Deep Learning based NLP Techniques In Text to Speech Synthesis for Communication Recognition

Article *in* Journal of Soft Computing Paradigm · December 2020

DOI: 10.36548/jscp.2020.4.002

1 author:

Edriss Eisa Babikir Adam
Huazhong University of Science and Technology
17 PUBLICATIONS **323** CITATIONS

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No.04
Pages: 209-215
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.4.002

Deep Learning based NLP Techniques In Text to Speech Synthesis for Communication Recognition

Eriss Eisa Babikir Adam
Assistant Professor / EEE,
Mainefhi College of Engineering and Technology,
Mainefhi, Eritrea.
bonzoga20@gmail.com

Abstract: The computer system is developing the model for speech synthesis of various aspects for natural language processing. The speech synthesis explores by articulatory, formant and concatenate synthesis. These techniques lead more aperiodic distortion and give exponentially increasing error rate during process of the system. Recently, advances on speech synthesis are tremendously moves towards deep learning process in order to achieve better performance. Due to leverage of large scale data gives effective feature representations to speech synthesis. The main objective of this research article is that implements deep learning techniques into speech synthesis and compares the performance in terms of aperiodic distortion with prior model of algorithms in natural language processing.

Keywords: Natural Language Processing, speech synthesis, deep learning

1. INTRODUCTION

Mainly speech is used for communication between the individuals. The people use speech in contract with other individuals. Synthesis in speech process is artificial production of human speech and stored in computer system which is dealing with speech synthesizer. This synthesized process can be produced many pieces of recorded speech that is store in database. This system converts natural language text into speech. Generally, this process helps visual impairments and reading disabilities people. Now a day, the Text To Speech (TTS) process becomes very easier with the many algorithms which are known as Natural Language Processing (NLP). The figure 1 shows the natural speech production model for human being.

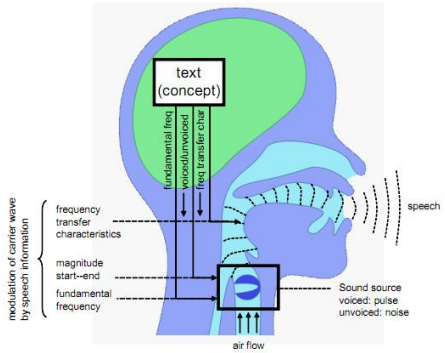


Figure 1: Natural speech production process

Generally, NLP comprises the text, phonetic and prosodic analysis. Articulatory synthesis will generate from human articulator behavior; the end frequencies of speech signal of speaking band gives formant synthesis to the model. Particularly, these syntheses are modified based on rule framing for speech synthesis [2].

ISSN: 2582-2640 (online)
Submitted: 25.11.2020
Accepted: 10.12.2020
Published: 18.12.2020

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No.04
Pages: 209-215
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.4.002

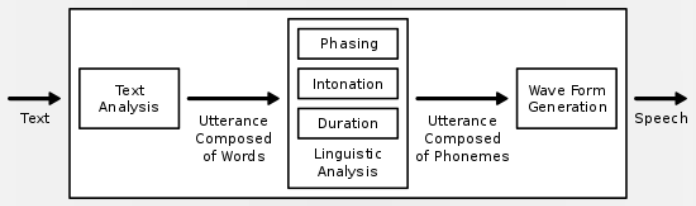


Figure 2: Overview of a typical TTS system

Finally concatenative synthesis gives speech by pre-recorded segment of speech such as "diphone" and "triphone" to construct the full bound. Many parameters in the speech synthesis is that solving unit boundaries problems during prosodic analysis. But this can solve by unit selection analysis of varying every unit of prosodies which is the patch of concatenative synthesis. Figure 2 shows the overview of typical TTS system, the waveform generation in the blocks deals with digital signal processing system. With the help of this basic concept, many algorithms are developed and implemented in natural language processing.

The feature based algorithm is evaluating a sentence in the input data on the set of rules framed by algorithm. Based on title feature, term weight, sentence length, sentence position, thematic word, the important of the sentence will be measured [7]. This feature based algorithm needs more modification in output result summary. For automatic text summarization, the template based algorithm will be very useful. This will be experimented with text pre-processing and information extraction.

