Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 18 No. 18 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijim.v18i18.51491

PAPER

Exploring a Mobile Technology-Driven Model for Intercultural Communication Education

Huajie Chen, Kehan Mei(✉)

School of Foreign languages & Cultures, Tibet University, Lhasa, China

mkh318@utibet.edu.cn

ABSTRACT

In the context of accelerating globalization, intercultural communication competence has become a crucial element for the success of individuals and organizations. The frequent international interactions have underscored the importance of enhancing methods for improving intercultural communication skills within the educational sector. The rapid advancement of mobile technology offers unprecedented opportunities for educational innovation. Its convenience and widespread use have made mobile device-based learning models highly attractive. Particularly, advancements in real-time speech processing technologies have provided new tools and methods for intercultural communication education. By leveraging mobile real-time speech detection and synthesis technologies, a more interactive and personalized learning experience can be achieved, thereby enhancing the efficiency and effectiveness of language learning. This study aims to explore a mobile technology-based model for intercultural communication education and is divided into three main parts: firstly, the investigation of mobile real-time speech detection technologies aimed at intercultural communication education to provide instant feedback and improvement suggestions; secondly, the exploration of mobile real-time speech synthesis technologies to generate high-quality speech samples for learners to practice against; and thirdly, the integration of the aforementioned technologies to develop a flexible, efficient, and highly interactive learning system based on mobile technology. This study is expected to not only improve the effectiveness of intercultural communication education but also provide significant references for the innovative application of educational technologies.

KEYWORDS

intercultural communication education, mobile technology, real-time speech detection, real-time speech synthesis, educational model innovation

1 INTRODUCTION

In the era of deepening globalization today, the ability to communicate across cultures has been identified as a crucial factor for the success of individuals and

Chen, H., and Mei, K. (2024). Exploring a Mobile Technology-Driven Model for Intercultural Communication Education. *International Journal of Interactive Mobile Technologies (iJIM)*, 18(18), pp. 62–75. https://doi.org/10.3991/ijim.v18i18.51491

Article submitted 2024-06-05. Revision uploaded 2024-07-24. Final acceptance 2024-07-31.

organizations [1–3]. With international exchanges becoming increasingly frequent, the effective enhancement of intercultural communication skills has emerged as an urgent issue within the field of education. The rapid development of mobile technology has provided new opportunities for innovation in education. The convenience and ubiquity of mobile devices have gradually drawn attention to learning models based on this technology [4–6]. Particularly, the advancement in real-time speech processing technologies offers new tools and methods for education in intercultural communication. These technologies enable more interactive and personalized learning experiences through real-time speech detection and synthesis on mobile devices, thereby improving the efficiency and effectiveness of language learning [7, 8].

A study indicates that language learning facilitated by mobile technology can break the constraints of time and space and can also be customized to meet the individual needs of learners, significantly enhancing both learning outcomes and learner engagement [9–11]. Moreover, intercultural communication education involves not only the mastery of language skills but also the cultivation of understanding and adaptability to different cultural backgrounds [12, 13]. The application of mobile technology promises to integrate these aspects, providing learners with a more comprehensive and profound education in intercultural communication.

However, existing study methods exhibit certain flaws and limitations. On the one hand, many models of intercultural communication education still rely on traditional classroom teaching and unidirectional language instruction, lacking interactivity and personalization [14–17]. On the other hand, although some studies have attempted to incorporate mobile technology, most remain at the basic level of language learning applications and fail to fully leverage the advantages of real-time speech detection and synthesis technologies [18, 19]. Additionally, existing studied often overlook the characteristics of speech communication across different cultural backgrounds, thus failing to provide effective training in intercultural communication skills.

This study aims to develop a mobile technology-based model for intercultural communication education, focusing on three key areas. First, it explores mobile real-time speech detection technologies to provide precise pronunciation feedback for learners. Second, it investigates how mobile real-time speech synthesis can produce high-quality samples for practice. Finally, the study aims to create a flexible and interactive learning system by integrating these technologies. This study enhances the effectiveness of intercultural communication education and offers valuable insights for innovative educational technology applications.

2 MOBILE REAL-TIME SPEECH DETECTION FOR INTERCULTURAL COMMUNICATION EDUCATION

To address the needs of real-time speech detection in mobile intercultural communication education, a compact native replay speech detection model integrating a convolutional feature extractor with a compact neural architecture was proposed in this study. The training principle aims to ensure the model's lightness and high accuracy on mobile devices. The model architecture is illustrated in Figure 1. Initially, a local data preprocessing module computes acoustic features, followed by the extraction of local dependency features using multi-layer convolution operations. The employed convolutional feature extractor consists of one-dimensional convolutional layers, batch normalization layers, and LeakyReLU non-linear transformation layers. A one-dimensional convolution block with a stride of 10 and a

kernel size of 3 × 128 × 6 generates a tensor of dimensions 10 × 128 × 6 and subjects it to batch normalization and non-linear transformation to enhance feature expressiveness. Owing to the convolution structure's limitations in capturing long-term contextual information, a structure of neural circuit policies (NCP) based on *CfC* neurons was incorporated to address this deficiency. Through its non-linear, time-varying synaptic transmission mechanism, this structure efficiently processes time-series data and expresses time-dependent features (FNCP). The NCP comprises a four-layer network topology, including sensory neurons V_T, intermediate neurons V_U, command neurons V_Z, and decision neurons V_l. This setup utilizes feedforward connections and highly recurrent connections to ensure network compactness and efficiency. Ultimately, the model undergoes non-linear mapping through a fully connected layer and optimizes client parameters using the cross-entropy loss function to ensure high accuracy. In the application scenario of intercultural communication education, the model is capable of real-time speech input detection on mobile devices, recognizing speech features from different cultural backgrounds, and providing immediate feedback and support. This promotes language learning and intercultural communication. By integrating a convolutional feature extractor with a compact neural architecture, not only does the model enhance the accuracy and speed of speech detection, but it also provides robust technical support for intercultural communication education, ensuring efficient operation on resource-limited mobile devices.

