Statistical Parametric Speech Synthesis for Punjabi Language using Deep Neural Network

Harman Singh, Parminder Singh, Manjot Kaur Gill

Department of Computer Science and Engineering, Guru Nanak Dev Engineering College Ludhiana, Punjab, India

Corresponding author: Harman Singh, Email: harmanghawaddi8@gmail.com

In recent years, speech technology gets very advanced, due to which speech synthesis becomes an interesting area of study for researchers. Text-To-Speech (TTS) system generates the speech from the text by using a synthesized technique like concatenative, formant, articulatory, Statistical Parametric Speech Synthesis (SPSS) etc. The Deep Neural Network (DNN) based SPSS for the Punjabi language is used in this research work. The database used for this research works contains 674 audio files and a single text file containing 674 sentences. This database was created at the Language Technologies Institute at Carnegie Mellon University (CMU) provided under Festvox distribution. Ossian toolkit is used as a front-end for text processing. The two DNNs are modeled using the merlin toolkit. The duration DNN maps the linguistic and duration features of speech. The acoustic DNN maps the linguistic and acoustic features. The subjective evaluation using the Mean Opinion Score (MOS) shows that this TTS system has good quality of naturalness that is 80.2%.

Keywords: TTS, SPSS, DNN, Punjabi, Speech Synthesis.

2021. In Raju Pal & Praveen K. Shukla (eds.), *SCRS Conference Proceedings on Intelligent Systems*, 431–441. Computing & Intelligent Systems, SCRS, India.
https://doi.org/10.52458/978-93-91842-08-6-41

1 Introduction

The primary and natural way of communication among humans is speech [1-2]. A speech synthesis system or Text-To-Speech (TTS) is the production of artificial speech from the text written in a language using a computer or a mechanical model [3]. In the last decades, the speech synthesis system has been used widely in many applications for the benefit of the users especially for visually impaired and illiterate people [4]. There arenumber of processing stages in TTS system. The two main phases of the TTS system are – text analysis and speech synthesis. In the text analysis phase, the raw text gets normalized and converted into its phonetic transcription and in the speech synthesis technique, the speech is produced from the phonetic transcription using a synthesis technique [5]. The TTS system has number of applications in our daily life such as for guiders of visually impaired people, video conferencing, telecommunication and multimedia, screen readers, and education [1, 6]. There are various stages for generating the pronunciation of each word in its context and for producing the synthetic speech of the sentence as shown in Fig. 1.

The steps performed for speech generation corresponding to the input text are described as follows.

- **Text Analysis and Normalization:** In text analysis, the input text of any language is converted into sentences, words, and letters. The input text may contain numerals, abbreviations and acronyms etc. which are needed to be converted into its intermediated form for generating speech corresponding to them. The text normalization transforms the text into its written form. The normalization of the input text into its pronunciation form depends upon its context of use which is a challenging task [7].

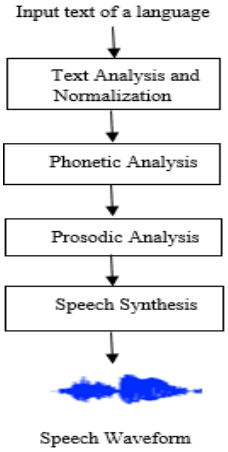

Fig. 1. General TTS System.

- **Phonetic Analysis:** The grapheme acts as the smallest unit in written language while phoneme is the smallest unit in spoken language. The TTS system needs phonetic transcription corresponding to graphemes for converting it into speech. This technique can be dictionary-based or rule-based. In dictionary-based technique, a dictionary is to be created that relates graphemes to corresponding phonemes. It generates very accurate phonetic transcription. But it requires a large memory to store all the graphemes and phonemes. It fails if there is no phoneme corresponding to the given phoneme in the

dictionary. In the rule-based technique, the set of rules are to be created for converting graphemes to their corresponding phonemes. It requires no dictionary to store [7].

