International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Applications of Speech Recognition Using Machine Learning and Computer Vision

Nandini Arjangi

Student, Rajam, Vizianagaram, 532127, India.

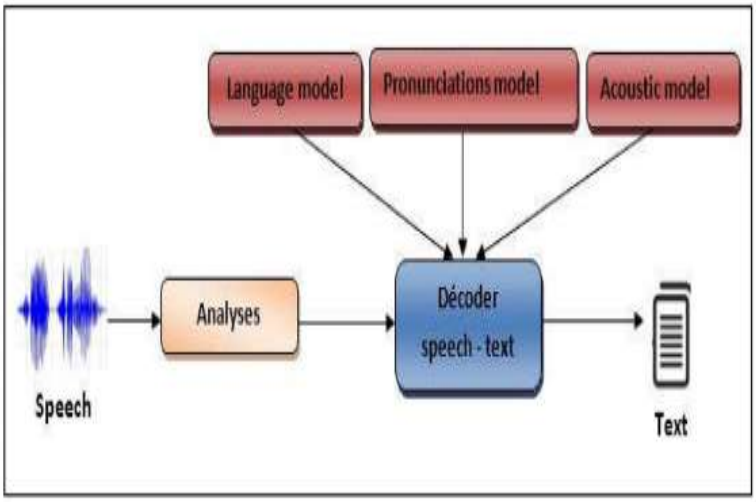
ABSTRACT

Speech is a straightforward and useful form of communication between people, but in the current world, human interaction goes beyond our interactions with the many pieces of technology we use on a daily basis. The computer is most important. Consequently, there are many ways for people and computers to communicate. This interaction takes place through interfaces, and the field known as "human computer interaction" studies it (HCI). This paper summarises the key definitions of Automatic Speech Recognition (ASR), a crucial field of artificial intelligence that should be taken into account during any pertinent research (Type of speech, vocabulary size... etc.). Some possible upgrades for future work are mentioned in the conclusion. Speech recognition plays a big part in the digital transformation. It is frequently used in many different.

Keywords: Predictive analysis, pandemic, linear regression, supervised learning, support vector regression.

1. Introduction

Speech Recognition simply is a process of converting spoken input to the text.The device you're speaking to creates a wave file of your words.The wave file is cleaned by removing background noise and normalizing volume .The resulting filtered wave form is then broken down into what are called phonemes.An example of speech recognition technology in use is speech-to-text platforms such as Speechmatics or Google's speech-to-text engine.Speech recognition software can translate spoken words into text using closed captions to enable a person with hearing loss to understand what others are saying. Speech recognition can also enable those with limited use of their hands to work with computers, using voice commands instead of typing.There are two types of speech recognition. One is called speaker–dependent and the other is speaker–independent.Speaker–dependent software is commonly used for dictation software, while speaker–independent software is more commonly found in telephone applications.



2. Literature Survey

In paper [1]. ***Chiu, C. C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., ... & Bacchiani, M. (2018, April). State-of-the-art speech recognition with sequence-to-sequence models. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)(pp. 4774-4778). IEEE.***

- Sequence-to-sequence models, which combine the many acoustic, pronunciation, and language models (AM, PM, LM) of a conventional ASR system into a single neural network, are becoming more and more popular in the automatic speech recognition (ASR) sector.
- merits: This project aims to explore potential structural and optimization improvements that will allow sequence-to-sequence models to significantly outperform a conventional ASR system on a voice search task.
- : Sequence-level loss functions, which are what attention-based systems strive for, have the following
- disadvantages: they are not directly correlated with the statistic that counts to us, namely word error rate.

In paper [2]. ***Kahn, J., Galibert, O., Quintard, L., Carré, M., Giraudel, A., Joly, P. (2012). A presentation of the REPERE challenge, Content Based Multimedia Indexing (CBMI), 2012 10th International Workshop on, IEEE, 1-6, 2012.***

- Finding individuals on video is a huge difficulty given that so much information comes from the Internet and from television. The REPERE Challenge aims to promote research on people recognition in multimodal situations. Without merging the details offered by the voice and the visual, it can be challenging to distinguish between the speakers and the persons in the video.
- Merits: The REPERE Challenge seeks to advance the creation of autonomous systems that can recognise individuals in a multimodal environment.
- By integrating the distinctive information from the voice and the video frames, the goal is to provide answers to those questions. Both supervised and unsupervised employees perform these tasks.
- Demerits: It's more challenging to find someone who just uses one type of transportation.

