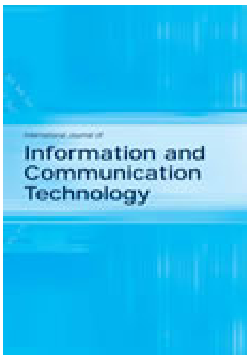


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

An automatic fluency evaluation method for broadcast hosting speech: autoregressive speech LLM

Bingyuan Li

DOI: 10.1504/IJICT.2025.10074862

Article History:

Received:	25 September 2025
Last revised:	22 October 2025
Accepted:	25 October 2025
Published online:	12 December 2025

Copyright © 2025 Inderscience Enterprises Ltd.

An automatic fluency evaluation method for broadcast hosting speech: autoregressive speech LLM

Bingyuan Li

Xiangshan Film and Television College,
Ningbo University of Finance and Economics,
Ningbo, 315175, China
Email: 18857460686@163.com

Abstract: Oral fluency is a key indicator for evaluating the professional skills of broadcast hosting. To address the current research gap in modelling deep semantic associations for spoken fluency, this paper first utilises Res2Net for multiscale feature extraction from broadcast hosts' speech. Subsequently, a pause prediction module is proposed. This module predicts multiple types of pause labels based on the original text. It then predicts a Gaussian mixture distribution for each phoneme and achieves diverse phoneme durations through random sampling. Finally, an autoregressive large language model and a discriminative module based on transformer are proposed. This module is applied at each time step of the autoregressive process and prevents misalignment phenomena via the transformer and judging mechanism. Experimental results show that the proposed model achieves an evaluation accuracy of 93.35% and a word error rate of 0.7%, enabling high-accuracy fluency evaluation for oral speech.

Keywords: spoken fluency assessment; feature extraction; Res2Net model; autoregressive large language model; transformer model.

Reference to this paper should be made as follows: Li, B. (2025) 'An automatic fluency evaluation method for broadcast hosting speech: autoregressive speech LLM', *Int. J. Information and Communication Technology*, Vol. 26, No. 43, pp.78–94.

Biographical notes: Bingyuan Li is a Professor in the Broadcasting Host Department at Ningbo Institute of Finance and Economics, China. He received his Bachelor's degree from Southwest University for Nationalities, China in 2005. His research interests are interviews with people in new media, spoken fluency assessment, spoken language and paralanguage.

1 Introduction

In the current context of rapid development in the media industry, broadcasting and hosting, as key carriers for information dissemination and cultural transmission, have their professional capabilities directly impacting communication effectiveness and audience experience. Whether it is the rigor and fluency of news broadcasting, the natural fluidity of variety show hosting, or the calm and coherent narration in special features, oral fluency plays an indispensable role (Shutian, 2024). Early automatic evaluation methods were primarily based on low-level features of audio signals, such as speaking

Copyright © The Author(s) 2025. Published by Inderscience Publishers Ltd. This is an Open Access Article distributed under the CC BY license. (http://creativecommons.org/licenses/by/4.0/)

rate, pause duration, and the number of filler words for analysis. Although these methods automated the evaluation process, these single-feature metrics struggled to comprehensively and accurately reflect the complex nature of oral fluency in broadcasting and hosting (Lv et al., 2022). Oral fluency in broadcasting and hosting involves not only surface-level aspects like speech rhythm and pause rationality, but also deeper dimensions such as logical coherence in language expression, semantic continuity, and contextual adaptability. Clearly, traditional methods fail to effectively capture and evaluate these dimensions (Liu et al., 2023). Therefore, developing a method that can comprehensively, accurately, and efficiently evaluate oral fluency in broadcasting and hosting has become a crucial issue to address in the current media industry and artificial intelligence field (Pakhomov et al., 2015).

Traditional research conducts automatic oral fluency assessment based on statistical models. Sharma et al. (2019) modelled the fluency between acoustic features of speech signals and corresponding text units based on hidden Markov models (HMMs) (Helske and Helske, 2019). Yu (2024) conducted automatic oral fluency assessment of English speech based on HMM and Gaussian mixture models (GMMs), but the evaluation accuracy was not high. Early oral fluency assessment methods typically used Markov models for acoustic modelling, but the GMM assumptions often did not align with actual situations, thus affecting model performance and fitting effects. Compared to traditional statistical methods such as Markov models, deep learning-based automatic oral fluency assessment methods leverage powerful feature learning, context modelling, and semantic understanding capabilities to break through the limitations of conventional approaches and better meet the core needs of oral fluency assessment. Sharma et al. (2023) proposed a Hindi fluency assessment model combining deep belief networks with HMM, but it was only applicable to speech recognition with small vocabularies. Alashban et al. (2022) combined convolutional neural network (CNN) with long short-term memory network (LSTM) to complete oral fluency assessment modelling, achieving an evaluation accuracy of 78.2%. HMM assumes that state transitions and observation generation follow a linear or simple nonlinear relationship, making it difficult to capture high-dimensional and nonlinear dynamics. Deep learning models directly learn acoustic features from the original audio waveforms and output text end-to-end. In noisy environments, their accuracy is significantly better than that of HMM models. In addition, speech recognition accuracy was significantly improved, but problems such as high model complexity and deployment difficulty remained, and the high latency made true real-time streaming speech recognition difficult to achieve.

End-to-end oral assessment is a speech recognition technology based on encoder-decoder architecture (Kang et al., 2024). It introduces a blank label, directly mapping feature sequences to phonemes or words, simplifying the training process and solving the problem of forced alignment. Garain et al. (2021) enhanced oral fluency assessment performance by employing an attention mechanism that helps the model attend to specific segments of the audio input when generating the corresponding text. By combining the local feature extraction of CNNs with the global feature capture of transformers, Song et al. (2022) developed the conformer model to leverage the advantages of both for assessing oral fluency. Li et al. (2024) utilised self-attention mechanisms to capture global dependencies and a convolutionally-gated multilayer perceptron (MLP) module to extract local relationships, thereby enhancing the prediction efficiency of the assessment model. Autoregressive large language models (LLMs) predict each word in generated text based on all previous content. This 'autoregressive generation' pattern makes the oral assessment results generated (such as fluency scores, issue localisation) more logically coherent (Wang et al., 2024). Chandrabanshi and Domnic (2024) designed a fully probabilistic and autoregressive oral assessment model based on the concept of PixelCNN, significantly improving the accuracy of evaluation. Dhahbi et al. (2025) proposed a method for oral fluency assessment that combines autoregressive generation models and attention mechanisms. This method directly models the mapping relationship between input text and acoustic feature Mel-spectrograms through a neural network. Zhao et al. (2024) designed a speaking fluency assessment method based on autoregressive encoders. Experiments showed that this method is effective. Noh and Park (2024) designed a spoken language assessment model based on CNN and autoregressive models. They extract spoken language features through CNN and achieve spoken language assessment through the autoregressive model, thus improving the assessment accuracy.

