SURVEY ON SPEECH-TO-TEXT CONVERSION

[1]Dhanush Kumar S, [2]Lavanya S, [3]Madhumita G,[4]Mercy Rajaselvi V
[1], [2], [3]-Student,[4] –Associate Professor
[1]Computer Science and Engineering,
[1]Easwari Engineering College, Chennai, India

Abstract : We are trying to use the speech to text engines that have been developed to create a new chatbot for the blind students to complete their examinations without depending on the volunteers who come forward to help them. This chatbot will work offline without any internet connection with the help of CMU Sphinx toolkit which runs on deep neural networks. This chatbot will change the examination scheme of the blind students community and help them excel in their examinations.

IndexTerms - Deep Neural Networks, HMM and Deep Learning.

1.INTRODUCTION

The main aim of our project is to automate the process of writing examinations by the blind students with the help of scribes. This project reduces the human resources required to help the blind students to write their examinations. We also have specific requirement to record the speech continuously for about 5 mins because of the type of answers that are required for each answer. The following survey contains the summary of the previously implemented models and systems to achieve speech to text recognition in high accuracy. We are also in search of a speech recognition engine with can work without internet.

The domain we've chosen which is artificial intelligence is the future of computing world and we're trying to impart automation to world in each and every way possible. Our work as computer engineers are only complete when we've created a system which acts and thinks like a human.

Several machine learning algorithms and models are being invented everyday. These machine learning concepts and algorithms take us one step closer in designing a complete intelligence system that can act and think like humans.Yet we have a lot more to go in creating the perfect AI and until then we have to move step by step in everything we face or else we won't be able to control the systems that we create.

Even in the systems that we create we have a lot of anomalies that we haven't looked into yet and hence we have taken up this survey so as to find out how far the technology has grown in speech recognition.

Another concept which deeply dealt with is Deep-learning networks. They are distinguished from the more commonplace single-hidden-layer neural networks by their depth; that is, the number of node layers through which data passes in a multistep process of pattern recognition.Earlier versions of neural networks such as the first perceptrons were shallow, composed of one input and one output layer, and at most one hidden layer in between. More than three layers (including input and output) qualifies as "deep" learning. So deep is a strictly defined, technical term that means more than one hidden layer.In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer 's output. The further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the previous layer. This is known as feature hierarchy, and it is a hierarchy of increasing complexity and abstraction. It makes deep-learning networks capable of handling very large, high-dimensional data sets with billions of parameters that pass through nonlinear functions.Above all, these nets are capable of discovering latent structures within unlabeled, unstructured data, which is the vast majority of data in the world. Another word for unstructured data is raw media;i.e. pictures, texts, video and audio recordings. Therefore, one of the problems deep learning solves best is in processing and clustering the world's raw, unlabelled media, discerning similarities and anomalies in data that no human has organized in a relational database or ever put a name to. We are trying to use the speech to text engines that have been developed to create a new chatbot for the blind students to complete their examinations without depending on the volunteers who come forward to help them. This chatbot will work offline without any internet connection with the help of CMU Sphinx toolkit which runs on deep neural networks. This chatbot will change the examination scheme of the blind students community and help them excel in their examinations. We are going to further induce the concepts of HMM-Hidden Markov Model into our system for acoustic speech recognition. Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (i.e. hidden) states. The hidden Markov model can be represented as the simplest dynamic Bayesian network. The mathematics behind the HMM were developed by L. E. Baum and co-workers. HMM is closely related to an earlier work on the optimal nonlinear filtering problem by Ruslan L. Stratonovich, who was the first to describe the forward-backward procedure. In simpler Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters, while in the hidden Markov model, the state is not directly visible, but the output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore, the sequence of tokens generated by an HMM gives some information about the sequence of states. The adjective hidden refers to the state sequence through which the model passes, not to the parameters of the model; the model is still referred to as a hidden Markov model even if these parameters are known exactly. Hidden Markov models are especially known for their application in reinforcement learning and temporal pattern recognition such as speech, handwriting, gesture recognition, part-of-speech tagging, musical score following, partial discharges and bioinformatics.

2. LITERATURE SURVEY

We have surveyed the following papers which are based on the concepts like HMM(Hidden Markov Model), Automatic speech recognition system(ASR),MFCC(Mel Frequency Cepstral Co-efficient) and DNN(Deep neutral network).Among all such we have found HMM and DNN to be the most commonly used concepts for Speech Recognition.