In paper [3].

2.3 Xu, H., Ding, S., & Watanabe, S. (2019, May). Improving end-to-end speech recognition with pronunciation-assisted sub-word modeling. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7110-7114). IEEE

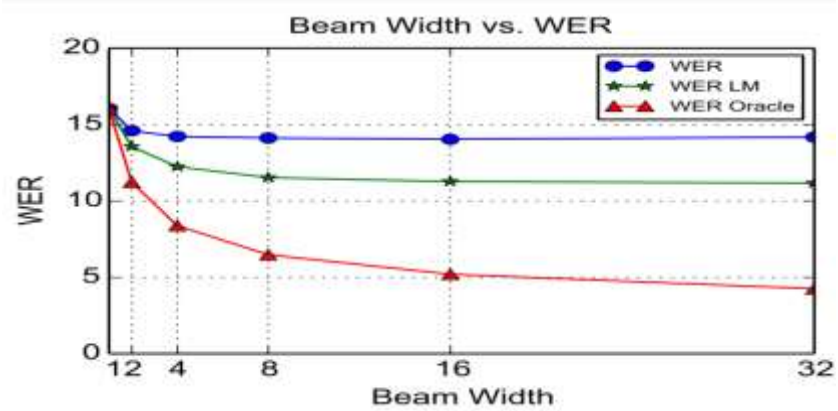
A collection of characters or subwords is how text is formally modelled in the majority of end-to-end speech recognition software. Current sub-word extraction methods solely consider character sequence frequencies, which occasionally results in unsatisfactory sub-word segmentation and potentially result in erroneous voice recognition results.

merits:

One of the different network typologies for end-to-end systems, the attention-based encoder-decoder mechanism, has demonstrated exceptional performance in a number of applications, including automatic speech recognition and neural machine translation.

Demerits: Due to the absence of a phonetic lexicon, the majority of end-to-end systems do not directly model words. Instead, they break down the output text sequence into smaller, usually character-sized, pieces.

In paper[4]:***Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016, March). Listen, attend and spell:A neural network for large vocabulary conversational speech recognition. In 2016 IEEE107 SEEU Review Volume 15 Issue 2International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4960-4964). IEEE.***



We present Listen, Attend and Spell (LAS), a neural network that learns to transform an audio sequence signal into a word sequence, one letter at a time, without the use of explicit language models, pronunciation models, HMMs, etc. LAS does not make any independent assumptions about how the probability distribution of the output character sequence will be made given the input auditory sequence.

merits:

The attention-based sequence-to-sequence learning paradigm is the foundation of this strategy [8, 9, 10, 11, 12, 13].

It consists of an encoder Recurrent Neural Network (RNN) called the listener and a decoder RNN called the speller. The listener is a pyramidal RNN that converts speech inputs into high level attributes.

Demerits: Rather than decoding with a language, we simply rescored the top 32 beams.

In paper [5]. ***Kubo, Y., & Bacchiani, M. (2020, May). Joint phoneme-grapheme model for end-to-end speech recognition.***

proposes methods to improve a commonly used end-to-end speech recognition model, Listen-Attend-Spell (LAS).

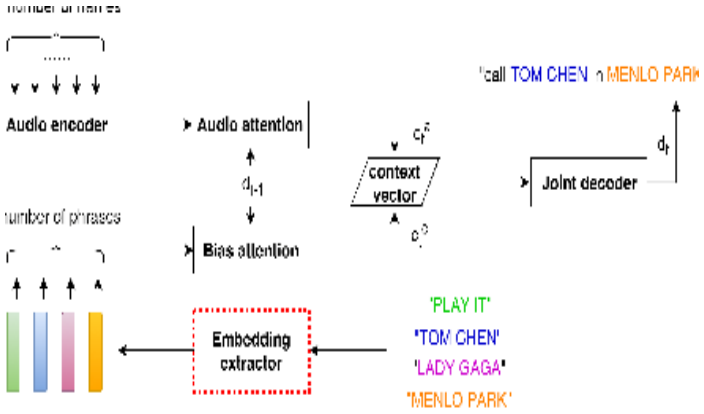


Figure 1: *The Framework of CLAS.*

Using phoneme identification as an auxiliary task can assist in directing the early stages of training to be more stable because phonemes are created to be an accurate description of the linguistic components of the speech signal.