According to the current research analysis, existing studies find it difficult to capture cross-sentence and cross-paragraph context dependencies when assessing the fluency of spoken language sequences for news anchors, leading to relatively low accuracy in spoken language fluency assessment. To address the above challenges, this paper proposes a news anchor spoken fluency assessment model based on autoregressive speech LLMs and transformers. First, spoken language signal analysis methods are used for feature extraction. The extracted spoken pronunciation features are adaptively matched to realise the mining of spoken pronunciation signals. A multi-wavelet decomposition approach is adopted to decompose the spoken pronunciation signal features, and the Res2Net model is used for multi-scale deep feature extraction. Then, evaluation modelling is carried out from two aspects: word-level pauses and phoneme duration. To enhance the diversity and accuracy of word-level pauses, a pause prediction module is proposed in the text front end. This module predicts multiple pause labels based on the original text, thereby providing an accurate reference for pause duration modelling in spoken fluency assessment. To improve the naturalness of phoneme duration, a duration prediction module is proposed. This module predicts a Gaussian mixture distribution for each phoneme and obtains diverse phoneme durations through random sampling. To enhance the stability of phoneme duration modelling in autoregressive models, an autoregressive LLM and transformer discriminative module are proposed. This module is applied at each time step of the autoregressive LLM and avoids alignment disorder through the transformer and judgement mechanism. Simulation experiments were conducted on public datasets. The results show that the evaluation accuracy of the suggested model is at least improved by 1.67% compared to the baseline model, and the word error rate is at least reduced by 0.9% compared to the baseline model, which can effectively enhance the stability of the spoken fluency assessment model, thus improving the model's evaluation performance.

2 Relevant theory

2.1 Autoregressive theory

In a speech or noise signal, the value of any given sample can be predicted from a linear combination of a number of its immediate predecessors, which reflects the autoregressive characteristic of the signal and can be described by an autoregressive model (Savchenko and Savchenko, 2024). As one of the most popular models for characterising signal properties, the autoregressive model provides a set of simple signal model parameters that accurately express the signal's spectral magnitude, making signal spectrum estimation more simple and effective. Due to the strong correlation between speech or noise signal samples, the current or future sample values can be approximated or predicted by the linear combination of past sample values. This time series regression model is called an autoregressive model (Aibinu et al., 2012). The core logic of the autoregressive model is to predict the value of a signal at a future moment using its past value, and its representation process directly models the dependency relationship of the signal in the temporal dimension. Compared with other signal representation models, the core advantage of the autoregressive model as a signal representation model lies in its explicit modelling ability for temporal dependencies and the interpretability of the generation process.

Assume $s(n)$ as the speech sample value at time n, and α_p denotes the p^{th} coefficient in the autoregressive model. The current sample value predicted by the past p sample values can be expressed as follows:

$$\hat{s}(n) = \sum_{i=1}^{p} \alpha_i s(n-i) \tag{1}$$

Then, the prediction error between the true value and the predicted value can be expressed as below:

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{i=1}^{p} \alpha_i s(n-i) \tag{2}$$

The key issue in autoregressive model analysis is how to obtain a set of autoregressive model coefficients α_i (i = 1, 2, ..., p) to accurately describe such an autoregressive characteristic of the signal. This set of autoregressive model coefficients is often referred to as linear prediction coefficients. In theory, the least mean square error $E[e^2(n)] = \sum_n e^2(n)$ criterion is usually adopted for solving. By minimising the mean square error between the predicted value and the true value, the following system of linear equations can be obtained, where $r(j)$ is the autocorrelation sequence of the speech signal, and E_p is the prediction error energy.

$$\begin{cases} r(j) - \sum_{i=1}^{p} \alpha_i r(j-i) = 0, & 1 \le j \le p \\ r(0) - \sum_{i=1}^{p} \alpha_i r(i) = E_p, & \text{others} \end{cases} \tag{3}$$

2.2 Transformer model

Transformer is a prominent deep learning architecture that addresses a wide range of natural language processing tasks. Unlike traditional recurrent neural network (RNN) and CNN, the transformer employs a self-attention mechanism that enables it to capture long-range dependencies within input sequences, allowing it to achieve higher efficiency and performance when processing long text sequences. The characteristics of the transformer have also enabled it to play an important role in speech recognition.

Currently, well-performing deep learning-based speech recognition models all incorporate the transformer model. The transformer operates as an end-to-end model through its encoder-decoder structure: the encoder first processes the input into a contextual representation, and the decoder then generates the output sequence from this representation (Bashiri and Naderi, 2024).

The transformer has many unique features. Since the transformer does not include position information of the input sequence, which is crucial for natural language processing, especially speech recognition (Song et al., 2022), the transformer introduces the concept of positional encoding. By adding positional encoding to the input sequence to preserve the position information of words in the sentence, the model is able to distinguish words at different positions.

This model pioneered the use of the self-attention mechanism, a defining characteristic that enables simultaneous attention to information across all positions in the input sequence, rather than processing sequentially in temporal order like RNNs, allowing it to process input sequences in parallel, which significantly improves computational efficiency. To enhance model performance, the transformer extends the self-attention mechanism into multiple parallel attention heads, with each head handling different representation vectors of the input sequence. The results from these attention heads are combined and then passed through a linear transformation to produce the ultimate output sequence. To mitigate gradient vanishing and exploding issues during training, the transformer includes residual connections and layer normalisation at each sub-level, such as the self-attention level and the feed-forward neural network level.