Andrew L.Maas et.al. [1] have proposed building DNN acoustic models for large vocabulary speech recognition. Understanding architectural choices for deep neural networks (DNNs) is crucial to improving state-of-the-art speech recognition systems. We investigate which aspects of DNN acoustic model design are most important for speech recognition system performance, focusing on feed-forward networks. We study the effects of parameters like model size (number of layers, total parameters), architecture (convolutional networks), and

training details (loss function, regularization methods) on DNN classifier performance and speech recognizer word error rates. On the Switchboard benchmark corpus we compare standard DNNs to convolutional networks, and present the first experiments using locally-connected, untied neural networks for acoustic modeling. Using a much larger 2100-hour training corpus (combining Switchboard and Fisher) we examine the performance of very large DNN models – with up to ten times more parameters than those typically used in speech recognition systems. The results suggest that a relatively simple DNN architecture and optimization technique give strong performance, and we offer intuitions about architectural choices like network depth over breadth. Our findings extend previous works to help establish a set of best practices for building DNN hybrid speech recognition systems and constitute an important first step toward analyzing more complex recurrent, sequence-discriminative, and HMM-free architectures.

In this paper they offer a large empirical investigation of DNN performance on two LVCSR tasks in an attempt to establish a set of best practices for building DNN acoustic models. To further make careful comparison possible they restrict our focus to feed-forward DNN models trained with cross entropy, because they are the foundation of neural network acoustic models. A deep understanding of these feedforward DNNs can help guide the emerging work utilizing a wide variety of new architectures based on recurrent neural networks (Graves et al., 2013; Li and Wu, 2015; Sak et al., 2014; Vinyals et al., 2012; Weng et al., 2014), sequence-discriminative training (Kingsbury et al., 2012; Vesel et al., 2013; Wiesler et al., 2015), and HMM-free neural network approaches (Graves and Jaitly, 2014; Maas et al., 2015). Further, we seek to understand which aspects of DNN training have the most impact on downstream task performance. This knowledge can guide rapid development of DNN acoustic models for new speech corpora, languages, computational constraints, and language understanding task variants.

Their work systematically explores several dimensions of DNN design. We study the role of model size, and the interaction between model size and the number of network layers, addressing questions like how many layers are useful for conversational LVCSR, or for a given number of parameters, is it better to have wider or deeper networks? We study network architecture: can convolutional networks improve performance and obviate the need for complex feature extraction? And we examine a number of other parameters of DNN training, including different training loss functions and different techniques for reducing over-fitting

All models use hidden units with the rectified linear nonlinearity. For optimization, we use Nesterov's accelerated gradient with a smooth initial momentum schedule which we clamp to a maximum of 0.95 (Sutskever et al., 2013). The stochastic updates are on mini-batches of 512 examples. After each epoch, or full pass through the data, we anneal the learning rate by half. Training is stopped after improvement in the cross entropy objective evaluated on held out development set falls below a small tolerance threshold.

In order to efficiently train models of the size mentioned above, we distribute the model and computation across several GPUs using the distributed neural network infrastructure proposed by Coates et al. (2013). The GPU cluster and distributed training software is capable of training up to 10 billion parameter DNNs. The ttenetion is restricted to models in the 30M–200M parameter range. In preliminary experiments we found that DNNs with 200M parameters are representative of DNNs with over one billion parameters for this task. We train models for this paper in a model- parallel fashion by distributing the parameters across four GPUs. A single pass through the training set for a 200M parameter DNN takes approximately 1.5 days.

D. Đorđe T. Grozdića et.al. [2] have proposed audio feature extraction by deep denoising autoencoder. For the audio feature extraction, we utilized a deep denoising autoencoder. Eleven consecutive frames of audio features are used as the short-time spectral representation of speech signal inputs. To generate audio input feature sequences, partially deteriorated sound data are artificially generated by superimposing several strengths of Gaussian noises to original sound signals. In addition to the original clean sound data, we prepared six different deteriorated sound data; the signal-to-noise-ratio (SNR) was from 30 to −20 dB at 10 dB intervals. Utilizing sound feature extraction tools, the following types of sound features are generated from eight variations of original clean and deteriorated sound signals. HCopy command of the hidden Markov model toolkit (HTK) is utilized to extract 39 dimensions of MFCCs. Auditory Toolbox Is utilized to extract 40 dimensions of log mel-scale filter bank (LMFB). Finally, the deep denoising autoencoder is trained to reconstruct clean audio features from deteriorated features by preparing the deteriorated dataset as input and the corresponding clean dataset as the target of the network. Among a 400-word dataset, sound signals from 360 training words (2.76 × 105 samples) and the remaining 40 test words

(2.91 × 104 samples) from six speakers are used to train and evaluate the network, respectively.

The denoised audio features are generated by recording the neuronal outputs of the deep autoencoder when 11 frames of audio features are provided as input. To compare the denoising performance relative to the construction of the network, several different network architectures are compared.