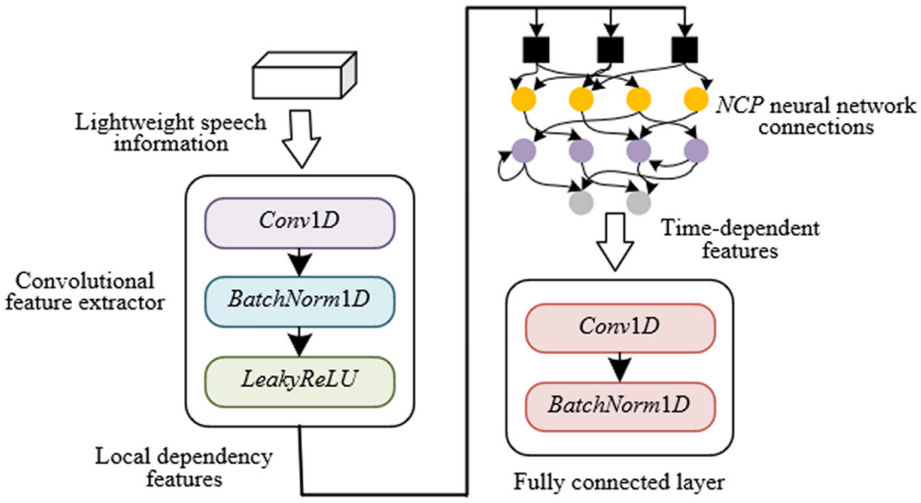

Fig. 1. Architecture of the mobile real-time speech detection model

In the constructed mobile real-time speech detection model, the NCP plays a pivotal role as a compact and efficient neural connection strategy. Initially, the convolutional feature extractor extracts local dependency features from the input speech and subsequently passes them to the sensory neurons of the NCP as perceptual input. Within the NCP, sensory neurons V_T receive features extracted by the convolution, D_Z, and transmit them to intermediate neurons V_U and command neurons V_Z for further processing, ultimately culminating in predictive outputs by the decision neurons V_l. The four-item unique connection strategy of NCP further optimizes the compactness and computational efficiency of the neural network: a) The basic network structure was constructed by incorporating V_T, V_U, V_Z, and V_l; b) Random synaptic connections, based on Bernoulli and binomial distributions, were inserted between every two consecutive layers, ensuring the network's

randomness and diversity; c) Supplementary connections were made for neurons without synaptic connections to enhance network connectivity; d) Recurrent connections were executed on the command neurons V_Z, further amplifying the propagation and integration of information within the network. To enhance detection speed, NCP utilizes closed-type *CfC* neurons for mapping. The time constant parameter vector is denoted by μ_π, the bias vector by X, the φ parameterized neural network by d, and the Hadamard product by *. The initial state of the *CfC* neurons can be determined by the following equation:

$$\frac{dA}{db} = -[\mu_\pi + d(A, D_{RE}, \varphi)] * A(s) + X * d(A, D_{RE}, \varphi) \tag{1}$$

The state update equation for the *CfC* neurons at the s-th time step is given by the following equation:

$$\begin{aligned} A(s) = {} & \delta(-d(-A, -D_{RE}; \varphi_d)s) * h(-A, -D_{RE}; \varphi_h) \\ & + [1 - \delta(-d(-A, -D_{RE}; \varphi_d)s)] * g(-A, -D_{RE}; \varphi_g) \end{aligned} \tag{2}$$

Building upon the integrated convolutional feature extractor and compact neural architecture previously described, the mobile real-time speech detection model further incorporates the concept of federated aggregation combined with differential privacy. This approach not only ensures the security of user speech data during local processing and transmission but also provides efficient and accurate speech detection services for practical applications in intercultural communication education. The framework of this model employs differential privacy algorithms to safeguard the security of local model parameters, denoted as φ_z^s, during transmission. Specifically, differential privacy protection is afforded at two critical stages: the upload phase and the broadcast phase. During the upload phase, the server initially initializes global parameters φ_0 and sets the privacy level parameters (γ, σ). After local model training, clients employ the Gaussian mechanism, denoted as H, to add noise λ_z^s to the trained parameters φ_z^s, ensuring that the magnitude of noise (δ) meets the (γ, σ)-*DP* conditions. These noised parameters, $\varphi_z^{\sim s}$, are then uploaded to the server for federated averaging aggregation. In the broadcast phase, the server updates the aggregated global parameters (φ_z^s) and, in compliance with differential privacy requirements, adds additive noise (λ_t) to obtain $\varphi_z^{\sim s}$. These new global parameters ($\varphi_z^{\sim s}$) are broadcast back to the clients, ensuring privacy protection during transmission. This method enables each round of federated training to progressively optimize the global model while safeguarding user privacy until the predetermined number of training rounds is reached, culminating in a comprehensive global speech detection model.