3 Analysis of oral signals and feature extraction in broadcasting and hosting

3.1 Noise reduction filtering preprocessing for spoken signals

To automatically evaluate the fluency of radio broadcasting speech, the purpose of spoken pronunciation signal analysis is to extract acoustic features. The extracted pronunciation features are dynamically weighted and combined to enhance the model's focus on the most relevant aspects of the speech signal. The multi-wavelet decomposition method decomposes spoken pronunciation signals into their constituent features. Utilising Res2Net (Gao et al., 2019), multi-scale feature extraction is performed based on the spoken pronunciation position, vowel categories, and spectral characteristics. In this study, audio sensors are used to collect spoken pronunciation signals from radio broadcasts. Let the initial input spoken pronunciation signal characteristic sequence be $x = [x(0), \ldots, x(N-1)]$, in which $x(n)$ represents a finite-length discrete spoken pronunciation signal, $0 \leqslant n \leqslant N-1$. The discrete Fourier transform (DFT) is applied to the spoken pronunciation feature sequence x as follows:

$$X(k) = \sum\nolimits_{n=0}^{N-1} x(n) \exp\left(-j\frac{2\pi}{N}nk\right) \tag{4}$$

where k represents the length of the spoken pronunciation signal. Let the signal $x(n)$ undergo discrete orthogonal wavelet transformation, and let $X = DFT\{x\}$ stand for the DFT of the finite-time sequence x for spoken pronunciation features, that is $X = [X(0), \ldots, X(N - 1)]$.

According to the wavelet decomposition process, the finite-length spoken pronunciation signal undergoes feature reconstruction, and different resolution $j = 0, 1, \ldots, M$ spoken pronunciation signals $E_j = \sum_k |C_j(k)|^2$ are reconstructed. For integer N_0, the voice recording signal N_1 at level $v_0(n)$, $v_1(n)$, has a length of $C_j(k) = [x(t), \varphi_{j,k}(t)]$.

The collection of spoken pronunciation signals is performed using audio sensors. The multi-layer wavelet transform is applied to spoken pronunciation signals for multi-scale feature decomposition and filtering (Popov et al., 2018). Modulated pulses are employed to decompose spoken pronunciation signals into wavelet-based features. Let the scale coefficients of signal decomposition be $N_0^{(j)} \approx \alpha^j N$, $N_1^{(j)} \approx \alpha^{j-1}\beta N$, N is the number of frames. The spoken pronunciation signal is segmented into frames. For Z_n, the length of the input signal of the j^{th} level filter bank is represented by $N_0^{(j)}$, $N_1^{(j)}$, which indicate the phonation duration of the spoken pronunciation. The impulse modulation variable R^N is introduced. Linear coding incorporates a signal component phase rotation technique. To achieve this, the rotational inertia of the output speech signal is derived as below:

$$angle\left(gX^N\right) = \left(angle\left(X^N\right) + \varphi_g\right)\text{mod}(2\pi) \tag{5}$$

The spoken pronunciation signal is recombined, and the multi-layer wavelet feature scale transformation method is used for noise reduction. Maximum likelihood detection is performed on the parameters φ_g, R^N and W^N using arithmetic coding. The positive correlation feature of the speech signal output is obtained as $gX^N = |g|R^N$. By combining with equation (5), $Z^N = |g|R^N + W^N$ is obtained.

3.2 Multi-scale feature extraction of spoken speech signals based on Res2Net

After noise reduction and filtering preprocessing of the spoken signals, the speech signals are segmented into frames and windowed, and the short-time Fourier transform is used to obtain a Mel-spectrogram. For the multi-scale audio feature extraction in the Mel-spectrogram, the Res2Net network is used. Res2Net is a variant of the ResNet network. Unlike the currently common practice of deepening and widening the network structure, Res2Net adopts a parallel branch structure. Res2Net replaces the single 3 × 3 convolution in the traditional ResNet with multiple sets of parallel small convolution kernels. The output of each set of convolution is superimposed with the input of the next set to form a hierarchical residual connection. This design enables the receptive field of each branch to expand step by step, forming a multi-scale feature representation from the local to the global. By simply modifying the residual blocks, Res2Net is able to perform multi-scale feature extraction of feature maps, thus capturing more detailed speech features.

For the Mel-spectrogram F_m, it is input into the Res2Net network to extract multi-scale spoken features. To adjust the feature dimensions and further enhance the feature representation ability, a one-dimensional convolutional layer is also added. The multi-scale feature extraction process is shown in equation (6), where F_m' represents the finally extracted multi-scale features of the Mel-spectrogram, and $Conv_{1D}$ represents the one-dimensional convolution.

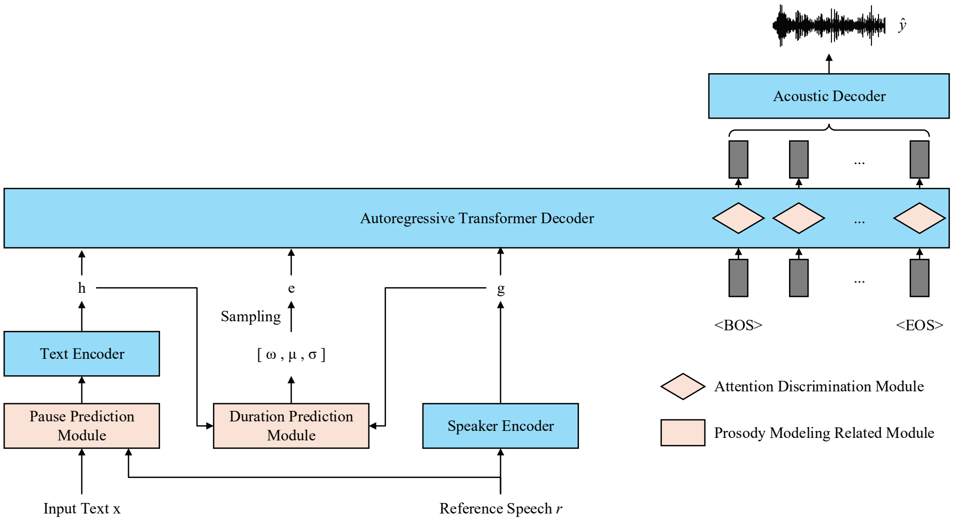
$$F_m' = Conv_{1D}\left(\text{Res2Net}\left(F_m\right)\right) \quad (6)$$

4 Automatic evaluation of spoken fluency in broadcasting and hosting based on autoregressive transformer

4.1 Overall structure of the model

Existing studies have difficulty capturing cross-sentence and cross-paragraph contextual dependencies when assessing the fluency of broadcasting spoken sequences, leading to low accuracy in fluency evaluation. To address this, this paper proposes an automatic assessment model for broadcasting spoken fluency based on autoregressive transformer. The entire framework of the model proposed in this article is shown in Figure 1. The main structure of the model can be classified into four parts: text encoder, autoregressive transformer decoder, acoustic decoder, and spoken fluency assessment. The prosodic modelling part mainly includes pause prediction module, duration prediction module, and attention discrimination module.