The acquired audio features are evaluated by conducting an isolated word recognition experiment utilizing a single- stream HMM. To recognize words from the audio features acquired by the deep denoising autoencoder, monophone HMMs with 8, 16, and 32 GMM components are utilized. While training is conducted with 360 train words, evaluation is conducted with 40 test words from the same speaker, yielding a closed-speaker and open-vocabulary evaluation. To enable comparison with the baseline performance, word recognition rates utilizing the original audio features are also prepared. To evaluate the robustness of our proposed mechanism against the degradation of audio input, partially deteriorated sound data were artificially generated by superimposing several strengths of Gaussian noises to original sound signals. In addition to the original clean sound data, we prepared 11 different deteriorated sound data such that the SNR was 30 dB to −20 dB at 5 dB intervals.

Gia-Nhu Nguyen et.al. [3] have proposed using exemplar based voice conversion to reduce over smoothness in HMM based speech synthesis . Speech synthesis (SS) is the artificial production of human speech, which is the core part of text-to-speech (TTS) systems that convert text content in a specific language into a speech waveform.

Among many kinds of SS that have been proposed, the state-of-the-art one is HMM-based SS (HMMSS). In this approach, spectral and prosodic features of speech are modeled and generated in an unified statistical framework using HMMs. HMMSS has many advantages that have been shown in the literature, such as the high intelligibility of synthesized speech, a small footprint, a low computational load, and the flexibility to change the voice characteristics. Although HMMSS has many advantages, the quality of its synthesized speech is still far from natural, which is mainly due to two reasons: buzziness and over-smoothness in synthesized speech. The former is a common issue with speech coding, which has recently been significantly improved, while the latter is caused by averaging in the statistical processing in HMMSS, which is still a remaining problem of HMMSS at present. In HMMSS, the structure of the estimated spectrum corresponds to the average of different speech spectra in the training database due to the use of the mean vector. In this case, the spectrum estimated by HMMs is an average approximation of all corresponding speech spectra in the training database. The detailed structure in the original speech may be missing in this kind of approximation. This characteristic in speech synthesized by HMMSS can be considered too medial or over-smooth in

synthesized speech. When synthesized speech is over-smooth, it sounds "muffled" and far from natural. Too smooth or too stable speech with slow movements cannot be efficient to represent some kinds of emotional speech with high movements of the tongue tip.

Although both the spectral and prosodic trajectories generated by HMMSS are over-smooth, the effect of over-smoothness in a spectral sequence is more serious since the structures of spectral features are more complex there have been many studies attempting to solve over-smoothness in HMMSS. However, these methods cause another problem with over-training due to the increased number of model parameters.Increase the complexity of HMMs and are not convenient in practical synthesis systems.

In this paper, a hybrid TTS between HMMSS and bilinear model NMF has been proposed to reduce temporal over-smoothness of HMMSS with significant improvements compared with the method in NMF is used to implement exemplar-based VC and is also used in the hybrid TTS proposed in this paper between HMMSS and exemplar-based VC.

The results indicate that Speech synthesized by HMMSS is most over-smooth; Both HMMSS + GV and HMMSS + VC can efficiently reduce the over-smoothness compared with HMMSS. However, HMMSS + VC is less over-smooth (more rough) and closer to the original speech; HMMSS + GV + VC reduce the over-smoothness compared with HMMSS + GV but do not reduce the over-smoothness compared with HMMSS + VC.

	Pairs: v1 and v2	v1%	v2 %
	HMMSS and HMMSS + GV	36.67	63.33
2	HMMSS and HMMSS + VC	31.33	68.67
3	HMMSS + GV and HMMSS + GV + VC	50.67	49.33
4	HMMSS + VC and HMMSS + GV + VC	51.33	48.67
	HMMSS + GV and HMMSS + VC	46.67	53.33

	Mel-cepstral of spectral magnitude (dB) distortion (MCD) (dB)	Average standard deviation
HMMSS	4.26	6.38
HMMSS + GV	7.43	4.76
HMMSS + VC	7.96	4.31
HMMSS + GV + VC	7.55	4.62

Table 2 shows the results of the subjective test for four pair choice. These results indicate that:

In pair 1, the preference for HMMSS + GV over HMMSS is clear with 63.33%.

In pair 2, the preference for HMMSS + VC over HMMSS is clear with 68.67%.

In pairs 3 and 4, there is no clear preference of HMMSS + GV + VC over HMM + GV or HMM VC. Therefore, there are no clear reasons to use GV and the VC together.