- **Prosodic Analysis:**Prosodic analysis identifies the different prosody features of the speech. Prosody features include pitch, duration, and stress. The accurate identification and controlof these features generate good quality of speech. The vocal features need to be accurately identified for developing good quality TTS system [7].
- **Speech Synthesis.**The synthetic speech is generated corresponding to the phonetic transcription by considering the prosodic features [8]. There are number of synthesis techniques like formant, articulatory, concatenative etc. The appropriate technique can be selected based onthe use of the application. Mostly corpus-based techniques are used for generating good quality speech that requires pre-recorded speech units in the database [7].

2 Related Work

There are very rare works on developing the speech synthesis for the Punjabi language. The initial attempts at developing Punjabi speech synthesis used concatenative techniques. The first TTS system for the Punjabi language was developed using a concatenative technique in 2012 [4]. Syllables were used for generating the speech corresponding to the text that retains the coarticulation effect within a unit. HTK (Hidden Markov Model based Toolkit) was used for developing speech synthesis for Punjabi language using Hidden Markov Model (HMM) model in 2012 [9]. Phonemes sounds were used for synthesizing speech using a trained HMM-based synthesizer. A formant-based TTS system was built for Punjabi language using eSpeak in 2016[10]. The formant-based techniques usednumber of rules for generating the speech corresponding to the phonemes. In 2017, an HMM-based TTS system was developed for the Punjabi language using the festival toolkit [11]. Diphones were used for generating the speech corresponding to the Punjabi text. A formant-based synthesis system was built using neural networks in 2019 [12]. Speech features corresponding to phonemes were extracted from the database and they are defined in the set of rules created for phonemes.

After the introduction of the Deep Neural Network (DNN), speech technology gets accelerated in many foreign languages. DNN based TTS system for the Japanese language by utilizing Non-Negative Matrix Factorization (NMF) was developed in 2019 [13]. The activity patterns are provided as acoustic features for modeling the relationship between linguistic features and activity patterns by using DNNs. In 2020, a DNN-based synthesis system was developed for the English language with additional post-filters to remove the demerits of the HMM-based synthesis system[14]. To remove the problem of the HMM-based system, DNN was used instead of a decision tree that modifies the lack of smooth parameters. The optimal post-filter algorithm was used for making greater naturalness. To filter the test speech, an optimal parameter post filter technique was applied to enhance the speech's formant. There are a few Indian languages that used DNN for acoustic modeling in speech synthesis.DNN based speech synthesis system for the Assamese language was built in 2019[15]. DNNs map the linguistic features to duration features and acoustic features for generating the synthetic speech using the world vocoder provided by the merlin toolkit. DNN based SPSS for the Bangla languagewas developed in 2019 [16]. It maps linguistic features to duration features and acoustic features using DNNs. It also used the world vocoder provided by the merlin toolkit.

3 Database

To build the text-to-speech synthesis system for the Punjabi language, the speech database constructed at the Language Technologies Institute at Carnegie Mellon University (CMU) as phonetically balanced,

single speaker databases, provided by the Festvox distribution, is used [17].The distribution contains 13 datasets for the Indian language that are Bengali (1), Gujrati (3), Hindi (1), Kannada (1), Marathi (2), Punjabi (1), Tamil (1), and Telugu (3). The distribution includes raw wave files and the corresponding native script. It also contains built synthesis voices from these databases using CMU clustergen statistical parametric speech synthesizer. The complete android voices for CMU Flite are voices built from these databases are available in the Google play store. These databases were collected and developed with help from the Hear2Read organization. This speech data can be downloaded for academic and research use.

Punjabi database contains sound files generated by a female speaker at a 16kHz sampling rate, 32-bit depth, and mono channel. This speech database of the Punjabi language contains 674 utterances. A single text file contains the text corpus comprising674 sentences. The 674 text files are created where each containing the sentence corresponding to one sound file.

4 Methodology

The basic TTS system consists of two parts front end and the back end. The front end of the TTS system analyses the input text and converts it into the intermediate form instead of directly using the text for speech generation. Ossian toolkit is used for text processing that is a language-independent tool. The input text is converted into its corresponding grapheme units. Linguistic and acoustic features are extracted from the database using the Ossian toolkit. Linguistic features are normalized using min-max normalization and acoustic features are to be normalized using zero mean and unit variance normalization, for training the DNNs. Duration DNN maps the linguistic features and the duration features by finding the relationship between linguistic and duration features. Acoustic DNN maps the linguistic and acoustic features by finding the relationship between the acoustic and linguistic features.WORLD vocoder provided by Merlin toolkit is used for generating the speech from parameters that generate good quality speech.The overall architecture of the proposed speech synthesis system for the Punjabi language is shown in Fig. 2.

