International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-19

https://doi.org/10.70917/ijcisim-2026-0004
Article

Deep Forgery Defense System for Live Agricultural E-Commerce—A Meta-Learning-Based Dynamic Verification Model for Anchor Voiceprints

Zhimin Liang *, Wenwen Liu and Shaocong Xu

School of Economics and Management, Jiangxi Agricultural University, Nanchang 330045, Jiangxi, China; LZHM202202@163.com

Abstract: With the booming development of agricultural products live e-commerce, the threat of deep forgery technology is becoming more and more prominent. In this paper, a deep forgery defense system for live agricultural products e-commerce is constructed to guarantee the authenticity and security of e-commerce. Based on MFCC features for voice feature extraction, and for the defects of ObamaNet synthesized anchor baseline system with large time overhead, a speech synthesis model for live agricultural products e-commerce based on mouth shape classification is proposed to match voice features with mouth shape. In addition, a semi-supervised voice verification model is proposed based on meta-learning, in which the meta-gradient computed by the model on the training set is used as a guide for gradient optimization on the test set, so that the model learns richer knowledge. The experiments show that there is a big difference between the synthesized speech and the prototype speech in terms of voiceprint features, and some syllables also have significant differences in the interval of 3kHz~4kHz, which indicates that the system in this paper can effectively distinguish the real speech from the forged speech, and enhances the reliability of the system, which plays an important role in the field of live e-commerce of agricultural products.

Keywords: voiceprint feature extraction; mouth shape classification; meta-learning; semi-supervised learning; voiceprint verification; deep forgery defense system

1. Introduction

With the rapid development of Internet technology, e-commerce live broadcast has become an important means of agricultural products sales [1]. According to statistics, in 2024, China's agricultural products e-commerce live streaming turnover exceeded 1.6 trillion yuan. Live agricultural e-commerce not only provides farmers with new sales channels, but also meets consumer demand for high-quality, fresh and green agricultural products, promotes farmers' economic growth, and contributes to the sustainability of agricultural development [2-4]. At the same time, due to the increasing development of artificial intelligence (AI) and voice synthesis and other technologies, the depth of forgery events in e-commerce live broadcasts, directly leading to more than 8 billion in economic losses, affecting the product and anchor reputation, becoming the e-commerce industry and its concerns [5-6]. The main manifestations include fraud by forging the information of well-known anchors, and misleading consumers to buy by synthesizing the voice of an anchor through voice synthesis technology to introduce a product. In addition, because the current regulatory technology is only in the static voice recognition, the dynamic change of voice recognition efficiency is extremely low, and the agricultural products live broadcast site noise interference and dialect interference is serious, voice feature recognition difficulties, further deteriorating the authenticity of the current live broadcast market [7-9].

Deep forgery techniques, which are developed based on some methods of deep learning. Generative Adversarial Network through the generator and discriminator this confrontation, the generator gradually

learns to generate high-quality deep forgery content, and its synthesized voice authenticity is close to the real human [10]. And by training on a large amount of real data, the autoencoder can learn the feature distribution of the data and can expand the voice of human expression [11]. Despite the good progress of voiceprint detection models, recent studies have shown that some of the models show vulnerability in the face of adversarial samples and small sample sizes that highly mimic natural speech, to the extent that they can easily lead to false recognition [12-13]. Meta-learning, on the other hand, enables machines to have the ability to self-adjust and self-optimize by learning from experience from multiple tasks [14]. It is applied in voiceprint recognition, which can quickly adapt to the dialect and continuously optimize the new attacks of deep forgery, providing support for the construction of a deep forgery defense system for live agricultural products e-commerce [15].

The purpose of this paper is to construct a deep forgery defense system for live agricultural products e-commerce, which contains two major basic models. First, the deep learning-based voice synthesis model of live agricultural products e-commerce extracts the vocal features through MFCC features, adopts a number of mouth shapes to represent the mouth states of different objects, and uses face reconstruction technology to complete the corresponding synthesis of mouth labels and portraits, which improves the synthesis speed of speech. The dynamic verification model of anchor voiceprint based on meta-learning uses the meta-gradient computed in the synthesis domain as a reference for the optimization of meta-gradient in the real domain, so that the model can adapt to the data in different domains and learn a wider range of knowledge to accurately distinguish between real and fake voices. Finally, relevant experiments are set up to confirm the feasibility and effectiveness of the system in this paper through voiceprint uniqueness and voiceprint verification accuracy experiments.

2. Deep Learning Based Speech Forgery Synthesis for Agricultural Products Live E-Commerce

2.1. Voiceprint Feature Extraction

Speaker recognition is a technology based on voice features, so the extraction of feature parameters is the most basic and important step in the speaker recognition system, and the extracted feature parameters characterize the personality information of the speaker. Therefore, in speaker recognition technology, the feature parameter extraction algorithm for speech has been the research hotspot in this field.

2.1.1. Feature Extraction Criteria

The feature parameters of speaker recognition are extracted from the speaker's speech signal, and the speaker recognition system is mainly based on the feature parameters. Research on speech perception has shown that both inborn and acquired factors may cause differences in speech signals. Innate conditions mean that different people are born with different physiological dimensions of the vocal organs, and therefore each person has his or her own unique voice; acquired conditions include each person's speaking habits, emotional state when speaking, cultural level, health status and so on. Under the combined effect of innate and acquired factors, the extracted feature parameters can maximally describe the individual's personality characteristics and thus differentiate them from those of other people. Under ideal conditions, the selected features should meet the following criteria:

(1) It can effectively characterize the speaker's personality information to determine his/her identity.

(2) They can be easily extracted from the speaker's voice.

(3) Not easily imitated by other people's speech features.

(4) Be as adaptable as possible to the environment and not easily change over time and space.

(5) Distinguish different speakers with a high degree of validity, and at the same time, the same speaker can ensure a relatively stable state when the acquired factors change, and there is a good independence between the dimensions of the feature parameters.

However, there are almost no feature parameters that satisfy the above criteria at the same time, and even some of the criteria are still the focus of the current speaker recognition research, such as time-varying robustness and speech mode robustness. Therefore, when selecting feature parameters, it is generally necessary to take a compromise measure according to the actual situation.

2.1.2. Introduction and Selection of Characteristic Parameters

The variety of feature parameters that can be used in speaker recognition is particularly ample, with the more common ones being fundamental frequency, fundamental period, resonance peak coefficient, linear prediction coefficient (LPC) [16], linear prediction cepstrum coefficient LPCC, vocal tract impact response, autocorrelation coefficients, MFCC features [17], perceptual linear prediction coefficients (PLP) [18], and difference cepstrum, among others. This section briefly describes several extremely

common and commonly used feature parameters:

(1) Linear Prediction Cepstrum Coefficient (LPCC).LPCC is actually the representation of LPC in the cepstrum domain. It presupposes an autoregressive signal-based speech signal, and the cepstrum coefficients are obtained by linear prediction analysis of this speech signal.

(2) Perceptual linear prediction coefficient (PLP). Similar to the linear prediction coefficients, it is a series of coefficients for the prediction polynomial of the full-polar model. However, the LPC parameters are analyzed for ordinary time-domain signals, whereas PLP is analyzed for speech signals that have been processed by the auditory model.

