Improving Quality of Text to Speech Using Neural Networks

[1]Neteti Aswani, [2] Dr Kunjam Nageswara Rao,[3]Dr G Sita Ratnam

[1]M.Tech Scholar, [2]Professor, [3] Associate Professor

Department of Computer Science & Systems Engineering (A), Andhra University & Adjunct Professor, IBCB, Visakhapatnam,

Department of CSE, LENDI Institute of Engineering and Technology, Vizianagaram.

Abstract: Most of the current literatures focus primarily on presence of single noise in corrupted speech which is far from real-world environments. In this paper the model is proposed for improving quality of text to speech using Convolutional Neural Networks (CNN) algorithm .It consists of three layers input layer, output layer and hidden layer. It takes text as input which is processed in the input layer now text is converted to speech using TTS(Text to Speech) system and noise is added to speech then CNN process this noisy speech using Mel cepstral coefficient(MCEP_CO) it measures noise error rate ,Short Term Objective Intelligibility(STOI)it represents the speed of reducing noise and Perceptual Evaluation of Speech Quality(PESQ) it gives the quality of speech these all are done in hidden layers of CNN then clear speech without noise is given by output layer of CNN .Quality of speech is 2.3dB using Recurrent Neural Networks (RNN) and using CNN it is 2.8dB,the higher value represents the better quality of speech , Speed of reducing noise always lies between 0 and 1 it is measured in milliseconds ,Noise error rate using RNN is 1.5 and using CNN is 1.4 and lower value indicates the better results. Based on the three parameters noise , speed of reducing noise and quality of speech CNN as shown better results than RNN.

Keyword: Text to Speech, Neural Networks, RNN, CNN.

I. INTRODUCTION

Text to Speech:

Text-to-Speech (TTS) is a useful technology that converts any text into a speech signal. It can be utilized for various purposes, e.g. car navigation, announcements in railway stations, response services in telecommunications, and e-mail reading. A typical speech sentence signal consists of two main parts: one carries the speech information, and the other includes silent or noise sections that are between the utterances, without any verbal information.

Neural networks:

A deep neural network is a neural network with a certain level of complexity. Deep neural networks use sophisticated mathematical modeling to process data in complex ways. Neural networks (NN) constitute both input & output layer, as well as a hidden layer .Convolutional neural networks(CNN) is part of deep neural networks.It uses a system much like a multilayer perceptron that has been designed for reduced processing requirements. The layers of a CNN consist of an input layer, an output layer and a hidden layer that includes multiple convolution layers, pooling layers, fully connected layers and normalization layers. The removal of limitations and increase in efficiency for noise reduction results in a system that is far more effective.

The proposed system used CNN for removing noise for enhancement of speech that can be used in applications like word recognition skills and vocabulary, reading comprehension, fluency, accuracy, information recall and learning memory enhancement and it is useful for blind people who can able hear speech by entering text.

II. RELATEDWORK

Although considerable progress has been made in the text-to-speech area, particularly in statistical parametric speech synthesis (SPSS), there is still little effort being put towards improving synthetic voices trained with lower quality recordings of speech. Most research projects and commercial systems are based on carefully recorded databases that contain very low levels of noise and reverberation. Although this is the case in many applications, there are some applications where other kinds of speech material is of a great interest [1]. For instance, the generation of personalised voices tend to rely on recordings from the end user over which we have limited control. Beyond the application driven scenario, improving quality of voices trained with lower quality data can potentially increase the amount of training material that can be used to create synthetic voices, particularly given the wealth of freely available speech data. The significant quality drop observed when the training data is noisy or reverberant can be compensated in a few different ways. Adaptation techniques have been shown to help but only to a certain extent [2]. Another way to improve quality is to discard data that is considered to be too distorted [3]. That becomes a bad strategy when there is not enough data, when distortion levels are too high or both. Alternatively, speech enhancement can be used to ‘clean’ the training data. In this paper we refer to speech enhancement as the process C. Valentini-Botinhao and of removing additive noise, often called noise suppression,as well as removing the effects of the room acoustics, i.e. dereverberation.

