Journal of Autonomous Intelligence (2024) Volume 7 Issue 3
doi: 10.32629/jai.v7i3.1060

ORIGINAL RESEARCH ARTICLE

An investigation and analysis on automatic speech recognition systems

R. V. Siva Balan[1,*], K. Vignesh[2], Teena Jose[1], P. Kalpana[1], Jothikumar R.[3]

[1] *Department of Computer Science, CHRIST (Deemed to be University), Bangalore 560029, India*

[2] *Department of Management Studies, Kumaraguru College of Technology, Coimbatore 641049, India*

[3] *Department of Computer Science and Engineering, Shadan College of Engineering and Technology, Hyderabad 500086, India*

*** Corresponding author:** R. V. Siva Balan, sivabalan.rv@christuniversity.in

ABSTRACT

A crucial part of a Speech Recognition System (SRS) is working on its most fundamental modules with the latest technology. While the fundamentals provide basic insights into the system, the recent technologies used on it would provide more ways of exploring and exploiting the fundamentals to upgrade the system itself. These upgrades end up in finding more specific ways to enhance the scope of SRS. Algorithms like the Hidden Markov Model (HMM), Artificial Neural Network (ANN), the hybrid versions of HMM and ANN, Recurrent Neural Networks (RNN), and many similar are used in accomplishing high performance in SRS systems. Considering the domain of application of SRS, the algorithm selection criteria play a critical role in enhancing the performance of SRS. The algorithm chosen for SRS should finally work in hand with the language model conformed to the natural language constraints. Each language model follows a variety of methods according to the application domain. Hybrid constraints are considered in the case of geography-specific dialects.

Keywords: speech recognition system; natural language; speech processing; language model; speech technology; ensemble methods

ARTICLE INFO

Received: 27 July 2023
Accepted: 16 October 2023
Available online: 3 January 2024

1. Introduction

Usage of new technology upon SRS[1–3] would help researchers in finding new ways and methods to apply new enhancements. To understand this specialization of SRS, a deep understanding of the fundamentals of SRS is mandated.

Computers are not designed to work in an analog environment. Digital computers are not capable of processing analog signals from analog environments. But the fundamental of SRS is the analog sound input obtained from an analog environment[4–6]. For this reason, the analog properties of a received sound signal should be pre-processed and converted to its corresponding digital form, so that, a computer can recognize and process those signals as digital data. To recognize and understand the information from the analog world peripheral devices such as microphones play a major role in recording and digitalizing these analog signals with the help of certain utility software and applications. During this activity of pre-processing, the quality of peripheral devices (e.g., microphone) used in this activity may influence the details of such digital data. This means that pre-processing activity should include the mechanisms to differentiate and isolate any unnecessary details added to the original input sound.

When the sound/voice is recorded unnecessary physical attributes[7] of the sound/voice are added in the digital form of it. But only useful information should be processed and digitized, for this, selective useful information will be traced.

Spectrogram is a mechanism that is used to find detailed insights into a recorded/received sound/voice[8–10]. The insights into the 'recorded sound' shall be obtained by creating a visualization of the digital features of the recorded sound/voice using a spectrogram. Understanding the idea behind a spectrogram is a must to get more details from the recorded sound/voice. To create a spectrogram of the given sound, three important steps should be followed:

1) Capture the sound.
2) Cluster the sound waves into blocks based on time units.
3) Apply fast Fourier transform.

For an illustration, consider **Figure 1** which represents 'three spoken words', and consider it is represented in analog format.

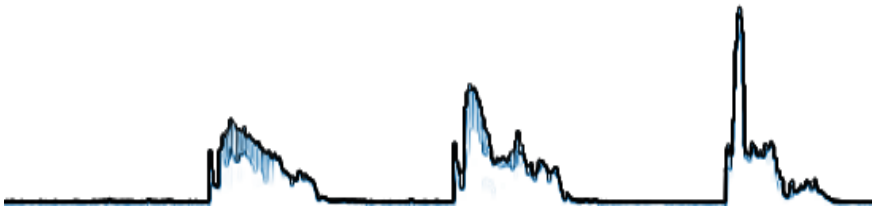

Figure 1. Analog model of speech wave.

The sound waves from the analog signals can be plotted "as-it-is", but, in a time-series format. The sound wave from the analog environment is represented in a time-divided graphical representation in the following **Figure 2**. Now it is easy to recognize the width of each block in time units and the height of the individual blocks as the amplitude of the soundtrack at a particular unit of time[11]. Now for each such state, a numerical value will be assigned. The essence of converting those analog units into a number system is to provide a digital representation to a set of features available from the input sound. And height of the blocks is considered for a numerical value and shall be considered as the digital representation[12,13] of the sound recorded at that specific time period. This means the conversion of analog sound is recorded in a digital format which can be traced for further details.

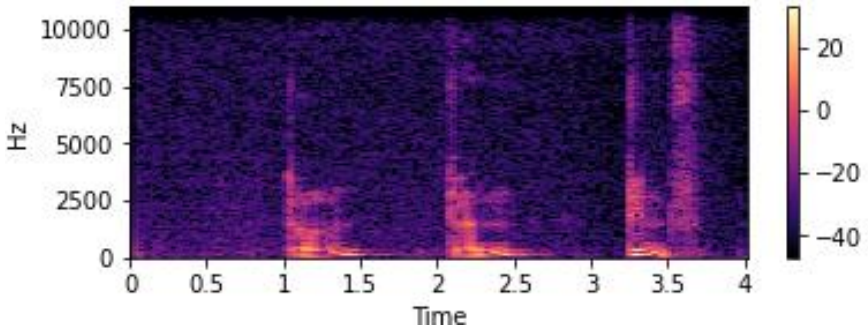

Figure 2. Time-divided speech wave in digital format.

