INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 392-397

REFINEMENT APPROACH FOR ENHANCING ARTIFICIAL INTELLIGENCE'S SPOKEN PRONUNCIATION PRECISION IN ONLINE READING

Dr K Pavan Kumar[1], Dr V V Subba Rao[2]

[1]Assoc Professor of English, Vasireddy Venkatadri Institute of Technology, Numbur, Guntur, A.P, India.

[2]Professor & Head, Dept. of BS & H, Chalapathi Institute of Technology, Mothadaka, Guntur, A.P, India.

ABSTRACT

The paper aims to enhance the accuracy of spoken English pronunciation in virtual English reading environments by integrating artificial intelligence (AI) technology. It constructs a correction model for improving pronunciation accuracy, utilizing intelligent speech technology for speech synthesis. Here combines AI technology to develop a correction model for improving the accuracy of spoken English pronunciation. This likely involves utilizing machine learning algorithms, such as neural networks, to analyze and correct pronunciation errors and analyzes the process of speech synthesis using intelligent speech technology. This involves examining how speech is generated synthetically and understanding the underlying mechanisms. This paper contributes to the development of AI-driven systems for improving spoken English pronunciation in virtual reading environments. It outlines a comprehensive approach that combines advanced technologies, such as AI and statistical modeling, to address the challenges associated with pronunciation correction. The experimental results provide evidence of the system's effectiveness in meeting the defined objectives.

Keywords- Artificial Intelligence, Pronunciation, Speech Synthesis Analysis

1. INTRODUCTION

The importance of AI virtual spoken English systems in communication and highlights the transition of AI virtual spoken English tools from theoretical research to practical applications, particularly in the domain of English teaching and pronunciation correction. It identifies three main modules in most English speech synthesis models: the pronunciation movement model, the cooperative pronunciation model, and the acoustic model. These modules simulate various aspects of the pronunciation process, and any inaccuracies in them can affect the quality of synthesized speech.

The paper aims to improve the accuracy of the pronunciation movement model, which approximates the morphological characteristics of the articulation organs to enhance pronunciation synthesis. It compares two mainstream modeling strategies: physiological models and geometric models. Physiological models use the finite element method to simulate the biomechanical properties of soft tissue, while geometric models directly control the shapes of vocal organs and vocal tracts using predefined parameters obtained through statistical analysis. The geometric model is preferred for its simplicity and reduced computational cost, making it suitable for applications like English speech animation where detailed analysis of internal organ structure is unnecessary.

Additionally, the paper discusses the importance of visual information in communication, particularly for individuals with hearing impairments who rely on lip reading and facial expressions to understand spoken language. It emphasizes the complementary nature of visual and auditory information in communication and suggests that visual cues can aid comprehension when auditory information is unclear.

Overall, the paper proposes combining intelligent voice technology to construct a correction system for improving the accuracy of spoken English pronunciation in AI virtual English reading environments. It aims to explore the effectiveness of the proposed model and enhance the correction effect of spoken English reading.

2. RELATED WORK

The speech-driven face modeling and animation technology in enhancing the teaching effectiveness of multimodal Mandarin pronunciation teaching systems. It provides an overview of various 3D speaker simulation technologies, which can be categorized into six main types:

Vector Graphics Animation: This technique uses simple vector graphics to illustrate the main facial articulation organs, such as the mouth, tongue, teeth, and soft palate. Raster Graphics System for Animation Rendering: Complex polygons are used to create human head models in this method, allowing for high rendering quality and realistic head models. However, it can be computationally expensive and time-consuming.

Data-Driven Synthesis: Digital image processing technology is employed to extract features from digital images. For example, a sound-to-speech reversal model based on generalized variable parameters-hidden Markov models (GVP-

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 392-397

HMM) can be used for 3D speaker modeling. Anatomically Modeling the Head: This approach involves modeling a 3D speaker based on the anatomy of the head. It includes proposing a muscle model to simulate facial expression animation. Deformation Algorithm: By capturing the displacement of facial control points, this method calculates the position of deformation points on the face. It can control both local and global deformations of the face. Machine Learning: Artificial intelligence techniques, such as machine learning, are used to learn the correspondence between speech/text and articulator movement or expression. This method allows for driving 3D head models without the need for extensive real-person data collection.

