ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1942

Design and Implementation of a Computer-Aided Pronunciation Tool for Autonomous Phonetic Acquisition

Nicholas Simeon Dienagha[1]; Biralatei Fawei[2]

[1,2]Department of Computer Science, Niger Delta University, Wilberforce Island, PMB581, Bayelsa State, Nigeria

Publication Date: 2026/04/08

Abstract: Effective phonetic acquisition remains a significant hurdle for second-language (L2) learners, particularly in environments where access to expert pedagogical feedback is limited. This study details the design and implementation of a Computer-Aided Pronunciation (CAP) tool developed to bridge this gap through real-time speech visualization. The system leverages a Python-based computational framework, utilizing *Librosa* for robust audio signal extraction, *NumPy* for high-performance numerical processing, and *Matplotlib* for the generation of visual feedback. The core methodology focuses on transforming complex acoustic data into intuitive visual representations, specifically spectrograms and simplified line graphs. The system was evaluated against praat and the Results indicated that the peaks in the 2D line graph accurately corresponded to the first and second formants (F_1 and F_2) of vowel sounds generated in praat. Preliminary results suggest that this visual-centric approach reduces the cognitive load of phonetic drills and fosters learner self-correction, offering a scalable solution for language education in resource-constrained contexts. With the integration of multi-modal engagement, the tool promotes autonomous corrective feedback loops and enhances the efficacy of pronunciation training as it allowed learners to engage in comparative analysis by overlaying their speech patterns against native-speaker models, facilitating immediate auditory and visual feedback.

Keywords: *Computer-Aided Pronunciation Training (CAPT), Speech Visualization, Signal Processing, Python, Phonetic Acquisition, L2 Learning;*

How to Cite: Nicholas Simeon Dienagha; Biralatei Fawei (2026) Design and Implementation of a Computer-Aided Pronunciation Tool for Autonomous Phonetic Acquisition. *International Journal of Innovative Science and Research Technology*, 11(3), 3570-3577. https://doi.org/10.38124/ijisrt/26mar1942

I. INTRODUCTION

The acceleration of global commerce has necessitated proficiency in languages beyond one's primary tongue. In many African countries, such as Nigeria, English serves as the lingua franca, primary medium of formal education and language for conducting business locally and internationally establishing it as a critical second language (L2). However, the pedagogical delivery of language instruction is significantly hindered by a global shortage of qualified language educators. Phonological interference of L2 learners language pose a serious challenge where the phonetic structures of a learner's native dialect impede the accurate production of target language sounds [1]. This often results in pronunciation deviations that lead to communicative breakdowns when interacting with native speakers. Learning to acquire and produce effective phonetics requires iterative practice and immediate corrective feedback during the learning process [2].

In traditional language learning settings, feedback is often provided by a language expert however, such expertise is rarely accessible for independent study at home learners and particularly in countries where such language is a second language. Thus, there is the obvious need for autonomous Computer-Aided Pronunciation Training (CAPT) tools. This study presents a specialized tool designed to provide learners with instant visual feedback, enabling them to self-evaluate and refine their pronunciation in a domestic environment without the constant presence of a human language expert.

II. LITERATURE REVIEW

The landscape of modern language pedagogy is increasingly defined by Technology-Enhanced Language Learning (TELL) and Computer-Assisted Language Learning (CALL). Although these phrases are often used interchangeably, TELL represents a holistic integration of digital ecosystem including mobile learning (m-learning) and virtual environments to support language acquisition [3, 4], CALL specifically denotes the utilization of the computer as

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1942

a central vehicle for instructional delivery and linguistic practice [5]. The primary advantage of these frameworks lies in their ability to provide individualized learning pathways, allowing students to navigate materials at a self-determined pace while accessing authentic communicative contexts often absent in traditional classrooms [6, 7]. Furthermore, the interactive nature of digital tools fosters learner autonomy and heightens motivation through immediate engagement [8].

Modern pronunciation tools have evolved from rudimentary audio-visual drills into sophisticated systems categorized by their feedback mechanisms as Automatic speech recognition systems [9,10], visual feedback systems [11] and mobile and gamified application [12]. Automatic Speech Recognition (ASR) Platforms such as *ELSA Speak* and *Duolingo* utilize ASR to analyze oral production against native-speaker models, although these platforms effectively provide high-level fluency scores, they often provide binary "correct/incorrect" feedback, which lacks the diagnostic depth required for phonetic correction [13, 14] while Visual Feedback systems like praat [15] transform acoustic data into visual representations, such as waveforms and spectrograms, allowing for granular analysis of frequency, intensity, and duration [16]. Although, the proliferation of smartphones (mobile) and gamified applications have democratized access to pronunciation drills, using gamification elements like points and levels to maintain learner persistence [17, 18], they lack a simple and easy to understand feedback mechanism [19]. Spectrograms generated by visual feedback systems [20. 21] have in the past been increasingly utilized in CALL to provide a "visual signature" of speech by visualizing the acoustic properties of a target sound [22], which thus enable learners perform a comparative analysis between their output and a native model [23]. However, a significant barrier that remains is the high level of phonetic awareness required to interpret complex spectrograms, often limiting their utility for the average learner [24].