Hendrik Meutzner et.al. [4] have proposed using HTK for improved audio CAPTCHAs based on audio perception and language processing. We have only taken the automatic speech recognition part from this journal which is used to get past the captchas in the websites. The authors have trained acoustic models to teach the system to understand the speech patterns. They used Hidden markov model toolkit (HTK) to train the acoustic models. For each state in the Hidden markov model, the output probability looks like

$b(om) = P(om|qm = i)$ $i = 1, 2,..., Q$,

where Q is the maximum number of states in the model and om represents the feature vector, referred to as the observation hereinafter. For training the speech recognizer, the goal is to estimate the set of model parameters λ. This can be achieved by optimizing a maximum likelihood criterion

$\lambda ML = \arg\max \lambda \{P(o1 \ldots oT |\lambda)\}$

After training, the individual segment models are combined into a larger compound HMM to represent a user-defined grammar. Then, the sequence of observed features o1 ... oT can be decoded by searching for the optimal state sequence q∗ 1 ... q∗ T through the compound HMM

[q1 ... qT] ∗ = arg max q 1...q T {P(q 1 ... q T |o1 ... oT ,λ)},

Results achieved:

		Word Accuracy [%]				Sentence Accuracy [%]			
		$\mathcal{D}=0$		$\mathcal{D}=1$		$\mathcal{D}=0$		$\mathcal{D}=1$	
Features	N_{Init} [%]	Direct	Rescore	Direct	Rescore	Direct	Rescore	Direct	Rescore
MFCC	25	74.37	75.32	77.97	78.85	11.11	14.14	19.19	21.72
	50	73.48	74.56	77.40	78.35	10.10	11.62	18.18	23.23
	75	74.94	75.51	78.60	79.04	14.14	16.67	22.22	23.23
	100	73.93	74.94	77.53	78.54	12.12	14.14	21.21	21.72
PLP	25	78.03	**79.55**	81.31	**82.64**	16.67	19.19	26.77	**29.80**
	50	**78.98**	79.29	**82.13**	82.45	**19.19**	19.19	**28.28**	29.29
	75	78.35	78.91	81.82	82.26	13.64	17.68	23.74	27.78
	100	77.97	78.79	81.31	81.94	18.18	**21.72**	26.26	**29.80**

Himangshu Sarma et.al. [5] have proposed a speech recognition systems for Assamese Language using HTK. The authors of this paper were building the entire corpus data by themselves so that the model they were trying to achieve would come out perfect. The speech database along with the transcripted files were fed into the system to recognize the speech patterns. The system was designed to impose the speech rules on the transcriptions and also on the incoming database. The recognizer was trained with all these speech data and the transcripted files to help it understand the assamese syllables well.

The achieved accuracy was very low for most of the alphabets and only a few letters had high accuracy. The lowest accuracy that was recorded being 0% for the pronunciation 'ou' and the highest accuracy that was recorded was 91.7% for the letter 'n'.

Kuniaki Noda et.al. [6] have proposed using deep learning for Audio-visual speech recognition system. The cautious selection of sensory features is very crucial for high performance and accuracy of the speech recognition system. This approach introduces a connectionist-hidden Markov Model (HMM) system for noisy robust Audio Visual speech recognition system(AVSR). By preparing the training data for the network with pairs of consecutive multiple steps of deteriorated audio features and the corresponding clean features, the network is trained to output denoised audio features from the corresponding features deteriorated by noise. Second, a convolutional neural network (CNN) is utilized to extract visual features from raw mouth area images. By preparing the training data for the CNN as pairs of raw images and the corresponding phoneme label outputs, the network is trained to predict phoneme labels from the corresponding mouth area input images. Finally, a multi-stream HMM (MSHMM) is applied for integrating the acquired audio and visual HMMs independently trained with the respective features. By comparing the cases when normal and denoised mel-frequency cepstral coefficients (MFCCs) are utilized as audio features to the HMM, our uni-modal isolated word recognition results demonstrate that approximately 65 % word recognition rate gain is attained with denoised MFCCs under 10 dB signal- to-noise-ratio (SNR) for the audio signal input. Moreover, our multimodal isolated word recognition results utilizing MSHMM with denoised MFCCs and acquired visual features demonstrate that an additional word recognition rate gain is attained for the SNR conditions below 10 dB. However, advances in deep learning research have led to recent breakthroughs in unsupervised audio feature extraction methods and exceptional recognition performance improvements. Advances in novel machine learning algorithms, improved availability of computational resources, and the development of large databases have led to self-organization of robust audio features by efficient training of large-scale DNNs with large-scale datasets.