The several studies and developments in the field of 3D speaker modeling and animation. These include: Developing FAP-driven facial animation Italian speaker head models based on the MPEG-4 standard. Creating virtual speakers that show the movement of tongue, teeth, and other pronunciation organs. Using electromagnetic pronunciation capture devices to collect control points for simulating developmental pronunciation. Developing visual pronunciation systems based on physiological models, deformation of muscle characteristics, and hidden Markov models for speech signal feature extraction. These developments aim to create realistic and effective systems for teaching pronunciation, utilizing advanced technologies such as motion capture, machine learning, and anatomical modeling.

Statistical Parametric Speech Synthesis Based on Hidden Markov Chain

The focus of the paper on parametric speech synthesis technology, highlighting its distinction from nonparametric methods and its reliance on data for training models. Here's a breakdown of the main points:

Parametric vs. Nonparametric Speech Synthesis: Parametric speech synthesis technology utilizes data to train models, allowing them to learn the mapping function from text to acoustic parameters. In contrast, nonparametric methods are not within the scope of the paper. Nonparametric synthesis methods often depend on databases for synthesizing speech.

Predictive Stage: Parametric synthesis technology enables direct synthesis of text into speech during the prediction stage, without relying on the dataset. This suggests that once the model is trained, it can generate speech directly from text input without needing additional data.

Statistical Parameter Speech Synthesis Based on Hidden Markov Chains: Among parametric speech synthesis models, statistical parameter speech synthesis based on hidden Markov chains is highlighted as the most popular technology. This approach involves three main modules: text analysis, acoustic, and vocoder synthesizer.

Process Overview: The process of statistical parameter speech synthesis based on hidden Markov chains is depicted in Figure 1. This figure likely illustrates the various stages involved in the synthesis process, such as text analysis, acoustic parameter generation, and speech synthesis using a vocoder.

Overall, an overview of parametric speech synthesis technology, emphasizing its reliance on data for training models and its ability to directly synthesize speech from text input during the prediction stage. It also mentions the popularity of statistical parameter speech synthesis based on hidden Markov chains among parametric synthesis methods.

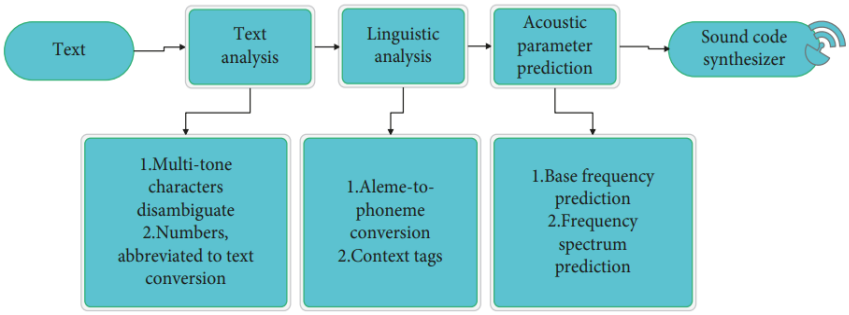

Figure 1: Statistical parametric speech synthesis model based on hidden Markov chain.

The importance of prosody in speech synthesis and the need to incorporate contextual information along with phonetic representations (phonemes) to enhance the naturalness of synthesized speech. Here's a breakdown of the main points: Character to Phoneme Conversion: This process involves converting words into phonetic representations, typically described by phonemes. Phonemes represent the basic sound units of speech.

Prosodic Units: Prosodic units are composed of adjacent phonemes and play a crucial role in conveying the speaker's mood and the mood of the sentence (e.g., declarative, interrogative, imperative). Prosody encompasses features such as pitch, length, and intensity, which contribute to the overall expression and naturalness of speech. Enhancing Naturalness: Adding prosodic information to the input helps enhance the naturalness of the synthesized voice. Without considering prosody, synthesized speech may sound unnatural or robotic.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 392-397

Contextual Information: To model context effectively, contextual information is often added to phoneme representations. This includes factors such as stress patterns and phoneme locations within words or sentences. Common Context Information: The common contextual information in English includes phoneme-related factors, stress-related factors, and location-related factors. These factors help capture the nuances of pronunciation and contribute to the overall quality of synthesized speech.

In summary, considering prosody and incorporating contextual information into phoneme representations are essential for generating natural-sounding synthesized speech. This approach ensures that synthesized speech reflects the nuances of human speech patterns and enhances the overall quality of the synthesized voice.