3 MOBILE REAL-TIME SPEECH SYNTHESIS FOR INTERCULTURAL COMMUNICATION EDUCATION

A mobile real-time speech synthesis model tailored for intercultural communication education was developed in this study, integrating pitch and energy predictors with a lightweight convolutional neural network (CNN). This model aims to facilitate smooth communication and interaction across various cultural and linguistic backgrounds through effective speech synthesis and detection techniques. It comprises four main components: a phoneme encoder, a duration predictor, pitch/energy predictors, and a decoder.

3.1 Lightweight convolution module

In the application scenario of intercultural communication education, the speech synthesis model should not only efficiently process and generate the target language's speech but also dynamically adapt to the phonetic features and rhythms of different languages. To ensure that phonetic features are effectively preserved and processed across various linguistic environments while considering the limited computational resources of mobile devices, a lightweight convolution module was introduced in this study. This module addresses the issue of missing context phoneme associations in the non-autoregressive model. The framework is depicted in Figure 2. Specifically, the lightweight convolution module in the model consists of a linear layer, a Gated Linear Unit (GLU), and lightweight convolution. The input to the module first undergoes processing by the linear layer, projecting and mapping the dimensions of the input from $V \times f$ to $V \times 2f$. This step primarily prepares for the subsequent gating mechanism. Then the output of the linear layer enters the GLU layer. The GLU layer enhances the convolution structure by incorporating a gating mechanism, where half of the input serves as the gating units and the remaining half as the input variables for the gating units, with point-wise multiplication being computed subsequently. Assuming the input variables to the gating units are denoted by A, the convolutional network by CNN, and the element-wise multiplication by matrices is denoted by ⊗, the operation equation for the GLU layer is provided as follows:

$$g_m(A) = A + CNN(A) \otimes CNN(A) \tag{3}$$

Assuming the convolution kernel is denoted by Q, the number of attention heads by G, the size of the convolution kernel by j, the input and output by A and P, respectively, the vector dimension of the input by V, the number of input channels by f, the number of output channels by z, the number of multi-head attention by G, the total number of output channels by f, which head the output belongs to by zG/f, and the percentage of channel number z in the total number of channels f by $\left[\frac{z}{f}\right]$. *LightConv* and *DeepwiseConv* are denoted by LC and DC, respectively. The output of the u-th phoneme in the lightweight convolution sequence can be obtained through the following equation:

$$LC\left(A, Q\left[\frac{zG}{f}\right], \ldots, u, z\right) = DC\left(A, \text{softmax}\left(Q\left[\frac{zG}{f}\right], \ldots\right), u, z\right) \tag{4}$$

3.2 Phoneme encoder module

In the application scenario of intercultural communication education, higher demands are placed on the speech synthesis model. The model is required not only to generate accurate speech but also to demonstrate good adaptability across different languages and cultural backgrounds. For this purpose, a phoneme encoder module was established in this study to transform input phonemes into phoneme hidden features enriched with semantic information, thereby laying a solid foundation for the subsequent speech generation process.

Specifically, the phoneme encoder utilizes an embedding layer to represent the input phonemes as a sequence of continuous one-hot vectors, transforming discrete

phonemes into continuous vector representations that can be processed by the model, thus ensuring the effective transmission of phoneme information. Then phoneme features are further extracted in conjunction with the lightweight convolution module. The role of this module is to utilize its efficient convolution operations to capture the local dependencies between phonemes, enhancing the model's generalization ability. The one-hot variables generated by the front-end network comprising the embedding layer and the lightweight convolution module subsequently enter the residual convolution module. The residual convolution module leverages the powerful feature extraction capabilities of CNNs to further extract phoneme hidden features rich in semantic information. The introduction of a residual structure not only enhances the depth of feature extraction but also effectively avoids the problem of gradient vanishing, allowing the model to capture more complex phoneme relationships. Following the front-end network and residual convolution module, the phoneme encoder's backend network consists of a lightweight convolution module and a normalization module. The design of this part aims to perform more detailed feature extraction, ensuring the robustness of the phoneme sequence representation. The normalization module plays a role in balancing feature distribution and accelerating model convergence, further enhancing the quality of phoneme representation. Assuming the input sequence is represented by $A = \{a_1, a_2, a_3, ..., a_V\}$, where a_u represents the u-th phoneme in the input text or text and V denotes the total length of the input phonemes. The output sequence of the phoneme encoder is represented by $C = \{c_1, c_2, c_3, ..., c_V\}$, and the phoneme encoding process can be characterized by the following equation:

$$C = EC(A) \tag{5}$$

3.3 Duration predictor module

Intercultural communication education involves a variety of languages and pronunciation habits, where the rhythm and phoneme duration may significantly differ across languages. For this reason, a duration predictor module was integrated into the model. The framework of this module is depicted in Figure 2. The core task of the duration predictor is to assign an appropriate number of Mel-spectrogram frames to each phoneme. By accurately predicting the duration of each phoneme, the generated speech is ensured to align more closely with the natural pronunciation patterns of the target language, thereby enhancing the naturalness and intelligibility of the synthesized speech.