After file classification from the text document, the pre-processing should be designed for the natural language process. The pre-processing can be constructed with the help of syntactic analysis, tokenization, semantic analysis, stop word removal, stemming evaluation. The first process syntactic analysis is used to determine the completion of the sentence with help of full stop symbol [6]. The better understanding of pieces of sentences includes numbers or special characters with the help of tokenization process. The role of every word understands by semantic process. Many repeated words are used often in the natural language text which meaning is very little; those words are labelled as "STOP WORDS" and are removed. Stemming process is that avoid the same words are used in different tenses and gives same meaning. The information will be extracted with the template based text summarization includes as following.

1. Dialogue control training and management
2. Determine the sentence knowledge
3. Intelligent approach for dialog conversation using deep learning approach

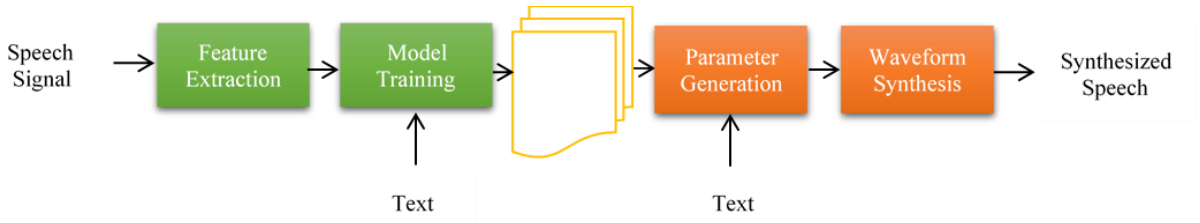


Figure 3: Overview of statistical parametric Speech synthesis

The figure 3 shows the overview of statistical parametric speech synthesis (SPSS). In order to create the knowledge base of the system, index terms, conversation person's name, place and important of communication will be measured and considered. This training can be leads the algorithm to store the input data into data base storage as a reference. This training module is managing the process which contains human and computer interaction. This trained data set is constructed and manages the user request, experience with reference data set and finally produce the answer which probably contains required information. Here knowledge based information means process of extracting intelligent information and stored it in unstructured text form. This storage structure will be used to store the different category data. Also it is reducing the search time and improves the performance of the algorithm. This template based summarization is the process of combined all these process and document in a compact format. It contains the specify trials, practised location and name of the person and many type and part of speech patterns [4].

ISSN: 2582-2640 (online)
Submitted: 25.11.2020
Accepted: 10.12.2020
Published: 18.12.2020

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No.04
Pages: 209-215
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.4.002

2. RELATED WORKS

Heiga Zen proposes MDN with SPSS in acoustic model for speech synthesis. It gives better result that statistical model. The work is based on optimisation of long short term memory – Recurrent Neural Network based SPSS on portable application devices. This work using multi frame inference reduced CPU used by 40% also with no significant difference in naturalness [13, 17]. Eliyahu gives bidirectional LSTM encoder to give effective approach for feature extraction. This Model is trained with parser to optimisation purposes. This model needs to improve in yielding very competitive parsing accuracy [14]. Zhou Yu introduces bidirectional LSTM with MLP output layer to improve to learn high level abstractions. This built model is constructed to improving in the speed of the process. The author is planning to improve the results with deep learning with more layers [15]. Bo Fan proposes deep bidirectional LSTM approach for video realistic talking head. This model will append sentimental analysis in future study [16, 18].

Suman K et al proposes the static feature extraction method with linear predictive coding. Author limited the spectral analysis due to resolution with limited frequency scale. Due to the technique the frequencies are in linear scale and inefficient [3]. The hidden markov models are contained only one layer nonlinear transformation factor to extract the features [7, 8].

Encapsulation the models of human vocal tract to provide a synthetic voice output for the given input text [3]. The storing of pre-recorded audio clips of all the words of language is bit difficult in practically in text-to-speech systems for speech synthesisers. Obviously, TTS doesn't have perfectly matched as natural due to audible anomalies. The extended architecture of bidirectional recurrent neural network replaces the hidden layers in their model with long short term memory blocks [4]. In order to produce the high quality speech waveform, the process approaches "wavenet" technique in speech synthesis. But it leads to slow and delay process from the model with the synthesis effect.