Current phoneme, first two phonemes, last two phonemes
The position of the current phoneme in the current syllable
The number of phonemes in the current syllable, the previous syllable, and the next syllable
The type of accent in the current syllable, the previous syllable, and the next syllable
Whether the current syllable, the previous syllable, and the next syllable are stressed
The position of the current syllable in the current word and current phrase
The number of syllables in the current phrase, the previous phrase, and the next phrase
The number of accented syllables in the current phrase, the previous phrase, and the next phrase
The number of syllables from the current position to the previous and next stressed syllable
Part of speech of the current word, the previous word, and the next word
The number of syllables in the current word, previous word, and next word
The position of the current word in the current phrase
The number of words before and after the current position in the current phrase
The number of words in the previous phrase and the next phrase
The number of syllables in the current, previous, and next phrase
The position of the current phrase in the main sentence
The distance from the current position to the stressed syllable
The number of phonemes, syllables, words, and phrases in the current sentence

Table 1: Context information.

In the statistical parameter speech synthesis model based on hidden Markov chains, the acoustic module plays a crucial role in converting the phoneme-level context sequence generated by the text analysis model into corresponding acoustic parameters. These acoustic parameters typically include Mel cepstrum coefficients, fundamental frequency, and vocalization sign.

During the training stage, various acoustic parameters such as cepstral coefficient sequences and fundamental frequency sequences are extracted using signal processing algorithms. Each different context corresponds to a different state (hidden variable) in the hidden Markov chain. Additionally, beginning and end substates are introduced in each state to provide further context. These states describe the context of prosody and linguistics.

The acoustic parameter sequence corresponds to the observation value of the state in the hidden Markov chain. The distribution of the observation value of each state is modeled using a multidimensional Gaussian mixture distribution.

Fundamental frequency information, which includes both the fundamental frequency itself and whether to speak or not, is another component of the acoustic model. The fundamental frequency is continuous, while the vocalization flag is discrete. Therefore, a multispace mixed distribution is employed to model these parameters.

It's important to note that, according to the hidden Markov model, the probability of the duration of the state sequence is also considered, although the specific details of how this is calculated are not provided in the given text.

In the context of speech synthesis, the Tacotron model is introduced. Tacotron addresses the challenge of aligning text with audio by representing speech signals using Mel-scale power values in a specified frequency range. Each frame consists of 1024 points, represented by 80 power values. The decoder in the Tacotron model predicts acoustic parameters based on the encoder's output and previous decoder output, using a cyclic neural network combined with an attention mechanism.

The attention mechanism in Tacotron's decoder allows the model to focus on relevant parts of the input sequence when predicting the current output. This attention mechanism involves splicing the context obtained from the encoder with the previous decoder output and projecting it to the specified dimension using a fully connected neural network. The mathematical operations involved in the attention mechanism are detailed in equations (7) to (11).

Overall, the Tacotron model leverages deep learning techniques, particularly attention mechanisms, to improve the alignment of text with audio and enhance the quality of synthesized speech. The detailed structure and operation of Tacotron's decoder, including the attention mechanism, are illustrated to provide insights into its functioning.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 392-397

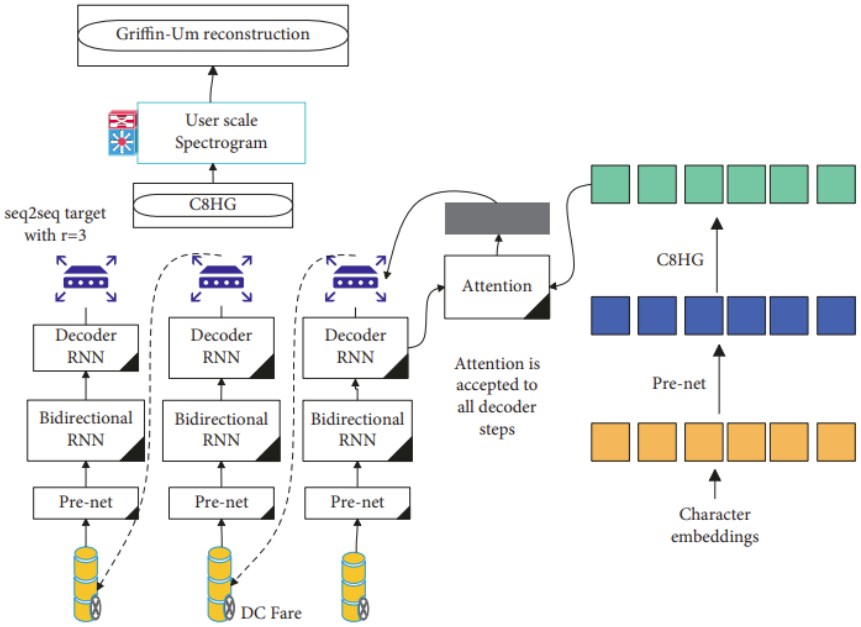

Figure 2: overall structure of the Tacotron model.