(3) Mel Frequency Cepstrum Coefficients (MFCC). In the study of human ear we found that the perceptual ability of different frequencies of speech signals to the human ear is different: when the frequency is less than 1000Hz, the perceptual ability is linearly related to the linear frequency; when the frequency is 1000Hz or more, they are logarithmic relationships. Based on this, research scholars proposed the concept of Mel frequency, because it can well reflect the perception characteristics of the human ear to speech. At the same time, the human ear cannot distinguish two tones with similar or identical frequencies, and if it wants to distinguish two tones, then the difference between their frequency components must be greater than or equal to a critical bandwidth.

According to the above analysis, among the three feature parameters, PLP and LPCC are basically based on the premise of the model of all-polar speech production, and both of them are strong in describing vowels, but weak in describing consonants. However, the MFCC feature can be said to be a good representation of the human ear's perceptual properties of speech, based on which it can make the recognition rate and robustness of the speaker recognition system improved. Therefore, the MFCC features will be selected for the experiments in the following sections.

2.1.3. MFCC Feature Extraction Process

After the advantages and disadvantages of multiple features are introduced above, MFCC features are selected as the object of study in the later paper, then its feature extraction process is shown in Fig. 1:

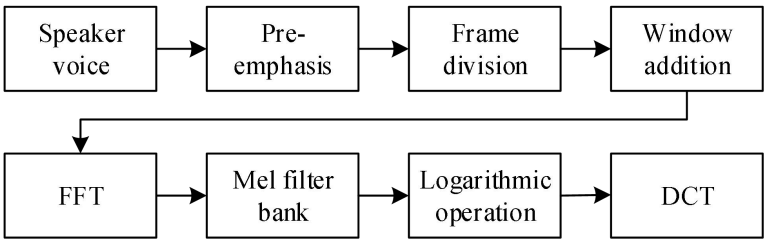


Figure 1. MFCC characteristics extract the detailed process.

1. Pre-emphasis. The essence of the pre-emphasis processing is to make the captured speech information pass through a high-pass filter: $H(Z)=1-\mu z^{-1}$. where μ is the pre-emphasis coefficient, which takes values between 0.9 and 1.0. The purpose of this step is to boost the high-frequency portion of the signal to reduce the effect of noise. At the same time, the effects due to the lips and vocal cords are eliminated to compensate for the high-frequency portion of the speech signal that is affected by the articulatory organs.

2. Segmentation. The voice signal is characterized by smoothness in a short time, based on this characteristic, the voice signal can be intercepted in segments with a length of 10ms~30ms per frame. At the same time, in order to preserve the correlation between frames when intercepting voice frames, so to make a smooth transition between frames, then finally use the method of overlapping segments to split frames, that is, there should be a frame shift when splitting frames.

3. Add window. This step reduces the edge effect of the speech frame and also increases the continuity of the ends of the speech frame so that the ends of the speech frame transition smoothly to 0. Currently, researchers are using Hamming window.

4. FFT. Because it is not easy to see the characteristics of the speech signal in the time domain, the speech signal is generally transformed from the time domain to the frequency domain, through the spectrogram, to observe its energy distribution, the specific practice is the Fast Fourier Transform.

5. Mel triangular filter bank. Speech spectrum through the triangular band-pass filter bank, in addition to the above analysis in order to characterize the human ear auditory properties, there are two other reasons: smoothing the spectrum and eliminate the role of harmonics, highlighting the resonance peaks in the speech; in addition, you can reduce the amount of arithmetic.

6. Logarithmic operations. Taking the logarithm of the output of the filter obtained in the previous step results in a set of logarithmic energy spectra.

7. DCT: Finally, the Mel frequency cepstrum coefficients are transformed from the frequency domain to the time domain by the discrete cosine transform, and the resulting MFCC features.

The above are the steps of MFCC feature extraction, but the actual speech signal contains not only static features, but also dynamic features, so it is necessary to respond to both features. Therefore, generally the final feature contains the MFCC feature (reacting to static features) and its first and second order derivatives (reacting to dynamic features).

2.2. Speech Synthesis Models and Methods for Live E-commerce of Agricultural Products

2.2.1. AgriLive E-Commerce Corpus

Since there is no open-source corpus of live agricultural product e-commerce in the Internet at present, this paper adopts the method of recording broadcasting video to expand the dataset. Firstly, the oral broadcast scripts of Chinese agricultural live e-commerce are collected and used as the read-aloud content of the dataset, with a total of 400 sentences and a combined word count of 25,000 words. Then a green screen studio was built, a male anchor was selected as the voice synthesized anchor image of this paper's agricultural products live e-commerce, and the video of the anchor reading the script was recorded, and a total of about 5 hours of original agricultural products live e-commerce voice was collected.

The collected live agricultural products e-commerce data were manually edited, cutting out invalid segments such as the beginning and the end, and splitting the video by sentence. The video resolution was compressed to 960 × 540 because the high resolution of the recorded video data would result in slow image processing and training. The final result is about 3 hours of valid live e-commerce data on agricultural products.

2.2.2. ObamaNet-Based Speech Synthesis Anchor System for E-commerce

ObamaNet [19] is a multi-module neural network architecture that uses text as input to generate realistic speaker videos with mouth shape synchronized with speech, and consists of three modules: a text-to-speech conversion network based on Char2Wav; a time-delayed LSTM network that generates key points of the mouth synchronized with speech; and a U-Net network based on pix2pix translation.

ObamaNet solves the problem of AI synthesized anchor speech by modifying the mouth region images of existing videos. In this paper, we propose a baseline system for AI synthesized anchor voice for live agricultural e-commerce based on ObamaNet. In the text-to-speech conversion module, since the input is the text of live agricultural products e-commerce, the existing researched live agricultural products e-commerce speech synthesis technology is used, and the live agricultural products e-commerce speech synthesis interface provided by it is called to get the corresponding live agricultural products e-commerce speech of the text.

Because the audio source of the voice of the live agricultural products e-commerce AI anchor video data collected in this paper is not the same person as the audio source provided by the live agricultural products e-commerce voice synthesis interface, the speed of speech and voice features differ greatly, and to learn the correspondence between the voice features and the mouth shape, if we do not carry out the feature alignment of the two kinds of voices and the feature conversion will lead to a large error in predicting the key point of the mouth from the synthesized voice. Therefore, in this paper, a speech feature conversion network is added to the model structure to convert the speech features of speech synthesis to the corresponding real human speech features of video data.

The ObamaNet-based AI synthesized anchor voice baseline system for live agricultural products e-commerce based on ObamaNet proposed in this paper firstly obtains the text of live agricultural products e-commerce, calls the voice synthesis interface of live agricultural products e-commerce, obtains the corresponding live agricultural products e-commerce voice, and extracts the features of the voice, and inputs the extracted voice features into the voice feature conversion model to get the converted features; secondly, inputs the voice features into the time-delayed LSTM network, the output data is reduced by the PCA model to get the mouth key point coordinates; then read the cache file of the basic background video, apply the rotation transformation matrix to the mouth key point coordinates, draw the whiteboard and key point lines on the portrait cache, input the image into the U-Net network, synthesize the portrait and then paste it back to the original background frame to get the complete portrait based on the coordinate cache; finally, add the subtitle and end credits, integrate the voice Finally, add subtitles and credits, and integrate the voice, to complete the synthesis process of the voice of the AI anchor of the live agricultural products e-commerce.

2.2.3. E-commerce Speech Synthesis Anchor Model Based on Mouth Classification

Due to the large time overhead of the baseline system video synthesis, which can not ensure fast and efficient video synthesis, this paper combines the idea of 2D animation production, and proposes a live agricultural products e-commerce AI synthesized anchor voice model based on mouth type classification.