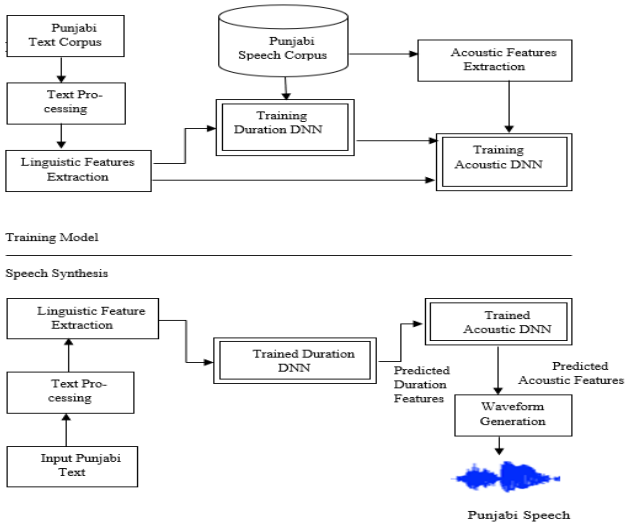

Fig. 2. The architecture of the TTS System.

4.1 Text Processing

The Ossian toolkit is used for processing the text in the UTF-8 format of the Punjabi language [18].Text is processed by tokenization, Part-Of-Speech (POS) tags, Letter-To-Sound (LTS) rule, and phrase breaks. In tokenization, the text string is tokenized into words by splitting them from whitespaces and punctuations. It also classifies the tokens as space and punctuation. The word vectors that describe the number of occurrences in terms of frequency in the whole text data, is used as POS tags. The letters are used as a substitute for phonemes in the LTS process. The pronunciation of each letter is represented by the corresponding letter. Whitespace is represented by the position of the pause tag. The silences are used for adding the prosodic phrase breaks.

4.2 Linguistic Feature Extraction

The HTS-style state-level labeling is created using the forced alignment method. The forced alignment method is applied by using the HTS toolkit. HTS-style labeling describes the identification of the quinphone, POS tag, positional information within a word, syllable, and phrase. HTS style labels are the linguistic specification of the utterance. The encoding of the linguistic features converts them into binary features. These features need to be up sampled by using duration information of frame sequence. The HTS style labeling produced a vector of 338 input features per phoneme and the 9 numerical features at frame-level are appended into it. A total of 347 linguistic features are generated which are binary and continuous numerical features. These features are normalized using min-max normalization to the range [0.01,0.99].

4.3 Acoustic Feature Extraction

The WORLD vocoder is used for acoustic feature extraction because it generates good quality speech and has fast processing speed as compared to other vocoders. It is ten times faster than other systems. It extracts the acoustic features of three types- spectral envelope, Fundamental Frequency (F_0), and aperiodicities. It manipulates the spectral envelope and F_0 but not the aperiodicities. Mel Generalized Cepstral (MGC) features are generated by transforming the full resolution spectra using Mel wrapping, which represents the spectral features. MGC features can be transformed back into the spectra of the frequency domain. MGC features are of 60-dimensional mel-cepstrum coefficients. Band Aperiodicity(BAP) and F_0 features are of a single dimension. The voicing decisions are also stored in the form of binary variable VUV that means voiced-unvoiced. The smooth trajectories are generated by using delta (Δ) and double delta ($\Delta\Delta$) features of all the vocoder features and are appended to the acoustic features. This generates a total of 187-dimensional features. The mean and variance normalization is performed on all features to reduce asymmetry and to bring features at the same scale. The normalization process speeds up the training of DNN by avoiding extra iterations. These features can be denormalized into the raw form at the synthesis stage [19].