We achieve additional benefits by developing a technique that can take advantage of label dependencies across many tasks, in addition to standard multi-task learning:

Performance of a conventional multi-task approach is compared to that of the joint model with iterative refinement.

Demerits: Drawbacks include the time-consuming and expensive construction procedures for conventional language and auditory models.

3.METHODOLOGY

(HMMs), Gaussian Mixture Models (GMMs), and several forms of Neural Networks (NNs), including Deep Learning, are used to combine the benefits of each technique to enhance speech recognition. One ASR paradigm that makes use of both the sequential modelling capabilities of HMMs and the strong representation learning power of DNNs is the combination of deep neural networks DNNs and HMMs (Yu and Deng,2015). The DNN-HMM hybrid system is the name of this paradigm. In this paradigm, DNNs are used to estimate observation probabilities while HMMs are used to describe the dynamics of the speech signal. The DNN trains each output neuron to estimate the posterior probability of the

Through the bottleneck, the internal layer builds a compressed continuous representation of the task-related information. To optimise the model parameters, an end-to-end system is trained with 40 mel-scale filter-bank features, delta and delta-delta features, and the AdaDelta algorithm (Zeiler, 2012). The DNN-HMM hybrid ASR system employs the convolutional neural network (CNN)-LSTM as an audio model (Tsoi, 2006). The softmax layer feeds the LSTM output. While the WFST-based decoder in the DNN-HMM hybrid ASR system (Hori et al., 2007). The hyper-parameters of end-to-end and DNN-HMM hybrid systems are the same in the joint DNN-HMM hybrid ASR systems, as shown in fig 3. The input features of the LSTM in the acoustic model are taken from the internal

In order to train on context-dependent (triphone) and context-independent (monophone) state space, GFCC and MFCC feature vectors (Wu and Cao, 2005) must first be retrieved.using the acoustic observations offered. Prior to training the triphone state, which aids in training the GMM-HMM model, the monophone state is trained. The SGMM-HMM technique makes additional use of this information state. The proposed system is assessed using Punjabi language with a medium vocabulary.

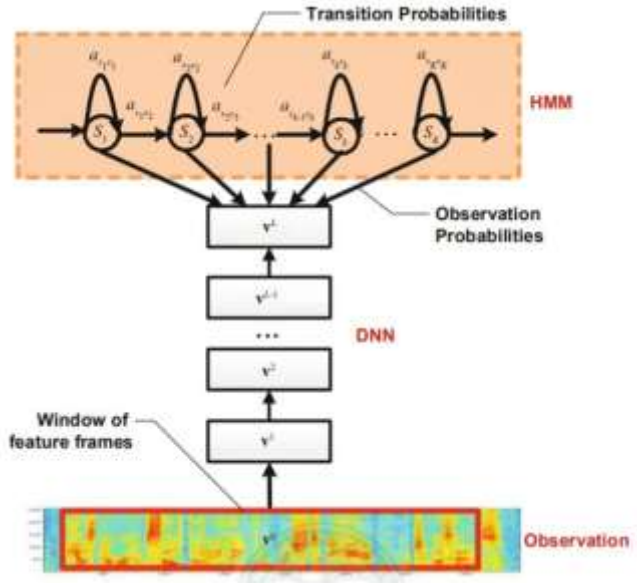
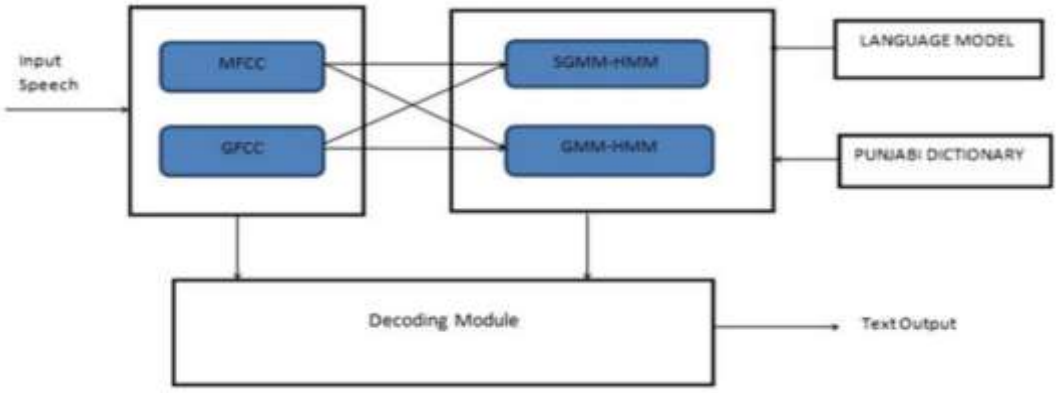