The structure of the AI synthesized anchor voice model based on mouth classification is shown in Figure 2, firstly, according to the text of the live agricultural e-commerce, the voice synthesis model is used to generate the voice of the live agricultural e-commerce, and the voice features are extracted, and the extracted voice features are input into the mouth synchronization network, and the mouth label sequence is predicted, then the corresponding candidate frames are selected according to the mouth label sequence, and finally the frame sequence is keyed to change the background, add title subtitles, credits, and merge voice files to complete the synthesis of the live agricultural products e-commerce AI anchor voice video.

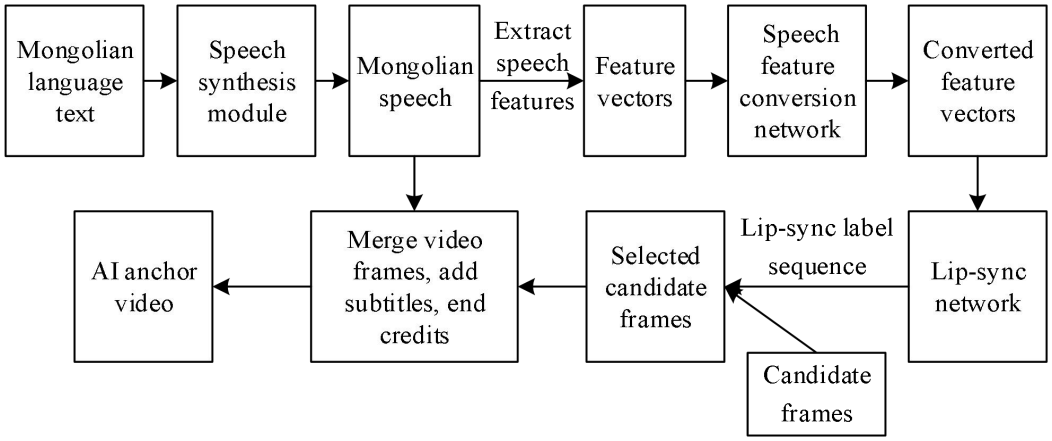


Figure 2. AI synthetic anchor voice model based on mouth type classification.

The method proposed in this paper uses mouth labels to represent mouth features, which belongs to the problem of classification from speech features to mouth labels.In this paper, mouth synchronization network is constructed based on DNN and Bi-LSTM models respectively to realize the generation of sequences from speech features to mouth labels.

(1) Mouth Label

Combined with the idea of 2D animation, six basic mouth types (labels: A-F) and three optional extended mouth types (labels: G, H, X) are used, and the mouth labels are shown in Fig. 3.

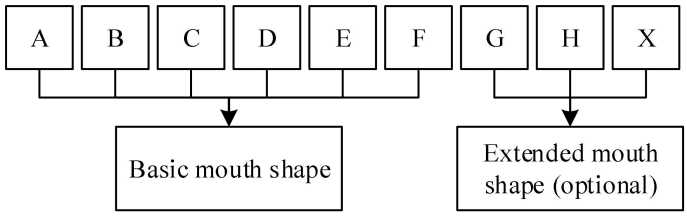


Figure 3. Mouth shapes tag.

(2) DNN based neural network for speech feature to mouth label sequence

The prediction of speech features to mouth labels is a multi-label classification problem and DNN is suitable for this scenario, where the number of mouth labels is the number of output nodes in the output layer of the network.

The DNN network uses Softmax as the activation function, which is commonly used in multi-classification problems, and the goal of Softmax during regression training is to minimize the cross-entropy between the probability distribution predicted by the model and the probability distribution of the real labels. Meanwhile, the cross-entropy loss is used when adjusting the model weights during training with the goal of minimizing the loss, which is defined as:

$$L_{CE} = -\sum_{i=1}^{n} t_i \log p_i \tag{1}$$

where: n denotes the number of classifications; t_i denotes the i th label; and p_i denotes the probability value that the data belongs to the i th class as calculated by Softmax.

(3) Bi-LSTM based neural network for speech feature to mouth label sequence

The RNN network structure is very practical in time series problems and compensates for the inability of fully connected DNN models to model changes in time series, so this paper proposes a Bi-LSTM based approach for speech feature to mouth label sequences.Bi-LSTM can retain future and past information and truly predict sequences based on context.

Bi-LSTM network is computed as a mapping between input sequences and output sequences such as $X = (X_1, X_2, \ldots, X_n)$ to $y = (y_1, y_2, \ldots, y_n)$. The formula is as follows:

$$f_{forget\ gate} = sigmoid\left(W_{fg} X_t + W_{hfg} h_{t-1} + b_{fg}\right) \tag{2}$$

$$i_{input\ gate} = sigmoid\left(W_{ig} X_t + W_{hig} h_{t-1} + b_{ig}\right) \tag{3}$$

$$o_{output\ gate} = sigmoid\left(W_{og} X_t + W_{hog} h_{t-1} + b_{og}\right) \tag{4}$$

$$\begin{aligned}(C)_t = (C)_{t-1} \otimes \left(f_{forget\ gate}\right) + \left(i_{input\ gate}\right) \\ \otimes \left(\tanh\left(W_c X_t + W_{hc} h_{t-1} + b_c\right)\right)\end{aligned} \tag{5}$$

$$h_t = o_{output\ gate} \tanh \otimes ((C)_{t-1}) \tag{6}$$

where $W_{fg}, W_{ig}, W_{og}, W_{hc}$ and $b_{fg}, b_{ig}, b_{og}, b_c$ denote the weights and bias variables of the three gates and one cell, respectively; X_t is the input at the current moment; h_{t-1} denotes the output of the previous moment; and $(C)_{t-1}$ is the last moment's Memory.

In order to solve the problem of slow synthesis speed of the baseline system, this paper synthesizes candidate frames corresponding to six base mouth labels for each portrait frame according to an anchor base background video in advance, so as to achieve the purpose of using the candidate frames directly when synthesizing the video. In view of the fact that the baseline system will have slight traces of splicing of images in the mouth region when the mouth images are translated, the use of whole-face translation avoids generating splicing traces around the mouth. Using the base background video to generate 6 base mouth images corresponding to each frame, the way to synthesize the face based on the face keypoint linkage images, the 6 mouth outlines are redrawn based on the 6 base mouth shapes, and then the 6 candidate frames corresponding to each frame are synthesized.

3. Meta-Learning-Based Dynamic Verification Model for Anchor Voiceprints

3.1. Semi-Supervised Anchor Voiceprint Dynamic Verification Framework

The idea based on adversarial training can improve the domain adaptation performance of dynamic verification models for anchor voiceprints. However, this approach does not fully utilize the easily available unlabeled data. The current mainstream semi-supervised anchor voiceprint dynamic verification methods are constructed through the mean teacher (MT), a classical framework in the semi-supervised domain. Its main purpose is to better utilize a large amount of strongly labeled synthetic data for supervision and unlabeled real data for label generalization by combining the student model with supervised learning and the consistency assumption, i.e., forcing the outputs of the student model and the teacher model to be the same for the same input samples, so as to efficiently utilize the unlabeled data to improve the model's generalization ability.

The teacher model updates the network parameters by calculating the exponential moving average (EMA) of the student model, which aims to find a more stable gradient update direction, reduce the gradient noise, and make the teacher model more stable for the input data, so as to better guide the student

model for generalization training.