One of the most successful applications of DNNs to ASR is the deep neural network hidden Markov model (DNN- HMM) , which replaces the conventional Gaussian mixture model (GMM) with a DNN to represent the direct projection between HMM states and corresponding acoustic feature inputs. The idea of utilizing a neural network to replace a GMM and construct a hybrid model that combines a multilayer perceptron and HMMs was originally proposed decades ago. However, owing to limited computational resources, large and deep models were not experimented with in the past, which led to hybrid systems that could not outperform GMM-HMM systems. A Japanese audiovisual dataset was used for the evaluation of the proposed models. In the dataset, speech data from six males (400 words: 216 phonetically-balanced words and 184 important words from the ATR speech database) were used. In total, 24000 word recordings were prepared (one set of words per speaker; approximately 1 h of speech in total). The audio-visual synchronous recording environment is shown in Fig. 1. Audio data was recorded with a 16 kHz sampling rate, 16-bit depth, and a single channel. To train the acoustic model utilized for the assignment of phoneme labels to image sequences, we extracted 39 dimensions of audio features, composed of 13 MFCCs and their first and second temporal derivatives. To synchronize the acquired features between audio and video, MFCCs were sampled at 100 Hz. Visual data was a full-frontal 640 × 480 pixel 8-bit monochrome facial view recorded at 100 Hz. For visual model training and evaluation, we prepared a trimmed dataset composed of multiple image resolutions by manually cropping 128 × 128 pixels of the mouth area from the original data and resizing the cropped data to 64 × 64, 32 × 32, and 16 × 16 pixels.

MODEL:

IN	OUT	LAYERS
429	429	300-150-80-40-80-150-300
429	39	300-150-80
429	429	300-300-300-300-300-300-300
429	429	300-300-300-300-300
429	429	300-300-300
429	429	300

Musab T. S. Al-Kaltakchi et.al. [7] have proposed a speaker identification system evaluation with and without fusion using three databases in the presence of noise and handset effects. Speaker identification is one important application of biometrics and forensics to identify speakers based on their unique voice pattern. Feature extraction within speaker identification should be less influenced by noise or the person's health. However, to improve the speaker identification accuracy (SIA), Mel frequency cepstral coefficients (MFCC) features were fused with inverse MFCC features (IMFCC), but the approach was limited by the number of GMM components.

However, only a limited number of studies have involved a handset, AWGN, and NSN types in conjunction with fusion strategies.This work extends the previous work with four combinations of features and their score fusion methods for the original recordings; and with AWGN, and three types of NSN: street traffic, bus interior and crowd talk, with and without the G.712 type handset at 16 kHz, to provide a wide range of environmental noise conditions. The authors emphasize that, although the GMM- UBM approach is well established, no previous study has comprehensively considered three databases, one of which only appeared in 2016, nor the effect of such a wide range of NSN and handset effects

In the work, to mimic human ear perception, MFCC features are used and combined with the corresponding power normalized cepstral coefficient (PNCC) features

Three methods to form a late fusion score were employed : weighted sum, maximum, and mean fusion. Combined normalization methods were employed to produce normalized MFCC features (FWMFCC and CMVN-MFCC). Likewise, normalized methods were used to form PNCC features (FWPNCC and CMVNPNCC). Four sets of score vectors could therefore be calculated and are denoted as:

f_1 = feature warping MFCC scores vector (FWMFCC),
f = CMVN MFCC scores vector,
g_1 = feature warping PNCC scores vector (FWPNCC) and
g_2 = CMVN PNCC scores vector.

On the basis of the evaluations of three databases without the noise and handset conditions, the best speaker identification method for all three databases used was weighted sum fusion.

Based on the three databases without the noise and handset conditions, the order for best SIA was NIST2008, TIMIT, SITW with 95.83, 95, and 82.5%, respectively, at mixture sizes 64, 512, and also 512, respectively. These SIAs were achieved by using weighted sum fusion with 90% from FWMFCC features and 10% from the corresponding CMVNPNCC features for both the TIMIT and NIST 2008 database. On the other hand, in the SITW database, 70% from FWMFCC features was fused with 30% from the corresponding CMVNPNCC features. The weighting should therefore be chosen as a function of the fidelity of the speech recordings.

On the basis of the results in this paper, the evaluations in noisy conditions suggest that mean fusion of four combinations of two types of features from (FWMFCC, CMVNMFCC, FWPNCC, and CMVNPNCC) is the most robust method for a practical speaker identification system, but there is not a consistent best pairing.

Tan Lee et.al. [8] have proposed using tone information in continuous speech recognition.Automatic speech recognition and voice/ unvoiced boundary segmentation methods were used to recognize the cantonese dialects and speech patterns. A different type of hidden markov model was used here and that is context dependent tone modeling.

By using Utterance wide normalization with context independent models they achieved an accuracy of 54% and with context dependent models they achieved an 60% accuracy.