Even after digitizing the sound the process of creating a spectrogram isn't completed. To complete the process three important properties of sound should be considered:

1) Frequency of the sound recorded.
2) Density of the sound.
3) Time taken to create this continuous sound.

Once these properties are specified and recorded, other properties of the recorded sound should be taken care of. The first and foremost such property is linguistics[14–16]. After constructing a spectrogram for an input, the linguistic properties of the sound wave should be considered. The basic linguistic property to be extracted from the sound wave is a phoneme. A phoneme is the fundamental element of a linguistic construct. A phoneme is a sound that is pronounced for 20–40 milliseconds and continuous data of phoneme are recognized as a word spoken. According to the SRS applications, and the considered linguistic of study, a phoneme shall be considered as a unit of sound that distinguishes each spelled word.

When phoneme changes, with a higher probability, the meaning of that spelled word may also change in accordance with phoneme variation. As an example, consider the three words, “turn”, “ten” and “tins” which are three different words distinguished by phonemes.

This article is organized to explain the methods and models in section (2), algorithm selection criteria is elucidated in section (3), an End-to-End Neural SRS system is pitched as an alternative model in section (4), and section (5) concludes.

2. Methods and models

Languages are spoken by different people with different physical properties such as age, gender, emotion, accent, and context that influence the variations in speech and pronunciation. Variations shall be recognized by interleaved hollow phonemes. These variations represent the uniqueness of natural human languages. Any technologies used for speech recognition should then include the facilities to identify any phonemes. Phonemes (**Table 1**) are very basic units in speech recognition, so any technologies used for SRS may use phonemes to recognize the words spelled from the analog environment. Any speech recognition system does it using two important techniques. The first one is the HMM[17,18], and the second one is the Artificial Neural Networks (ANN)[19].

Table 1. Three words input listed with the corresponding phoneme sequences.

Word	Phoneme
Turn	/tɜ:(er) n/
Ten	/tɜ:(eh) n/
Tins	/tɜ:(ih) n/z/

2.1. The role of HMM

It is used to reconstruct the right phrase by putting the right phonemes at the correct places of that phrase under construction. HMM does it using statistical probabilities; statistical probability decides on the sequence of phonemes, particularly about a phoneme in the sequence which precedes or succeeds the other one. HMM does it in three layers:

Layer-1: The first layer of the model checks the statistical probabilities at the acoustic level, for whether the correct phoneme is heard or not. The factors behind this are the physical properties including the human emotions of the speaker[20]. Here, the influence of such physical properties needs to be understood.

Layer-2: In the second layer of the model, HMM makes sure that the phonemes are recognized in proper order. These sequences of phonemes are considered time-series data[21,22] and ensured each phoneme is in its correct position. If the sequence tends to provide any meaningless words, then the alternative phonemes are reconsidered for presumably misplaced phonemes[23].

Similarly, it is possible to evaluate any linguistic properties including the grammar of the language. Once a word is constructed out of a sequence of phonemes it should be ensured that this new word is meaningful in the phrase under construction.

2.2. Artificial neural networks

The fundamental idea[24] of a neural network is based on the "simulation of a human brain", such that, it contains several neurons which are organized into several layers, and nodes from each adjacent layer are connected to each other, shown in **Figure 3**. The above figure represents a complete model which takes the audio signals as continuous input signals and turns them into a sequence of meaningful words, perfectly aligned, as per the requirement of the context. The model above includes convolutional layers at the initial level to learn the features, then, a set of dense layers to take in the learned features of convolution layers. Then, LSTM layers will be employed to learn the continuous sequence of words according to the context. Finally, dense layers will be used to predict time-sequenced words. Varieties of connections are demonstrated to produce different versions of network structures in correspondence with the requirement of any application domain. Here, the speech recognition model is a hybrid combination[25] of HMM and ANN.

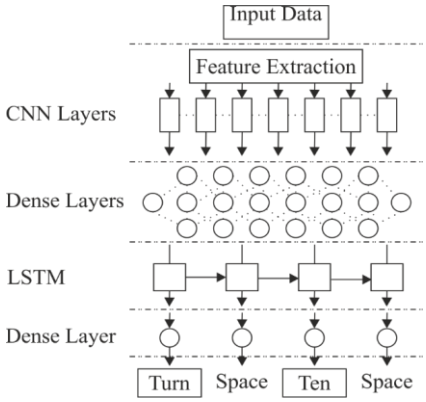

Figure 3. ANN model for SRS system.

2.3. Dealing with the physical properties of speech

The acoustic model will be a neural network that takes in the speech waves as input and provides a sequence of text. In order to do this, this neural network model should be trained with a proper and precise dataset according to the application domain. The construction of the model depends on the type of problem for which this solution is required. As it is understood that the speech waves are recognized in sequential structure, it requires a capable neural network to process this time series data. A recurrent Neural Network (RNN) is the best-known solution for such problems which involve time-series data[26].