Figure 1 The structure of the automatic assessment model for the fluency of oral communication in broadcasting and hosting (see online version for colours)



For the input speech text sequence x, the pause prediction module is first designed using the spoken pronunciation features extracted in the previous chapter to predict inter-word pauses, and the output pause labels are inserted into x. Then, a pretrained word-to-phoneme model is used to convert the text with inserted pauses into a phoneme sequence, which is then input into the transformer-based text encoder to obtain the text vector h. For r, it is used as the input to a pretrained broadcasting host encoder to obtain the voiceprint vector g. Then, h and g are used as inputs to the duration prediction module to output the duration distribution $[w, \mu, \sigma]$. Subsequently, sampling $[w, \mu, \sigma]$ is performed to obtain the duration vector e. Finally, h, g and e are input into the transformer-based autoregressive decoder, which autoregressively generates the acoustic representation sequence. During pre-training, the attention mechanism is used to achieve implicit speech-text alignment. At the same time, decoding efficiency and performance can be adjusted by dynamically setting the masking length, thereby achieving automated assessment of broadcasting spoken fluency.

4.2 Pause prediction module

In automated assessment of spoken fluency, models typically model pauses based on punctuation marks in the input text sequence, namely punctuation-induced pauses (PIPs). However, this approach neglects non-punctuation pauses (RPs) on one hand, and lacks classification for pause durations on the other. Based on the above observations, this paper divides pauses into categories according to different durations and separately predicts multi-class labels for both PIPs and RPs at the word level.

The proposed pause prediction module is a classification language model. x is the original text composed of words and punctuation marks. Since pause prediction is based on the word level, the training data cannot cover all words, so it is necessary to use a tokenisation tool to process x, breaking down complex words in x into subwords substitution. For example, bookshelf is divided into book and shelf, and then shelf substitutes bookshelf. Through tokenisation operations, the requirement for vocabulary coverage in training data can be significantly reduced. To improve the model's ability to understand contextual information in language, this paper uses a pre-trained bidirectional encoder representations from transformer (BERT) model and a bidirectional gated recurrent unit (GRU) to process the text sequence. A broadcasting host modulation module is inserted between them, with the purpose of providing the model with personalised habit information of the target speaker using the reference speech r of the current speaker. The output of the bidirectional GRU is passed through a softmax to obtain the final classification probability vector p. Traditional multimodal fusion methods ignore temporal information, resulting in feature jumps between frames. GRU retains the multimodal feature states of historical frames through a loop structure, ensuring that the fusion features of the current frame maintain temporal continuity with those of the previous frame. In the tracking of fast-moving targets, GRU can utilise the depth information of the previous frame to correct the positioning deviation of the RGB features in the current frame, reducing target loss.

During training, a loss function is calculated between p and the ground-truth pause labels. This paper uses the Montreal Forced Alignment model (MFA) (Mahr et al., 2021) to obtain all silence positions and durations in the real audio and classify them into different categories based on the stated criteria, thereby obtaining the ground-truth pause labels. The number of different pause categories is imbalanced, so this paper adopts a weighted cross-entropy loss function (WCE) (Wu et al., 2024), which dynamically adjusts the weights of different categories according to the sample count of each category during training, to prevent the model from ignoring the categories with few samples. WCE is expressed as follows, where u is the ground-truth label vector, $\hat{u}$ is the forecasting probability vector, w_k is the weight vector, C is the vector dimension.

$$L_{WCE}\left(u,\ \hat{u}\right) = -\sum_{0}^{C-1} w_k u_k \ln\left(\hat{u}_k\right) \tag{7}$$

4.3 Duration prediction module

Duration prediction is typically a subtask of automatic spoken language fluency assessment, with the goal of predicting the duration of each phoneme based on the input phoneme sequence. Traditional duration prediction modules usually predict a set of definite phoneme durations, which greatly limits the diversity of prosodic patterns. Based on the above observations, this paper improves the diversity of duration modelling by predicting a mixture of Gaussian distributions for phoneme durations.

The text vector *h* is first input into two-layer feature extraction networks, with each layer consisting of a convolutional layer, conditional layer normalisation, and dropout. Among these, conditional layer normalisation introduces the spoken language pronunciation feature vector g, using the announcer's information to guide the model to predict a more personalised duration distribution in line with pronunciation habits. *g* is used to obtain the weights *s* and biases *b* of the conditional layer normalisation through convolutional and linear layers, and *s* and *b* are used to modulate the current features h_i to obtain the output features h_o, where u_i and v_i are the mean and standard deviation of the current characteristics, respectively.

$$h_o = s * \frac{h_i - u_i}{v_i} + b \tag{8}$$

In order to better handle the context dependencies in sequences, a bidirectional GRU further processes the output of the previous network, resulting in the vector group [α, *m*, *n*]. Since the sum of the component weights in the Gaussian mixture distribution needs to be 1, and each component variance must be a positive value, a nonlinear transformation is needed for [α, *m*, *n*]. Based on the characteristics of the softmax function and the exponential function, the nonlinear transformation is expressed as follows, where w_i, μ_i and σ_i^2 are the weights, mean, and variance of the *i*th Gaussian component, respectively, and *G* is the number of Gaussian components.