Meta-learning is an important research method in the field of machine learning, and its core goal is to let the model learn how to learn, a process that aims to enable the model to adapt quickly in the face of different tasks, thus improving the model's generalization ability.

3.2. Semi-Supervised Anchor Voiceprint Dynamic Verification Based on Meta-Learning

3.2.1. Semi-Supervised Anchor Vocal Dynamics Verification Problem Description

The training dataset $\{D_s, D_{wc}, D_u\}$ for the semi-supervised anchor voiceprint dynamic verification method contains three sub-datasets: the strongly labeled synthetic sample set $D_s = \{(x_s^1, y_s^1), \cdots, (x_s^i, y_s^i), \cdots, (x_s^N, y_s^N)\}$, the weakly labeled real sample set $D_{wc} = \{(x_{wc}^1, y_{wc}^1), \cdots, (x_{wc}^i, y_{wc}^i), \cdots, (x_{wc}^M, y_{wc}^M)\}$, and unlabeled true sample set $D_u = \{x_u^1, \cdots, x_u^i, \cdots, x_u^K\}$. Where $x^i \in \mathbb{R}^{T \times F}$ denotes the FBank spectral feature extracted from the i th audio sample with feature dimension F and spectral feature sequence length T ; N, M, K represents the number of samples in each of the 3 subdatasets; $y_{wc}^i \in \mathbb{R}^C$ is a weak label of real data with dimension C , which represents the event category information (without time information) appearing in the i th segment of real audio, and C is the number of sound event categories; $y_s^i \in \mathbb{R}^{T \times C}$ is a strong label of synthesized data, which contains the event categories appearing in the i th segment of synthesized audio as well as the corresponding start and end time stamp information. For an audio sequence x , it is input into the anchor voiceprint dynamic validator f to obtain the prediction result $f_\theta(x) \in \mathbb{R}^{T \times C}$, where θ denotes the model parameters. In strong labeling, each sample contains both event categorization information and localization timestamps, so frame-level loss computation is possible, while in weak labeling, each sample contains only event categorization information, and only sentence-level event categorization loss computation is possible.

For labeled samples x_s and x_{wc} , the supervised classification loss L_{cls} is computed by binary cross entropy (BCE), which is shown in equation (7):

$$\mathrm{L}_{cls} = y \log f_\theta(x) + (1 - y) \log(1 - f_\theta(x)) \tag{7}$$

where L_{cls} denotes the cross-entropy loss function computed from the student model output $f_\theta(x)$ with label y , $(x, y) \in \{D_s, D_{wc}\}$. Next, the unsupervised comparison loss, L_{con} , is computed over the entire dataset, D , by means of the consistency regularization technique, using the mean square error, as shown in equation (8):

$$\mathrm{L}_{con} = \frac{1}{N} \sum_{i=1}^{N} (f_{\theta_s}(x) - f_{\theta_t}(x))^2 \tag{8}$$

where L_{con} denotes the consistency regular loss function computed between the student model output $f_{\theta_s}(x)$ and the teacher model output $f_{\theta_t}(x)$, $x \in D$, and N denotes the total number of samples; θ_s and θ_t denote the parameters of the student model and the teacher model, respectively, which have the same network structure.

Finally, the student model is updated according to equation (9). Namely:

$$\mathrm{L}_{total} = \mathrm{L}_{cls} + \lambda(t)\mathrm{L}_{con} \tag{9}$$

where L_{total} denotes the loss function that is ultimately used to update the student model, and $\lambda(t)$ denotes the tuning weights that grow slowly with the number of training iterations. The teacher model is then updated by an exponential moving average (EMA) of the student model, as shown in equation (10):

$$\theta_t \leftarrow \alpha\theta_{t-1} + (1-\alpha)\theta_s \tag{10}$$

where α is the decay rate defined in the EMA algorithm.

3.2.2. Dynamic Verification of Anchor Voiceprints with Meta-Gradient Optimization

The core of using meta-learning to improve the generalization performance of a model lies in explicitly placing the model in an environment with domain differences so that it perceives such domain differences and learns to be more generalizable. The key lies in dividing the training data into non-overlapping meta-training and meta-testing sets to simulate the domain differences, and reducing the testing error of that model on the meta-testing set by optimizing the meta-gradient computed by the model on the meta-training set. Specifically, data is first randomly sampled from the N visible domains that constitute the training set $D_{train} = \{D_1, \cdots, D_N\}$, and a task set containing multiple tasks is constructed as the meta-training set $D_{meta\text{-}train}$ $\left(D_{meta\text{-}train} \subset D_{train}\right)$, and training on the meta-training set enables the model to learn generic features or patterns from multiple tasks; secondly, a meta-test set $D_{meta\text{-}test} = D_{train} - D_{meta\text{-}train}$ is constructed, which is used to test the generalization ability of the model updated by meta-training on the meta-test task, so as to enable the model to adapt to a new tasks or domains. In the semi-supervised anchor voiceprint dynamic verification task, the construction process of the task set is simplified by using the synthetic data as the meta-training set for meta-training and the real data as the meta-testing set for meta-testing within the same training step.

The synthetic data x_s of the meta-training set is fed into the student model $f_{\theta_s}(x_s)$ and the teacher model $f_{\theta_t}(x_s)$, respectively, and supervised training is carried out by utilizing the strong labeling information y_s of the synthetic data to independently update the parameters θ_s and θ_t of the student and teacher networks. The computational formula is shown in Eq. (11):

$$\begin{aligned} \theta_s' &= \theta_s - lr \cdot \nabla_{\theta_s} \mathrm{L}_{train\_s} \\ \theta_t' &= \theta_t - lr \cdot \nabla_{\theta_t} \mathrm{L}_{train\_t} \end{aligned} \tag{11}$$

where L_{train_s} and L_{train_t} are the classification losses of the student and teacher models, respectively, $\nabla_{\theta_s} \mathrm{L}_{train\_s}$ and $\nabla_{\theta_t} \mathrm{L}_{train\_t}$ denote the gradient of the student and teacher model parameters regarding the classification loss in the current training step, respectively. The parameters of the student and teacher network models are updated by gradient descent to obtain θ_s' and θ_t' , respectively. It is important to note that θ_s' and θ_t' obtained on the meta-training set are only used to allow the model to learn the knowledge of the synthetic data, and ultimately used to test the model's performance on the meta-testing set.

After the meta-training is completed and the student model and the teacher model learn the knowledge in the meta-training set (synthetic data), the next step is to test the performance of the model on the meta-test set (real data). Specifically, the weakly labeled real data x_{wc} from the meta-test set is fed into the student model $f_{\theta_s'}(x_{wc})$, and the weakly labeled information y_{wc} from a small amount of real data is utilized to compute the student model classification loss L_{test_s} ; at the same time, in order to effectively utilize the unlabeled real data x_u in the meta-test set, at this time, the student model $f_{\theta_s'}(x_u)$ and teacher model $f_{\theta_t'}(x_u)$ between the consistency loss L_{con} . Combining the classification loss and consistency loss, the student model parameters θ_s are updated:

$$\mathrm{L}_{total} = \beta \mathrm{L}_{train\_s} + (1-\beta)\mathrm{L}_{test\_s} + \lambda(t)\mathrm{L}_{con} \tag{12}$$

where β is the training weight super-parameter. In this step, the student model parameters θ_s are actually updated by L_{total} as shown in equation (13):

$$\theta_s \leftarrow \theta_s - lr \cdot \nabla_{\theta_s} \mathrm{L}_{total} \tag{13}$$

After the completion of training the model using the training data enters the evaluation phase to assess the performance of the model on the test data. The test data is independent of the training data and consists of strongly labeled real data.

