but also reduce the reverberation effect. It can be seen that homomorphic filtering can improve the anti-reverb ability of the signal. The homomorphic filter decomposition method of the signal is shown in the following figure 2.

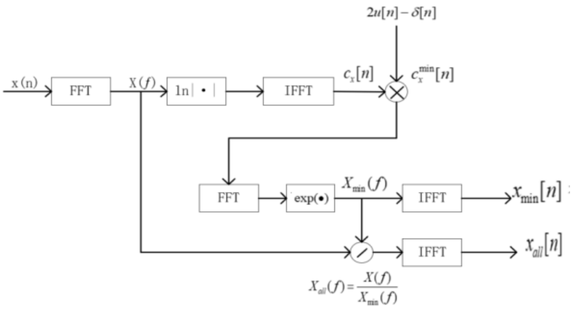


Figure 2 Decomposition of the minimum phase and full-pass components using homomorphic filtering.

The specific steps of the homomorphic filtering technology are shown in the figure above, and the received signal is homomorphic filtered by using the signal all-pass component to have the characteristics of anti-reverberation. After that, the all-pass components of the signal are cross-correlated to achieve the purpose of de-reverberation. However, since the minimum phase component is discarded during processing, the correlation between the signals is reduced, so consider using the secondary intercorrelation [6] to process the filtered signal to improve the correlation.

3.3 Hilbert transform

The Hilbert transform is to convert the correlation function of the even symmetry into odd symmetry, and the detection of the peak in the generalized mutual correlation is converted into the detection of the corresponding zero crossing point, and when the noise and reverberation interference cause the cross-correlation peak to become wider and flat, the delay can be estimated to a certain extent, and the noise immunity is certain. However, there is still a problem in judging the zero crossing point, and the influence of noise and other interferences makes it possible to fluctuate near the zero crossing point, resulting in multiple zero crossings, which delays the judgment on time, resulting in an increase in estimation error. In addition, when the signal sequence is too long, there are often multiple zero crossings, and in order to avoid this, other auxiliary algorithms are usually needed to detect it. However, this processing leads to a higher complexity of the peak detection algorithm.

Considering that the peak of the cross-correlation function corresponds to the zero crossing point of the Hilbert transform, a delay estimation algorithm for the Hilbert difference based on the intercorrelation is further proposed, that is, the difference between the cross-correlation function and the absolute value of the intercorrelation function after the Hilbert transformation[7]:

$$\mathrm{R}(\tau) = \mathrm{R}_{12}(\tau) - |\,hilbert(\mathrm{R}_{12}(\tau))\,| \qquad (3\text{-}1)$$

This algorithm can not only keep the value at the delay estimation of the cross-correlation function basically unchanged, but also suppress other pseudo-peaks of adjacent main peaks, solve the problem that adjacent pseudo-peaks and main peak amplitudes are similar, and also have better noise immunity. On the waveform, the peak of the intercorrelated waveform after the difference is sharper, which acts as a sharpening of the main peak and reduces the delay estimation error.

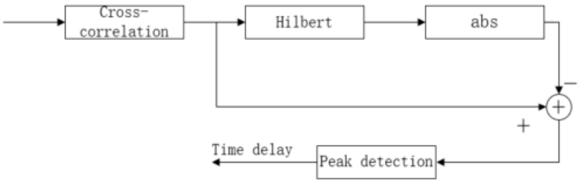


Figure 3 Schematic of Hilbert's delay estimation of the difference method.

The definition of the Hilbert transform is

$$\overset{\%}{R}(t)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{R(t)}{t-\tau}d\tau \qquad (3\text{-}2)$$

3.4 New algorithm

This paper improves the algorithm block diagram as shown in the figure 4, the receiver receives the signal for inverted spectral domain filtering, to obtain the full-pass component of the signal, the full-pass component signal for secondary cross-correlation, before the cross-correlation function peak detection, the cross-correlation function for Hilbert difference processing, that is, the cross-correlation function and the Hilbert transformation and take the absolute value after the cross-correlation function subtraction, in order to achieve the effect of sharpening the main peak. Finally, the peak detection is performed to estimate the delay.