4.4 Duration DNN

The duration model is trained using a simple feed-forward DNN. The Duration DNN maps the linguistic features to the duration features by learning the relationship between them. The normalized binary and continuous linguistic feature vectors and five state duration features are provided for training the duration DNN. The 347-dimensional binary and numerical valued linguistic feature vectors are provided as input. Root Mean Square Error (RMSE) is used to measure the deviation between the original and the predicted values of duration. It is the square root of the mean of the square of all the errors and can be calculated as

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(Si-0i)^2}{n}} \quad (1)$$

where o_i and s_i are the original and predicted values of a variable respectively and n is the total number of observations. RMSE values for duration DNN with a different number of layers and neurons are represented in Table 1 and it is analyzed that RMSE value is smaller for the DNN of 6 hidden layers with 1024 neurons in each layer than other configurations. Therefore, a duration DNN of 6 hidden layers with 1024 layers in each layer is used for training the duration features.

Table 1.RMSE values of duration prediction.

Numberof hidden layers	Number of neurons in each layer	RMSE value of phone duration prediction (frames)
5	512	4.598714
5	1024	4.581620
6	512	4.594193
6	1024	4.581479
7	512	4.601959
7	1024	4.584396

Duration DNN contains 6 hidden layers where each layer contains 1024 neurons that map the linguistic and duration features.Twenty-five epochs are used for training this duration DNN.The hyperbolic tangent function is used as an activation function for the hidden layer. This DNN model is trained using a stochastic gradient descent learning algorithm with a learning rate of 0.002. It takes approximately 2 hours to train this DNN. The architecture of the duration DNN is shown in Fig. 3.

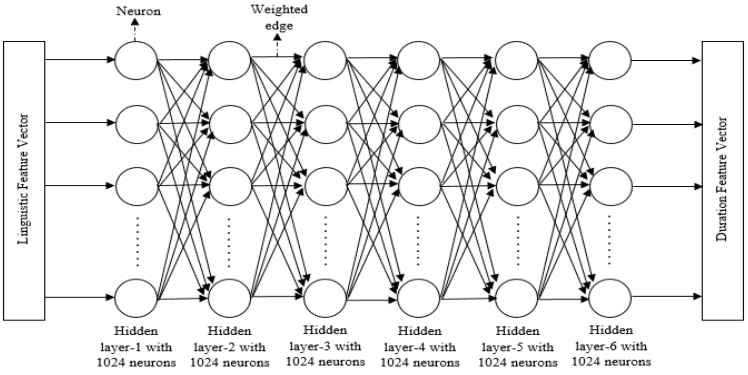

Fig. 3. The architecture of the duration DNN.

At the synthesis stage, normalized binary and continuous linguistic feature vectors are provided as input to trained duration DNN. The Duration DNN predicts the corresponding duration features of the linguistic features. These duration features are used to form the frame-level linguistic features that are to be used as input to the acoustic DNN.

4.5 Acoustic DNN

The acoustic model is trained using a simple feed-forward DNN. Acoustic DNN maps the linguistic features and the acoustic features. The 347-dimensional binary and numerical valued linguistic feature vector computed at frame level are provided as input. RMSE is used to measure the deviation between the original and the predicted values of F_0. RMSE values for acoustic DNN with a different number of layers and neurons are represented in Table 2 and it is analyzed that RMSE value is smaller for the DNN of 6 hidden layers with 1024 neurons in each layer than other configurations. Therefore, an acoustic DNN of 6 hidden layers with 1024 layers in each layer is used for training the acoustic features.

Table 2.RMSE values of F_0 prediction.

Numberof hidden layers	Number of neurons in each layer	RMSE value of phone duration prediction (frames)
5	512	167.842181
5	1024	168.484575
6	512	168.368769
6	1024	167.722292
7	512	167.768832
7	1024	168.343571

Acoustic DNN contains 6 hidden layers where each layer contains 1024 neurons that map linguistic and acoustic features. This DNN model is trained using a stochastic gradient descent learning algorithm with a learning rate of 0.002. The hyperbolic tangent function is used as an activation function for the hidden layer. Ten epochs are used for training this acoustic DNN. It takes approximately 10 hours to train this DNN.The architecture of the acoustic DNN is shown in Fig. 4.