4. E-Commerce Anchor Deep Voice Forgery Defense and Dynamic Verification Results

4.1. Experimental Arguments for Acoustic Uniqueness

We will compare the spectrograms of speech signals (resonance peaks, fundamental frequency, impulses, tone intensity, etc.) by looking at them and arguing for the uniqueness of the voiceprint from them. First a brief introduction to these factors of speech signals. Usually we call the plot of the results of a time-series Fourier analysis, a speech spectrogram. The spectrogram is a three-dimensional graph with time as the horizontal coordinate, frequency as the vertical coordinate, and intensity as the sound intensity, from which we can observe the changes of the spectrum of the speech signal on the time axis. Spectrogram combines the characteristics of time-domain waveform and spectrogram, which contains a large amount of information related to the characteristics of the speech signal, and dynamically shows the specific changes in the spectrum. In the spectrogram, there are many different shapes of stripes: the stripes parallel to the horizontal coordinate axis (usually called horizontal bar stripes) are the resonance peaks of the speech signal, from which we can conveniently observe the bandwidth and frequency values of the resonance peaks at various moments; the stripes perpendicular to the horizontal coordinate axis (usually called impulse stripes) are different impulse stripes corresponding to different fundamental tones of each tone, from which we can easily observe the bandwidth and frequency values of the resonance peaks. From the distance and density of these stripes, we can observe the specific changes of the fundamental period and frequency. The following two sets of experiments will be used to observe the speech signal and to demonstrate the uniqueness of the acoustic pattern. The two sets of experiments are: the same person says the same words, and different people say the same words.

Firstly, we conduct the first set of experiments to observe the same person speaking the same words. We take the speech files of three boys and three girls (each person says the same sentence three times) as samples. Draw their speech spectrograms, and then make observation and comparison. Since the results of the experiments are similar, only the speech spectrograms of one boy's and one girl's speech files are discussed below.

Figures 4 and 5 show the two recordings of Boy A on the phrase "Hello Little One", and the spectrum depicted by the two recordings. Figures 6 and 7 show the two recordings of Girl A on the phrase "Hello Little One", depicting the spectrum.

From the observation and comparison of the spectral map, we can find that the spectral map described by the voice of the same person with the same words has a high degree of similarity. The alignment of formants, intensity, fundamental frequency, etc., are all very similar, so we can derive this characteristic of voiceprint stability.

In addition, we can find that even if different people say "hello little one", the spectrum described by their corresponding voices is still very different. There are obvious differences in the direction of the stripes such as formants, sound intensity, and fundamental frequency, so we can conclude that the voiceprints of different people have different characteristics.

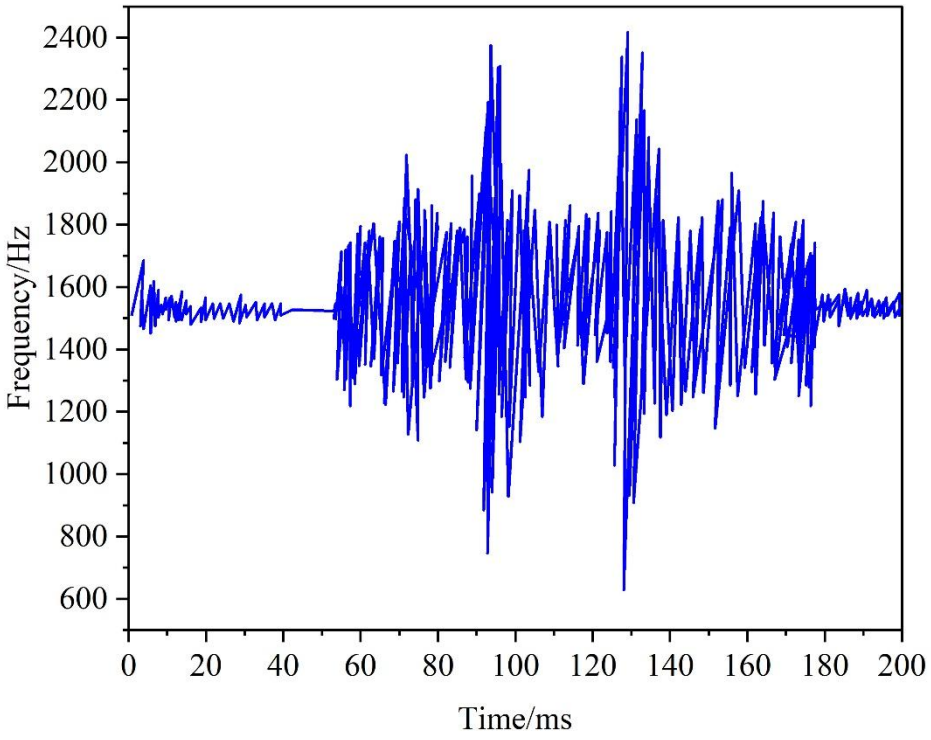


Figure 4. First recording.

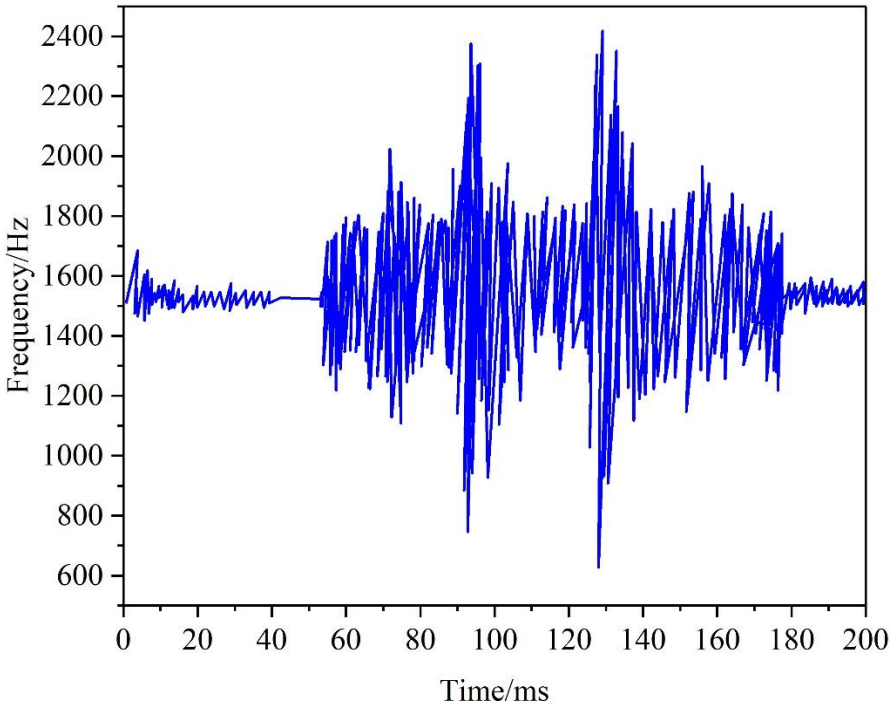


Figure 5. Second recording.