There are a great variety of noise suppression methods in the speech enhancement literature. Methods that are based on statistical models have been shown to produce higher quality speech than methods such as spectral subtraction, Wiener filter and subspace-based ones [4]. An alternative methodology whose popularity is growing is to use neural networks to map acoustic parameters extracted from noisy speech to parameters describing the underlying clean data [5],[6,][7],[8],[9]. It is hard to compare results across studies as the choice of evaluation metrics is inconsistent, and highly application specific. Often no subjective evaluation is performed or results are shown in terms of automatic speech recognition (ASR) performance. We mention here a selection of neural network based noise suppression studies that illustrate some of the challenges and techniques used in this area. In the work described in train a feed-forward neural network with noise aware training and global variance estimation using more than 100 different noise conditions and both techniques seem to improve results[6][10]. Investigated the use of additional input features derived from the underlying spoken text and found that spectral distortion decreases when text-based features are included[7]. In these studies around eleven frames (at least 220 ms) of acoustic features are used as the network input. Alternatively, the use of only one acoustic frame as the input to a recurrent neural network (RNN) composed of long short-term memory (LSTM) units. They reported improvements with regards to ASR performance[8],[9].

III. METHODOLOGY

Speech Signal Analysis

A typical speech sentence signal consists of two main parts: one carries the speech information, and the other includes silent or noise sections that are between the utterances, without any verbal information. The verbal (informative) part of speech can be further divided into two categories: (a) The voiced speech and (b) unvoiced speech. Voiced speech consists mainly of vowel sounds. It is produced by forcing air through the glottis, proper adjustment of the tension of the vocal cords results in opening and closing of the cords, and a production of almost periodic pulses of air. These pulses excite the vocal tract. Psychoacoustics experiments show that this part holds most of the information of the speech and thus holds the keys for characterizing a speaker.

Unvoiced speech sections are generated by forcing air through a constriction formed at a point in the vocal tract (usually toward the mouth end), thus producing turbulence. Being able to distinguish between the three is very important for speech signal analysis.

Design And Implementation

Figure1 shows a schematic of the TTS training framework adopted in this work. In this framework speech enhancement takes place prior to acoustic model training, acting like a pre-processing stage. The speech enhancement is done at a frame level and on parameters extracted from the magnitude spectrum. The magnitude spectrum is derived from the complex short-term Fourier transform (STFT). From the N length magnitude spectrum we extract M Mel cepstral coefficients, where M<N via truncation. We refer to these coefficients as MCEP-DFT. An RNN is used to generate enhanced MCEP-DFT from the distorted ones. The generated coefficients are then converted to magnitude spectrum via a warped discrete cosine transform. The enhanced magnitude spectrum and the phase spectrum obtained from the input waveform are combined and using the inverse discrete Fourier transform we obtain the enhanced waveform signal. For the purpose of acoustic model training, this signal is once again analysed, this time using a TTS-style vocoder, which in this work is the STRAIGHT vocoder. The extracted features are then used to train the acoustic model together with the linguistic features that are extracted from the underlying text in the signal. The linguistic features are initially aligned using acoustic features derived from the enhanced waveform.