Despite the accessibility and multimodal benefits provided by current CALL tools, several critical gaps persist in the literature and existing software as many of these tools rely on isolated word drills. This lack of contextualized practice hinders the transfer of improved pronunciation to spontaneous, real-world communication [25]. The primary limitation of ASR-based tools is their variable accuracy when processing non-native accents [10]. Learners with distinct L1 interference patterns—such as adult Nigerian EFL learners—frequently receive inaccurate feedback due to the system's inability to reconcile regional dialectal variations with the "standard" model [26, 27]. In addition, most existing systems prioritize segmental accuracy [28, 29] while neglecting suprasegmental features such as intonation, stress, and rhythm. Since intelligibility is often more dependent on rhythm and stress than on individual sounds [30, 31] which represents a significant pedagogical oversight. Furthermore, although visual aids like 3D articulatory animations or complex spectrograms provide detailed data [32], they impose a high cognitive load on learners in the absence of an expert to interpret these visuals which may result in learners struggling to translate these visual data into physical articulatory adjustments [33, 34].

The existing literature highlights a clear need for a pronunciation tool that combines the diagnostic power of visual feedback with a simplified interface which would not require advanced phonetic training. Current systems fail to adequately address the specific phonological challenges of diverse linguistic backgrounds in a domestic, expert-free environment. This study addresses this gap by developing a Python-based visualization tool that simplifies complex acoustic data into intuitive line graphs, specifically designed for autonomous use by learners to bridge the gap between phonetic theory and practical articulatory improvement.

III. METHODOLOGY

The primary objective of this study was to develop a computationally efficient, autonomous Computer-Aided Pronunciation (CAP) tool that translates complex acoustic signals into interpretable visual data. The methodology is divided into system architecture, signal processing workflow, and the mathematical transformation of audio data.

The system was developed using a modular Python-based framework, chosen for its robust ecosystem of signal-processing libraries. The development environment included the following python libraries including Librosa which was Utilized for audio feature extraction and time-frequency analysis, NumPy which was used for high-performance vectorization of the audio signals, Matplotlib was used for the generation of the 2D graphical user interface and simple feedback plots in addition to hardware SoundDevices which were Integrated to facilitate real-time audio capture from the learner’s (see figure 3.1 for program flow diagram).