This acoustic neural network model[27], should come with computations of low complexities and with real-time capabilities. And should be trained with a dataset in which a wide variety of data details should be included, to improve the evaluation accuracy and to improve real-time prediction. Whatever a neural network learns during training is dependent on the variations in the given data (gender, age, accent, noise, emotions, etc.).

3. Algorithm selection criteria

The algorithms used in this form of technology include PLP features, Viterbi search, deep neural networks, discrimination training, WFST framework, etc.[28]. Google's recent publications[29,30] on their new inventions in speech are very exciting. The algorithms used by Google are available in open-source format. It would be better if machine learning communities are using speech recognition along with voice synthesis to bring in the power of input recognition for the better.

3.1. Acoustic model approach

Acoustic model output = language model + rescoring algorithm. The rescoring algorithm may be a CTC Beam Search algorithm.

Ensemble methods:

1) Use a CNN model for feature extraction and with a few final dense layers.
2) CNN output from dense layers is taken as inputs to another RNN layer to work on phonemes (go to Layer-1 and -2 of HMM).

The implementation model: Again, in **Figure 4**, from the second module, the sub-module of comparison and selection of words is done based on the phoneme sequences produced by the first module. Now the selection of words is done with the help of a rescoring algorithm[31].

Probability dictionaries are used to find the difference between the word spelled and its alternative words (with a similar probability of being pronounced). CMU.DICT is such a dictionary[32]. Most language models support 20–60 phonemes. The above diagram depicts the architecture to extract the acoustic features which are extracted from a raw speech signal; the expected interim output from the acoustic model is likely a phoneme sequence that corresponds to the particular speech utterances.

HMM is a paradigm that is used to learn this mapping between speech utterances and phoneme sequences. Each of these acoustic feature (frame) vectors extracted from speech utterances corresponds to a speech frame[33].

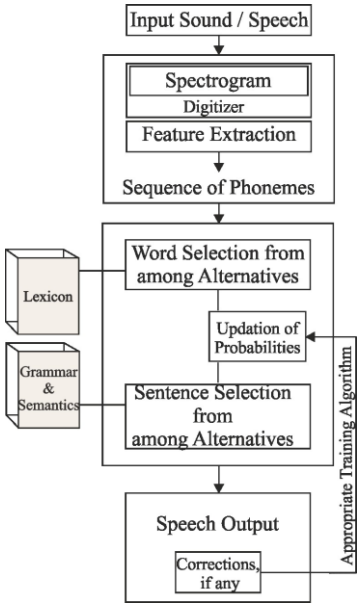

Figure 4. Architecture of an SRS system.

The number of frames that correspond to a particular phoneme depends on the chains of hidden states in the HMM[34]. **Figure 5** given below represents such dependencies as a graph and the probabilistic values as the weights on the arcs. All the weights on available arcs are probabilities. Here a single chain represents just one sequence of these phonemes[35]. But the problem here is that—At each point, probabilistic mapping is done for each phoneme that is going to appear next.

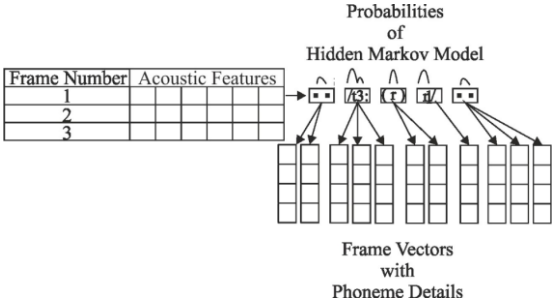

Figure 5. Mapping of acoustic features to phoneme sequences.

The entire model is probabilistic, and in **Figure 6** then is simplified by showing a single chain of phonemes. And then it provides estimates that forecast "the initial few frames" most likely corresponds to a certain phoneme and the transition probability.

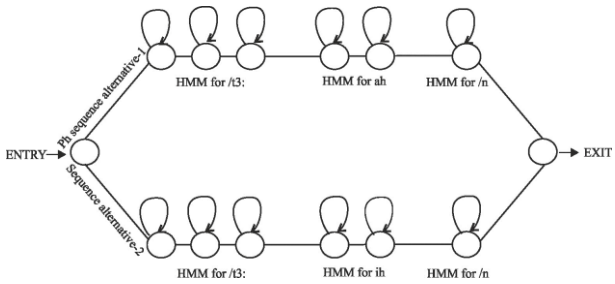

Figure 6. Search graph working on phoneme sequences.

It is the probability of transition from one phoneme to another and thus determines "how many frames" are to be formed in correspondence to each phoneme. And then once a particular STATE of transition is attained, then there are probabilities for generating each of these frame vectors. These probabilities are derived from well-known Gaussian mixture models[36]. So, it is again a probability distribution that determines a particular speech vector that could be generated with a certain probability. And provides insights and understandings about these probabilities that are learned from training data.

Now, this is a high-level idea of having a probabilistic model which maps a sequence of feature vectors to a sequence of phonemes. So, HMM works at the heart of a speech model for a long period in historical reviews on SRS systems[37]. Currently, deep neural networks are used for a similar mapping, with the available speech signal, fixed windows of speech frames are extracted and a lookup is performed at a fixed window of frames around it, then all of those generated features are put together. And those features are input to a Deep Neural Network (DNN), for this, the output is "the most likely phonemes" about to be produced, given for a particular set of speech frames[38]. But this is only a posterior probability of the phonemes. In other words, it means an estimate is derived on the most likely phoneme next in the speech frame. This also means a probability distribution for all the phonemes[39].