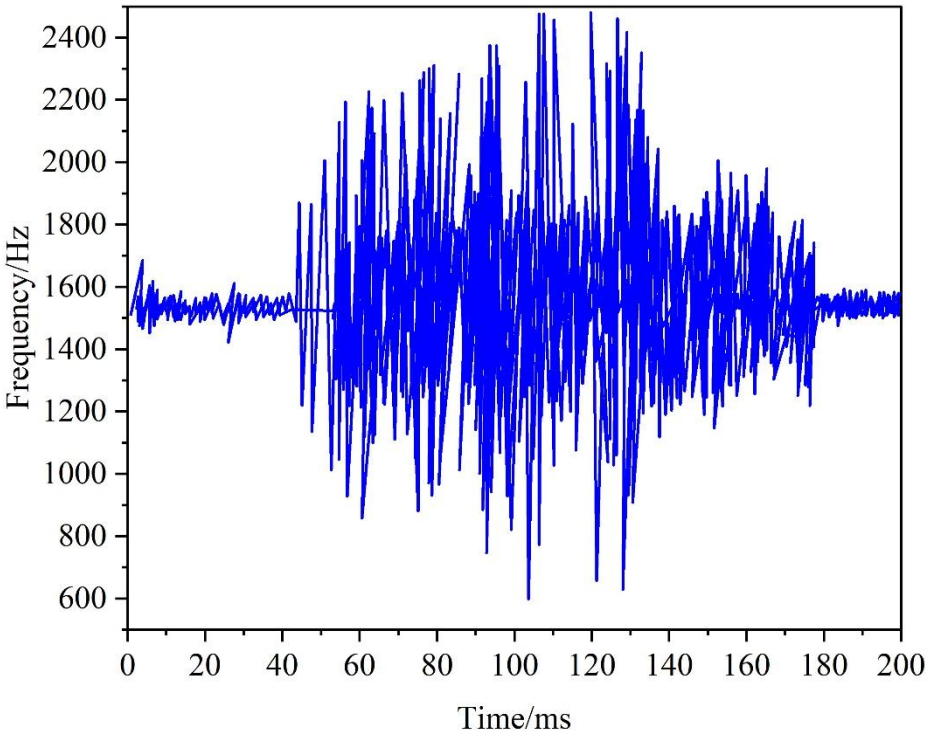


Figure 6. First recording.

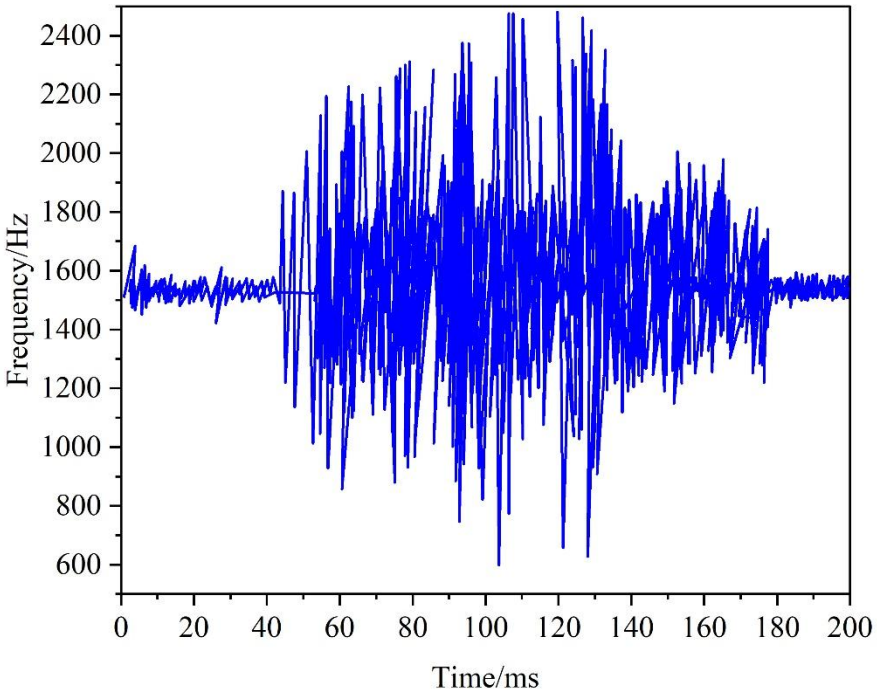
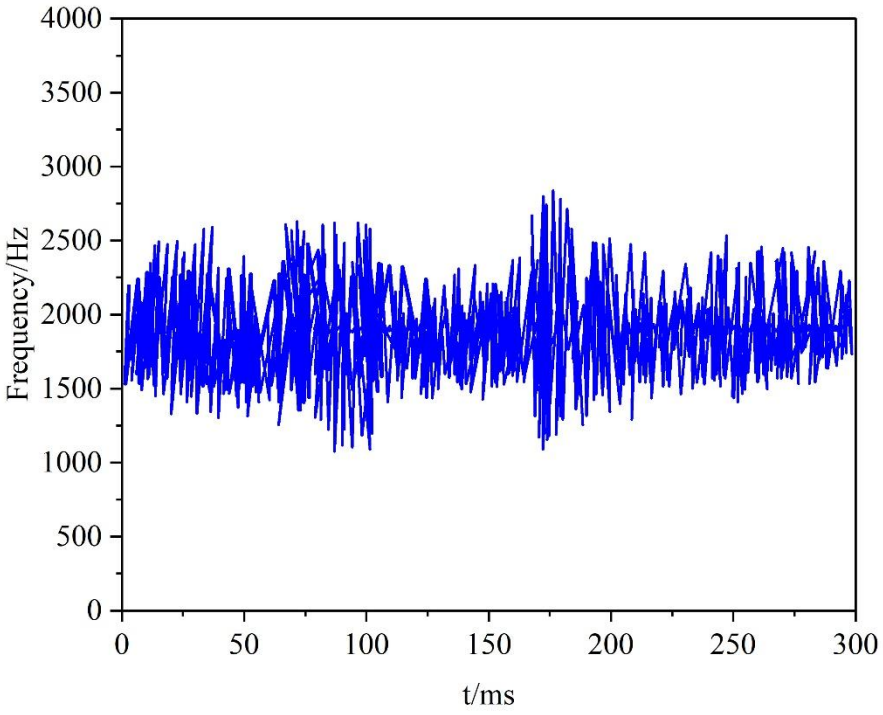


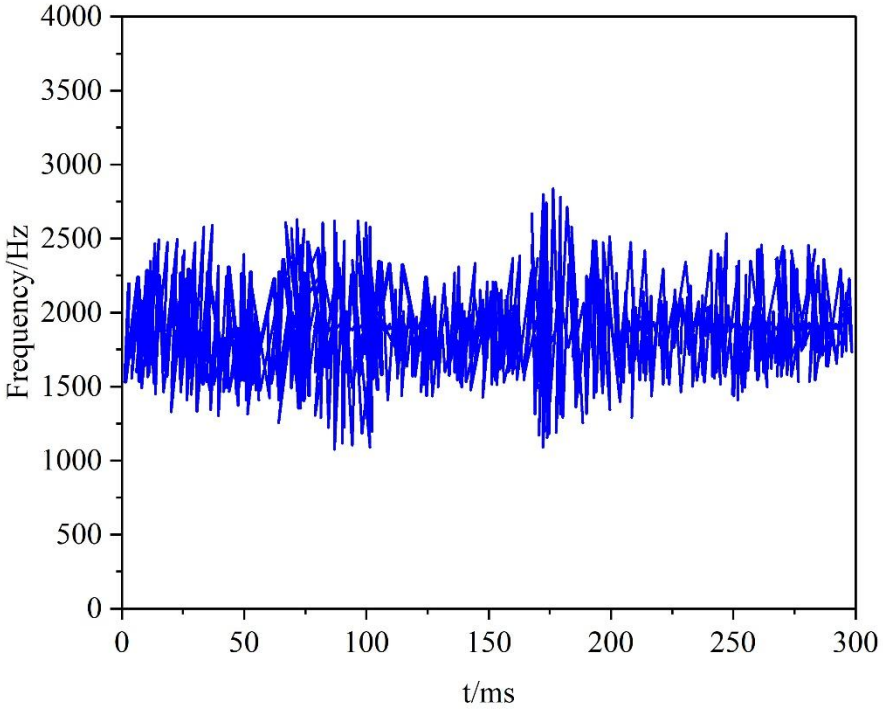
Figure 7. Second recording.