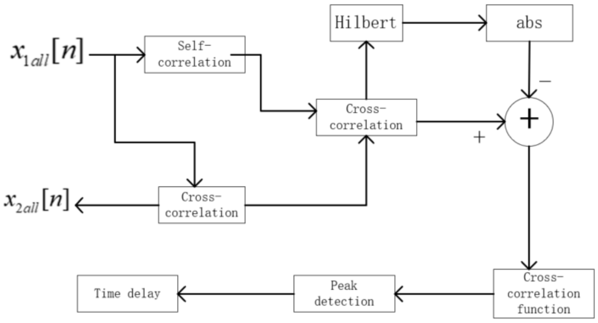


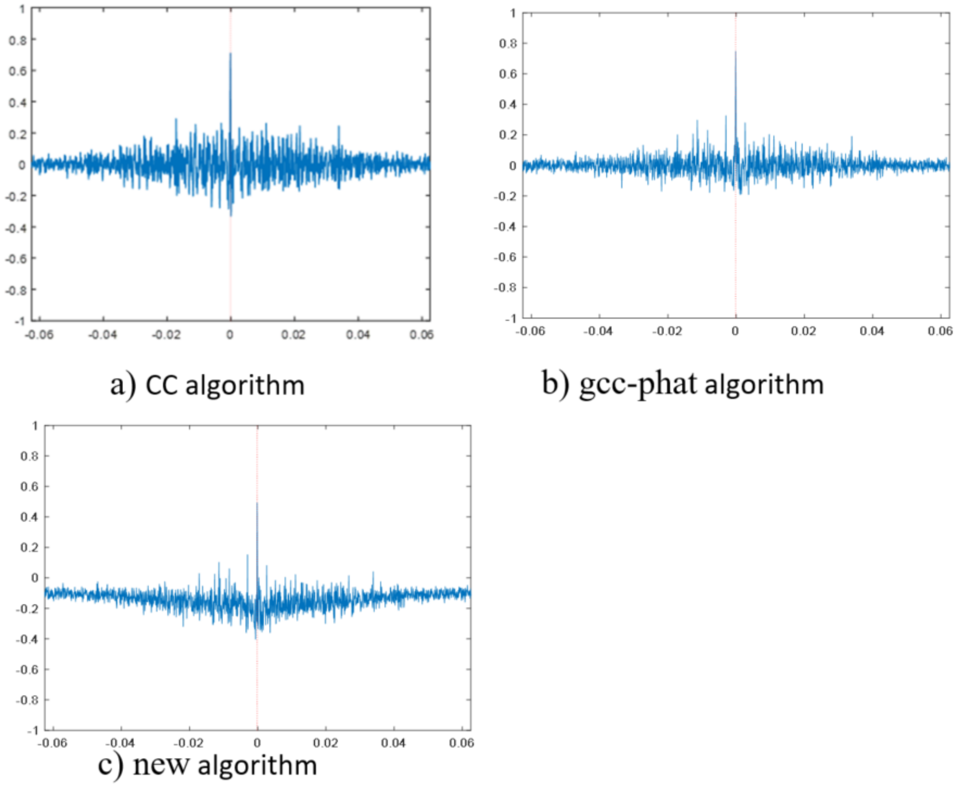
Figure 4 Block diagram of the new algorithm.

3.5 Simulation and analysis of results

In the simulation experiment, the length, width and height of the room are 10m, 5m and 3m, respectively, in the delay estimation algorithm, the delay difference between the two channels is estimated, and only the delay estimation of the signal received by the two microphones is required to verify the effectiveness of the algorithm.

During the simulation, a pure voice from the TIMT voice dataset is used, and the frequency range of the voice is 300Hz to 3400Hz, and the sampling frequency is 16kHz. The IMAGE method is used to simulate the pulse response of the room, simulating the reverberation caused by the absorption reflection of walls and objects in the room. In addition, there is the sound of tapping the keyboard, the hum of the air conditioner, and the boiling water of the water dispenser. These ambient noises are uncorrelated noises, which are simulated by adding Gaussian white noise with different signal-to-noise ratios to the signal.