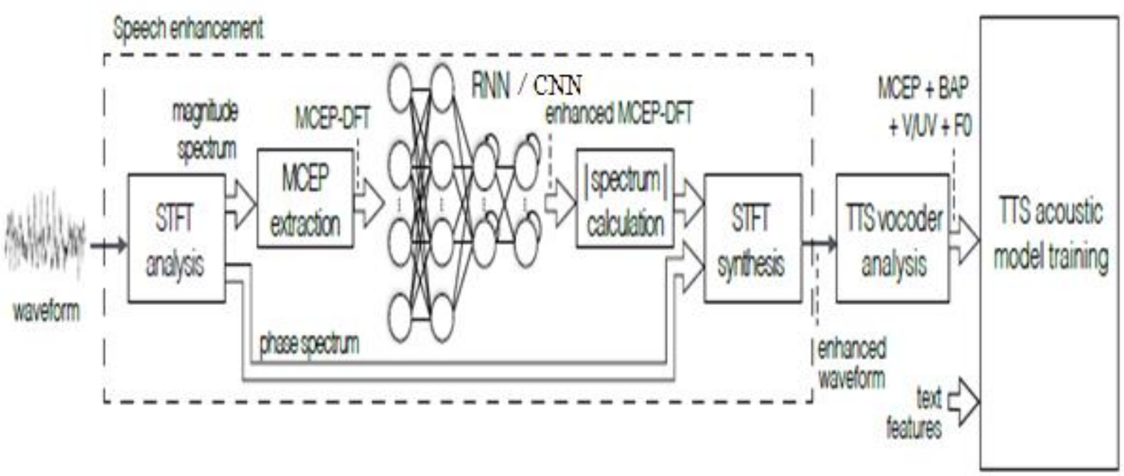


Figure 1: Proposed framework for training TTS acoustic models using an RNN-based speech enhancement method to pre process the speech waveform prior to acoustic model training.

In this framework the RNN used for speech enhancement is previously trained with a parallel database of MCEP-DFT extracted from clean and distorted speech in order to minimize the error between generated features and features extracted from clean speech data. In this work the distortion can be either additive noise, reverberation (via convolution with a room impulse response) or both.

STFT Analysis:

The short-time Fourier transform (STFT), is a Fourier-related transform used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time. In practice, the procedure for computing STFTs is to divide a longer time signal into shorter segments of equal length and then compute the Fourier transform separately on each shorter segment. This reveals the Fourier spectrum on each shorter segment. One then usually plots the changing spectra as a function of time.

MECP Extraction:

In the feature extraction part of the system we extract the features we think will be a good representation for our signal. A speech signal is non-stationary so when working with a signal like this we more or less always do the actual processing frame-wise. We can look at the signal as short-time stationary, so when considering one frame we say that the signal is stationary within that time. Usually the frames will have a length of about 10-30 msec and usually have some overlap between our frames in order to catch variations better. It is from these frames we extract the features, and we end up with a feature vector per frame.

Usually, lower MCD implies better synthesis system. A typically good synthesis system will have an MCD in range 4-5 dB. The MCEP parameters that we extract from the speech represent the speech signal in 25ms frames at 5ms shifts. The AFs which are extracted from the MCEPs also are of the same time intervals. Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware. A text-to-speech (TTS) system converts normal language text into speech. A text-to-speech system (or "engine") is composed of two parts: a front-end and a backend. The frontend has two major tasks. First, it converts raw text containing symbols like numbers and abbreviations into the equivalent of written-out words. This process is often called text normalization, pre-processing, or tokenization. The front-end then assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text-to-phoneme or grapheme-to-phoneme conversion. Phonetic transcriptions and prosody information together make up the symbolic linguistic representation that is output by the frontend. The back-end often referred to as the synthesizer, then converts the symbolic linguistic representation into sound.

Recurrent Neural Networks

A recurrent neural network (RNN) is a class of artificial neural network where connections between nodes form a directed graph along a sequence. This allows it to exhibit dynamic temporal behaviour for a time sequence. Using the knowledge from an external embedding can enhance the precision of your RNN because it integrates new information (lexical and semantic) about the words, information that has been trained and distilled on a very large corpus of data. The pre-trained embedding we'll be using is GloVe (Global Vector for word representation). GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. Although they're complex to understand, they're quite interesting. RNN is a sequence of neural network blocks that are linked to each others like a chain. Each one is passing a message to a successor.