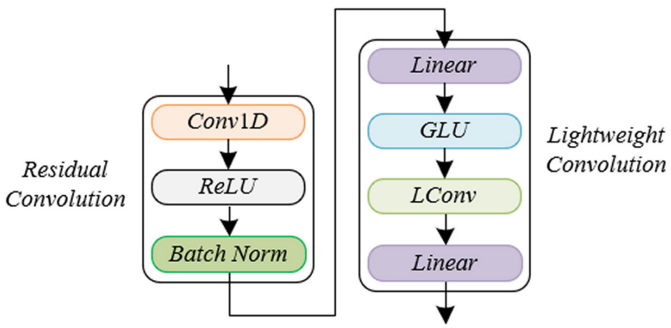

Fig. 2. Framework of the duration predictor module

The module comprises a lightweight convolution module and a residual convolution module. The former is responsible for the initial extraction of temporal features from the input phonemes, capturing the temporal dependencies between phonemes through efficient convolution operations. The latter further processes the temporal features extracted by the lightweight convolution module. The introduction of a residual structure significantly enhances the depth and complexity of the model, enabling it to capture more detailed temporal patterns. The input and output of the duration predictor are the phoneme sequence outputted by the phoneme encoder and a sequence of phoneme durations, respectively. Assuming the output phoneme duration sequence is represented by $F = \{f_1, f_2, f_3, ..., f_V\}$, where f_u represents the predicted duration for the u-th phoneme in the input sequence. The actual phoneme duration sequence is denoted by $S = \{s_1, s_2, s_3, ..., s_V\}$, where s_u represents the actual duration of each phoneme. The set of errors between the predicted and actual phoneme durations is denoted by $LOSS_{DU}$, where each error between the predicted and actual values for each phoneme is represented by m_k. The expression for the phoneme duration prediction process is provided in the following equations:

$$LOSS_{DU} = \{m_1, m_2, m_3, ..., m_v\} \tag{6}$$

$$m_k = \begin{cases} 0.5 \times (f_k - s_k)^2 & if \left| f_k - s_k \right| < 1 \\ \left| f_k - s_k \right| & othersize \end{cases} \tag{7}$$

3.4 Pitch/energy predictor module

In the application scenario of intercultural communication education, learners from different linguistic and cultural backgrounds are required to adapt to various pronunciation habits and prosodic features. The pitch and energy predictors set up in the model enrich the generated speech with more prosodic information, addressing the issue of incoherent audio generation. This enhancement makes the speech more natural in tone, stress, and rhythm, aligning it with the phonetic characteristics of the target language. The framework is depicted in Figure 4. The computation process of the pitch and energy predictors begins with the extraction of hidden sequence features through a one-dimensional convolution layer. This is followed by a non-linear transformation via a Relu activation layer. The activated hidden variables are then normalized through a normalization layer, which not only accelerates the convergence of the model but also enhances its stability. To further increase the robustness of the model, a dropout method was introduced during the computation to effectively prevent overfitting. Finally, the results are transformed into the same dimensions as the expanded hidden sequence through a linear mapping by the linear layer. Assuming the expanded phoneme hidden sequence is denoted by $J = \{j_1, j_2, j_3, ..., j_M\}$, where M represents the maximum Mel scale value of the Mel spectrogram. The output sequence of the pitch predictor is represented by $O = \{o_1, o_2, o_3, ..., o_M\}$, and the output sequence of the energy predictor is represented by $R = \{r_1, r_2, r_3, ..., r_M\}$. The pitch predictor is denoted by PP, and the energy predictor by EP. The expressions for the processes of pitch and energy prediction are given by:

$$O = PP(J) \tag{8}$$

$$R = EP(J) \tag{9}$$

3.5 Decoder module

During intercultural communication education, learners from various linguistic and cultural backgrounds should handle complex pronunciation and prosodic features. In the mobile real-time speech synthesis model designed for intercultural communication education, the decoder module combines pitch and energy information with the phoneme hidden sequence. The generated Mel-spectrogram accurately reflects the prosodic and tonal characteristics of the target language. This not only aids learners in better mimicking and mastering foreign language pronunciations but also enhances their understanding and perception of language prosody and rhythm. In the implementation process, the decoder initially receives outputs from the pitch and energy predictors, which contain the pitch and energy information for each frame. Simultaneously, the decoder receives the expanded phoneme hidden sequence that includes the temporal dynamics needed for speech generation. By integrating this input data, the decoder is capable of generating a Mel-spectrogram that conforms to the characteristics of the target language. Assuming the output of the decoder is represented by $L = \{l_1, l_2, l_3, ..., l_{80}\}$, the expanded phoneme hidden sequence by J, and the decoder itself by DE. The variable concatenation operation is denoted by CAT. The means of the Gaussian functions fitted after comparing both actual and predicted values are denoted by ω_h and ω_l, while the variances are denoted by δ_h and δ_l. The structural similarity loss function is denoted by $LOSSss$, the mean absolute error function by $LOSS_{m1}$, and the overall model loss by $LOSS_{ME}$. The decoding process can be characterized by the following equations:

$$L = DE\ (CAT\ (J, O, R)) \tag{10}$$

$$LOSS_{SS} = \frac{(2\omega_h\omega_l + c_1)(2\delta_{hl} + z_2)}{(\omega_h^2 + \omega_l^2 + z_1)(\delta_h^2\delta_l^2 + z_2)} \tag{11}$$

$$LOSS_{m_1} = \frac{\sum_{u=1}^{v}\left|d(a_u) - b_u\right|}{v} \tag{12}$$

$$LOSS_{ME} = LOSS_{SS} + LOSS_{m_1} + LOSS_{DU} \tag{13}$$

4 EXPLORATION OF A MODEL FOR INTERCULTURAL COMMUNICATION EDUCATION BASED ON MOBILE TECHNOLOGY

Based on the mobile real-time speech detection and synthesis model developed for intercultural communication education, a new educational model was explored in this study. This model fully utilizes mobile technology and speech processing techniques to provide a flexible, efficient, and interactive learning experience. Through the mobile real-time speech detection and synthesis model, learners can practice languages at any time and any place. Real-time speech detection captures the learners' pronunciations and promptly provides feedback, identifying pronunciation errors and suggesting improvements. The speech synthesis module can generate standard speech samples for learners to practice against. This instant feedback mechanism significantly improves learners' pronunciation accuracy and learning efficiency.