There are two ways in which DNN shall be used in an acoustic model. First, with an HMM, the states are derived and then probability distributions are calculated to conclude on "mapping of speech vectors to a particular state". Second, instead of the probability distribution from Gaussian mixtures, the probability distribution of DNN is used. By doing this, these HMM and DNN models can be combined and put together within an acoustic model. Thus, the mixed values may be the output from such acoustic models. Now, the resultant observation probability shall be either extracted out of the Gaussian mixture model, or from the scaled posteriors of the DNN. Once the phoneme sequence and corresponding word are identified, now it is time to identify the word sequence.

3.2. To identify the word sequence

Till now, all those sub-modules are put together to produce intermediate results which are just identified as a correct sequence of phonemes. From the speech utterances, the model is trying to identify the sequence of words, and in the due course, it identified the correct sequence of phonemes which are actually intermediate representations. Now the question is "how to move further to identify the word sequences?" The answer is "pronunciation dictionaries"[40,41].

Large open-source pronunciation dictionaries are available (**Table 2**) for usage, and by involving such dictionaries, it is possible to extend the model to create a link between phoneme sequence and the expected words. And this is the only sub-module in the whole SRS system which has not been learned from the training data. This sub-module is connected to a language model, where this new language model itself is also a learned

model or trained model. The pronunciation model is an expert-driven model or knowledge-derived dependency model. There are several significant works available on these pronunciation models[42].

Table 2. Meta Data on Existing Datasets.

Dataset	Meta Data			
	Open Source	**Sampling**	**Speech**	**Annotation**
Voforge	YES	8 KHz and 16 KHz	130	Words
Libri speech	YES	16 KHz	1000	Words
Tedlium	YES	16 KHz	118	Words
Switch board	NO	8 KHz	300	Words, Phonemes, Sentiments

Switchboard and other related works[43], provide a hindrance that the pronunciation model is a restricted representation constructed using "sequence of phonemes". Linguistic datasets are annotated with a very detailed level of information to work on transcribed phonetic sequences. Such models provide information about 'linguistic pronunciation' regardless of the standard pronunciation mapping available in benchmark dictionaries[44].

3.3. An evaluation model

Articulatory models are very significant models to track and identify pronunciation variations[45]. Even though this articulatory model is not considered a part of the SRS studied, this model shall be used as a tool to validate the results of a sub-system that is used to produce words from an identified sequence of phonemes. This model is actually a simulation of the human articulatory sub-system which considered speech as an output of continuous data produced by articulatory movements. Even though this is used as a validation model here, it can also be used as a substitute for the dictionary-based restricted model. Articulatory models work better in the case of finding a larger deviation in pronunciation[46].

In this model, pronunciation is represented as a stream of features. In the human articulatory system, certain features may not synchronously move due to certain physical malfunctions; misspelled phonemes may be produced without intention. To identify this error, any data with overlapping details shall be probed. In that case, if a wobble sound is expected after a nasal sound in sequence but if the nasal sound continues to appear with an overlapped wobble sound, then this frame of data can be probed for errors. This error may be caused due to the malfunction of the nasal tube. Actually, this framework of the articulatory model works perfectly in identifying the pronunciation variation. The combination of a generalized HMM and Dynamic Bayesian Network[47] shall be used to implement this model with variable attributes and constraints set up for this model.

3.4. The final component (language model)

The final component of this SRS system is a language model[48]. Again, it is a probabilistic model used to identify the "most likely next word" to appear in the sequence of words that is under construction. The previous sub-module (pronunciation model) tries to find correct words with the help of a sequence of the phoneme, whereas, this final module tries to identify the correct sequence of words (one by one). For each probable word, the sequence is validated in compliance with the scope of context. The context of the sequence will be tried on the basis of the language chosen to work with. To represent a model in a finite state machine these models[49,50] will be useful.

Estimating word probabilities using the N-Gram language model

In a working word context, it is concentrated on the counts of particular words in a large text corpus/corpus, for a particular natural language. It is done in order to find the frequency of a particular set of words (N-Grams) appearing in the corpus. And a relative count of occurrences of such a sequence (N-Grams)

is computed to get the probability of occurrence of this sequence (N-Grams) in such an application context. For example, Probability

$$(\text{“tins”}/\text{“Turn ten\_A-GRAM\_”}) = \boldsymbol{\pi}\ (\text{“Turn ten tins”})/\boldsymbol{\pi}\ (\text{“Turn ten”}) \quad (1)$$

But this N-Gram-based model has its own disadvantage. It is possible that a machine/component deployed to predict the occurrences of a word/sequence might not be trained to predict a certain occurrence. That is, in training data the sub-module under training might not see a sequence of words (“Turn ten tins”) to appear, if it is so, then “how can the same component predict an unseen combination?” This is the disadvantage of such an un-smoothed N-Gram model.