In order to demonstrate this characteristic of voiceprint stability more powerfully, we take two recordings of the phrase "Hello Little One" by Boy A and one recording of Girl A on the phrase "Hello Little One" as an example to further observe and compare the corresponding formant frequency values. Figure 8 shows a comparison of the first formant depicted in two recordings of the phrase "Hello Little One" by Boy A and one recording of the phrase "Hello Little One" by Girl A. Figure 9 is a comparison of the second formant depicted by the two recordings, and the (a)~(c) in Figure 8 and Figure 9 are the first and second recordings of boy A and the recording of girl A, respectively, and the formant of speech is

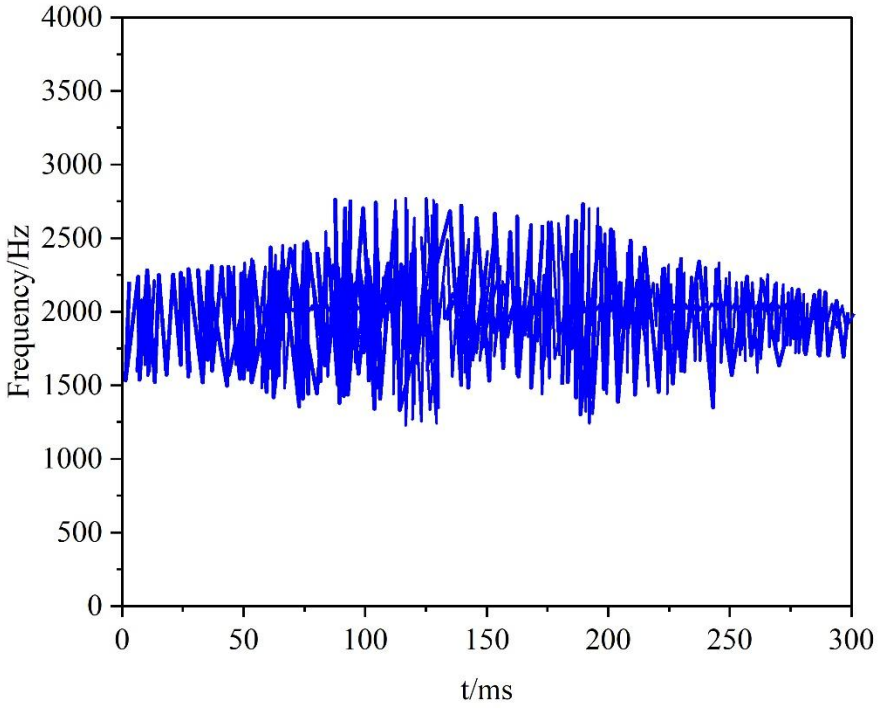
marked as F1, F2, F3, F4 and F5 respectively. For the sake of description, the first sentence of Boy A is A1, the second sentence is A2, and the recording of Girl A is A3.