Each learner has a unique language learning background and needs. The speech model based on mobile technology can collect and analyze learners' speech data

to provide personalized learning content and recommendations. The system can customize pronunciation exercises, intonation adjustments, and prosodic training based on the learner's pronunciation characteristics, helping them to master the target language more effectively. With speech synthesis technology, learners can engage in simulated intercultural speech interactions. The system can synthesize speeches with different accents, speeds, and tones, simulating real intercultural communication scenarios. Learners can interact with these simulated speeches to practice various vocal communication strategies and enhance their capabilities in real intercultural interactions.

Mobile devices enable learners to study languages anytime and anywhere, unrestricted by time and place. Combined with the real-time speech detection and synthesis model, learners can utilize fragmented time for efficient language practice, greatly enhancing learning flexibility and convenience. Additionally, the mobile platform can integrate social features, allowing learners to join intercultural learning communities, share learning experiences, practice together, and correct each other's pronunciation errors. Through community interaction, learners can access more learning resources and support, creating a conducive learning atmosphere and promoting sustained and in-depth language learning.

Incorporating various media forms such as speech, text, images, and video, a rich array of learning content and diverse methods are provided. Learners can enhance their understanding of language usage and customs across different cultural backgrounds by watching videos with speech synthesis explanations and listening to standard pronunciations. This multimodal learning experience aids in deepening learners' comprehensive understanding of language and culture. By collecting speech data and learning behaviors through mobile applications, the system can analyze this data, track learners' progress, identify areas of weakness, and provide targeted learning suggestions. Learners can review their progress at any time, understand areas that require improvement, and thus formulate more effective learning plans.

This educational model not only focuses on enhancing language skills but also emphasizes the cultivation of intercultural communication abilities. By simulating real intercultural communication scenarios and providing cultural background knowledge, learners are enabled to understand and respect the communication habits and customs of different cultures, thereby enhancing their adaptability and effectiveness in intercultural interactions. The intercultural communication education model based on the mobile real-time speech detection and synthesis model leverages the convenience of mobile technology and the advancements in speech processing technology to offer a personalized, interactive, and flexible learning experience. This model not only effectively improves learners' language abilities but also nurtures their intercultural communication skills, preparing them for intercultural interactions in a globalized era.

5 EXPERIMENTAL RESULTS AND ANALYSIS

According to the data presented in Figure 3, diverse models exhibit varying performances in terms of equal error rate (EER) and the normalized tandem detection cost function (t-DCF). Initially, traditional models such as dynamic time warping (DTW) and Hidden Markov Model (HMM) show relatively poorer performance, with EERs of 13.6 and 11.2, and t-DCFs of 0.31 and 0.25, respectively. In contrast, technologies based on the Gaussian Mixture Model (GMM), particularly the GMM-universal

background model (UBM), demonstrate superior performance with an EER of 5.2 and a t-DCF of 0.14. Support vector machines (SVM) and deep neural networks (DNN) also perform well, with EERs of 6.2 and 5.7, and t-DCFs of 0.15 and 0.17, respectively. Notably, the model proposed in this study shows the best performance on both metrics, achieving the lowest EER of 5.1 and t-DCF of 0.13, significantly outperforming the other models. The experimental results indicate that deep learning and hybrid model techniques have clear advantages in mobile real-time speech detection. Particularly, the GMM-UBM and the proposed model excel in both EER and t-DCF, demonstrating their superiority in accuracy and cost-effectiveness. Traditional models such as DTW and HMM, with their higher EERs and t-DCFs, are gradually being replaced by more advanced machine learning and deep learning models. The outstanding performance of the proposed model highlights its efficiency and reliability in handling mobile speech detection tasks and is well-suited for application in real-time speech detection and synthesis technologies for intercultural communication education.