Fig 3.1: A ITS structure

3.2 Workflow of SGMM-HMM-based Punjabi ASR system

In order to train on context-dependent (triphone) and context-independent (monophone) state space, GFCC and MFCC feature vectors (Wu and Cao, 2005) must first be retrieved.



Using the acoustic observations offered. Prior to training the triphone state, which aids in training the GMM-HMM model, the monophone state is trained. The SGMM-HMM technique makes additional use of this information state. The proposed system is assessed using Punjabi language with a medium vocabulary.

4. Results and Discussion

Language model rescoring

We use the open-source end-to-end speech recognition tools ESPnet to carry out our experiment. Our baseline is the traditional character-based method, which employs location-based attention, bidirectional LSTMs with projection layers as the encoder, and an LSTM decoder with a CTC-weight of 0.5 during training. We publish the 1st pass numbers directly rather than performing language model rescoring in order to properly see the impact of sub-word techniques.

	Baseline	PASM	BPE
dev93	20.7	**18.5**	19.5
eval92	15.2	**14.3**	15.6

Num-BPEs	50	108	200	400
dev93	20.7	**19.5**	21.3	24.6
eval92	15.2	**15.6**	17.7	20.0

	Baseline	BPE	PASM
dev-clean	23.8	29.5	**21.4**
dev-other	52.8	53.1	**50.7**
test-clean	23.2	29.5	**21.3**
test-other	54.8	55.3	**52.8**

We also compare our systems with BPE baselines. The BPE follows the algorithmprocedure

5. Conclusion

For the purpose of recognising speech from sequence to sequence, we created an attention-based model. The model eliminates the need for a vocabulary or a separate text normalisation component by combining acoustic, pronunciation, and language models into a single neural network. For the purpose of enhancing the model, we investigated several structural and optimization mechanisms. WER increased by 11% as a result of structural improvements (WPM, MHA), 27.5 % as a result of optimization improvements (MWER, SS, LS, and synchronous training), and 3.4% as a result of language model rescoring. We achieve a WER of 5.6% when applied to a Google Voice Search task, whereas a hybrid HMM-LSTM system yields a WER of 6.7%.Implementing this model with a streaming attention-based model, like Neural Transducer, is thus a crucial next step.

REFERENCES

1. 1.Chiu, C. C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., ... & Bacchiani, M. (2018, April). State-of-the-art speech recognition with sequence-to-sequence models. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)(pp. 4774-4778). IEEE.
2. 2.Kahn, J., Galibert, O., Quintard, L., Carré, M., Giraudel, A., Joly, P. (2012). A presentation of the REPERE challenge, Content_x0002_Based Multimedia Indexing (CBMI), 2012 10th International Workshop on, IEEE, 1-6, 2012.
3. Xu, H., Ding, S., & Watanabe, S. (2019, May). Improving end-to-end speech recognition with pronunciation-assisted sub-word modeling. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7110-7114). IEEE
4. Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016, March). Listen, attend and spell:A neural network for large vocabulary conversational speech recognition. In 2016 IEEE107 SEEU Review Volume 15 Issue 2International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4960-4964). IEEE.
5. 5..Kubo, Y., & Bacchiani, M. (2020, May). Joint phoneme-grapheme model for end-to-end speech recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6119-6123). IEEE.