Problems with un-smoothed N-Gram estimate maximize[51] the likelihood of the observed data overfitting the training data. To overcome this disadvantage smoothing techniques like “Good Turing”[52] are used. And such smoothing techniques reserve some probability mass to N-Gram that doesn’t occur in the training corpus. Based on the technique of distribution of the probability mass, there are a variety of smoothing techniques available (Equation (1) is used in smoothing). Actually, the efficiency of those algorithms is based on the technique used for the distribution of mass. And the probabilities for unseen N-Grams are provided by the trade-off and priority assumed by those smoothing algorithms. There are a wide variety of ways to do this. Even though these language models, following N-Grams are estimated as not as enough faster to perform their task. Because of these limitations of N-Gram-based language models, there exists a need to build more efficient language models. This need shifted the focus of the research community on Artificial Neural networks (ANN) and is exploring the options with Recurrent Neural Network (RNN) based models. One of such better works identified comes with the combination of a ‘rescoring algorithm’ and an RNN-based model.

3.5. Next sub-module, the decoder

Until here, the components of the SRS system discussed will do their part to estimate the sequence of phonemes, and then will estimate the sequence of words (N-Grams). Now, the task assigned with a decoder is to identify “the most likely word sequence”[53] according to the input speech utterances. This means that the results of the previous components are put together to find the most accurate sequence of words from the available solution space. At this stage, the estimation problem turns over into a searching problem.

Figure 7 illustrates a naïve search graph with a particular starting point and leads to provide alternatives. Assume that the decoder has to choose between two words while forming a sequence; for example, “ten” and “tin”, then it will have to work on a transition process to select any of the alternatives (words) by considering the sequence of phonemes extracted from the speech utterances (a look-up to the output of previous sub-modules). In this context, the words, “ten” or “tin” refer to the phonemes and each phoneme refers to its respective HMM module which calculates their probabilities. From the given illustration it is understood that for just two alternative words this model is large enough; and when more alternatives and sequences are queued in consideration, then this process of construction of a graph and searching through the network of the graph itself will become highly time-complex. Normally, this kind of component works with more than 20,000 words[54], and in such scenarios, the computations involved in this process of graph construction, monitoring, and searching will become a task of exponential time complexity.

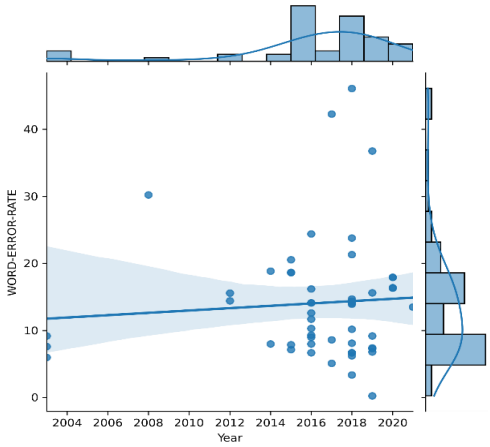

Figure 7. Increased WER obtained in recent works.

Network of words—The part of decoder

This graph model can grow very rapidly and then the mapping process[55,56] will get into a much more complex task to result in exponential order of time of computations. The graph actually will represent millions of states at a time. So, it is almost impossible to search throughout the entire graph for a perfect (combination) solution.

As a resolution to this kind of problem, a suitable "approximate search technique"[57] should be employed. Another sub-domain of research in this field is to study the varieties of such useful approximate-search algorithms. These algorithms are efficiently made used in the decoder component of an SRS.

4. An alternative (end-to-end neural SRS system) and analysis

Although this system under study has a great scope for further research, development, and enhancement, an alternative perception in the development of SRS systems is moving towards End-to-End Neural SRS systems[58]. End-to-End systems are prone to word errors. Those word errors are measured on the basis of word error rate (WER)[59]. A study on WER is done, and the data based on WER is shown in **Table 3**. The trend of recent research works in handling the WER is shown in **Figures 7** and **8**, and error margins in assessing WER are depicted in **Figure 9**. **Figure 10** demonstrates the quality of error-free work done on specific natural languages. Regardless of WER, SRS systems are the current trend and, in such systems, the individual letters (characters) are learned to be mapped to the acoustic features extracted from the input speech utterances. This means it is a component to directly produce character sequences from the input speech vector without working on the intermediate search space of "Network of Words". This leads to a character language model which rescores the character sequence by bypassing the pronunciation model. In the principle of these systems, they don't bother whether these words are there in the vocabulary/dictionary or not, because the system not predicting the sequence of words, but instead, the system is working on a sequence of characters.

Table 3. Mitigated word error rate year-wise data.

Language	Corpus	Model	(%) Word-error-rate	Year
Arabic	TARIC (Tunisian Arabic Railway Interaction Corpus)	Rule based ASR	8	2014
English	Wall Street Journal (WSJ) corpus	DNN-HMM	7.14	2015
Multi-Linguistic	Wall Street Journal (WSJ) corpus and Corpus of Spontaneous Japanese (CSJ)	CTC-BLSTM-MAP	6.7	2016
English	2000 Switchboard evaluation set	Confusion network combination + LSTM rescoring + ngram rescoring + backchannel penalty	5.1	2017

Table 3. (*Continued*).

Language	Corpus	Model	(%) Word-error-rate	Year
English	Wall Street Journal (WSJ) corpus	Pruned algorithm for lattice-rescoring with RNNLMs	3.38	2018
English	Google voice-search traffic	Bi-LAS + MWER	6.2	2018
English	Google's voice search and dictation traffic	RNN-T + Word Piece	6.8	2019
English	Microsoft data	Recurrent Neural Network Transducer (RNN-T)	16.3	2020
German	Swiss German multi-dialect dataset	Supervised acoustic pre-training (wav2vec)	13.5	2021

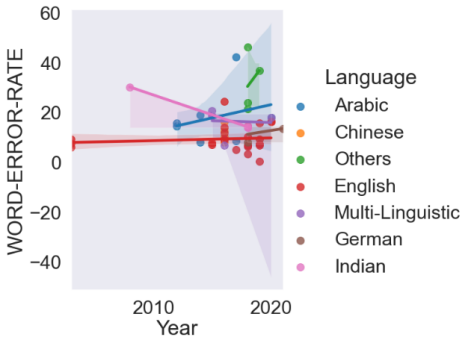

Figure 8. Trends of WER in natural languages.