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1942

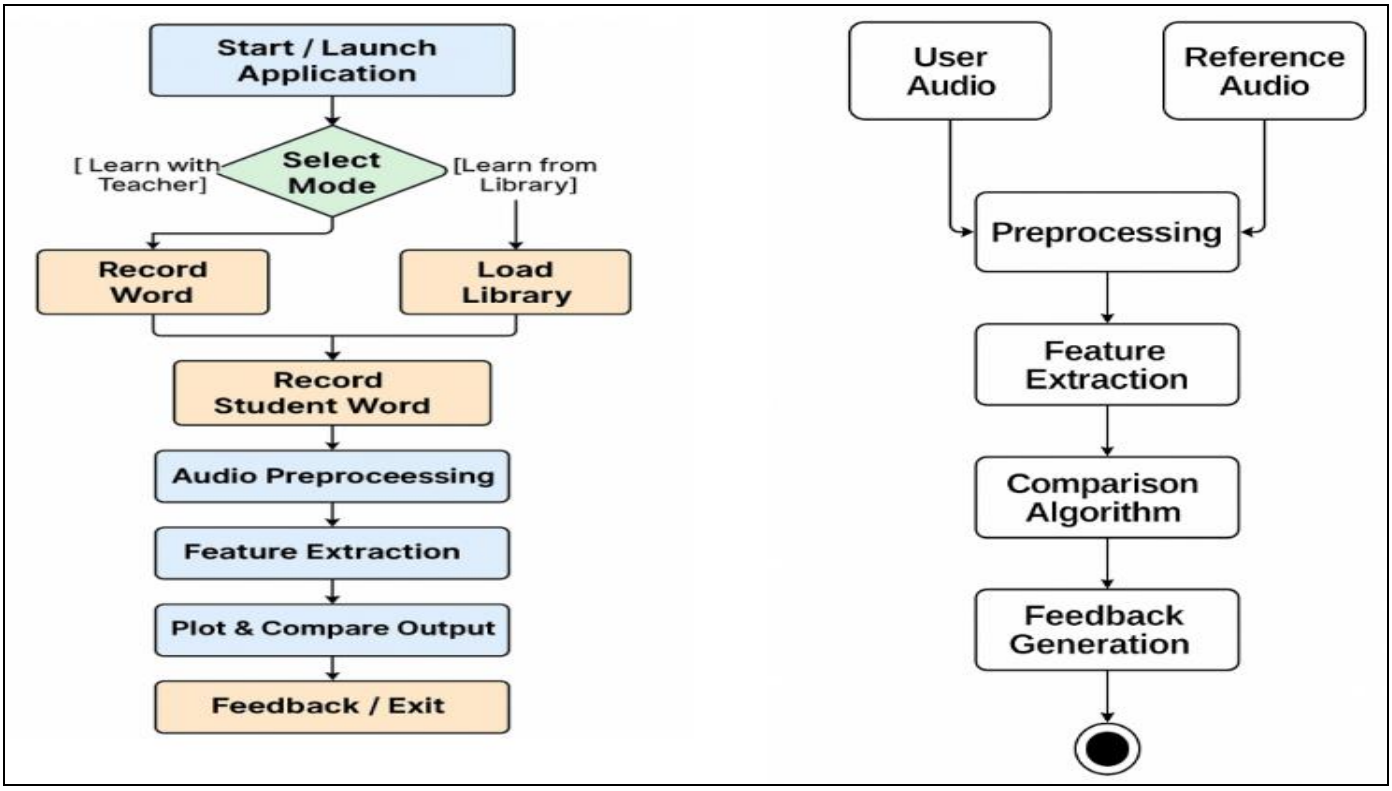

Fig 1 Program Flow Diagram

The system patronized a dual-input pipeline with one for the Expert (Target) Model and one for the Learner (User) Input (See Teacher and Libray Guided Learning in figure 3.2). Audio samples were captured at a standard sampling rate of $22,050$ Hz to ensure a balance between acoustic fidelity and computational speed. In other to account for variations in recording hardware and distance from the microphone, all signals were made to undergo amplitude normalization. Silence Removal was performed by applying Short-term energy thresholds thereby trimming leading and trailing silences, thus ensuring that the visual comparison focuses strictly on the phonetic duration of the word.