The template based algorithm is a superficial than feature based algorithm due to lag of database. So no training can be taken up for the natural language processing in feature based algorithm. These machine learning algorithms are non-updated training data set which is taking to lag of system accuracy and overall performance [5]. The deep learning methods promises existing natural language systems replaces that can achieve appropriate or better performance. Also deep learning methods are superior based on the real outputs and its developments. There many unanswered challenging problems are in natural language processing. Especially, there are more powerful demonstration tools in speech recognition by machine learning methods. The DL based neural network model is answering many new achievements and welcoming the challenging tasks of speech synthesis problems in natural language processing.

3. MOTIVATION

In Hidden Markov Model (HMM), spectrum modelling is quite hard because of high dimensionality and strong correlation. Also feature and classifier will not be possible in old algorithm. Representation of complex dependencies in acoustic features are inefficient by exist algorithm [8, 10]. This motivates us to do in speech spectrum as acoustic features as well as classifier training. Deep learning is the choice to satisfy those problems in speech synthesis. It gives possible for better representation ability with less number of factors. Also it is novel research direction of machine language which can effectively capture the many features and model the process. It can be more efficient than fragmented representation of exist one [14]. Overall, the acoustic features are over smoothed and drop in synthesized speech for some input text by hidden markov model.

4. ABSTRACT ANALYSES

Text To Speech (TTS) synthesis is having good flexibility to change acoustics features with statistical parametric model which trains deep neural network in order to get good knowledge in a sentence [19]. Mainly Recurrent Neural Network (RNN) based acoustic mapping architecture uses for gathering many speech signals simultaneously [20]. The RNN structure can be replaced by transformer network multi head attention mechanism that improves accuracy. But it fails in auto regressive error accumulation. Also it suffers from slow inference between the frames [21]. Theoretically, appending continues features of acoustics to the training network with the help of bidirectional LSTM for speech synthesis is easier [23]. Our DL based training model of speech synthesis decreases the aperiodic distortion and voice / unvoiced error rate is minimized when increases the number of units per layer with the help of mixture density.

ISSN: 2582-2640 (online)
Submitted: 25.11.2020
Accepted: 10.12.2020
Published: 18.12.2020

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No.04
Pages: 209-215
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.4.002

4.1.1 Deep Learning based Synthesis

Researchers proposed many numerous models for speech synthesis in a long tradition of studies. There are many background information are playing important role during parameter prediction. The acoustic feature contains the hierarchical structure which uses to convert from context information to speech waveform [9]. With the help of this concept the deep neural network (DNN) based methods are doing framework in predicting acoustic feature parameters for speech synthesis.

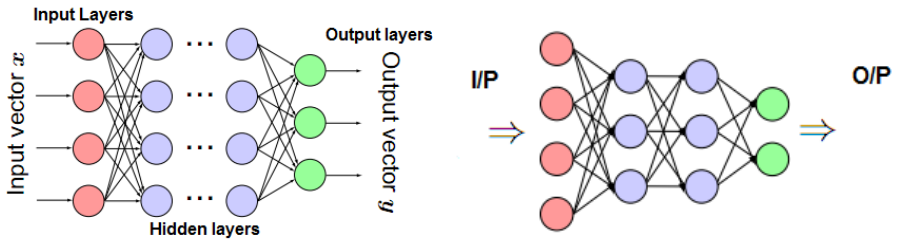
Generally the linguistic features will be mapped into probability densities of speech synthesis for various decision trees. Deep learning is quite fast to train the model for fully end-to-end speech synthesis. Template based algorithm for speech synthesis fails in high natural voice due to voice glitches. Because of some limitation presents continues in features extraction for speech synthesis.

