



# A Review on Enhancing Naturalness in Text-to-Speech: The Role of Breathing Sound and Emotion Modulation

Aishwarya Chandrakant Dindore<sup>1</sup>, Dr.Nilesh Chaudhari<sup>2</sup>

<sup>1</sup> PG student, Research Scholar :  [0009-0002-9656-067X](https://orcid.org/0009-0002-9656-067X)  
Godavari College of Engineering, Jalgaon, Maharashtra, 425001

<sup>2</sup> Assistant Professor, Professor:  [0009-0002-2928-6201](https://orcid.org/0009-0002-2928-6201)  
Department of Computer Science & Engineering, Godavari College of Engineering, Jalgaon 425001

Email of Corresponding Author: [aishwarya.dindore07@gmail.com](mailto:aishwarya.dindore07@gmail.com)

Received on: 05 May,2025

Revised on: 04 June,2025

Published on: 07 June,2025

**Abstract** – The development of speech-processing technology has greatly enhanced communication between humans and computers. With the combination of deep learning, natural language processing (NLP), and artificial intelligence (AI), text-to-speech (TTS) and speech-to-text (STT) systems have been developed. The article also emphasizes how Python libraries help implement these technologies, increasing their usability and effectiveness. One essential kind of communication is speech. By bridging the gap between spoken and written language, TTS and STT technologies enable automatic transcription, virtual assistants, accessibility, and other applications. A rapidly developing area of artificial intelligence, emotion detection is essential to sentiment analysis, psychological research, and human-computer interaction. This review examines several approaches to emotion recognition, such as speech-, text-, and facial expression-based methods. In the domains of human-computer interaction, speech synthesis, and healthcare, breathing sound detection has drawn interest. It is essential for increasing the accuracy of Speech-to-Text (STT) systems and the naturalness of Text-to-Speech (TTS) systems. This paper examines several breathing sound detection techniques, such as deep learning, machine learning, and signal processing. Converting PDFs to audio has become a useful tool in assistive technology, education, and accessibility. This analysis examines several approaches for leveraging Text-to-Speech (TTS) technologies to turn text from PDF documents into speech. The study explores how machine learning, deep learning, and natural language processing

(NLP) techniques might improve the precision and naturalness of synthesized speech.

**Keywords:** Text-to-Speech (TTS), Automatic Speech Recognition (ASR), Speech Synthesis, Breathing Sound Simulation, Signal Processing.

## INTRODUCTION

Text-to-speech (TTS) technology facilitates human-computer interaction in a variety of applications by converting textual input into spoken sounds. TTS was first created using concatenative and formant synthesis techniques, but as deep learning and artificial intelligence have advanced, it has improved speech naturalness, intelligibility, and prosody considerably. Sequence-to-sequence architectures are used by machine learning models like Tacotron, WaveNet, and FastSpeech, which are used in modern TTS systems to produce speech that sounds human. These models can be used in virtual assistants, assistive technology, and audiobook creation since they incorporate intonation, pitch modulation, and emotional expression. Emotion detection aims to use computational methods to identify and interpret human emotions. Understanding emotions has become essential for applications in healthcare, customer service, and entertainment as artificial intelligence becomes more and more integrated into

daily life. The process of recording and examining respiratory sounds in order to find patterns suggestive of either normal or aberrant breathing is known as breathing sound detection. Natural breathing sounds are included in TTS to improve speech realism and humanize synthetic speech. By lowering background noise interference, the detection and filtering of breathing sounds in STT increases the accuracy of speech recognition. As the number of digital documents has increased, the ability to convert PDFs to audio has become more important for those who are blind or visually impaired, multitasking, or looking for a different way to consume text-based content. It is now feasible to produce natural-sounding, high-quality voice from digital text because of developments in speech synthesis and TTS technologies.

## LITERATURE REVIEW

### 2.1 text-to-speech (TTS)

Formant synthesis and concatenative synthesis were used in early TTS systems, although these methods frequently resulted in robotic and artificial speech. Although unit selection synthesis increased naturalness, it was constrained by the size of speech databases.

Hidden Markov Model (HMM)-based Statistical Parametric Speech Synthesis (SPSS): This method reduced storage requirements and increased flexibility, but it had problems with prosody over-smoothing (Zen et al., 2009).

TTS using Deep Learning: TTS performance has greatly improved with the introduction of Transformer-based models (Tacotron, WaveNet, FastSpeech), Deep Neural Networks (DNNs), and Recurrent Neural Networks (RNNs). WaveNet and Tacotron are renowned for their expressiveness and human-like intonation (Hinton et al., 2012; Shen et al., 2018).