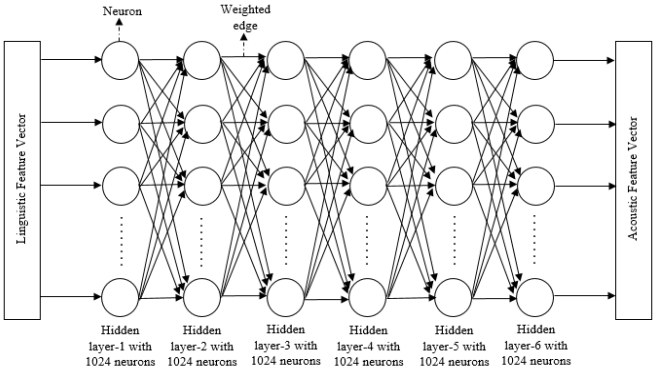

Fig. 4. The architecture of the acoustic DNN.

At the synthesis stage, normalized binary and continuous linguistic feature vectors at frame level are provided as the input to the trained acoustic DNN. The Acoustic DNN predicts the corresponding acoustic features of the linguistic and duration features. The predicted acoustic features are to be used for speech generation using the WORLD vocoder [20].

4.6 Speech Generation

To synthesize the time waveform using the acoustic features estimated by the DNN, it needs a vocoder that will generate speech corresponding to the speech parameters. In this TTS system,the WORLD

vocoder provided by the Merlin toolkit is used [21]. The acoustic features provided by acoustic DNN need to be denormalized into raw vocoder features before using them in speech generation by the vocoder. Then WORLD vocoder generates speech corresponding to the raw acoustic features. The WORLD vocoder generates vocal cord vibration by convoluting the minimum phase response and the exciting signal. It has also fewer convolutions that's why it has less computational cost compared to other vocoders. The block diagram of the speech synthesis for the Punjabi language is shown in Fig. 5.

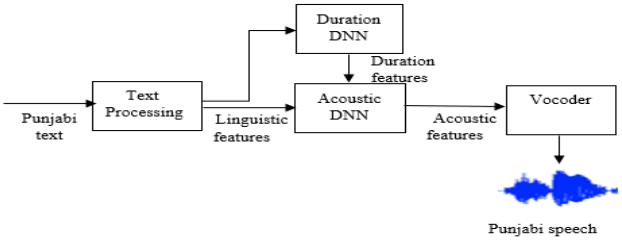

Fig. 5. Block Diagram of Speech Synthesis

5 Results and Discussion

5.1 Graphical Analysis

The visual graphical representations of an original and synthesized sentence generated by using the Praat tool are presented in Fig. 6 and Fig. 7 respectively [22]. In the graphical representation

- Blue color represents Pitch (Range: 75 – 500 Hz),
- Black color represents Spectrograms (Range: 50 – 100 dB)
- Yellow color represents Intensity (Range: 0 – 5000 Hz)
- Red color represents Formant I to Formant IV (Range: 0 – 5000 Hz)