Research on Correction Method of Spoken Pronunciation Accuracy of AI Virtual English Reading

Basis of the Correction System: The correction system is built upon previous algorithm improvements, suggesting an iterative approach to enhancing the accuracy of spoken pronunciation.

Core Framework: The core framework of the AI virtual English reading system is depicted in Figure 7. This framework likely outlines the key components and interactions within the system.

Process of AI Virtual English Reading: The process of AI virtual English reading is illustrated in Figure 8. It involves creating subprocesses to handle various tasks and facilitating data transfer through anonymous pipes.

Anonymous Pipe Usage: The system utilizes anonymous pipes to facilitate communication between the main process and subprocesses. These pipes allow for the transfer of data, enabling efficient interaction between different components of the system.

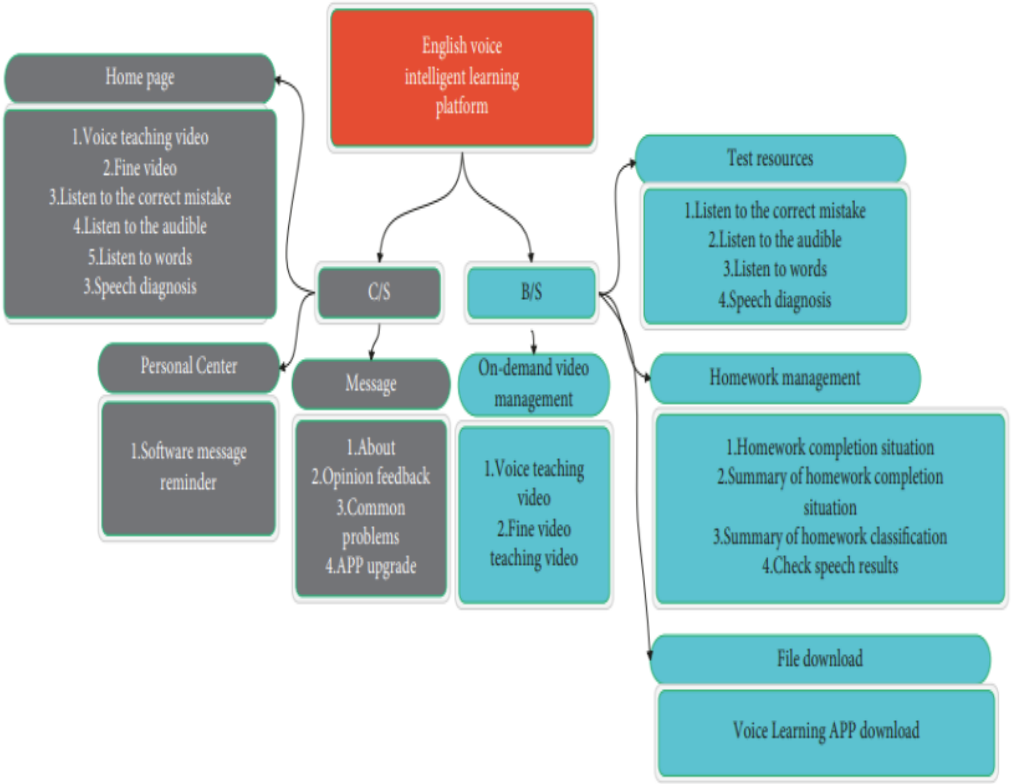

Figure 3: Core framework of AI virtual English reading system.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 392-397

Impact Factor: 5.725

3. RESULT ANALYSIS

On the basis of the above research, the effect of pronunciation correction in English reading is evaluated, and the results are shown in Figure 4 and 5.

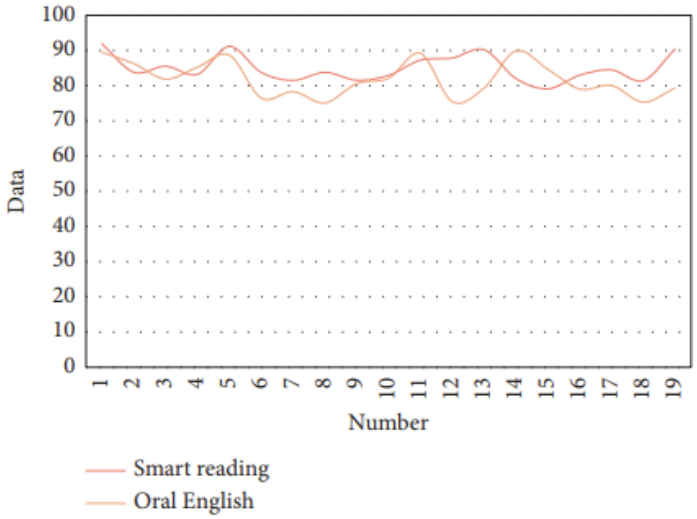

Figure 4: Statistical diagram of the evaluation of English reading and spoken pronunciation.