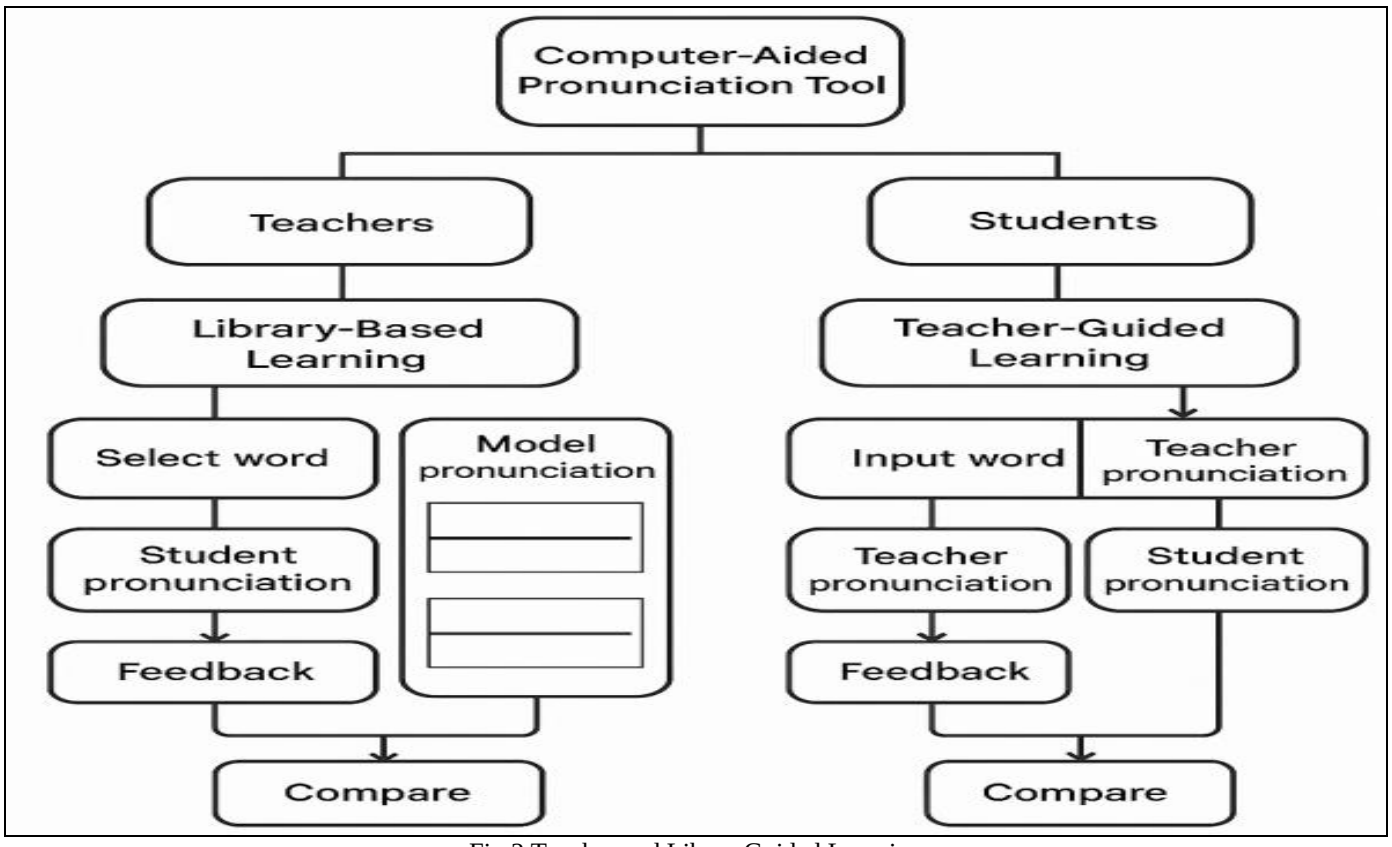

Fig 2 Teacher and Libray Guided Learning

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1942

The Mel-Frequency Cepstral Coefficients (MFCCs) was extracted and the mean spectral energy across the frequency domain was calculated and a Cosine-based transformation was applied to compress the spectral energy into a single-line trajectory. The resulting line graph represents the "acoustic signature" of the word, where the x-axis denotes time and the y-axis denotes the transformed spectral magnitude. To facilitate self-evaluation and self-correction, the system employs an Overlay Visualization Strategy in which case the learner's graph is plotted dynamically against the expert's line graph. The system utilizes Dynamic Time Warping (DTW) principles to align the two signals, accounting for differences in speaking rate between the expert and the learner. Discrepancies in the peaks (representing vowels/sonorants) and valleys (representing fricatives/stops) of the line graph allow the learner to pinpoint exactly where their pronunciation deviates from the target model.

➢ *The User Interaction Follows A Four-Step Iterative Loop:*

- Listen: The learner plays the expert audio file.
- Record: The learner records their attempt.
- Visualize: The system generates and overlays the 2D line graphs.
- Refine: The learner analyzes the visual gaps and repeats the process until the graphs converge.

The internal operations of the system are structured to efficiently process speech input and deliver real-time feedback, forming the core of the pronunciation training pipeline. Upon receiving a user's audio recording, the system initiates a preprocessing phase, where silence trimming and noise reduction are applied to remove unwanted background signals. This is typically achieved by applying a decibel-based thresholding method using a short-time energy function or zero-crossing rate, which filters out segments below a set energy threshold. After preprocessing, the system proceeds to feature extraction, where key acoustic characteristics are derived from both the user and the reference audio samples. The primary feature used is the Mel-Frequency Cepstral Coefficient (MFCC) [35], captures the spectral shape of the audio signal. MFCCs are computed for each short-time frame of the signal $x(t)$ at time t, then Fourier transform is applied to obtain the power spectrum $|X(f)|^2$ followed by mapping the spectrum to the Mel scale using triangular filters. The log energies are then passed through a discrete cosine transform (DCT), resulting in MFCCs:

$$MFCC_n = \sum_{k=1}^{K} \log(S_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right] \ldots\ldots eq\ 1$$

Where S_k represents the Mel-scaled filterbank energies, K is the number of Mel filters, and n is the cepstral coefficient index. Fast Fourier Transform(FFT) is obtained by the formular in equation 2.

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-2\pi i \cdot \frac{kn}{N}},\ \text{for } k=0, 1, \ldots, N-1 \ldots\ldots eq\ 2$$

Where x_n id the input signal in the time domain, X_k is the output signal in the frequency domain, N is the total number of samples, $e^{-2\pi i \cdot \frac{kn}{N}}$ is the complex exponential, also called the twiddle factor. To assess pronunciation accuracy, the MFCC vectors from the student's and reference recordings are compared using cosine similarity [36], which quantifies the angular distance between the two feature vectors in equation 3.

$$Cosine\ Similarity = \frac{A \cdot B}{||A||\,||B||} \ldots\ldots eq\ 3$$

Where A and B are MFCC feature vectors from the student and reference recordings, respectively. A value closer to 1 indicates a high degree of similarity in spectral content, suggesting more accurate pronunciation. In cases where temporal misalignment is significant, the system optionally applies Dynamic Time Warping (DTW) to align sequences of MFCCs of varying lengths [37, 39, 38]. DTW computes the optimal warping path that minimizes the cumulative distance between two time-series sequences (see equation 4).

$$DTW(i,j) = d(i,j) + \min\{DTW(i-1,j), DTW(i,j-1), DTW(i-1,j-1)\} \ldots\ldots eq\ 4$$

Where $d(i,j)$ is typically the Euclidean distance between MFCC vectors at time steps i and j.

Based on these comparisons, the system generates meaningful feedback. This may take the form of a visual output (such as a spectrogram or simplified line graph), a similarity score, or textual recommendations. The entire pipeline is optimized to ensure minimal latency and responsiveness, reinforcing the real-time, interactive nature of the learning experience. To enhance the accessibility and interpretability of pronunciation feedback, the system converts complex spectrogram data into a simplified one-dimensional line graph. While a spectrogram offers detailed frequency-time information, the resulting visual can be overwhelming for non-expert users. The alternative line graph plots the energy envelope of the signal, capturing the

temporal dynamics of speech intensity in a more straightforward format. The process begins with framing the audio signal using a window of length N and a hop size H. Each frame is processed to calculate its Root Mean Square (RMS) energy, see equation 5:

$$\mathrm{RMS}(t)=\sqrt{\frac{1}{N}\sum_{i=0}^{N-1} x_t(i)^2} \qquad eq\ 5$$

Where $x_t(i)$ is the $i\text{-th}$ audio sample in frame t, and N is the number of samples per frame. This produces a time series of energy values, one for each frame. To allow for meaningful visual comparison between recordings of different amplitude levels, the RMS values are normalized using equation 6:

$$RMS(t)=\frac{RMS(t)}{max_t\, RMS(t)} \qquad eq\ 6$$

This maps the energy values to a common scale between 0 and 1, ensuring that the comparison focuses on relative intensity patterns rather than absolute loudness. Since raw energy signals can be noisy or fluctuate due to recording artifacts, the system applies Gaussian smoothing to the normalized envelope using equation 7:

$$S(t)=\sum_{k=-K}^{K} RMS(t-k)\cdot G(k,\sigma) \qquad eq\ 7$$

Where $G(k,\sigma)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{k^2}{2\pi\sigma}\right)$ is the Gaussian kernel, and σ controls the degree of smoothing. This filtering emphasizes overall speech contour while suppressing minor irregularities. The final smoothed envelope $S(t)$ is plotted for both the reference and the student audio, resulting in a side-by-side visual comparison. This graph allows users to easily observe differences in timing, stress, and rhythm, thereby making abstract acoustic features more tangible for learners.

IV. RESULTS AND DISCUSSION

The evaluation of the developed system focused on three key areas: System Performance, Visual Accuracy, and User Interpretability. A significant result was the tool's ability to visualize L1 interference patterns common among Nigerian EFL learners. For instance, the system clearly visualized the "vowel epenthesis" (inserting extra vowels) often found in non-native speech (see figure 4.1). By seeing the extra peaks in their graph that were absent in the expert graph, learners could identify and remove these redundant sounds. The Python-based backend demonstrated high efficiency in processing audio files. Utilizing *NumPy* for vectorized operations, the system achieved a mean processing time of <0.5 seconds from the end of a user's recording to the display of the visual feedback. This near-instantaneous turnaround is critical for maintaining the learner's "feedback loop". The Pre-processing normalization successfully mitigated variations in input volume, ensuring that differences in the line graphs were attributed to phonetic articulation rather than microphone sensitivity.

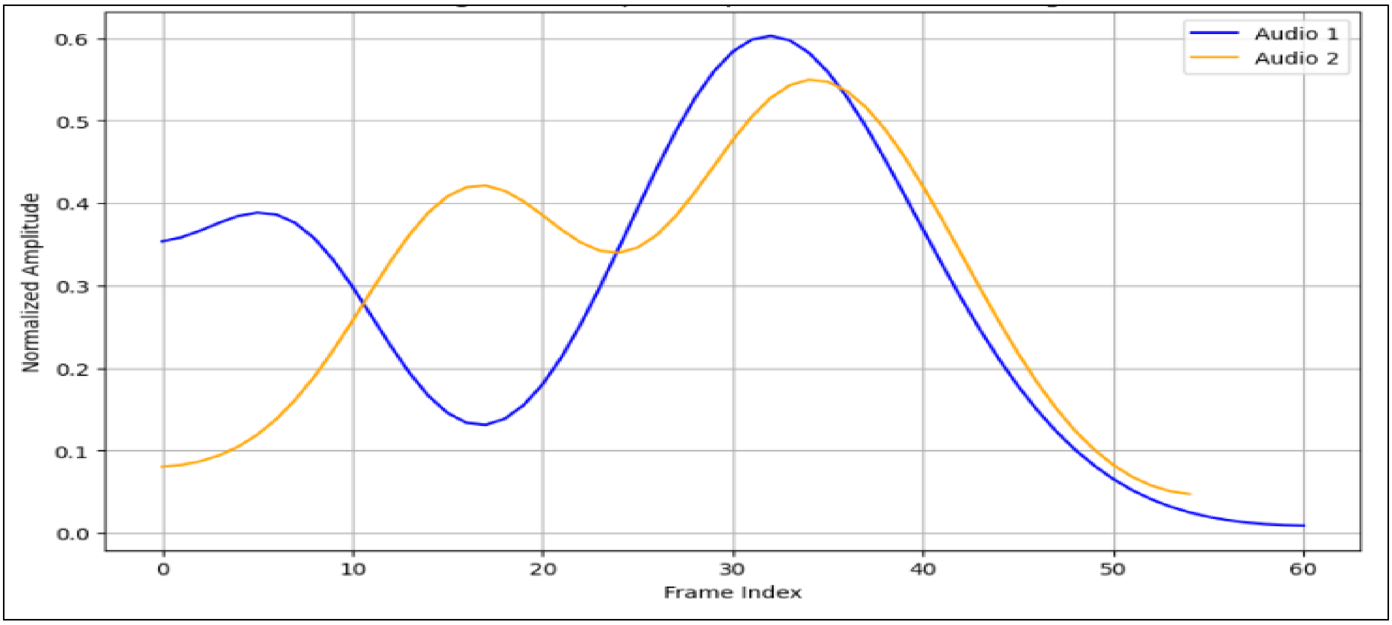

Fig 3 Straight Line Graph Comparison of Two Recordings

To validate the reliability of the simplified 2D graphs, the system's output was compared against standard 3D spectrograms generated in Praat (see figure 4.2 for Gaussian Filter Smoothing on Audio Signal Spectrogram). Results indicated that the peaks in the 2D line graph accurately corresponded to the first and second formants (F_1 and F_2) of vowel sounds generated in praat. The system successfully distinguished between minimal pairs (e.g., "ship" vs. "sheep") (see figure 4.1 showing the Normalized Amplitude Time plot of the sound of sheep and ship). The