### 2.2. Speech-to-text(STT)

From HMM-GMM (Gaussian Mixture Models) techniques to deep learning models, Speech recognition developed. Traditional STT: Gaussian Mixture Models (GMMs) and Hidden Markov

Models (HMMs) were used in early systems to model acoustics (Jelinek, 1976;

Rabiner&Juang,1993).Using RNNs and Transformer models, Deep Learning for STT: Whisper (OpenAI) and Deep Speech (Mozilla) have significantly increased recognition accuracy (Hinton et al., 2012; Graves et al., 2013).

### 2.3 Emotion detection in speech

AI-driven interactions that are more expressive are made possible via voice emotion recognition. Conventional Methods: Lexicon-based and rule-based approaches were used in early emotion detection to link speech characteristics to predetermined emotions (Ekman,1992; Picard,1997).Machine Learning for Emotion Recognition: Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and spectral energy are some of the variables that were employed to identify emotions using classical models like Support Vector Machines (SVMs) and Random Forests (Kim et al.,2013).Advances in Deep Learning: CNNs and LSTMs performed better in identifying spatial and temporal correlations in speech data (Schuller et al., 2018). Emotion recognition in speech and text was further improved using transformer models such as BERT and GPT (Devlin et al., 2019).

### 2.4 Breathing Sound Simulation in Speech

The incorporation of breathing sounds into synthesized speech improves naturalness and comfort.

Conventional Methods: TTS outputs were manually modified to include pre-recorded breathing sounds. Nevertheless, this strategy was not flexible enough. Signal Processing for Breathing Synthesis: To produce realistic breath sounds, methods like Low-Pass Filtering and Pink Noise Generation are employed (Chang et al., 2012).

## METHODOLOGY

### Methodologies Used in Text-to-Speech and Speech-to-Text

#### 3.1.Text-to-Speech (TTS) Methodologies Concatenative Synthesis:

concatenates pre-recorded speech samples to create words and phrases. Although it lacks flexibility, this approach produces speech of excellent quality.

**Formant Synthesis:**

uses mathematical models of human vocal tract sounds and rules to create artificial speech. Although it sounds robotic, it has a high level of intelligibility.

**Statistical Parametric Synthesis:**

provides more flexibility than concatenative synthesis by generating speech using statistical models like Hidden Markov Models (HMMs).

**Neural TTS:**

Tacotron and WaveNet, two deep learning-based models, produce speech that is more expressive and natural than that of a human.

**Prosody Modelling:**

improves speech quality by modifying stress, pitch, and tone to produce speech that sounds more natural.

**3.2 Speech-to-Text (STT) Methodologies**

**Acoustic Modeling:**

uses Deep Neural Networks (DNNs) and Gaussian Mixture Models (GMMs) to display the connections between linguistic units and auditory signals.

**Language Modeling**

uses models like n-grams and Recurrent Neural Networks (RNNs) to improve recognition accuracy by predicting word sequences in spoken language.

**Feature Extraction:**

Raw speech signals are transformed into features that can be used in recognition models using methods such as spectrogram analysis and Mel-Frequency Cepstral Coefficients (MFCCs).

**Deep Learning-based STT**

By transforming speech into text immediately, end-to-end neural network models like DeepSpeech and Whisper do away with the requirement for independent parts.

**Speaker Adaptation:**

uses adaptive training approaches to adapt STT models to the voices of individual speakers for increased accuracy.

**3.3 Methodologies for Emotion Detection**

**Text-Based Emotion Detection**

Textual input is analyzed using Natural Language Processing (NLP) techniques to determine emotions. Techniques like lexicon-based algorithms, machine learning classifiers, and deep learning models (such as LSTMs and transformers) have been used to categorize emotions from written language.

**Speech-Based Emotion Detection**

Emotions in speech can be recognized by acoustic characteristics like pitch, intensity, and rhythm. Emotional speech is often processed and classified using machine learning and deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

**Facial Expression-Based Emotion Detection**

To categorize emotions, facial recognition technology uses traits including mouth curvature, eyebrow placement, and eye movements. The accuracy of identifying emotions from facial expressions has greatly increased because of deep learning techniques, particularly CNNs.

**3.4 Methodologies for Breathing Sound Detection**

**Signal Processing Techniques**

To find breathing patterns, traditional techniques use spectral analysis, feature extraction, and filtering. Respiratory sounds are frequently analyzed using methods like the Fourier Transform, Wavelet Transform, and Mel-Frequency Cepstral Coefficients (MFCCs). In TTS and STT applications, these methods aid in differentiating speech components from breathing noises.

**Machine Learning Approaches**

Based on features that have been retrieved, breathing patterns are classified using machine learning models like Support Vector Machines (SVM), Decision Trees, and Random Forests.

These methods increase the accuracy of identifying speech pauses from breaths in STT systems and

distinguishing between normal and abnormal breathing sounds.