Tejas Godambe et.al [9] have developed a unit selection voice given audio without corresponding text. Unit selection speech synthesis is one of the techniques for synthesizing speech, where appropriate units from a database of natural speech are selected and concatenated. Unit selection synthesis can produce natural-sounding and expressive speech output given a large amount of data containing various prosodic and spectral characteristics. As a result, it is used in several commercial text-to-speech (TTS) applications today.

Building a new general- purpose (non-limited domain) unit selection voice in a new language from scratch includes a huge overhead of data preparation, which includes preparing phonetically balanced sentences, recording them from a professional speaker in various speaking styles and emotions in a noise-free environment, and manually segmenting or correcting the automatic segmentation errors. All of it is time consuming, laborious, and expensive, and it restricts rapid building of synthetic voices. A free database such as CMU ARCTIC has largely helped to rapidly build synthetic voices in the English language. But CMU ARCTIC is a small database, contains only a few speakers data, and is not prosodically rich (contains short declarative utterances only).

Now, the questions to be asked are whether we can readily use such data to build expressive unit selection synthetic voices and will the synthesis be good? In this paper, there is an attempt to answer these questions. There are a few problems related to it.
The audio files are generally long and audio-text alignment becomes memory intensive; Precise corresponding transcriptions are unavailable; Often, no transcriptions are available, and manually transcribing the data from scratch or even correcting the imprecise transcriptions is laborious, time consuming, and expensive; The audio may contain bad acoustic (poorly articulated, dis-fluent, unintelligible, inaudible, clipped, noisy) regions as the audio is not particularly recorded for building TTS systems; and if we obtain automatic transcripts using a speech recognition system, the transcripts will not be error free.

First, the ASR system accepts the audio data and produces corresponding labels. Then, data pruning using confidence measures takes place. The pruned audio and label data form the unit inventory for the TTS system. During synthesis time, the TTS system accepts normalized text, takes into account the duration and phrase break information predicted by a statistical parametric speech synthesizer trained using the same audio data and hypothesized transcriptions, and chooses an appropriate sequence of units that minimizes the total of the target and concatenation costs. The output of the TTS system is an audio file.

Decoding of the audiobooks was done in two passes. In the first pass, lattices containing competing alternative hypothesis were generated, while in the second pass, Viterbi decoding was applied to find the 1-best hypothesis. While decoding the Olive and lecture data with the ASR systems trained on themselves, the same 3-gram language model was used for both lattice generation and 1-best Viterbi decoding.

The results are that voices built using an audiobook seem to be more natural than those built using lecture speech.The MOS is almost the same for the first and second rows except the case of the last column where a noticeable improvement is observed in MOS.The MOS decreases as we move down rows as it becomes difficult to find units having a duration close to predicted duration and which can also maintain continuity in terms of energy, F_0, and MFCCs.

Yochay R. Yeminy et.al. [10] have proposed using diffusion based hidden markov models for single microphone speech recognition. Single-channel speech separation (SCSS) is one of the most challenging tasks in speech processing, where the aim is to unmix two or more concurrently speaking subjects, whose audio mixture is acquired by a single micro-phone. The goal is therefore to decompose the single input signal into multiple output channels, each dominated by a single speaker. The core obstacle in such tasks is the lack of spatial information, and the common statistical characteristics of the mixed signals.

First, we derive a novel non-iterative speech separation approach based on the diffusion framework, to compute HMM and FHMM models

Second, by analysing the asymptotics of Markov random walks the proposed scheme allows to directly estimate the states of the HMMs and FHMMs, without having to assume any underlying observation model, nor to apply EM-based iterative training. Hence, the estimation of the Markov states and transitions is decoupled from the estimation of the emission p.d.f.s, and their corresponding parametric model. Thus, the authors propose two FHMM-based approaches that estimate the underlying HMM in the diffusion domain. The first, provides a direct extension of the FHMM, where the underlying HMMs are computed in the diffusion domain, and the Gaussian observation models utilize the log-max approach applied in the original domain. This approach denotes a hybrid FHMM (HFHMM), as it utilizes both the temporal and diffusion domains. In the second approach, denoted dual FHMM (DFHMM), the emission p.d.f.s are estimated in the diffusion domain, without assuming an explicit emission p.d.f. model. The underlying HMMs are computed similarly to the HFHMM approach.

Last,an approach to utilize the diffusion embedding as a nonlinear projection of the input mixture onto the manifolds spanning each of the speakers was discussed. Thus, the aim is to utilize the diffusion embedding as a manifold-adaptive projection operator, where the states of each speaker are detected by an FHMM in its manifold. The HFHMM is experimentally shown to compare with previous results, and is shown to outperform DFHMM. The latter requires low computational cost, and can be applied alongside other approaches in the low-dimensional space to further reduce the computational complexity.