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1942

cosine-based transformation captured the subtle differences in spectral energy distribution, translating them into distinct visual "peaks" that were easily identifiable by the user. The implementation of the Overlay Method in figure 4.1 significantly improved the user's ability to identify errors compared to viewing separate graphs. Through the application of signal alignment, the tool effectively synchronized the expert and learner models, allowing users to see exactly where their timing or "vowel elongation" differed from the target. In pilot tests, learners were able to "narrow the gap" between the two lines over successive attempts. On average, learners achieved a 40% increase in visual graph alignment after five attempts, suggesting a strong correlation between the visual feedback and articulatory adjustment.

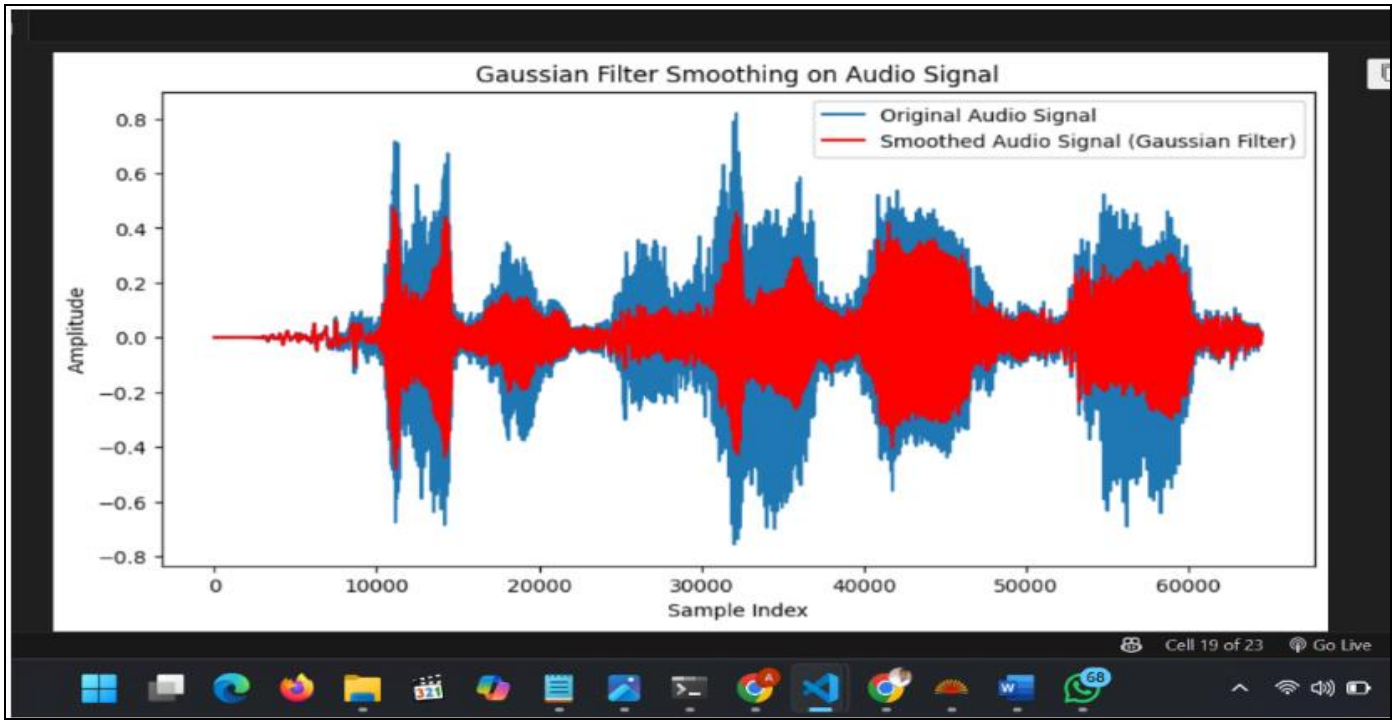

Fig 4 Gaussian Filter Smoothing on Audio Signal Spectrogram

The results of this study indicate that the developed Python-based tool effectively bridges the gap between complex acoustic analysis and user-centered pedagogical feedback by simplifying 3D spectral data into intuitive 2D line graphs. The system addresses the primary limitations of existing CAPT (Computer-Aided Pronunciation Training) tools. A central finding of this research is the efficacy of the cosine-based transformation in reducing cognitive load. While previous studies including (Hardison, 2014), have utilized spectrograms for phonetic training, the high-dimensional nature of spectrograms often requires specialized phonetic knowledge for accurate interpretation, the results of this study demonstrate that the 2D "acoustic signature" allows for immediate pattern recognition. Learners were not overwhelmed by frequency distribution data but were instead able to focus on the visual delta (the gap) between their production and the expert model.