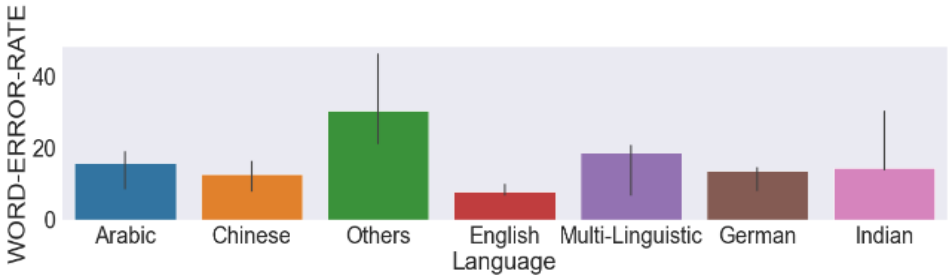

Figure 9. Error margins for exclusive Languages.

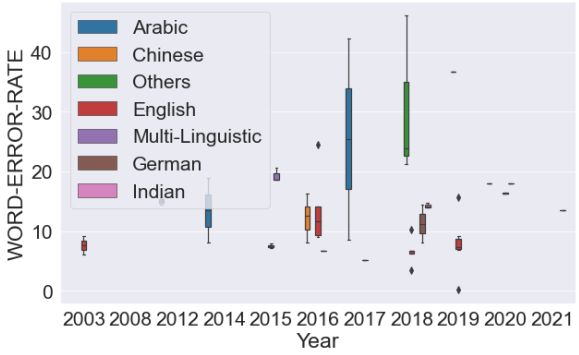

Figure 10. Performances of individual research works based on WER for natural languages.

Here, each wave corresponds to various phonemes and averages all the character probabilities corresponding to those particular phonemes.

This uses a very popular paradigm, the sequence-to-sequence model which is known as encoder-decoder networks with attention[60,61]. But End-to-End systems are not considered to be in the standard stream. Researchers may bridge the gap between this end-to-end system and the whole research stream[62].

Along with the trial to understand the speech sound, here the efforts are taken to understand the spellings from the speech utterances. For a global language like English for which aortography is not so clear it is a

critical job to implement such end-to-end systems.

5. Conclusion

While studying a complete SRS system and a few alternative components within SRS systems, many flaws in the existing systems are identified. These flaws include the systems' inability to adopt physical attributes of speakers such as age, accent, and ability. Efficiency should be improved in handling noisy real-life settings with many speakers. Better algorithms should be designed to handle pronunciation variability. Algorithms should be able to learn any cues in speech and should disambiguate the utterances in the speech translation part. There are "N" solutions available for SRS systems. But each of those solutions can be achieved using "N" different methods. This means a combination of "N × N" possibilities shall be probed to find an efficient solution among the probabilities. The other area which needs to be concentrated on is the resource constraints for the solution, which includes the ability of real-time decoding using limited computational power. To improve the efficiency of SRS systems, its ability to reduce duplicated effort across the domain should also be considered.

Author contributions

Conceptualization, RVSB; methodology, RVSB; validation, RVSB; formal analysis, RVSB; investigation, KV and JR; resources, KV and JR; data curation, KV and JR; writing—original draft preparation, TJ; writing—review and editing, TJ; visualization, PK; supervision, PK; project administration, PK. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Ockph T. Fundamental technologies in modern speech recognition. IEEE Signal Processing Magazine, 2012.
2. Xiong W, Wu L, Alleva F, et al. The Microsoft 2017 Conversational Speech Recognition System. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. doi: 10.1109/icassp.2018.8461870
3. Pratap V, Hannun A, Xu Q, et al. Wav2Letter++: A Fast Open-source Speech Recognition System. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. doi: 10.1109/icassp.2019.8683535
4. Boll S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1979, 27(2): 113-120. doi: 10.1109/tassp.1979.1163209
5. Barlindhaug G. Analog sound in the age of digital tools. The story of the failure of digital technology.
6. Di Rosario G. Electronic poetry: understanding poetry in the digital environment (No. 154). University of Jyväskylä; 2011.
7. Layher W. Sound, Voice, and Vox: The Acoustics of the Self in the Middle Ages. Queenship and Voice in Medieval Northern Europe. 2010, 29-52. doi: 10.1057/9780230113022_3
8. Oppenheim AV. Speech spectrograms using the fast Fourier transform. IEEE Spectrum. 1970, 7(8): 57-62. doi: 10.1109/mspec.1970.5213512
9. Jeong J, Williams WJ. Mechanism of the cross-terms in spectrograms. IEEE Transactions on Signal Processing. 1992, 40(10): 2608-2613. doi: 10.1109/78.157305
10. Zeng Y, Mao H, Peng D, et al. Spectrogram based multi-task audio classification. Multimedia Tools and Applications. 2017, 78(3): 3705-3722. doi: 10.1007/s11042-017-5539-3
11. Banerjee A, Pandey S, Hussainy MA. Separability of Human Voices by Clustering Statistical Pitch Parameters. 2018 3rd International Conference for Convergence in Technology (I2CT). 2018. doi: 10.1109/i2ct.2018.8529762
12. McDonald JC. The Analog Termination. Fundamentals of Digital Switching. 1990, 237-284. doi: 10.1007/978-1-4684-9880-6_7
13. Story BH, Bunton K. Formant measurement in children's speech based on spectral filtering. Speech Communication. 2016, 76: 93-111. doi: 10.1016/j.specom.2015.11.001
14. Harris ZS. Structural linguistics.