$$\mu_i = m_i \tag{9}$$

$$\sigma_i^2 = \exp(n_i) \tag{10}$$

$$w_i = \frac{\exp(\alpha_i)}{\sum_{k=1}^{G} \exp(\alpha_k)} \tag{11}$$

In the inference stage, the model samples a duration vector e from the predicted Gaussian mixture distribution [*w*, μ, σ], providing duration references for the automatic spoken language fluency assessment using the autoregressive transformer. In the training stage, the duration prediction module is trained in a supervised manner with external alignment information, as shown in Figure 4. First, a pre-trained MFA model processes the real speech y, obtaining elements in the phoneme duration sequence as scalars. To meet the requirements of supervised training [*w*, μ, σ], variational data augmentation is needed to upsample *d*. This paper uses a 4-layer WaveNet (Du et al., 2021) as the model structure for variational data augmentation. This model takes d as conditional input and generates a high-dimensional vector group v from Gaussian noise *N* (where *d* and *v* have the same temporal resolution). Then, *d* is concatenated with v along the feature dimension to obtain the duration vector *e*. This paper uses the negative log-likelihood loss function (Li et al., 2023) to constrain the training of the duration prediction module, expressed as follows:

$$L_d = \sum_{i=1}^{K} -\ln\left(\sum_{k=1}^{G} w_{k,i} N\left(e_i ; \mu_{k,i}, \sigma_{k,i}^2\right)\right) \tag{12}$$

where $w_{k,i}$, $\mu_{k,i}$ and $\sigma_{k,i}^2$ represent the weight, mean, and variance of the *k*th Gaussian component corresponding to the *i*th phoneme, e_i represents the *i*th random variable.

4.4 Automatic evaluation of spoken fluency based on autoregressive transformer

Autoregressive speech synthesis models perform prosody alignment in light of the association between input characteristics at each time step and the text (phoneme) vector, allowing autoregressive models to better focus on the context and sequence dependency. However, in practice, training struggles to achieve perfect generalisation, which means that during inference, alignment disorder may occur. Based on the above observations, this paper addresses the alignment disorder issue by using an attention discrimination module.

Figure 2 demonstrates the architecture of a single time step within the autoregressive transformer decoder, including the attention discrimination module (within the dashed box) and the transformer decoding module. The automatic spoken fluency evaluation module has three inputs: the input representation at the *k*th time step $i_k \in R^{C_1}$, the text vector $h \in R^{T \times C_2}$, and the duration vector $e \in R^{T \times C_3}$.

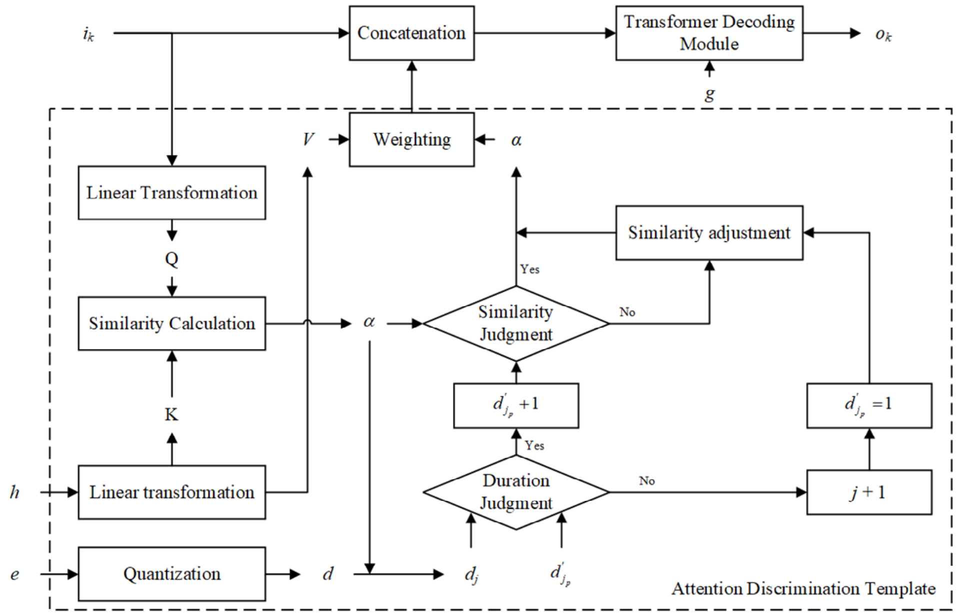
First, perform linear transformation to convert i_k and *h* into three different feature domains $Q = i_k W_q$, $K = hW_k$ and $V = hW_v$. Then calculate the similarity to obtain the attention weights α of i_k on *h*, as shown below, where *C*′ is the feature dimension.

$$\alpha = \text{Softmax}\left(\frac{QK^T}{\sqrt{C'}}\right) \tag{13}$$

Next, the attention discrimination module performs duration judgement. The duration vector e, sampled from a mixed Gaussian duration distribution, is quantised to obtain the phoneme duration sequence *d* (each element is a scalar). Then, based on the phoneme index j with the highest similarity in α, d_j is selected from *d*, and d_j is the reference duration at the current time step. In addition, the model constantly maintains a cumulative duration d'_{j_p}, representing the cumulative duration of the most similar phoneme j_p in the previous step. The duration judgement has two criteria. One is to determine whether the most similar phoneme *j* at the current time step is the same as the most similar phoneme j_p at the previous time step. The second is to determine whether d'_{j_p} is less than d_j. Only when $j = j_p$ and $d'_{j_p} < d_j$ is satisfied will the duration judgement pass. This means that the content information at the present time step is correct and meets the requirement of not exceeding the reference duration. After passing, $d'_{j_p} + 1$ is executed, and the similarity judgement continues; if the duration judgement fails, the most similar phoneme

$j_p + 1$ will be updated, and d'_{j_p} will be reset to 1 $(d'_{j_p} = 1)$, then the similarity evaluation will proceed.

Figure 2 Autoregressive transformer decoder



The similarity judgement is only executed if the duration judgement passes. This paper sets a minimum threshold β for maximum similarity to constrain autoregressive generation, while the similarity judgement criterion is: to determine whether the maximum similarity α_j in α is greater than β. If $\alpha_j > \beta$, the judgement passes, and α will become the phoneme vector weight α' at the current time step. If the similarity judgement fails, a similarity adjustment will be conducted. Similarity adjustment relies on β. By introducing threshold β and similarity adjustment, similarity judgement controls the weight of the most similar phoneme to a higher value, avoiding errors in pronunciation fluency evaluation. In addition, the relative weight relationships of other phonemes are preserved, maintaining context-dependent relationships.