The system will process text data, which is a sequence type. The order of words is very important to the meaning. Hopefully RNNs take care of this and can capture long-term dependencies. We will try to tackle the problem by using recurrent neural network and attention based LSTM encoder. By using LSTM encoder, we intent to encode all the information of text in the last output of Recurrent Neural Network before running feed forward network for classification. The sequential model is used to identify the text frame which is converted as speech and get processed

Convolutional Neural Network (CNN)

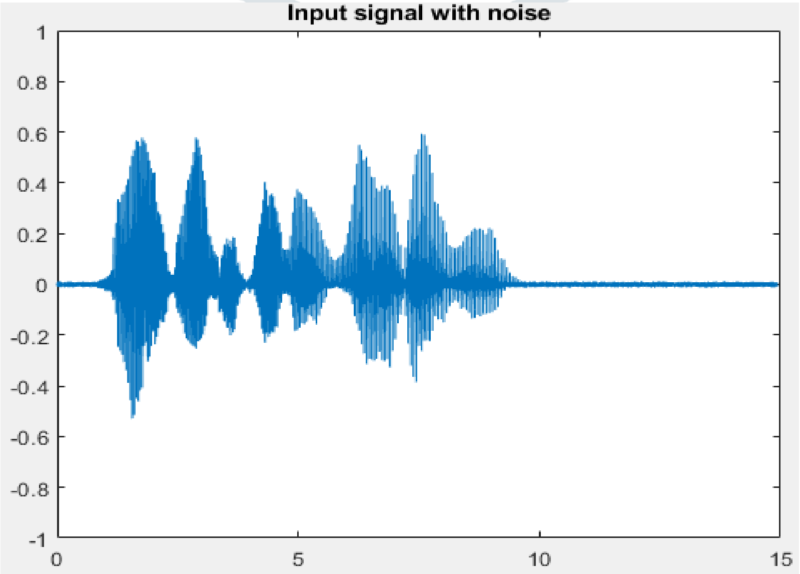
CNN is a class of deep, feed-forward artificial neural networks (where connections between nodes do not form a cycle) & use a variation of multilayer perceptrons designed to require minimal pre-processing. These are inspired by animal visual cortex.The result of each convolution will fire when a special pattern is detected. By varying the size of the kernels and concatenating their outputs, you're allowing yourself to detect patterns of multiples sizes (2, 3, or 5 adjacent words).Patterns could be expressions (word grams) like “I hate”, “very good” and therefore CNNs can identify them in the sentence regardless of their position. Convolutional Neural Networks (CNN) are everywhere. It is arguably the most popular deep learning architecture. The recent surge of interest in deep learning is due to the immense popularity and effectiveness of convnets. CNN is now the go-to model on every speech related problem. In terms of accuracy they blow competition out of the water. It is also successfully applied to recommender systems, natural language processing and more. The main advantage of CNN compared to its predecessors is that it automatically detects the important features without any human supervision. For example, given many pictures of cats and dogs it learns distinctive features for each class by itself.

CNN-based speech enhancement

To train the speech enhancement neural network we extracted MCEP-DFT features using a hamming window of16 ms and a 4 ms shift. From each windowed speech frame we extracted a DFT of 1024 size and from its magnitude value we extracted 87 Mel cepstral coefficients, which we refer here as MCEP-DFT features. We chose this value as it matches the overall number of parameters extracted using the STRAIGHT vocoder, a comparison point for our feature experiments with additive noise. The input of the network is the frame level MCEP-DFT extracted from the lower quality speech signal and the target output is the MCEP-DFT extracted from the underlying clean speech signal of that particular rframe. A different network was trained with only the noisy data (RNN-N), the reverberant data (RNN-R) and the noisy and reverberant data (RNN-NR).