### 3.5 Deep Learning Approaches

Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have shown better performance in classifying respiratory sounds. By automatically extracting hierarchical features from unprocessed audio signals, these models eliminate the requirement for human feature engineering. Deep learning models help in the synthesis of realistic breathing patterns in TTS and improve the segmentation of speech and non-speech components in STT.

### 3.6 Methodologies for PDF to Audio Conversion Text Extraction from PDF

The initial step in the conversion process is to extract text from PDFs. Typical methods include of: Scanned document optical character recognition(OCR)

Text extraction with PyPDF2 and pdfminer libraries Preprocessing methods to enhance readability and tidy up retrieved text.

### Text-to-Speech (TTS) System

TTS systems translate the retrieved text into speech. TTS strategies consist of Rule-Based Synthesis: Produces speech by using predetermined phonetic rules. Concatenative Synthesis: Produces output by combining pre-recorded voice parts. Neural TTS: Produces incredibly natural speech by using deep learning models like WaveNet, Tacotron, and FastSpeech.

### Enhancing Speech Quality

The following strategies are used to enhance the listening experience: Prosody modeling is used to modify stress, pitch, and tone. Synthesis based on emotions to improve expressiveness. For authenticity, incorporate breathing sounds and reduce background noise.

## CONCLUSION

Human-computer interface has been transformed by TTS and STT technologies. As AI and machine learning continue to progress, these technologies will become increasingly effective and extensively used in a variety of industries. Applications for emotion recognition in AI-driven interactions are numerous. Even though there has been a lot of development, more precise and trustworthy emotion identification systems will be possible if issues like ambiguity, data limits, and ethical considerations are resolved. Speech synthesis and recognition depend heavily on the detection of breathing sounds. The accuracy of detection has been greatly increased by developments in deep learning, machine learning, and signal processing. The precision and caliber of synthetic speech are continuously being enhanced by developments in OCR, NLP, and deep learning. Overcoming current obstacles, however, will improve its usability and uptake much more.

## REFERENCES

- [1] J. Allen, S. Hunnicutt, and D. Klatt, "Text-to-Speech: The MITalk System," Cambridge University Press, 1987.
- [2] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. Eurospeech*, 1997.
- [3] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Proc. Eurospeech*, 1997.
- [4] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis using decision trees," *Speech Communication*, vol. 51, no. 11, pp. 1039-1064, 2009.
- [5] G. Hinton et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, 2012.
- [6] D. Klatt, "Review of Text-to-Speech Conversion for English," *Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 737-793, 1987.
- [7] H. Zen, K. Tokuda, and A. W. Black, "Statistical Parametric Speech Synthesis Using Decision Trees," *Speech Communication*, vol. 51, no. 11, pp. 1039-1064, 2009.
- [8] F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532-556, 1976.
- [9] L. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition," Prentice Hall, 1993.
- [10] G. Hinton et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, 2012.
- [11] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in *Proc. ICASSP*, 2013.
- [12] W. Chan et al., "Listen, Attend and Spell," in *Proc. NeurIPS*, 2016.

*International Journal of Innovations in Engineering and Science*, <https://ijies.net/>

- [13] P. Ekman, "Facial Expressions of Emotion: An Old Controversy and New Findings," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3449-3457, 2009.
- [14] R. Picard, *Affective Computing*, MIT Press, 1997.
- [15] J. Pennebaker, R. Booth, and M. Francis, "Linguistic Inquiry and Word Count (LIWC): A Tool for Textual Emotion Analysis," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24-54, 2010.
- [16] S. Kim, P. Georgiou, and S. Narayanan, "Emotion Recognition Using Speech Features," in *Proc. ICASSP*, 2013.
- [17] S. Poria, E. Cambria, and A. Hussain, "Multimodal Sentiment Analysis: Addressing Key Issues and Setting Future Directions," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 132-145, 2015.
- [18] I. Goodfellow et al., "Challenges in Representation Learning: Facial Expression Recognition Challenge," in *Proc. ICML*, 2013.
- [19] B. Schuller et al., "Speech Emotion Recognition with Deep Learning: Results and Comparisons," in *Proc. Interspeech*, 2018.
- [20] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019.
- [21] A. Zadeh et al., "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph," in *Proc. ACL*, 2018.
- [22] A. A. Katsis et al., "Analysis of Respiratory Sounds Using STFT," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 4, pp. 623-630, 2005.
- [23] J. C. Smith et al., "Wavelet-Based Feature Extraction for Respiratory Sound Classification," in *Proc. IEEE ICASSP*, 2010.
- [24] M. Y. Chang et al., "MFCCs for Automated Breathing Sound Analysis," *IEEE Transactions on Signal Processing*, vol. 60, no. 3, pp. 1052-1062, 2012.