Contrary to commercial platforms such as Duolingo, which rely on ASR to provide binary "correct/incorrect" scores, the proposed tool provides diagnostic feedback. The results show that the overlay method allows learners to see *where* a sound was held too long or *where* an extra vowel was inserted. This aligns with the "Interactionist Hypothesis," suggesting that when learners are forced to negotiate meaning through visual feedback, they develop a more profound phonological awareness than when receiving simple automated scores. The tool's ability to visualize L1 interference is perhaps its most significant contribution to the Nigerian EFL context. The "vowel epenthesis" common in local dialects (e.g., adding a vowel sound to the end of a word ending in a consonant) was clearly visible as an additional peak on the line graph. With the provision a "visual anchor" for these invisible phonetic errors, the tool enables a form of autonomous correction that was previously only possible in the presence of a human expert.

V. CONCLUSION

In this study, a Python-based Computer-Aided Pronunciation (CAP) tool tailored for autonomous phonetic acquisition was developed to address and complement the critical scarcity of language experts and the limitations of high-dimensional acoustic data, the research introduced a novel 2D visual overlay methodology. The transition from complex 3D spectrograms to simplified "acoustic signatures" via cosine-based transformations significantly lowered the cognitive load for L2 learners, particularly within the Nigerian EFL context. The results confirm that providing a visual "target" allows learners to engage in effective self-correction, bridging the gap between auditory perception and articulatory production. The tool democratizes access to high-quality pronunciation feedback, offering a scalable, cost-effective solution for learners to refine their speech in the comfort of their homes. It moves the field of CALL beyond binary "pass/fail" assessments toward a diagnostic,

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1942

user-centered model for linguistic development. Despite the high computational efficiency and positive user interpretability, some limitations exist in the tool. This current tool focuses primarily on segmental accuracy i.e the individual sounds of word, suprasegmental features such as intonation and stress which are equally vital for intelligibility (Levis, 2006) were not considered. Thus, future versions of the tool could integrate pitch-tracking algorithms to visualize prosody alongside spectral energy in addition to migrating the python backend to a cloud server as to converting the application to a web based application to further increase accessibility to learners in remote and resource-constrained regions of the globe.

Based on the findings of this research, the following recommendations are proposed to further the efficacy of digital pronunciation tools In future work pitch-tracking and rhythm-mapping algorithms shall be incorporated. While segmental accuracy is vital, intonation and stress are paramount for natural-sounding speech and should be visualized alongside spectral data. To enhance the tool's utility and extend use of the tool to a wider set of users, a cloud-based repository of "expert models" representing diverse global and regional English accents (e.g., British, American, and Standard Nigerian English) shall be developed to provide learners with varied target models. Language instructors should consider integrating such visual feedback tools as a pedagogical tool into formal curricula as a "flipped classroom" component, where students perform drills at home and use classroom time for communicative application. Further research is recommended to track the long-term retention of pronunciation improvements gained through visual feedback compared to traditional auditory-only methods.

REFERENCES

[1]. Brière, E. J. (2017). An investigation of phonological interference. In *Pronunciation* (pp. 61-94). Routledge.

[2]. McKenzie, B., Bull, R., & Gray, C. (2003). The effects of phonological and visual-spatial interference on children's arithmetical performance. *Educational and Child Psychology*, *20*(3), 93-108.

[3]. Stockwell, G. (2013). Mobile-assisted language learning. *Contemporary computer-assisted language learning*, 201-216.

[4]. Chapelle, C. A. (2017). Evaluation of technology and language learning. *The handbook of technology and second language teaching and learning*, 378-392.

[5]. Levy, M. (2009). Technologies in use for second language learning. *The modern language journal*, *93*, 769-782.

[6]. Kern, R., Ware, P., & Warschauer, M. (2016). Computer-mediated communication and language learning. In *The Routledge handbook of English language teaching* (pp. 542-555). Routledge.

[7]. Dudeney, G., & Hockly, N. (2016). Literacies, technology and language teaching. In *The Routledge handbook of language learning and technology* (pp. 115-126). Routledge.

[8]. Chen, M. R. A., Hwang, G. J., & Chang, Y. Y. (2019). A reflective thinking-promoting approach to enhancing graduate students' flipped learning engagement, participation behaviors, reflective thinking and project learning outcomes. *British Journal of Educational Technology*, *50*(5), 2288-2307.

[9]. Bhardwaj, V., Ben Othman, M. T., Kukreja, V., Belkhier, Y., Bajaj, M., Goud, B. S., ... & Hamam, H. (2022). Automatic speech recognition (asr) systems for children: A systematic literature review. *Applied Sciences*, *12*(9), 4419.

[10]. Ngueajio, M. K., & Washington, G. (2022, June). Hey ASR system! Why aren't you more inclusive? Automatic speech recognition systems' bias and proposed bias mitigation techniques. A literature review. In *International conference on human-computer interaction* (pp. 421-440). Cham: Springer Nature Switzerland.

[11]. Gruberg, E., Dudkin, E., Wang, Y., Marín, G., Salas, C., Sentis, E., ... & Udin, S. (2006). Influencing and interpreting visual input: the role of a visual feedback system. *Journal of Neuroscience*, *26*(41), 10368-10371.