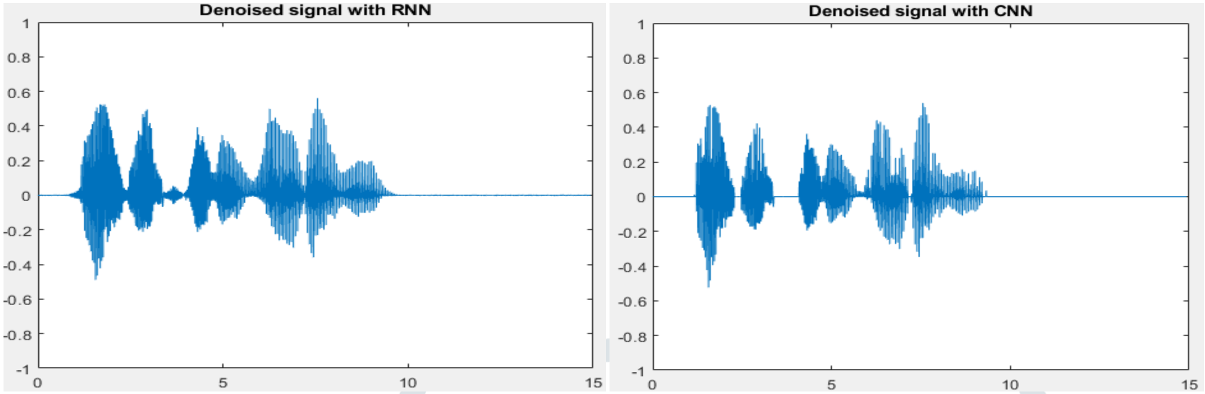
IV. RESULTS AND DISCUSSIONS

Entering the text “ Hi Aswini How are You”



Plot 1: Converted Speech signal with noise

For the given text the speech signal is generated by converting text to speech using Text-to-Speech(TTS)and additive noise is added to the converted speech signal this is shown in the plot1 with the 8ridges where each ridge represents the frequency of speech with noise.



(a) Denoised Signal using RNN **(b)** Denoised Signal with CNN

Plot 2: Comparison between denoised signal using RNN and CNN

Plot 2(a) shows the denoised signal with RNN where 3rd ridge represents reduction in noise when compared to 3rd ridge in plot 1. Plot 2(b) shows the denoised signal with CNN here 3rd ridge is completely disappeared when compared to plot 1 and plot 2(a) which is giving the better result than RNN.

Table 1.Parametrics Comparison of RNN and CNN.

Parametres	RNN	CNN
MCEP_DFT(dB)	1.545	1.449
STOI(ms)	0.8805	0.8815
PESQ(dB)	2.3697	2.858

In the above table 1 MCEP_DFT, STOI , PESQ are the three parameters used to compare RNN and CNN. MCEP_DFT(Mel cepstral coefficients) is a measure serves as an indication of the errors that vocoder makes when extracting parameters from natural speech that is not clean. It is measured in ‘dB’ and lower values indicate better results. STOI (Short term Objective intelligibility) it is the measure of linear correlation coefficient between time-frequency representation of clean and normalized time – frequency of noisy speech averaged over time frames. It is measured in millisecond(ms) and it values are always between zero and one. PESQ(Perceptual Evaluation of Speech Quality) it is a measure for predicting the quality of speech signals transmitted over telecommunication and it is measured in ‘dB’. For both STOI and PESQ, higher score represent better results.

V. CONCLUSION

The proposed system used recurrent neural network and convolutional neural network to remove additive noise and reverberation of speech material used for training a text-to-speech system. The previous work presented a series of objective and subjective evaluations on the quality of vocoded and synthetic speech created from clean (studio recordings), distorted (noisy and/or reverberated recordings) and speech that has been enhanced using convolutional neural networks. To train the network we extracted the magnitude Fourier transform of clean and distorted speech and use it as target and input of the network respectively. It has been found that synthetic speech quality can be significantly improved by simply improving the quality of the recordings used for training the voices. The most challenging scenario, where both additive noise and reverberation are present, was judged worst by listeners. However, enhancement was still significantly beneficial. The experimental results are compared then CNN method considered to be best in removing noise compared to RNN.