After obtaining α, α is weighted with V to obtain the weighted phoneme vector at the current time step. This vector serves as content information for autoregressive generation. It is concatenated with i_k as the input to the transformer decoding module, finally generating the output representation o_k at the current time step.

For each audio of spoken language and its corresponding reference text r, if the goal is to obtain the likelihood score of spoken fluency, this work masks a certain phoneme in the pronunciation, retaining the remaining phonemes and audio data. The objective function is defined as follows:

$$L\left(y_i \middle| X, Y_i; \theta\right) = \log P\left(y_i \middle| X, Y_i; \theta\right) \tag{14}$$

By calculating the above equation, the score of the evaluated phoneme can be obtained. To obtain the scores of all phonemes in the entire sentence, duration and pause prediction can be performed for each different phoneme. By providing all phonemes in the reference text except for the phoneme to be predicted, the mask token can automatically capture the parts that should align with the acoustic features, which are then used for prediction. Since the proposed model adopts a parallel decoding strategy based on the autoregressive transformer model, the proposed evaluation model can achieve a trade-off between decoding efficiency and effectiveness. Specifically, it is possible to mask a continuous segment of text instead of a single phoneme, allowing the evaluation for these masked tokens to be decoded in parallel, thereby improving decoding efficiency.

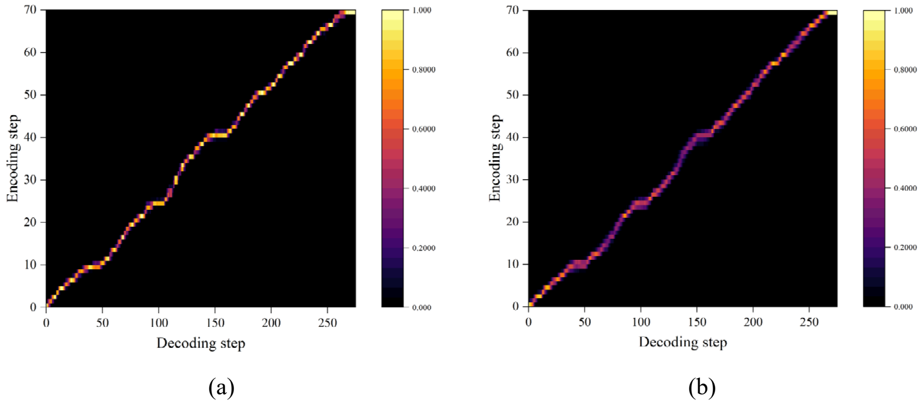
5 Experimental results and performance analysis

5.1 The impact of the number of attention heads on evaluation results

The experiments were conducted on a sub-set of the AISHELL-1 dataset, consisting of broadcast hosting speech with a total duration of 178 hours. The dataset comprises recordings from 400 broadcasters representing various Chinese dialect regions. All recordings were conducted in a quiet indoor environment with a 16 kHz sampling rate. The recorded texts span five domains: finance, sports, technology, entertainment, and current news, with transcription accuracy exceeding 95%. The dataset was partitioned into a training set, a validation set, and a test set. Specifically, the training set comprises 120,098 data samples, amounting to 150 hours in total; the validation set contains 14,326 data samples, totaling 18 hours; and the test set includes 7,176 data samples, summing up to 10 hours. The server used for the experiment is equipped with an Intel Xeon Gold 6230 CPU, 256 GB of RAM, and four NVIDIA Tesla V100 (32 GB) GPUs, running on the CentOS 7.5 operating system.The deep learning framework used is PyTorch 1.10. During training, the batch size was set to 32, the AdamW optimiser was used, the initial learning rate was set to 2×10^{-4}, and the exponential decay rate was set to 0.999. The experiment was based on the PyTorch deep learning framework and used the DeepSpeed acceleration framework to train for 60 epochs on an NVIDIA A100.

In this paper, the number of attention heads in the autoregressive transformer is set to 3 and 5, and attention maps under different numbers of attention heads are compared. The attention maps show the distribution of attention weights at each decoding step across all encoding steps. The brighter the color of a point in the map, the higher the corresponding weight value. Good attention alignment forms a clear and bright line along the diagonal of the attention map. In other words, the more concentrated and closer the attention weights are to the diagonal, the better the alignment, and the fewer cases of pronunciation errors or missing sounds in the spoken output. The attention weights of the multi-head attention mechanism are obtained by calculating the arithmetic mean of the weights from each attention head. A text with rare and repeated characters is selected from the test set, and the attention alignment graphs of the models during the evaluation of spoken fluency are shown in Figures 3(a) and 3(b). As shown in Figure 3, when there are only three attention heads, the attention weights are relatively scattered. When the number is increased to five attention heads, the attention weights become more concentrated along the diagonal, resulting in better alignment. This demonstrates that multi-head attention achieves better alignment than ordinary dot-product attention. Further increasing the number of attention heads leads to slightly better alignment, but the improvement is not significant. In addition, increasing the number of attention heads increases the number of model parameters. Considering both the alignment performance and the parameter count, using five attention heads in the attention module is more appropriate.

Figure 3 Attention maps under different numbers of attention heads (see online version for colours)



(a)

(b)

5.2 *Performance comparison of spoken fluency assessment*

For the goal of fully validating the efficiency of the proposed spoken fluency evaluation model, this paper selects common quantitative metrics, including accuracy (ACC), mean squared error (MSE), perceptual evaluation of speech quality (PESQ), word error rate (WER), to conduct comparative experiments between the proposed model OURS and benchmark models. The selected benchmark models are SAM-CGRU (Li et al., 2024), AR-PIX (Chandrabanshi and Domnic, 2024), AR-SAM (Dhahbi et al., 2025), and CNN-AR (Noh and Park, 2024). The ACC, MSE, PESQ, and WER of different models are compared in Table 1. The ACC and PESQ of OURS are 93.35% and 96.12, respectively, which are improved by 9.78% and 10.09 over SAM-CGRU, by 8.33% and 6.07% over AR-PIX, by 5.18% and 2.85% over AR-SAM, and by 1.67% and 1.04% over CNN-AR, respectively. The MSE of SAM-CGRU, AR-PIX, AR-SAM, CNN-AR, and OURS are 0.1908, 0.1587, 0.1361, 0.1052, and 0.0839, respectively. The MSE of OURS is reduced by 50.89%, 43.58%, 33.47%, and 11.87% compared to SAM-CGRU, AR-PIX, AR-SAM, and CNN-AR, respectively. Further comparing the WER, OURS reduces the WER of the other four models by 0.9–6.4%.