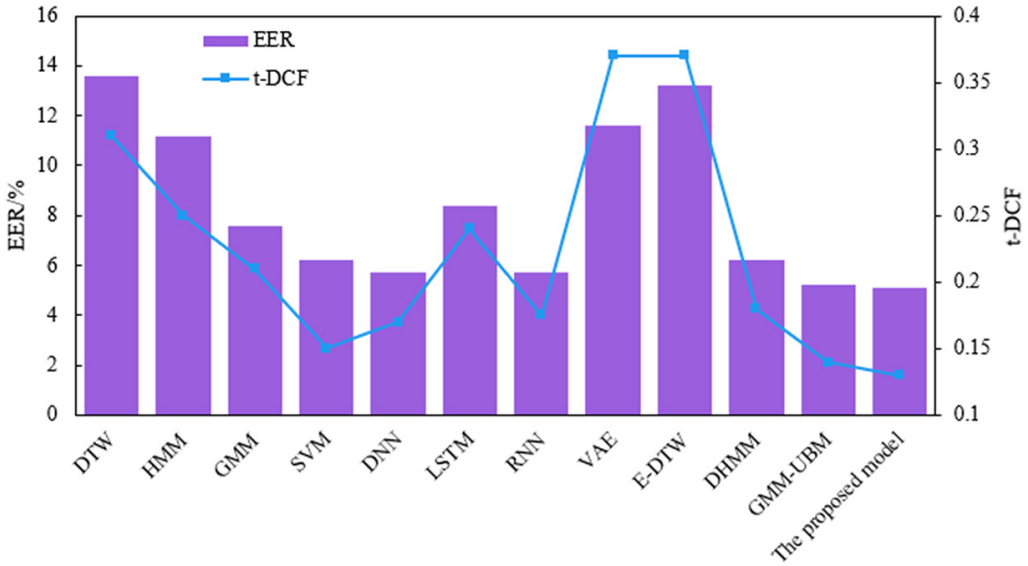

Fig. 3. Performance comparison of different mobile real-time speech detection models

The data from Table 1 shows significant differences in performance among various mobile real-time speech synthesis models in terms of the mean opinion score (MOS) and Mel spectrum distortion (MSD). Specifically, real audio achieves the highest MOS on both the test and validation sets, reaching 4.45 and 4.36, respectively. The traditional statistical parametric speech synthesis (SPSS) model scores 4.05 and 4.22 in MOS on the two datasets, respectively, with MSD scores of 5.45 and 5.78. WaveNet scores 3.74 on the test set in MOS, but significantly lower on the validation set at 2.89, with MSD scores of 5.23 and 5.56. The CycleGAN voice conversion model scores 3.87 and 3.26 in MOS on the two datasets, respectively, with MSD scores of 5.18 and 5.58. The attention-based synthesis model scores higher in MOS at 4.08 and 4.25 on the test and validation sets, respectively, with MSD scores of 5.26 and 5.89. The proposed model performs excellently, with scores of 4.12 and 4.06 in MOS and 5.25 and 5.78 in MSD on the test and validation sets, respectively. The experimental results indicate outstanding performance of the proposed model in the task of mobile real-time speech synthesis, especially noted by the MOS of 4.12 on the test set,

slightly higher than the attention-based synthesis model at 4.08. This suggests that the proposed model has a strong capability of generating high-quality speech samples, closely approaching the effect of real audio. Moreover, the model also shows relatively good performance in MSD, with scores of 5.25 and 5.78, indicating lower acoustic distortion. In contrast, the WaveNet and CycleGAN voice conversion models score lower in MOS on the validation set, particularly WaveNet at 2.89, revealing issues with adaptability across different datasets. Overall, the model proposed in this study demonstrates balanced performance across different datasets, considering both speech quality and acoustic consistency, providing reliable technical support for intercultural communication education, and enabling the generation of high-quality speech samples for learners to practice, thereby enhancing learning outcomes.

Table 1. Performance comparison of different mobile real-time speech synthesis models

Model	MOS		MSD	
	Test Set	Validation Set	Test Set	Validation Set
Real audio	4.45 ± 0.05	4.36 ± 0.05	*N/A*	*N/A*
SPSS	4.05 ± 0.07	4.22 ± 0.06	5.45 ± 0.06	5.78 ± 0.06
WaveNet	3.74 ± 0.06	2.89 ± 0.07	5.23 ± 0.06	5.56 ± 0.06
CycleGAN Voice Conversion	3.87 ± 0.06	3.26 ± 0.07	5.18 ± 0.06	5.58 ± 0.06
Attention-based Synthesis	4.08 ± 0.06	4.25 ± 0.06	5.26 ± 0.06	5.89 ± 0.06
The proposed model	4.12 ± 0.06	4.06 ± 0.06	5.25 ± 0.06	5.78 ± 0.06

In the experiments, a detailed analysis was conducted on the fundamental frequency variations of the generated audio by introducing a lightweight convolution module. The results demonstrate that the model equipped with the lightweight convolution module exhibits a higher similarity in prosodic variations, aligning more closely with the fundamental frequency trends of the actual audio. Specifically, the results displayed in Figure 4 indicate that, at the end of the audio, the fundamental frequency trend of the audio generated by the model with the lightweight convolution module significantly approximates that of the real audio, whereas the model without the lightweight convolution module deviates markedly at the same points. This indicates that the lightweight convolution module possesses significant advantages in capturing and reproducing subtle vocal prosody changes. The experimental results adequately validate the effectiveness of the lightweight convolution module in enhancing audio generation quality. The audio produced by the model is highly consistent with the real audio in terms of fundamental frequency changes, indicating that the model can more accurately simulate the prosodic features of real speech during the generation process. This finding holds significant implications for intercultural communication education, as high-quality speech synthesis can provide learners with more authentic and natural speech samples, aiding them more effectively in pronunciation comparison and improvement. Furthermore, the instant feedback functionality combined with high-quality speech synthesis technology provides a solid technical foundation for constructing a flexible, efficient, and interactive intercultural communication learning system, showcasing the vast potential and application prospects of this mobile technology in the educational sector.