4.1.2 Proposed Speech Synthesis Method

Based on the system, the words are mapping to the vector of real numbers. Each word are representing with real numbers 0 to N-1. These indices of the word are corresponded as a vector of length N which is directly used in network model. The text sequences synchronized like as skip gram model [11]. Each word is represented as vectors that used to calculate the conditional probability. It gives central target word with the help of softmax operation on the vector as given in equation 1 below,

$$C_p(w_o \mid w_t) = \frac{\exp(u_0^{\top} v_c)}{\sum_{i \in v} \exp(u_i^{\top} v_c)} \quad (1)$$

Where v is index set. v = $\{0,1, \ldots . . |v| - 1\}$ T is denoted as length of the sequence. The context word will be in the dictionary (index set). During the approximate training, binary tree structure constructs the loss function based on the route between one node to another node. The training process contains gradient computational which has logarithm of the dictionary size. The data sets are pre-processing with negative sampling for set the serial data. The model is creating for word embedding which can find the similarity between the two words like “Hi” & “Hai”. The morphology function is not used for word to vector instead of used different vector. The skip gram model has been implemented. We specified the specified length for extracted sub words. But dictionary format size is non-predefined one. We have used byte pair encoding technique for compression [1]. Based on the speech quality, the Convolution Neural Network (CNN) can be trained very fast. The Boltzmann machine is using for speech synthesis in deep learning algorithm. In order to create spectral envelope for the speech parameters, Boltzmann machine is used. The input layers accept the input vectors which can obey the probability distributions. Also it will calculate the hidden layers parameters. The hidden layer waits until reach convergence data into it. The output layer will interpolate the acoustic parameters based on the length of extracted subwords. These parameters are trained in the same network which will not create fragmentation problems in future [12]. Also the acoustic features are correlated with different dimensions of the same frame to reduce the over smooth phenomenon [22].



(a) (b)

Figure 4: a) Three layers in training model b) I/O in the layer for DNN

In figure 4 b shows the back propagation for supervised fine tuning in the input. From the output side the acoustic features are classified by deep neural network. During modelling of acoustic feature parameter, the single modality of the objective function couldn’t predict the variance.

ISSN: 2582-2640 (online)
Submitted: 25.11.2020
Accepted: 10.12.2020
Published: 18.12.2020

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No.04
Pages: 209-215
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.4.002

4.1.3 Modified Mixture Density Networks

This modified mixture density networks gives joint probability density function (J-PDF) of output given by input parameters [1]. The J-PDF is expressed as follows,

$$P_J(output, M_c) = \sum_{m=1}^{M_c} w_m(x).N\left(y; \mu_m(x), \sigma_m^2(x)\right) \quad (2)$$

Where $w_m(x)$,$\mu_m(x)$, $\sigma_m^2(x)$ are Gaussian parameters such as mixture weight, mean and variance respectively. The maximum likelihood estimator becomes with logarithmic function as follows in the equation 3,

$$\widehat{M_d} = \underbrace{\arg\max}_{M_d} \sum_{n=1}^{N} \quad \sum_{t=1}^{T(n)} logp\ (y_t^{(n)} \mid \ x_t^{(n)}, M) \quad (3)$$

We consider number of sentences and frames with separate ensemble in the training model.

5. RESULTS AND DISCUSSIONS

In order to rectify prediction the variance problems in the model, modified deep mixture density network patching algorithm can be appending on those extra prediction parameters. Recurrently, the output layer is predicting the probability distribution of the output features based on the input parameters. The processes are recurrence model in order to get the fine-tuned output. Because of modified mixture density function the system,

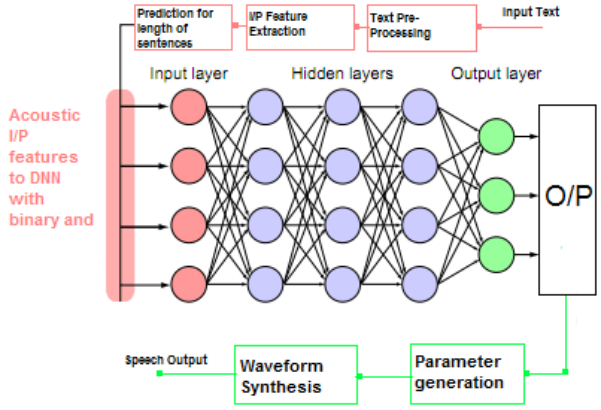


Figure 5: Proposed Neural Networks for Speech Synthesis

Our proposed model is easy to integrate the acoustic features extraction and modelling. Also, no fragmentation problems arise in the training network.