The reverberation time in the indoor environment is usually 0.3 to 0.7s[], the following figure 5 is in the room small noise (signal-to-noise ratio 20dB), strong reverberation (the longer the reverberation time, the stronger the reverberation), the received signal mutual correlation obtained the result graph. where the abscissa is time and the ordinate is the normalized amplitude. The simulation diagram is, from top to bottom, the basic cross-correlation algorithm, the generalized cross-correlation algorithm and the cross-correlation function diagram of the new algorithm, and the main peak local diagram. By comparing function waveforms, you can compare the strengths and weaknesses of the new algorithm.



a) CC algorithm

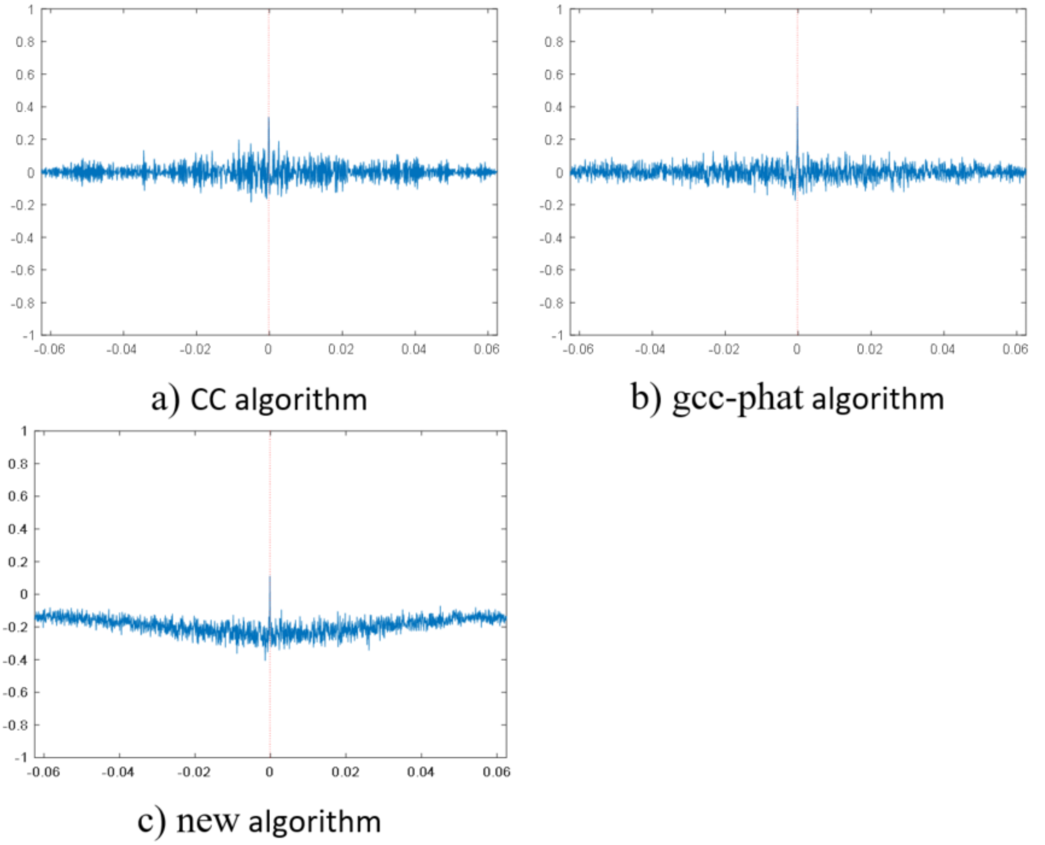
b) gcc-phat algorithm

c) new algorithm

Figure 5 Algorithm comparison waveform plot in reverberation environment.

Comparing Figure 5, it can be seen that in the case of large reverberation intensity, there are more glitches near the peak of the basic cross-correlation function, and the peak is high, and the position of the main peak is error. The correlation algorithm after inverted spectral domain filtering reduces the number of spectral peaks near the main peak, and the correlation function waveform processed by the algorithm in this paper makes the main peak sharper, and the difference between the main peak and the nearby spectral peak increases, which is conducive to improving the delay estimation accuracy.

Similarly, Figure 6 is a simulation performed in an environment with high ambient noise interference (signal-to-noise ratio of -5dB) and small reverberation times. It can be seen that when the ambient noise is large, the main peak of the correlation function is very close to the noise, and the main peak is submerged in the noise. The new algorithm makes the peak of the main peak sharp and has a certain resistance to noise.



a) CC algorithm

b) gcc-phat algorithm

c) new algorithm

Figure 6 Algorithm comparison waveform plot in noisy environments.

In summary, the new algorithm has a good effect on anti-reverb and a slight improvement in anti-interference.

Figure 7 a is a graph of the mean squared error curve of delay estimation for each algorithm at different signal-to-noise ratios under low reverberation conditions (RT60=300ms).