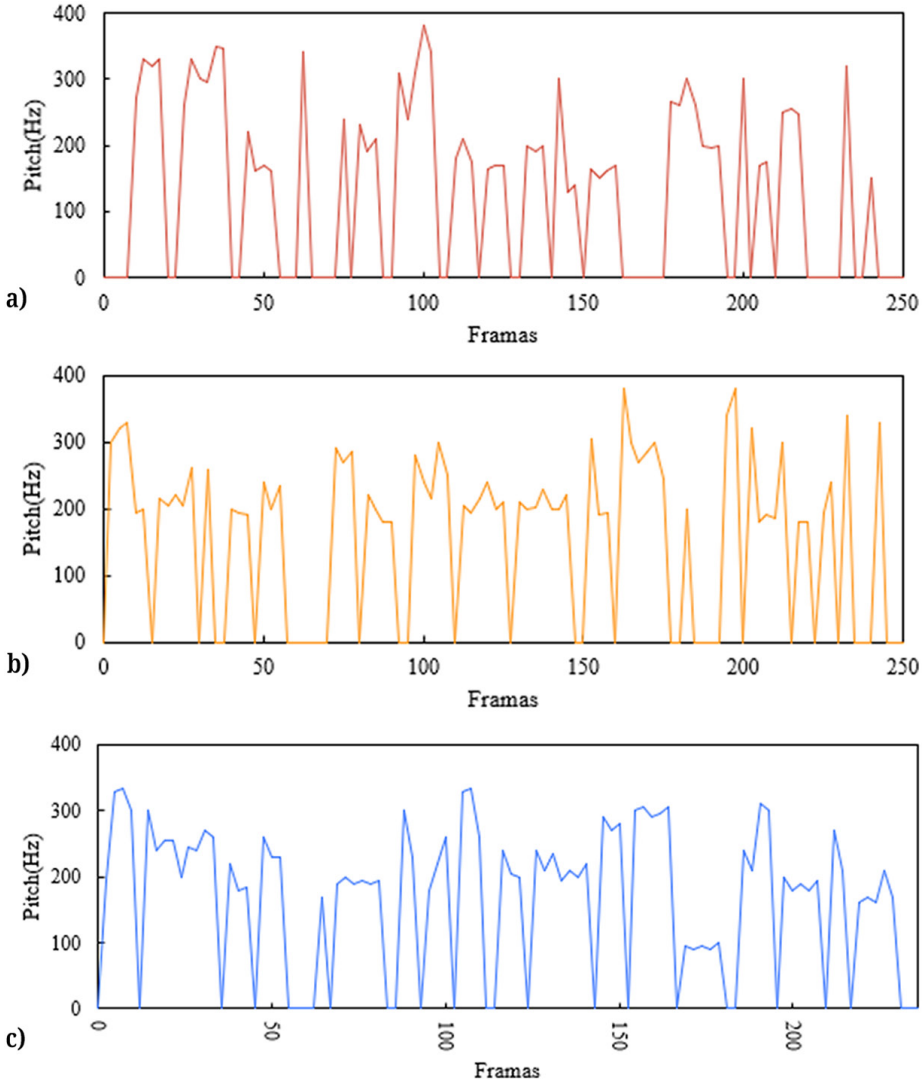

Fig. 4. Fundamental frequency variation in (a) real audio, (b) a model without a lightweight convolution module, and (c) a model with a lightweight convolution module

6 CONCLUSION

This study aims to explore an intercultural communication education model based on mobile technology, with the study primarily divided into three parts: mobile real-time speech detection and synthesis technologies, and the construction of an intercultural communication education model based on these technologies. Initially, in real-time speech detection, learners' pronunciations were accurately captured and analyzed through mobile devices, with immediate feedback and suggestions for improvement provided. Subsequently, in speech synthesis, high-quality speech samples were generated for learners to practice against. Finally, by integrating these technologies, a flexible, efficient, and interactive intercultural communication education system was constructed. The experimental results displayed the performance comparison between different mobile real-time speech detection and synthesis models. Particularly in the speech synthesis experiments, by incorporating the lightweight convolution module, the generated audio exhibited a fundamental

frequency trend that closely matched the real audio. Specific data demonstrated that the model with the lightweight convolution module significantly approximated the real audio's fundamental frequency trend at the end of the audio segment, whereas the model without this module deviated markedly. These findings validate the effectiveness of the lightweight convolution module in enhancing the quality and accuracy of generated audio.

This study introduced advanced mobile technology into intercultural communication education, providing an innovative educational model. This model not only improved learning efficiency but also enhanced interactivity and flexibility, with significant practical value. Additionally, the effectiveness of the lightweight convolution module in speech synthesis was successfully validated, offering new directions and ideas for future speech technology development. Despite the significant achievements, there are some limitations to this study. For instance, the diversity and scale of the experimental datasets were limited, which may affect the model's generalizability. A future study could expand the scale and diversity of datasets to further verify the model's robustness and applicability. Moreover, more advanced speech processing technologies, such as the integration of deep learning and natural language processing, could be explored to further enhance the precision and naturalness of speech detection and synthesis. Ultimately, the study could further incorporate practical teaching applications, conduct large-scale user testing, and collect feedback to continuously optimize and refine the educational model.

7 REFERENCES

[1] N. Mykytenko, M. Fedorchuk, O. Ivasyuta, N. Hrynya, and A. Kotlovskyi, "Intercultural communicative competence development in journalism students," *Advanced Education*, vol. 9, no. 20, pp. 121–131, 2022. https://doi.org/10.20535/2410-8286.261521

[2] H. R. F. Rahmah, N. Wahyuningtyas, N. Ratnawati, R. Marsida, M. K. A. W. Mufid, and M. H. Ibrahim "Development of 'ARCIL' media in prehistoric culture materials for junior high school students," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 16, no. 22, pp. 176–190, 2022. https://doi.org/10.3991/ijim.v16i22.36155

[3] P. Zarkova, "Methodological activities of developing intercultural communicative competence of students in Bulgarian language teaching at first high school stage," *Bulgarski Ezik I Literatura-Bulgarian Language and Literature*, vol. 61, no. 1, pp. 55–68. 2019.