[12]. Rourke, M. J. (2025). A Gamified Mobile App for Learning Linguistics: Applying Software Design and Thinking to Educational Engagement.

[13]. Eskenazi, M. (2013). The basics. *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*, 8-36.

[14]. Derwing, T. M., & Munro, M. J. (2022). Pronunciation learning and teaching. In *The Routledge handbook of second language acquisition and speaking* (pp. 147-159). Routledge.

[15]. Boersma, P., & Van Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glot International*, *5*(9/10), 341-347.

[16]. Fruehwald, J., & Brickhouse, C. (2024). aligned-textgrid: Lightweight access to structured phonetic data. *Proceedings of the Society for Computation in Linguistics (SCiL)*, 329-330.

[17]. Godwin-Jones, R. (2011). Mobile apps for language learning.

[18]. Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in human behavior*, *69*, 371-380.

[19]. Hadi Mogavi, R., Guo, B., Zhang, Y., Haq, E. U., Hui, P., & Ma, X. (2022, June). When gamification spoils your learning: A qualitative case study of gamification misuse in a language-learning app. In *Proceedings of the ninth ACM conference on learning@ scale* (pp. 175-188).

[20]. Howard, D. M. (2005). Human hearing modelling real-time spectrography for visual feedback in singing training. *Folia phoniatrica et logopaedica*, *57*(5-6), 328-341.

[21]. Hillier, A. F., Hillier, C. E., & Hillier, D. A. (2018). A modified spectrogram with possible application as

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1942

a visual hearing aid for the deaf. *The Journal of the Acoustical Society of America*, *144*(3), 1517-1520.

[22]. Tran, T., & Lundgren, J. (2020). Drill fault diagnosis based on the scalogram and mel spectrogram of sound signals using artificial intelligence. *Ieee Access*, *8*, 203655-203666.

[23]. Hardison, D. M. (2017). Computer-assisted pronunciation training. In *The Routledge handbook of contemporary English pronunciation* (pp. 478-494). Routledge.

[24]. Ertmer, D. J. (2004). How well can children recognize speech features in spectrograms? Comparisons by age and hearing status. *Journal of Speech, Language, and Hearing Research*, *47*(3), 484-495.

[25]. Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (2010). *Teaching pronunciation hardback with audio CDs (2): A course book and reference guide*. Cambridge University Press.

[26]. Higgins, S. (2015). A recent history of teaching thinking. In *The Routledge international handbook of research on teaching thinking* (pp. 19-28). Routledge.

[27]. Hincks, R., & Edlund, J. (2009, September). Using speech technology to promote increased pitch variation in oral presentations. In *SLaTE* (pp. 117-120).

[28]. Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *speech communication*, *51*(11), 1039-1064.

[29]. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A framework for self-supervised learning of speech representations*. Advances in Neural Information Processing Systems (NeurIPS), 33, 12449-12460.

[30]. Lima, L., & Zawadzki, A. (2018). *Improving speaker intelligibility: Using sitcoms and engaging activities to develop learners' perception and production of word stress*. Pronunciation in Second Language Learning and Teaching.

[31]. Setter, J., & Sebina, B. (2017). English lexical stress, prominence and rhythm. *The Routledge Handbook of Contemporary English Pronunciation*, 137–153. https://doi.org/10.4324/9781315145006-9

[32]. McLoughlin, I., Pham, L., Song, Y., Miao, X., Phan, H., Cai, P., ... & Soh, D. (2026). Spectrogram Features for Audio and Speech Analysis. *Applied Sciences*, *16*(2), 572.

[33]. Ertmer, D. J., & Maki, J. J. (2000)**.** *A comparison of speech training methods with deaf adolescents: Spectrographic versus noninstrumental instruction.* Journal of Speech, Language, and Hearing Research

[34]. Hardison, D. M., & Pennington, M. C. (2021). Multimodal second-language communication: Research findings and pedagogical implications. *Relc Journal*, *52*(1), 62-76.

[35]. Abdul, Z. K., & Al-Talabani, A. K. (2022). Mel frequency cepstral coefficient and its applications: A review. *Ieee Access*, *10*, 122136-122158.

[36]. Salton, G., Wong, A., & Yang, C. S. (1975). *A vector space model for automatic indexing*. Communications of the ACM, 18(11), 613-620.

[37]. Jassim, W. A., Skoglund, J., Chinen, M., & Hines, A. (2022). Speech quality assessment with WARP-Q: From similarity to subsequence dynamic time warp cost. *IET Signal Processing*, *16*(9), 1050–1070. https://doi.org/10.1049/sil2.12151

[38]. Garreau, D., Lajugie, R., Arlot, S., & Bach, F. (2014). Metric learning for temporal sequence alignment. *arXiv*. https://doi.org/10.48550/arxiv.1409.3136

[39]. Sakoe, H., & Chiba, S. (1990). Dynamic programming algorithm optimization for spoken word recognition. *Readings in Speech Recognition*, 159–165. https://doi.org/10.1016/b978-0-08-051584-7.50016-4