15. Caravolas M, Volín J, Hulme C. Phoneme awareness is a key component of alphabetic literacy skills in consistent and inconsistent orthographies: Evidence from Czech and English children. Journal of Experimental Child Psychology. 2005, 92(2): 107-139. doi: 10.1016/j.jecp.2005.04.003
16. Davidson A. Writing: the re-construction of language. Language Sciences. 2019, 72: 134-149. doi: 10.1016/j.langsci.2018.09.004
17. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. 1989, 77(2): 257-286. doi: 10.1109/5.18626
18. Young S. HMMs and Related Speech Recognition Technologies. Springer Handbook of Speech Processing. 2008, 539-558. doi: 10.1007/978-3-540-49127-9_27
19. Hussain S, Nazir R, Javeed U, et al. Speech Recognition Using Artificial Neural Network. Lecture Notes in Networks and Systems. 2021, 83-92. doi: 10.1007/978-981-16-2422-3_7
20. Mary L. Significance of Prosody for Speaker, Language, Emotion, and Speech Recognition. Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition. 2018, 1-22. doi: 10.1007/978-3-319-91171-7_1
21. Dennis D, Acar DAE, Mandikal V, et al. Shallow RNN: accurate time-series classification on resource constrained devices. Advances in Neural Information Processing Systems, 2019, 32.
22. Han K, et al. Transformer in transformer. Advances in Neural Information Processing Systems 2021, 34: 15908-15919.
23. Roodenrys S, Miller LM, Josifovski N. Phonemic interference in short-term memory contributes to forgetting but is not due to overwriting. Journal of Memory and Language. 2022, 122: 104301. doi: 10.1016/j.jml.2021.104301
24. Rafay A, Hasan Y, Iqbal A. Recognition of Fingerprint Biometric System Access Control for Car Memory Settings Through Artificial Neural Networks. Advances in Information and Communication Networks. 2018, 385-397. doi: 10.1007/978-3-030-03405-4_26
25. Bhatt S, Jain A, Dev A. Continuous Speech Recognition Technologies—A Review. Lecture Notes in Mechanical Engineering. 2020, 85-94. doi: 10.1007/978-981-15-5776-7_8
26. Bird A. Multi-task dynamical systems: Customising time series models. Journal of Machine Learning Research 2022, 23, 1-52.
27. Ciaburro G, Iannace G, Ali M, et al. An artificial neural network approach to modelling absorbent asphalts acoustic properties. Journal of King Saud University - Engineering Sciences. 2021, 33(4): 213-220. doi: 10.1016/j.jksues.2020.07.002
28. Deekshitha G, Mary L. Multilingual spoken term detection: a review. International Journal of Speech Technology. 2020, 23(3): 653-667. doi: 10.1007/s10772-020-09732-9
29. Tran DC, Nguyen DL, Ha HS, et al. Speech Recognizing Comparisons Between Web Speech API and FPT.AI API. Proceedings of the 12th National Technical Seminar on Unmanned System Technology 2020. 2021, 853-865. doi: 10.1007/978-981-16-2406-3_64
30. Martins M, Mota D, Morgado F, et al. ImageAI: Comparison Study on Different Custom Image Recognition Algorithms. Trends and Applications in Information Systems and Technologies. 2021, 602-610. doi: 10.1007/978-3-030-72651-5_57
31. Xu H, Chen T, Gao D, et al. A Pruned Rnnlm Lattice-Rescoring Algorithm for Automatic Speech Recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. doi: 10.1109/icassp.2018.8461974
32. Březinová P. Computational Analysis and Synthesis of Song Lyrics. Ústav formální a aplikované lingvistiky; 2021.
33. Norouzian A, Mazoure B, Connolly D, et al. Exploring Attention Mechanism for Acoustic-based Classification of Speech Utterances into System-directed and Non-system-directed. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. doi: 10.1109/icassp.2019.8683565
34. Dighe P, Asaei A, Bourlard H. On quantifying the quality of acoustic models in hybrid DNN-HMM ASR. Speech Communication. 2020, 119: 24-35. doi: 10.1016/j.specom.2020.03.001
35. Gwilliams L, King JR, Marantz A, Poeppel D. Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content. Nat Commun 2022, 13, 6606. doi: 10.1038/s41467-022-34326-1
36. Yuan W, Eckart B, Kim K, et al. DeepGMR: Learning Latent Gaussian Mixture Models for Registration. Lecture Notes in Computer Science. 2020, 733-750. doi: 10.1007/978-3-030-58558-7_43
37. Karhila R, Smolander AR, Ylinen S, et al. Transparent Pronunciation Scoring Using Articulatorily Weighted Phoneme Edit Distance. Interspeech 2019. 2019. doi: 10.21437/interspeech.2019-1785
38. Kreyssig F. Deep learning for user simulation in a dialogue system. University of Cambridge; 2018.
39. Banerjee T, Rao Thurlapati N, Pavithra V, et al. Few-Shot learning for frame-Wise phoneme recognition: Adaptation of matching networks. 2021 29th European Signal Processing Conference (EUSIPCO). 2021. doi: 10.23919/eusipco54536.2021.9616234
40. Wang YH, Lee HY, Lee LS. Segmental Audio Word2Vec: Representing Utterances as Sequences of Vectors with Applications in Spoken Term Detection. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. doi: 10.1109/icassp.2018.8462002