The proposed system involves training an HMM model per speaker using the diffusion framework. The evaluation starts with the performance of the HFHMM scheme. The results for the male-female case is given. It indicates that for the male speaker, 30 dimensions yield the best score, whereas for the female speaker D = 50 is better than D = 30 by 0.7 dB. For the same gender mixture a similar trend is observed. The results for the training procedure that incorporates the Nyström extension are less satisfying. Consequently, they are not extensively presented due to space limitations. For the male-female mixtures, the results related to each gender separately are reported, and for male-male and female-female pairs, both speakers are extracted and the results are averaged. It follows that the HFHMM-S approach outper-forms the HFHMM-H formulation in both the SDR and SAR figures-of-merit for most mixtures, although a degradation in the SIR is encountered. The results indicate a performance gap between the HFHMM-H-N, HFHMM-S-N and the HFHMM-H-E, HFHMM-S-E, in favor of the latter. Consequently, we conclude that the Nyström extension leads to performance degradation. The results of applying a soft mask to the male-female mixture are similar to those of the HFHMM. As expected, the SAR and SDR measures indicate that the DFHMM-S yields lower distortion levels as compared with the DFHMM-H scheme. However, the SIR measure deteriorates. This can be attributed to the higher level of residual interference, being a consequence of the softer mask.

Serial no	Paper	Technique	Result	Issues
1.	Building DNN acoustic models for large vocabulary speech recognition-Andrew L. Maas, Peng Qi, Ziang Xie, Awni Y. Hannun, Christopher T. Lengerich, Daniel Jurafsky, Andrew Y. Ng	DNN(Deeply-neutral network)	The DLUNN architecture proved to be more efficient and can hide more weights irrespective of the number of features.	The use of normal DNN was an issue in the first which was solved with the replacement of DLUNN.
2.	Whispered speech recognition using deep denoising autoencoder - D. Đorđe T. Grozdića, Slobodan T.Jovičića,Miško Subotić	Hidden Markov model(HMM)	The ASR could recognize the transcriptions with 30db to -20db at 5db intervals.	The transcription files were not audible enough for the tone recognizer to detect the patterns in the voice.
3.	Reducing over smoothness in HMM based speech synthesis using exemplar based voice conversion- Gia-Nhu Nguyen and Trung-Nghia Phung	A hybrid TTS between HMMSS and bilinear model NMF	The preference for HMMSS + GV over HMMSS is clear with 63.33%,HMMSS + VC over HMMSS is clear with 68.67% and	The subjective preference for HMMSS + VC over HMMSS + GV is not very clear

			HMMSS + VC over HMMSS + GV is 53.33%	
4.	Towards improved audio CAPTCHAs based on audio perception and language processing- Hendrik Meutzner, Santhosh Gupta, Viet-Hung Nguyen, Thorsten Holz and Dorothea Kolossa	Hidden Markov model(HMM)	Sentence accuracy 29.80 and 21.72 for PLP and MFCC respectively	The recognizer was hard to train for each full transcription file and the transcription files were huge for each instance.
5.	Development and analysis of speech recognition systems for Assamese Language using HTK- Himangshu Sarma, Navanath Saharia and Utpal Sarma	Hidden Markov model(HMM)	The lowest accuracy that was recorded being 0% for the pronunciation 'ou' and the highest accuracy that was recorded was 91.7% for the letter 'n'	The accuracy achieved from the transcription file was much less than the syllabification which in real time was the biggest issues because transcriptions contained the voice signals.
6.	Audio-visual speech recognition using deep learning-Kuniaki Noda,Yuki Yamaguchi,Kazuhiro Nakadai,Hiroshi G. Okuno,Tetsuya Ogata	HMM, CNN and MSHMM	It is found that we can improve the efficency of the system by using MSHMM for an AVSR task.	Over-fitting may occur in the validation process and it is inevitable due to the requirement of huge datasets.
7.	Evaluation of a speaker identification system with and without fusion using three databases in the presence of noise and handset effects- Musab T. S. Al-Kaltakchi, Wai L. Woo, Satnam Dlay and Jonathon A. Chambers	FWMFCC, CMVNMFCC, FWPNCC, and CMVNPNCC	The order for best SIA was NIST2008, TIMIT, SITW with 95.83, 95, and 82.5%, respectively, at mixture sizes 64, 512, and also 512, respectively	The relative sensitivities of the various methods to the environments
8.	Using tone information in cantonese continuous speech recognition- Tan Lee, Wai Lau, Y. W. Wong and P. C. Ching	A different type of hidden markov model was used that is based on context dependent tone modelling	They achieved an accuracy of 54% and with context dependent models they achieved an 60% accuracy	Incorrect syllable detections was an huge issue and the tone independent inviduals voices cant be recognized properly.
9.	Developing a unit selection voice given audio without corresponding text-	Data pruning, Lattice generation and 1-best Viterbi	The WER further reduces when more units based on	Unavailability of transcriptions or availability