[4] R. Arumugam and N. Md Noor, "Mobile apps based on Keller personalized system of instruction to promote English vocabulary acquisition," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 15, no. 23, pp. 4–17, 2021. https://doi.org/10.3991/ijim.v15i23.27227

[5] I. Zulaeha, Subyantoro, C. Hasanudin, and R. Pristiwati, "Developing teaching materials of academic writing using mobile learning," *Ingénierie des Systèmes d'Information*, vol. 28, pp. 409–418, 2023. https://doi.org/10.18280/isi.280216

[6] A. Trifunović, S. Čičević, T. Ivanišević, S. Simović, and S. Mitrović, "Education of children on the recognition of geometric shapes using new technologies," *Education Science and Management*, vol. 2, no. 1, pp. 1–9, 2024. https://doi.org/10.56578/esm020101

[7] C. Lytridis *et al.*, "Audio signal recognition based on intervals' numbers (INs) classification techniques," in *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 2019, pp. 1–4. https://doi.org/10.1109/IISA.2019.8900749

[8] P. Arisaputra and A. Zahra, "Indonesian automatic speech recognition with XLSR-53," *Ingénierie des Systèmes d'Information*, vol. 27, pp. 973–982, 2022. https://doi.org/10.18280/isi.270614

[9] C. Puebla, T. Fievet, M. Tsopanidi, and H. Clahsen, "Mobile-assisted language learning in older adults: Chances and challenges," *ReCALL*, vol. 34, no. 2, pp. 169–184, 2022. https://doi.org/10.1017/S0958344021000276

[10] R. Ramadania, Y. Hartijasti, B. B. Purmono, D. M. N. Haris, and M. Z. Afifi, "A systematic review on digital transformation and organizational performance in higher education," *International Journal of Sustainable Development and Planning*, vol. 19, pp. 1239–1252, 2024. https://doi.org/10.18280/ijsdp.190402

[11] M. Hawamdeh and E. Soykan, "Systematic analysis of effectiveness of using mobile technologies (MT) in teaching and learning foreign language," *Online Journal of Communication and Media Technologies*, vol. 11, no. 4, 2021. https://doi.org/10.30935/ojcmt/11256

[12] S. Xue, "A conceptual model for integrating affordances of mobile technologies into task-based language teaching," *Interactive Learning Environments*, vol. 30, no. 6, pp. 1131–1144, 2022. https://doi.org/10.1080/10494820.2019.1711132

[13] W. Pengnate, "Students' attitudes and problems towards the use of mobile-assisted language learning (MALL)," in *2018 5th International Conference on Business and Industrial Research (ICBIR)*, 2018, pp. 590–593. https://doi.org/10.1109/ICBIR.2018.8391266

[14] Q. Xu, D. Sun, and Y. Zhan, "Embedding teacher scaffolding in a mobile technology supported collaborative learning environment in English reading class: Students' learning outcomes, engagement, and attitudes," *International Journal of Mobile Learning and Organization*, vol. 17, no. 1–2, pp. 280–302, 2023. https://doi.org/10.1504/IJMLO.2023.128340

[15] A. Krasulia and K. Saks, "Students' perceptions towards mobile learning in an English as a foreign language class," in *2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT)*, 2020, pp. 238–240. https://doi.org/10.1109/ICALT49669.2020.00078

[16] G. Dzhumayov, "Attitudes of learners towards the use of technology in language learning," *Chuzhdoezikovo Obuchenie-Foreign Language Teaching*, vol. 48, no. 5, pp. 477–485, 2021. https://doi.org/10.53656/for21.53nagl

[17] M. M. Elaish, M. H. Hussein, and G. J. Hwang, "Critical research trends of mobile technology-supported English language learning: A review of the top 100 highly cited articles," *Education and Information Technologies*, vol. 28, pp. 4849–4874, 2023. https://doi.org/10.1007/s10639-022-11352-6

[18] A. Togaibayeva, D. Ramazanova, M. Yessengulova, A. Yergazina, A. Nurlin, and R. Shokanov, "Effect of mobile learning on students' satisfaction, perceived usefulness, and academic performance when learning a foreign language," *Front. Educ.*, vol. 7, 2022. https://doi.org/10.3389/feduc.2022.946102

[19] L. Bradley, L. Bartram, K. W. Al-Sabbagh, and A. Algers, "Designing mobile language learning with Arabic speaking migrants," *Interactive Learning Environments*, vol. 31, no. 1, pp. 514–526, 2023. https://doi.org/10.1080/10494820.2020.1799022

8 AUTHORS

Huajie Chen is an Associate Professor and Master's Supervisor in the School of Foreign Languages and Cultures at Tibet University. His research focuses on TESOL and intercultural studies (E-mail: hjchen@utibet.edu.cn; ORCID: https://orcid.org/0009-0005-8148-6077).

Kehan Mei holds a Ph.D. in literature from the University of Texas at Dallas and currently serves as a Lecturer in the School of Foreign Languages and Cultures at Tibet University. Her research areas encompass Asian American literature, comparative literature, and Tibetan literature (E-mail: mkh318@utibet.edu.cn; ORCID: https://orcid.org/0000-0001-9128-0375).