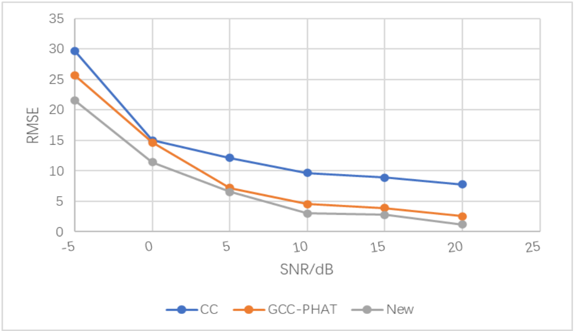


Figure7 The relationship between signal-to-noise ratio and delay estimation mean squared error.

According to Figure a, the CC algorithm and the GCC-PHAT algorithm are sensitive to noise, when the signal-to-noise ratio is less than 10dB, the delay estimation mean squared error of the CC algorithm is greater than 10%, in comparison, the noise immunity of the GCC-PHAT algorithm is slightly stronger than that of the CC algorithm. Compared with the traditional first two algorithms, the new algorithm has better noise resistance performance than high signal-to-noise ratio at low signal-to-noise ratio. However, on the whole, the new algorithm has a high signal-to-noise ratio, and the noise resistance performance is improved by 1% to 2%.

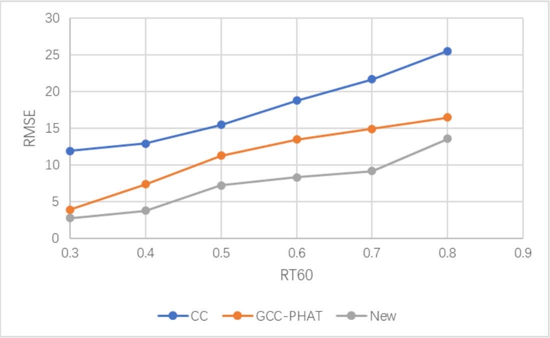


Figure8 Reverberation time and delay estimate the relationship between mean squared error.

4. CONLUSION

In this paper, after homomorphic filtering of the received signal, the correlation of the signal is weakened, which affects the accuracy of the delay estimation, and a new delay algorithm is proposed. Homomorphic filtering processing signal, the signal is divided into full pass component and minimum phase component, the signal all-pass component contains more direct sound part, and the minimum phase component contains more reverberation components, so the reverberation has almost no effect on the all-pass component, only the all-pass component part is mutually correlated to achieve the purpose of anti-reverb. The quadratic cross-correlation is used to improve the correlation degree of the signal, and the Hilbert difference is introduced to sharpen the cross-correlation peak and improve the delay estimation accuracy. Simulation results show that compared with the basic cross-correlation algorithm, the proposed algorithm reduces the rms error of delay estimation by 13% and the GCC-PHAT algorithm by 5%, which can weaken the influence of room reverb on delay estimation. In terms of noise immunity, the proposed algorithm is also improved compared with the traditional algorithm, which improves the noise immunity of the delay estimation algorithm.

5. ACKNOWLEDGMENTS

This work is partially supported by the Sichuan Science and Technology Program (2021YFG0149).

6. REFERENCES

[1] So H C , Ching P C . Comparative study of five LMS-based adaptive time delay estimators[J]. IEE proceedings. Radar, sonar and navigation, 2001, 148(1):p.9-15.

[2] Hinich, M. J , Wilson, et al. Time delay estimation using the cross bispectrum[J]. Signal Processing, IEEE Transactions on, 1992, 40(1):106-113.

[3] Chen J . Time delay estimation in room acoustic environments : an overview[J]. EURASIP J. Applied Sig. Proc, 2006, 2006.

[4] Knapp C H . The generalized correlation method for estimation of time delay[J]. IEEE Trans. Acoust. Speech and Signal Processing, 1976, 24.

[5] Mosayyebpour S , Lohrasbipeydeh H , Esmaeili M , et al. Time delay estimation via minimum-phase and all-pass component processing[C]// Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[6] Tang J , Xing H Y . Time Delay Estimation Based on Second Correlation[J]. Computer Engineering, 2007. Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[7] Cabot R . A note on the application of the Hilbert transform to time delay estimation. IEEE, 2003. Spector, A. Z. 1989. Achieving application requirements. In Distributed Systems, S. Mullender