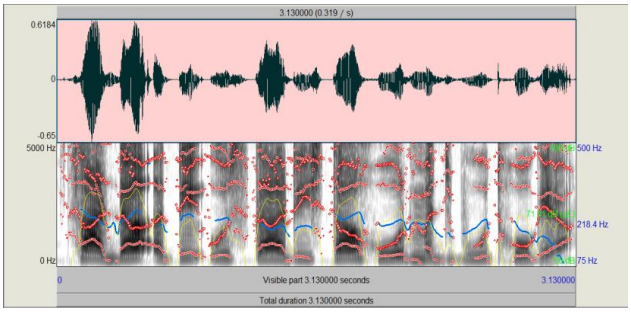

Fig. 6. Spectrogram of Original Sentence

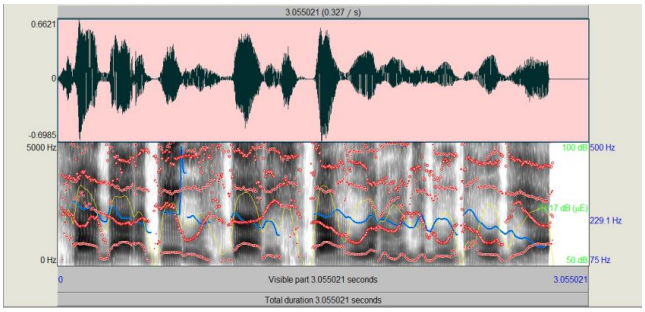

Fig. 7. Spectrogram of Synthesized Sentence

The graphical analysis of the 10 synthesized and original sentences is summarized in Table 3. It can be analyzed that there is a very small difference between the pitch and intensity values of the original and synthesized sentences as compared to the formant values. The pitch values difference is in the range of 0.04 to 22 Hz and the intensity values difference is in the range of 0.4 to 7 dB for the original and generated speech. In the second, fourth, and ninth sentences, the pitch and intensity values are almost the same as compared to formant values.

Table 3. Summary of the Graphical Analysis of 10 sentences.

Sentence No.	Original				Synthesized			
	Pitch (Hz)	Intensity (dB)	Formant I (Hz)	Formant II (Hz)	Pitch (Hz)	Intensity (dB)	Formant I (Hz)	Formant II (Hz)
1	218.43	71.01	757.94	1815.98	229.12	73.16	710.83	1785.37
2	225.43	69.51	736.56	1907.18	226.79	70.19	654.97	1861.56
3	224.17	70.67	720.06	2045.07	230.14	73.79	626.54	1949.13
4	230.61	70.55	750.11	2119.57	231.63	72.75	670.30	2024.00
5	246.66	66.45	864.43	2047.23	225.91	71.75	677.53	1985.82
6	255.43	73.43	811.36	2299.00	238.61	76.59	723.46	2170.05
7	238.63	68.23	864.68	2076.72	226.24	72.33	688.42	1987.25
8	234.13	70.73	725.06	1980.89	228.37	73.39	685.22	1964.76
9	236.28	71.17	673.14	2217.18	236.24	71.57	681.15	2156.49
10	213.54	70.85	725.06	1907.08	223.48	76.16	639.73	1906.00

5.2 Performance Evaluation

To assess the naturalness of this TTS System the subjective evaluation based on Mean Opinion Score (MOS) is conducted. Ten random synthetic utterances are selected for the MOS test. The MOS is provided by the 55 native Punjabi speakers of age from 20 to 40 years.Listeners listened to 20 wave files, 10 of the original audio, and 10 of synthetic speech. A rating from 1 to 5 is provided by the listeners. The rating parameters are specified in Table 4.

Table 4. Rating Meanings

Rating	1	2	3	4	5
Meaning	Bad	Poor	Fair	Good	Excellent

A high score means better naturalness. The average of all scores needs to be conducted for predicting naturalness. The bar chart is used for showing the average score for each sentence. The average score of each sentence is shown in Fig. 8.

Fig. 8. Average Ratings of sentences

So, to find the naturalness of the TTS system the following formula can be used, where the sum of average rating is the sum of the average score for each sentence.

$$Naturalness = \frac{\text{Sum of average rating}}{\text{Number of sentences} * \text{Total number of rating factor}} \quad (2)$$

$$= (4.0 + 4.1 + 3.9 + 4.0 + 3.9 + 4.0 + 4.0 + 4.1 + 4.2 + 3.9) / (10)*(5)$$
$$= (40.1) / 50$$
$$= 0.802$$

Therefore, the naturalness of this TTS system is 0.802 or 80.2 %.

6 Conclusion and Future Scope

In this research work, the speech synthesis system for the Punjabi language is developed that uses deep neural networks to map the linguistic, duration, and acoustic features. The dataset of 674 sentences of Punjabi language is used for development. The subjective evaluation using the MOS test shows that the naturalness of the TTS system is 0.802 or 80.2%. The value of naturalness depicts that this TTS system is of very good quality.

This TTS system can be extended further by adding more text normalization techniques. So that it can also synthesis numbers and abbreviations provided as input. The naturalness can also be improved by adding more hidden layers or neurons. The emotions and music can also be introduced for further using it for converting text into poetic speech. By adding the emotions, the vocally challenged people can express their feelings with others. The naturalness of this TTS system can also be enhanced by using good quality large speech dataset.