Table 1 Comparison of spoken fluency evaluation metrics across different models

Model	*ACC (%)*	*MSE*	*PESQ*	*WER (%)*
SAM-CGRU	83.57	0.1708	86.03	7.1
AR-PIX	85.02	0.1487	90.05	5.5
AR-SAM	88.17	0.1261	93.27	2.3
CNN-AR	91.68	0.0952	95.08	1.6

SAM-CGRU implements spoken fluency evaluation through convolutional gates and self-attention mechanisms; however, when processing ultra-long sequences, its attention weights may become ineffective due to issues such as gradient vanishing or gradient exploding, making it hard for the model to effectively learn long-range dependencies. In spoken fluency evaluation, fluency is not only related to local speech features (such as pauses and repetitions) but is also influenced by global semantic coherence. If a model cannot capture long-range dependencies, it may misclassify reasonable pauses caused by semantic shifts as disfluent expressions. AR-PIX combines PixelCNN and autoregressive models for spoken fluency evaluation. However, the convolutional kernels of PixelCNN can only capture local regions, and although stacking layers expands the receptive field, it still struggles to directly model ultra-long-range dependencies. AR-SAM maintains contextual information through recursive state passing, but gradient vanishing or exploding issues commonly occur during long-sequence training, leading to the loss of distant dependency information. In spoken fluency evaluation, fluency may be affected by global semantic coherence, and the model might misjudge if it fails to capture long-range dependencies. CNN-AR maintains context through recursive passing of hidden states, but gradient vanishing often occurs during long-sequence training, making it difficult for distant information to be effectively transmitted. OURS not only extracts multi-scale features of speech using Res2Net, but also designs an autoregressive transformer capable of effectively addressing alignment disorders in autoregressive generation, thereby improving the stability of spoken fluency evaluation models.

Figure 4 The proportion of identification errors for sentences of different lengths (see online version for colours)

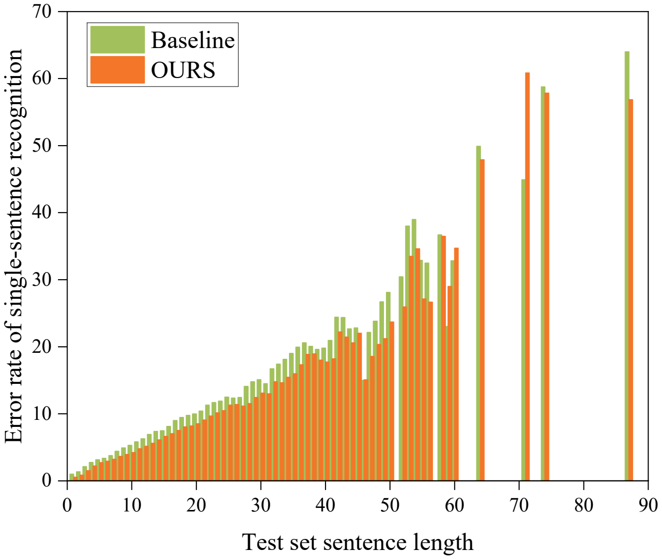


Figure 4 shows the proportion of identification errors for sentences of different lengths in the test set. The *x*-axis represents the duration of a single sentence, and the *y*-axis represents the average identification error rate for sentences of that length. Blue represents the identification error rate of the baseline, and light green represents the identification error rate after training with context concatenation methods. The baseline model is CNN-AR. From the figure, the identification error rates of OURS are low for short sentences but high for long sentences, and the error rate is positively correlated with sentence length. Training with context concatenation improves identification for sentences of all lengths, particularly significantly for medium to long sentences. Although the error rate is even higher than the baseline for some individual lengths, this is because the number of sentences of that length is extremely small (only one sentence), making it an individual case. Based on the above analysis, the OURS model can more comprehensively evaluate spoken fluency.

6 Conclusions

To address the current research issue of difficulty in capturing cross-sentence contextual dependencies during fluency evaluation of broadcast hosting speech sequences, leading to low fluency evaluation accuracy, this paper proposes a broadcast hosting spoken fluency evaluation model based on autoregression and transformer. First, speech signal analysis methods are used for spoken signal feature extraction, and adaptive matching is performed on the captured spoken pronunciation characteristics to realise the mining of spoken pronunciation signals. A multiwavelet decomposition approach is adopted for feature decomposition of English spoken pronunciation signals, and the Res2Net model is used for multi-scale deep feature extraction. Then, evaluation modelling is carried out from two aspects: word-level pauses and phoneme duration. To enhance the diversity and accuracy of word-level pauses, a pause prediction module is proposed in the text frontend. This module predicts multiple pause labels based on the original text, providing an accurate reference for modelling pause duration in spoken fluency evaluation. To improve the naturalness of phoneme duration, a duration prediction module based on a mixture of Gaussian distributions is proposed to enhance the naturalness of phoneme durations. This module predicts a mixture of Gaussian distributions for each phoneme and obtains diverse phoneme durations through random sampling. To improve the stability of phoneme duration modelling in autoregressive models, an autoregressive LLM and transformer discriminative module are proposed. This module is applied at each time step of the autoregressive LLM and avoids voice-text alignment disorders during the evaluation process through transformer and decision mechanisms. Experimental results show that the evaluation accuracy of the proposed model is improved by 1.67% and 9.78% compared to baseline models, significantly outperforming various baseline models. This study not only provides a reference for spoken fluency evaluation but also opens up new avenues for the application of speech LLMs in fine-grained speech quality assessment tasks.

Declarations

This work is supported by one of the research achievements of the Ningbo Urban Image Construction and Promotion Research Base.