	Tejas Godambe, Sai Krishna Rallabandi, Suryakanth V. Gangashetty, Ashraf Alkhairy and Afshan Jafri	decoding. CMU ARCTIC is used as database	duration are pruned (even when only 30 % units are retained)	of imprecise transcriptions and the presence of speech and non-speech noises
10.	Single microphone speech recognition using diffusion based hidden markov models- Yochay R. Yeminy, Yosi Keller and Sharon Gannot	HMM and FHMM	The SAR and SDR measures indicate that the DFHMM-S yields lower distortion levels as compared with the DFHMM-H scheme	Informal listening tests demonstrate the insufficiency of the current solution to fully recover the two speakers

3.CONCLUSION

The aim of our approach is to implement the speech recognition using CMU sphinx toolkit. From the literature survey, we've come to a conclusion that hidden markov models and deep neural networks are the best bet to implement speech recognition and the toolkit we've chosen have been designed with the same hidden markov models and deep neural networks concept. The acoustic model for Indian accent is already available for use. We need to create the speech recognition module and test it with the syllabus of blind student's examination and see how the whole exam process happens. We may also need to change the configurations to enable the engine to listen to the students for a long time because there are also some long answers in the question paper. Since we're running the engine in java, it is easy to read the questions file and write the answers to another file without any malfunctions.

The toolkit has a liveSpeechRecognizer to start the recognition and the time limit can be set between each intervals the engine should start the recording and the toolkit has classes to convert the recorded .wav files to text format for printing it in the console, The converted string format can be compared with several cases and the user can be asked to repeat the command again. Once the voice is processed properly, the modules for conversion of the entire answer to a text file can be called subsequently one after another.

We will have to test the engine by speaking to it with long answers and see if its converting the voice exactly as spoken by the user.

4. REFERENCES

[1] D. Đorđe T. Grozdića, Slobodan, T.Jovičića, Miško Subotić. MARCH 2017. Whispered speech recognition using deep denoising autoencoder ,ELSEVIER Journal on Engineering Applications of Artificial Intelligence, Volume 59, No. 4.

[2] Kuniaki Noda · Yuki Yamaguchi · Kazuhiro Nakadai · Hiroshi G. Okuno · Tetsuya Ogata. JUNE 2015. Audio-visual speech recognition using deep learning , Springer Science+Business Media Journal, Volume 42, No. 4.

[3] Andrew L. Maas, Peng Qi, Ziang Xie, Awni Y. Hannun, Christopher T. Lengerich, Daniel Jurafsky, Andrew Y. Ng. JANUARY 2017. Building DNN Acoustic Models for Large Vocabulary Speech Recognition ,ELSEVIER Journal on Computer Speech and Language, Volume 41, No. 4.

[4] Hendrik Meutzner, Santhosh Gupta, Viet-Hung Nguyen, Thorsten Holz and Dorothea Kolossa, NOVEMBER 2016. Towards improved audio CAPTCHAs based on audio perception and language processing Association of Computing Machinery Transactions on Privacy and Security. Volume 19, No. 4, Article 10.

[5] Yochay R. Yeminy, Yosi Keller and Sharon Gannot, Single microphone speech separation by diffusion-based HMM estimation ,EURASIP Journal on Audio, Speech and Music Processing.

[6] Himangshu Sarma, Navanath Saharia and Utpal Sarma. OCTOBER 2017. Development and analysis of speech recognition systems for Assamese Language using HTK , Association of Computing Machinery Transactions on Asian low Resource Language Information Processing. Volume 17 No. 1, Article 7.

[7] Tan Lee, Wai Lau, Y. W. Wong and P. C. Ching. MARCH 2002. Using tone information in cantonese continuous speech recognition ,Association of Computing Machinery Transactions on Asian low Resource Language Information Processing. Volume 1. No 1.

[8] Musab T. S. Al-Kaltakchi, Wai L. Woo, Satnam Dlay and Jonathon A. Chambers, Evaluation of a speaker identification system with and without fusion using three databases in the presence of noise and handset effects, EURASIP Journal on Advances in Signal Processing.

[9] Tejas Godambe, Sai Krishna Rallabandi, Suryakanth V. Gangashetty, Ashraf Alkhairy and Afshan Jafri. Developing a unit selection voice given audio without corresponding text , EURASIP Journal on Audio, Speech and Music Processing.

[10] Gia-Nhu Nguyen and Trung-Nghia Phung, Reducing over-smoothness in HMM-based speech synthesis using exemplar-based voice conversion ,EURASIP Journal on Audio, Speech and Music Processing.