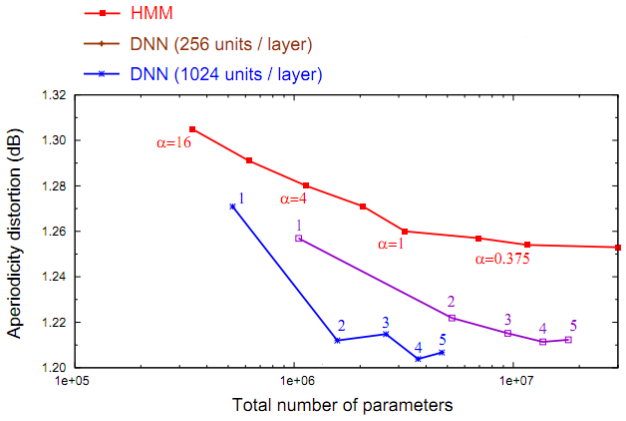


Figure 6: Aperiodicity distortion measurement comparison of proposed model

ISSN: 2582-2640 (online)
Submitted: 25.11.2020
Accepted: 10.12.2020
Published: 18.12.2020

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No.04
Pages: 209-215
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.4.002

The aperiodicity distortion is measured and plotted in the graph is shown figure 6. When increases the number of units per layer, distortion decreases. Hidden markov model is more aperiodic distortion during the training and it takes longer time to complete. Our process constructed details shows in the figure 5. This framework consists of "T" frames which are including binary and numeric features at input side of the network. The statistical report are collected includes mean and variance of the speech parameter vector sequence at output of the network.

Our DL model tags on with distributed representation to replace the gathering process during training model of the context decision tree in HMM. Because of our proposed model use multiple hidden layers are mapping context features for high dimensional acoustic features. This is creating the quality in synthesised speech better than conventional methods.

6. CONCLUSION

The old concatenation speech synthesis methods are very low intelligibility and strange from the regular. For the sensitive speech synthesis, the context decision trees technique in HMM is not sufficient to attain the enough target. Our proposed model can complete those requirements for speech synthesis and change the complications. The proposed model is very suitable to communication recognition model due to its very low aperiodic distortion.

Due to various audible glitches in the speech synthesis, our proposed technique is containing as natural human speaking capabilities. Also the implementation of sentiment analysis in text classification of natural language processing is good challenging work still. Because of country to country the passion, emotions are poles apart. Obviously models need more and more hidden layers and nodes to increase the parameters in the modified mixture density network in order to get better results in speech synthesis. Our proposed algorithm needs some better network framework and optimisation techniques to evaluate the model in a best. We are strongly believes that the readers and beginner researchers can get better understand the development process of deep learning method after red this article and test with some sample is given in http://www.ai1000.org/samples/index.html. The model is getting over fitting problems while using less number of training data. The DL based approach requires more memory to store the input parameters.

ACKNOWLEDGEMENT

We would like to thank the Mr.Karunakaran, Senior instructor, Bahrain Training Institute, Bahrain for his data collection for this research article.

REFERENCES

[1] Zen, H "A Deep Mixture density networks for acoustics modeling in statistical parameteric speech synthesis" In proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, 4-9 May 2014;pp. 3844-4848

[2] Li, R.N et al " Multi task learning of structured output layer bidirectional LSTMs for speech synthesis" In proceedings of the 42nd IEEE International Conference on Acoustics, speech and signal processing, New Orleans, LA, USA, 5-9 March 2017; pp. 5510-5514.

[3] Suman K. Saksamudre, P.P. Shrishrimal, R.R. Deshmukh, A Review on Different Approaches for Speech Recognition System, International Journal of Computer Applications (0975 8887) Volume 115 No. 22, April 2015.

[4] Seltzer, et al " Multi-task learning in deep neural networks for improved phoneme recognition" in proceedings of the 38th IEEE international conference on Acoustics, speech and signal processing, Vancouver, BC, Canada, 26-31 May 2013;pp.6965-6969.

[5] Prashant G. Desai , Saroja Devi H ,Niranjan N. Chiplumkar, "A Template Based Algorithm for Automatic Summarization and Dialogue Management for Text Documents", Proceedings of International Journal of Research in Engineering and Technology, Vol. 04 Issue: 11, pp334-340, 2015.