VI. REFERENCES

[1] J. Yamagishi, C. Veaux, S. King, and S. Renals, “Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction,” J. of Acoust. Science and Tech., vol. 33, no. 1, pp. 1–5, 2012.

[2] R. Karhila, U. Remes, and M. Kurimo, “Noise in HMM-Based Speech Synthesis Adaptation: Analysis, Evaluation Methods and Experiments,” J. Sel. Topics in Sig. Proc., vol. 8, no. 2, pp. 285–295, April 2014.

[3] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. A. J. Clark, J. Yamagishi, and S. King, “Tundra: a multilingual corpus of found data for tts research created with light supervision,” in interspeech, 2013, pp. 2331–2335.

[4] Y. Hu and P. C. Loizou, “Subjective comparison of speech enhancement algorithms,” in Proc. ICASSP, vol. 1, May 2006, pp. I–I.

[5] Y. Wang and D. Wang, “A deep neural network for time-domain signal reconstruction,” in Proc. ICASSP, April 2015, pp. 4390–4394.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE Trans. on Audio, Speech and Language Processing., vol. 23, no. 1, pp. 7–19, Jan 2015.

[7] K. Kinoshita, M. Delcroix, A. Ogawa, and T. Nakatani, “Text-informed speech enhancement with deep neural networks,” in Proc. Interspeech, Sep. 2015, pp. 1760–1764.

[8] F. Weninger, J. Hershey, J. Le Roux, and B. Schuller, “Discriminatively trained recurrent neural networks for single-channel speech separation,” in Proc. GlobalSIP, Dec 2014, pp. 577–581.

[9] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Roux, J. R. Hershey, and B. Schuller, Proc. Int. Conf. Latent Variable Analysis and Signal Separation. Springer International Publishing, 2015, ch. Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR, pp. 91–99.

[10] T. Toda and K. Tokuda, “A speech parameter generation algorithm considering global variance for HMM-based speech synthesis,” IEICE Trans. Inf. Syst., vol. E90-D, no. 5, pp. 816–824, 2007.

[11] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A shorttime objective intelligibility measure for time-frequency weighted noisy speech,” in Proc. ICASSP, Dallas, USA, March 2010, pp. 4214–4217.

[12] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in Proc. ICASSP, vol. 2, Salt Lake City, USA, May 2001, pp. 749–752.

[13] Method for the subjective assessment of intermediate quality level of coding systems, ITU Recommendation ITU-R BS.1534-1, International Telecommunication Union Radiocommunication Assembly, Geneva, Switzerland, March 2003.

[14] IEEE, “IEEE recommended practice for speech quality measurement,” IEEE Trans. on Audio and Electroacoustics, vol. 17, no. 3, pp. 225–246, 1969.

[15] S. Pascual, A. Bonafonte, and J. Serr`a, “SEGAN: speech enhancement generative adversarial network,” in Proc. Interspeech, Stockholm, Sweden, 2017.

[16] D. Rethage, J. Pons, and X. Serra, “A Wavenet for Speech Denoising,” CoRR, vol. abs/1706.07162, 2017.

[17] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florˆencio, and M. Hasegawa- Johnson, “Speech enhancement using Bayesian Wavenet,” in Proc. Interspeech, Stockholm, Sweden, 2017.

[18] S. W. Fu, T. y. Hu, Y. Tsao, and X. Lu, “Complex spectrogram enhancement by convolutional neural network with multi-metrics learning,” in Proc. MLSP, Sept 2017, pp. 1–6.

[19] S. R. Park and J. W. Lee, “A fully convolutional neural network for speech enhancement,” in Proc. Interspeech, Stockholm, Sweden, 2017.

[20] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, “Neural network-based spectrum estimation for online WPE dereverbertion,” in Proc. Interspeech, Stockholm, Sweden, 2017.