References

[1] Himmy and Sharma, D. D. (2017). Punjabi Text to Speech using Phoneme Concatenation. *International Journal of Advanced Research in Computer Engineering & Technology*, 6:1189–1192.

[2] Nwakanma, I., Oluigbo, I. and Izunna, O. (2014). Text–To–Speech Synthesis (TTS). *International Journal of Research in Information Technology*, 2:154–163.

[3] Mishra, P. and Shukla, J. (2013). Research Proposal Paper on Sanskrit Voice Engine: Convert Text-to-Audio in Sanskrit/Hindi. *International Journal of Computer Applications*, 70:30–34.
[4] Singh, P. and Lehal, G. S. (2012). Punjabi Text-To-Speech Synthesis System. In *Proceedings of COLING 2012: Demonstration Papers*, 409–416.
[5] Kaur, H. and Singh, P. (2019). Text to speech synthesis system for punjabi language using statistical parametric speech synthesis technique. *International Journal of Innovative Technology and Exploring Engineering*, 8:268–272.
[6] Priya and A. K. Gahier (2016). Text to speech conversion in Punjabi-a review. *International Journal of Control Theory and Applications*, 9:373–379.
[7] Panda, S. P., Nayak, A. K. and Rai, S. C. (2020). A survey on speech synthesis techniques in Indian languages. *Multimeia Systems*, 26:453–478.
[8] Joshi, M. et al. (2019). Text to Speech Synthesis for Hindi Language using Festival Framework. *International Research Journal of Engineering and Technology*, 6:630–632.
[9] Bansal, D., Goel, A. and Jindal, K. (2012). Punjabi Speech Synthesis System using HTK. *International Journal of Information Sciences and Techniques*, 2:57–69.
[10] Kaur, R. and Sharma, D. (2016). An Improved System for Converting Text into Speech for Punjabi Language using eSpeak. *International Research Journal of Engineering and Technology*, 3:500–504.
[11] Nandwani, P. and Sharma, D. V. (2017). Speech Synthesis for Punjabi Language Using Festival. *IOSR Journal of Computer Engineering*, 19:14–21.
[12] Kaur, G. and Singh, P. (2019). Formant Text To Speech Synthesis Using Artificial Neural Networks. in *2nd International Conference on Advanced Computational and Commununicational Paradigms*, 1–6.
[13] Goto, S., Saito, D. and Minematsu, N. (2019). DNN-based Statistical Parametric Speech Synthesis Incorporating Non-negative Matrix Factorization. in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 148–153.
[14] Dong, S., Li, C. and Zhang, H. (2020). An Improved Speech Synthesis Algorithm with Post filter Parameters Based on Deep Neural Network. *Communications, Signal Processing and Systems*. *Lecture Notes in Electrical Engineering, Springer Singapore*, 517:233–243.
[15] Deka, A. et al. (2019). Development of Assamese Text-to-speech System using Deep Neural Network. In *National Conference on Communications*, 1–5.
[16] Raju, R. S. et al. (2019). A Bangla Text-to-Speech System using Deep Neural Networks. In *International Conference on Bangla Speech and Language Processing*, 1–5.
[17] Punjabi Database, http://festvox.org/h2r_indic/ cmu_indic_pan_amp.tar.bz2, last accessed 2021/01/05
[18] Ossian toolkit, https://github.com/CSTR-Edinburgh/Ossian, last accessed 2021/01/05
[19] Ronanki, S. et al. (2016). DNN-based Speech Synthesis for Indian Languages from ASCII text. In *9th ISCA Speech Synthesis Workshop*, 70–75.
[20] Morise, M. and Ozawa, K. (2016). WORLD : A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Transactions on Information and Systems*, E99.D:1877–1884.
[21] Wu, Z., Watts, O. and King, S. (2016). Merlin: An Open Source Neural Network Speech Synthesis System. In *9th ISCA Speech Synthesis Workshop*, 202-207.
[22] Paul, B. and Weenink, D. Praat: doing phonetics by computer [Computer program], http://www.praat.org/.