[6] Archana Garg, Vishal Gupta and Manish Jindal, "A Survey of Language Identification Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 6, No. 4, pp. 388-400, November 2014.

[7] Vishal Gupta, "A Survey of Natural Language Processing Techniques", International Journal of Computer Science & Engineering Technology, pp.14-16, Vol. 5 No. 01, 2014.

ISSN: 2582-2640 (online)
Submitted: 25.11.2020
Accepted: 10.12.2020
Published: 18.12.2020

Journal of Soft Computing Paradigm (JSCP) (2020)
Vol.02/ No.04
Pages: 209-215
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2020.4.002

[8] Tokuda et al " Speech parameter generation algorithms for HMM based speech synthesis" in proceedings of the 2000 IEEE International Conference on Acoustics,Speech and Signal Processing, Istanbut, Turkey, 5-9 June 2000, Volume 3, pp. 1315-1318.
[9] Yang, J.A. et al. "Deep Learning theory and its application to speech recognition" Commun.Countermeas. 2014,33, 1-5
[10] Zen H, et al. " The HMM based speech synthesis system version 2.0" In proceeding of the ISCA Workshop on speech synthesis, Bonn, Germany, 22-24 August 2007; pp. 294 – 299.
[11] Zen, H "Acoustic modeling in statistical parametric speech synthesis from HMM to LSTM-RNN' In proceedings of the first international workshop on Machine learning in Spoken Language processing, Aizu, Japan, 19-20, September 2105.
[12] "HMM/DNN based Speech synthesis system (HTS)" Available online at http://hts.sp.nitech.ac.jp/.
[13] Heiga Zen " Deep mixture density networks for acoustic modelling in statistical parametric speech synthesis" In proceeding of ICASSP,, May 2014.
[14] Eliyahu Kiperwasser et al "Simple and Accurate Dependency Parsing using Bidirectional LSTM feature Representatios" Transaction of the Association for Computational Linguistics Vol. 4, pp. 313-327, 2016.
[15] Zhou Yu et al. "Using Bidirectional LSTM Recurrent Neural Networks to Learn High Level Abstractions of Sequential Features for Automated Scoring of Non-Native Spontaneous Speech" published in IEEE workshop on Automatic Speech Recognition and understanding, Dec 2015 DOI: 10.1109/ASRU.2015.7404814.
[16] Bo Fan, "A Deep bidirectional LSTM approach for video-realistic talking head" Published in Springer, Multimed Toos Appl, Sep 2015 DOI: 10.1007/s11042-015-2944-3.
[17] Heiga Zen et al "Fast, Compact, and High Quality LSTM-RNN Based Statistical Parametric Speech Synthesizers for Mobile Devices" In Proceeding Inter speech, San Francisco, CA, USA (2016), pp. 2273-2277.
[18] Bo Fan "Photo Real Talking Head with Deep Bidirectional LSTM" Published in conference proceeding ICASSP, April 2015, DOI: 10.1109/ICASSP.2015.7178899.
[19] Yuchen Fan " TTS Synthesis with Bidirectional LSTM based Recurrent Neural Networks" Published in Fifteenth Annual Conference of the International Speech Communication, Jan 2014.
[20] Santiago Pascual et al "Multi-output RNN-LSTM for multiple speaker speech synthesis and adaptation" published in 24th European Signal Processing Conference (EUSIPCO), 2016.
[21] Naihan Li et al "Neural Speech Synthesis with Transformer Network" Published in The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19),
[22] Tom Young et al "Recent Trends in Deep Learning Based Natural Language Processing" published in Proceedings of the AAAI Conference on Artificial Intelligence 33:6706-6713, July 2019. DOI: 10.1609/aaai.v33i01.33016706.
[23] M. S. Al-Radhi, T. Gábor Csapó and G. Németh, "RNN-based speech synthesis using a continuous sinusoidal model," *2019 International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, 2019, pp. 1-8, doi: 10.1109/IJCNN.2019.8852253.

ISSN: 2582-2640 (online)
Submitted: 25.11.2020
Accepted: 10.12.2020
Published: 18.12.2020