The author declares that he has no conflicts of interest.

References

Aibinu, A.M., Salami, M-J.E. and Shafie, A.A. (2012) 'Artificial neural network based autoregressive modeling technique with application in voice activity detection', *Engineering Applications of Artificial Intelligence*, Vol. 25, No. 6, pp.1265–1276.

Alashban, A.A., Qamhan, M.A., Meftah, A.H. and Alotaibi, Y.A. (2022) 'Spoken language identification system using convolutional recurrent neural network', *Applied Sciences*, Vol. 12, No. 18, pp.81–103.

Bashiri, H. and Naderi, H. (2024) 'Comprehensive review and comparative analysis of transformer models in sentiment analysis', *Knowledge and Information Systems*, Vol. 66, No. 12, pp.7305–7361.

Chandrabanshi, V. and Domnic, S. (2024) 'A novel framework using 3D-CNN and BiLSTM model with dynamic learning rate scheduler for visual speech recognition', *Signal, Image and Video Processing*, Vol. 18, No. 6, pp.5433–5448.

Dhahbi, S., Saleem, N., Bourouis, S., Berrima, M. and Verdu, E. (2025) 'End-to-end neural automatic speech recognition system for low resource languages', *Egyptian Informatics Journal*, Vol. 29, pp.15–23.

Du, H., Tian, X., Xie, L. and Li, H. (2021) 'Factorized WaveNet for voice conversion with limited data', *Speech Communication*, Vol. 130, pp.45–54.

Gao, S-H., Cheng, M-M., Zhao, K., Zhang, X-Y., Yang, M-H. and Torr, P. (2019) 'Res2Net: a new multi-scale backbone architecture', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 2, pp.652–662.

Garain, A., Singh, P.K. and Sarkar, R. (2021) 'FuzzyGCP: a deep learning architecture for automatic spoken language identification from speech signals', *Expert Systems with Applications*, Vol. 168, pp.41–56.

Helske, S. and Helske, J. (2019) 'Mixture hidden Markov models for sequence data: the seqHMM package in R', *Journal of Statistical Software*, Vol. 88, pp.1–32.

Kang, B.O., Jeon, H.B. and Lee, Y.K. (2024) 'AI-based language tutoring systems with end-to-end automatic speech recognition and proficiency evaluation', *ETRI Journal*, Vol. 46, No. 1, pp.48–58.

Li, D., Yang, Z., Liu, J., Yang, H. and Wang, Z. (2024) 'Emotion embedding framework with emotional self-attention mechanism for speaker recognition', *Expert Systems with Applications*, Vol. 238, pp.12–18.

Li, Q., Zhao, S., Zhao, S. and Wen, J. (2023) 'Logistic regression matching pursuit algorithm for text classification', *Knowledge-Based Systems*, Vol. 277, pp.61–75.

Liu, J., Wumaier, A., Fan, C. and Guo, S. (2023) 'Automatic fluency assessment method for spontaneous speech without reference text', *Electronics*, Vol. 12, No. 8, pp.17–25.

Lv, C., Lan, H., Yu, Y. and Li, S. (2022) 'Objective evaluation method of broadcasting vocal timbre based on feature selection', *Wireless Communications and Mobile Computing*, Vol. 8, No. 1, pp.70–86.

Mahr, T.J., Berisha, V., Kawabata, K., Liss, J. and Hustad, K.C. (2021) 'Performance of forced-alignment algorithms on children's speech', *Journal of Speech, Language, and Hearing Research*, Vol. 64, No. 6, pp.2213–2222.

Noh, H-K. and Park, H-J. (2024) 'A light-weight autoregressive CNN-based frame level transducer decoder for end-to-end ASR', *Applied Sciences*, Vol. 14, No. 3, pp.13–20.

Pakhomov, S.V., Marino, S.E., Banks, S. and Bernick, C. (2015) 'Using automatic speech recognition to assess spoken responses to cognitive tests of semantic verbal fluency', *Speech Communication*, Vol. 75, pp.14–26.

Popov, D., Gapochkin, A. and Nekrasov, A. (2018) 'An algorithm of daubechies wavelet transform in the final field when processing speech signals', *Electronics*, Vol. 7, No. 7, pp.120–135.

Savchenko, V.V. and Savchenko, L.V. (2024) 'A method for the asynchronous analysis of a voice source based on a two-Level autoregressive model of speech signal', *Measurement Techniques*, Vol. 67, No. 2, pp.151–161.

Sharma, U., Maheshkar, S., Mishra, A.N. and Kaushik, R. (2019) 'Visual speech recognition using optical flow and hidden Markov model', *Wireless Personal Communications*, Vol. 106, No. 4, pp.2129–2147.

Sharma, U., Om, H. and Mishra, A.N. (2023) 'HindiSpeech-Net: a deep learning based robust automatic speech recognition system for Hindi language', *Multimedia Tools and Applications*, Vol. 82, No. 11, pp.16173–16193.

Shutian, Z. (2024) 'A study on the path of language innovation in broadcasting and hosting in the new media era', *Academic Journal of Humanities & Social Sciences*, Vol. 7, No. 8, pp.157–162.

Song, Q., Sun, B. and Li, S. (2022) 'Multimodal sparse transformer network for audio-visual speech recognition', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 34, No. 12, pp.10028–10038.

Wang, S., Shen, Y., Feng, S., Sun, H., Teng, S-H. and Chen, W. (2024) 'Alpine: unveiling the planning capability of autoregressive learning in language models', *Advances in Neural Information Processing Systems*, Vol. 37, pp.119662–119688.

Wu, Y-X., Du, K., Wang, X-J. and Min, F. (2024) 'Misclassification-guided loss under the weighted cross-entropy loss framework', *Knowledge and Information Systems*, Vol. 66, No. 8, pp.4685–4720.

Yu, J. (2024) 'Online learning system for English speech automatic recognition based on hidden Markov model algorithm and conditional random field algorithm', *Entertainment Computing*, Vol. 51, pp.72–89.

Zhao, J., Li, R., Tian, M. and An, W. (2024) 'Multi-view self-supervised learning and multi-scale feature fusion for automatic speech recognition', *Neural Processing Letters*, Vol. 56, No. 3, pp.16–28.