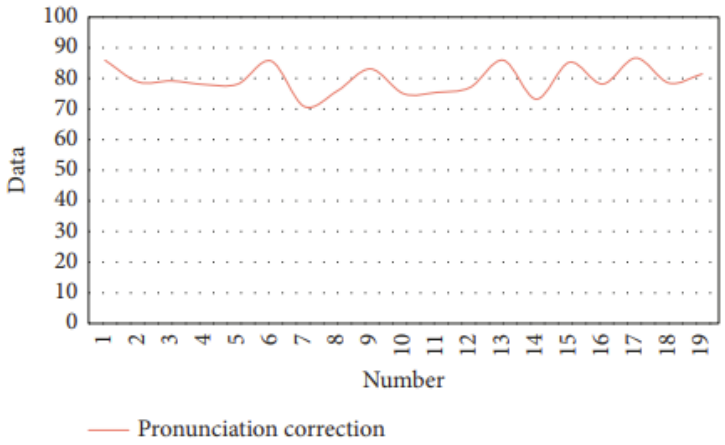

Figure 5: Statistical diagram of evaluation of pronunciation correction effect of spoken English reading

4. CONCLUSION

The integration of intelligent voice technology to improve the accuracy of spoken pronunciation in AI virtual English reading systems. It underscores the importance of experimental research in validating the proposed correction system and ensuring its effectiveness in meeting the objectives outlined in the paper.

5. REFERENCES

[1] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 7, pp. 1315–1329, 2016.

[2] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," Applied Intelligence, vol. 42, no. 4, pp. 722–737, 2015.

[3] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," IEEE/ ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 12, pp. 2263–2276, 2016.

[4] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 745–777, 2014.

[5] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: a survey," Speech Communication, vol. 56, no. 3, pp. 85–100, 2014.

[6] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, “Hybrid CTC/attention architecture for end-to-end speech recognition,” IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1240–1253, 2017.

[7] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, “An analysis of environment, microphone and data simulation mismatches in robust speech recognition,” Computer Speech & Language, vol. 46, no. 3, pp. 535–557, 2017. Advances in Multimedia 11

[8] P. Swietojanski, A. Ghoshal, and S. Renals, “Convolutional neural networks for distant speech recognition,” IEEE Signal Processing Letters, vol. 21, no. 9, pp. 1120–1124, 2014. [20] K. Angell and E. Tewell, “Teaching and un-teaching source evaluation: questioning authority in information literacy instruction,” Comminfolit, vol. 11, no. 1, pp. 95–121, 2017.

[9] Sai Srinivas Vellela, M Venkateswara Rao, Srihari Varma Mantena, M V Jagannatha Reddy, Ramesh Vatambeti, Syed Ziaur Rahman, "Evaluation of Tennis Teaching Effect Using Optimized DL Model with Cloud Computing System", International Journal of Modern Education and Computer Science(IJMECS), Vol.16, No.2, pp. 16-28, 2024. DOI:10.5815/ijmecs.2024.02.02

[10] Biyyapu, N., Veerapaneni, E.J., Surapaneni, P.P. et al. Designing a modified feature aggregation model with hybrid sampling techniques for network intrusion detection. Cluster Comput (2024). https://doi.org/10.1007/s10586-024-04270-4

[11] Vellela, S.S., Balamanigandan, R. Optimized clustering routing framework to maintain the optimal energy status in the wsn mobile cloud environment. Multimed Tools Appl 83, 7919–7938 (2024). https://doi.org/10.1007/s11042-023-15926-5

[12] S Phani Praveen, Sai Srinivas Vellela, Dr. R. Balamanigandan, "SmartIris ML: Harnessing Machine Learning for Enhanced Multi-Biometric Authentication”, Journal of Next Generation Technology (ISSN: 2583-021X), 4(1), pp.25-36 . Jan 2024.

[13] Vellela, S. S., & Balamanigandan, R. (2023). An intelligent sleep-awake energy management system for wireless sensor network. Peer-to-Peer Networking and Applications, 16(6), 2714-2731.