41. Rebello S, Yu H, Ma L. An integrated approach for system functional reliability assessment using Dynamic Bayesian Network and Hidden Markov Model. Reliability Engineering & System Safety. 2018, 180: 124-135. doi: 10.1016/j.ress.2018.07.002
42. Li Y, Shao Y, Zhao Y. Construction of a General Lexical-Semantic Knowledge Graph. Chinese Lexical Semantics. 2021, 464-472. doi: 10.1007/978-3-030-81197-6_39
43. Mamyrbayev O, Alimhan K, Zhumazhanov B, et al. End-to-End Speech Recognition in Agglutinative Languages. Lecture Notes in Computer Science. 2020, 391-401. doi: 10.1007/978-3-030-42058-1_33
44. Forkel R, List JM, Greenhill SJ, et al. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. Scientific Data. 2018, 5(1). doi: 10.1038/sdata.2018.205
45. Ramanarayanan V, Tilsen S, Proctor M, et al. Analysis of speech production real-time MRI. Computer Speech & Language. 2018, 52: 1-22. doi: 10.1016/j.csl.2018.04.002
46. Wang S, Cao L, Wang Y, et al. A Survey on Session-based Recommender Systems. ACM Computing Surveys. 2021, 54(7): 1-38. doi: 10.1145/3465401
47. Wang W, Wang G, Bhatnagar A, et al. An Investigation of Phone-Based Subword Units for End-to-End Speech Recognition. Interspeech 2020. 2020. doi: 10.21437/interspeech.2020-1873
48. Li K, Xu H, Wang Y, et al. Recurrent Neural Network Language Model Adaptation for Conversational Speech Recognition. Interspeech 2018. 2018. doi: 10.21437/interspeech.2018-1413
49. Liu M, Ho S, Wang M, et al. Federated learning meets natural language processing: A survey. arXiv preprint arXiv:2107.12603
50. Doval Y, Gómez-Rodríguez C. Comparing neural- and N-gram-based language models for word segmentation. Journal of the Association for Information Science and Technology. 2018, 70(2): 187-197. doi: 10.1002/asi.24082
51. Erdmann A, Elsner M, Wu S, et al. The Paradigm Discovery Problem. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. doi: 10.18653/v1/2020.acl-main.695
52. Avasthi S, Chauhan R, Acharjya DP. Processing Large Text Corpus Using N-Gram Language Modeling and Smoothing. Proceedings of the Second International Conference on Information Management and Machine Intelligence. 2021, 21-32. doi: 10.1007/978-981-15-9689-6_3
53. Sterpu G, Harte N. Taris: An online speech recognition framework with sequence to sequence neural networks for both audio-only and audio-visual speech. Computer Speech & Language. 2022, 74: 101349. doi: 10.1016/j.csl.2022.101349
54. Raj PP, Reddy PA, Chandrachoodan N. Reduced Memory Viterbi Decoding for Hardware-accelerated Speech Recognition. ACM Transactions on Embedded Computing Systems. 2022, 21(3): 1-18. doi: 10.1145/3510028
55. Pan B, Yang Y, Zhao Z, et al. Bi-Decoder Augmented Network for Neural Machine Translation. Neurocomputing. 2020, 387: 188-194. doi: 10.1016/j.neucom.2020.01.003
56. Cui Y, Che W, Yang Z, et al. Interactive Gated Decoder for Machine Reading Comprehension. ACM Transactions on Asian and Low-Resource Language Information Processing. 2022, 21(4): 1-19. doi: 10.1145/3501399
57. Wang J, Shen J. Fast spectral analysis for approximate nearest neighbor search. Machine Learning. 2022, 111(6): 2297-2322. doi: 10.1007/s10994-021-06124-1
58. Gurunath Shivakumar P, Narayanan S. End-to-end neural systems for automatic children speech recognition: An empirical study. Computer Speech & Language. 2022, 72: 101289. doi: 10.1016/j.csl.2021.101289
59. Wang YY, Acero A, Chelba C. Is word error rate a good indicator for spoken language understanding accuracy. In: Proceedings of the 2003 IEEE Workshop on Automatic Speech Recognition and Understanding. pp. 577-582.
60. Subakan C, Ravanelli M, Cornell S, et al. Attention Is All You Need in Speech Separation. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021. doi: 10.1109/icassp39728.2021.9413901
61. Wang D, Wang X, Lv S. An Overview of End-to-End Automatic Speech Recognition. Symmetry. 2019, 11(8): 1018. doi: 10.3390/sym11081018
62. Furnon N, Serizel R, Essid S, et al. DNN-Based Mask Estimation for Distributed Speech Enhancement in Spatially Unconstrained Microphone Arrays. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021, 29: 2310-2323. doi: 10.1109/taslp.2021.3092838