(a) Boys for the first time

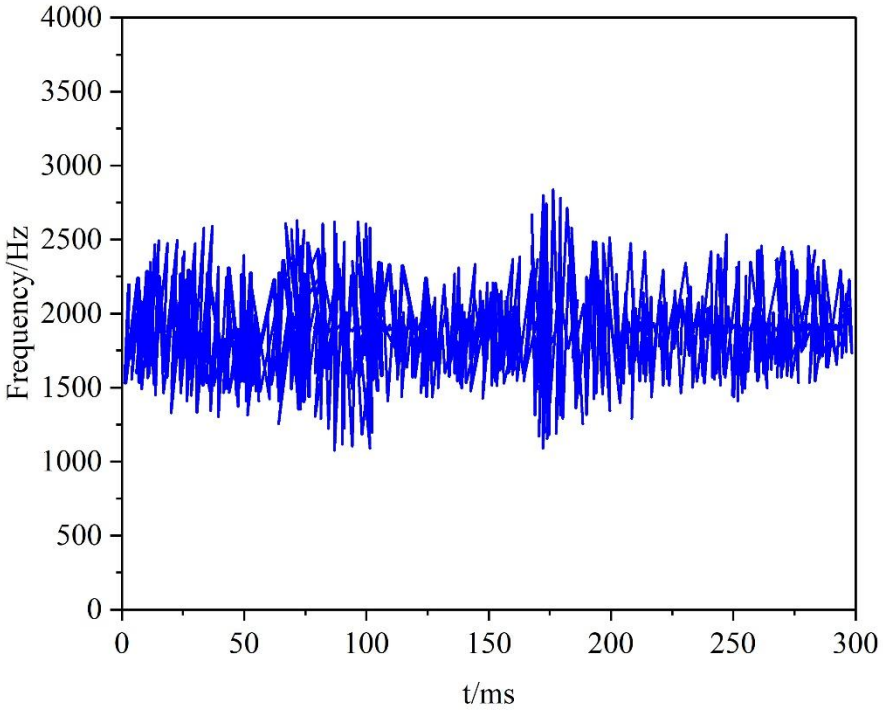


(b) Boys for the second time

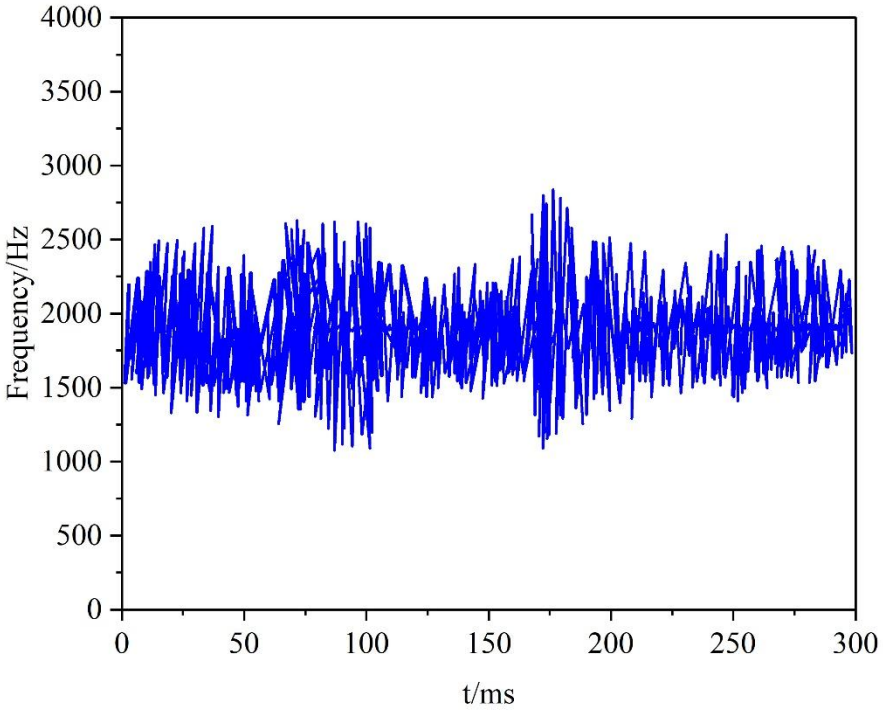


(c) Girls for the first time

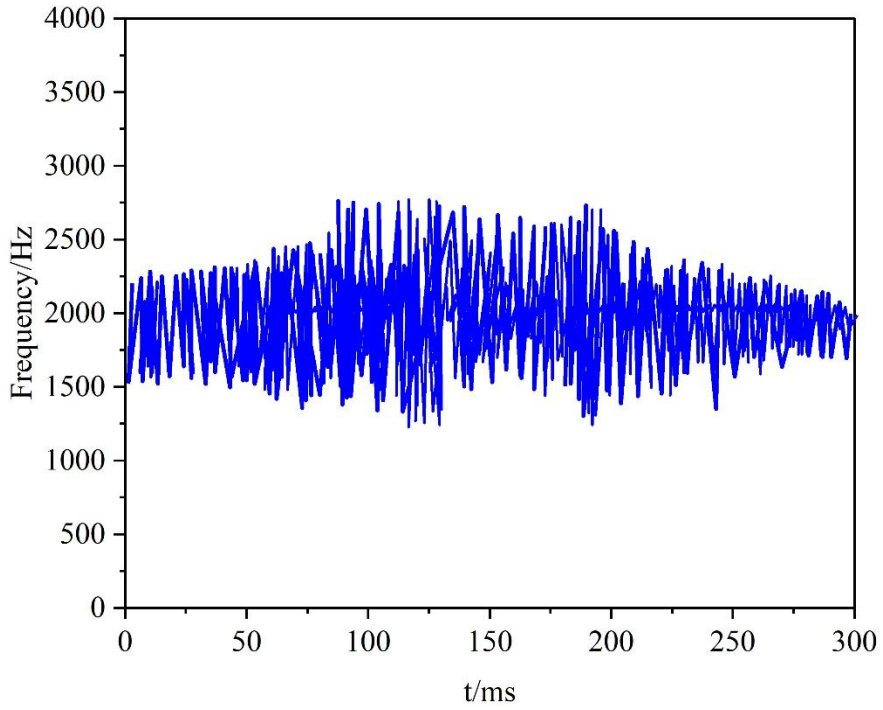
Figure 8. The first total vibration peak contrast diagram.



(a) Boys for the first time



(b) Boys for the second time



(c) Girls for the first time

Figure 9. The second resonance peak comparison diagram.

In order to facilitate the observation and comparison, we fill in the table with the values of resonance peaks F1, F2, F3, F4 and F5 corresponding to each of these three speech segments, as shown in Table 1.

From the above resonance peak comparison chart and the comparison of the resonance peak values in the above table, we can find that the resonance peak frequency directions of the two speech segments A1 and A2 of the same person, boy A, are generally consistent with each other, with great correlation, while the resonance peak frequency directions of the speech of the other person, girl A, as well as the peak

resonance peak frequency values and the differences between boy A and boy A are relatively large, and the resonance peak value of F5 and the boy's recording with the A1 differ by 1630 Hz, which is clearly not relevant. Thus, the stability and uniqueness of the voiceprint is more strongly argued. The above comparison shows that voiceprint identification technology has great applicability in the field of deep forgery defense of agricultural live e-commerce.

Table 1. Resonance peak (Hz).

Resonance peak	F1	F2	F3	F4	F5
A1	598.4	1865	2684	3654	4651
A2	549	1804	2815	3598	4659
A3	325.56	1055	2700	2541	4021

4.2 Sound Verification Results

4.2.1. Auditory Perception

After audition, the synthesized voices of the two AI virtual anchors have no significant differences in sound quality, purity, pitch, etc., compared with the voices of their respective prototypes; their Mandarin pronunciation is standardized without dialectal features, and they have already mastered a certain speech-flow phonological variation of prosodic patterns.

The differences between the synthesized speech and its prototype speech are mainly manifested in two aspects:

(1) The naturalness of the synthesized speech is not enough, and the overall sense of speech mechanics is stronger. The intonation and rhythm of the platitudes are uniform and raw and unnatural.

(2) There are errors in the expression of synthesized speech. First, the pause is inappropriate, such as the synthetic voice will be “Organization of Petroleum Exporting Countries commitment” broadcast break for “oil exporting countries / organization commitment”; Second, the polyphonic words tone misreading.

In terms of naturalness and emotionality, the lack of naturalness is mainly due to the difference in rhyme. Rhythm is the phenomenon of stress, rhythm and intonation in discourse. Synthesized speech rhythm is flat, lack of variation, difference between light and heavy sounds, and lack of personal speech emotion. Taking the aforementioned prose as an example, although the synthesized voice does not have intense emotional expression in news broadcasting, it still has a gap in emotional expression compared to the prototype voice, even if it is the same low emotional arousal voice in the news category. Emotional synthetic speech should be based on synthetic speech technology, adding rich rhyming control, so that the synthetic speech can express the speaker's emotions, the current need to strengthen the research on the link between rhyming features and emotional expression, which is also an important reference dimension to distinguish synthetic speech.

As for the expression error, it shows that the synthesized speech has no understanding of the semantic understanding of the text and the logic of expression, which is specifically manifested in the synthesized speech's inaccurate grasp of the rhyming boundary. In addition, the two prototypical hosts have standardized pronunciation of Mandarin, which itself does not carry dialectal accent characteristics. If dialectal Mandarin is used as the prototype for extracting speech features, it remains to be studied to what extent different synthesized speech methods can reflect the characteristics of dialectal Mandarin.

In summary, the synthesized speech is closer to its prototype in terms of sound quality, purity, pitch, etc., but still needs to be improved in dealing with the rhyme problem, which makes the synthesized speech lack of emotion and naturalness, and shows broadcasting errors and unrealistic characteristics.

4.2.2. Results of Speech Spectrum Analysis

As the key speech spectral feature of voiceprint testing, this paper focuses on the formant characteristics of synthesized speech. In the video used, the audio sample rate was 44.1kHz. 20 groups of the same pronunciation syllables (10 groups for female voices and 10 groups for male voices) were selected from the virtual e-commerce anchors and their respective prototype voices. Under the condition of spectral range of 4 and 8 kHz, the 20 groups of the same syllable voiceprints were observed and measured.

In terms of the spectral characteristics of the spectral range of 4 kHz, the spectral characteristics of 15 groups of the same tone are basically the same, and the differences of the remaining 5 groups of the same tone are mainly manifested in the high-frequency part of the 4 kHz spectral range (i.e., above 3 kHz). If

the spectral range is adjusted to 8 kHz, all 15 sets of the same tone are in the spectrum in the range of 4 kHz to 8 kHz, there is a significant difference. Taking the two syllables "jie and mei" as an example, the main difference between the broadband spectrum of the "jie" syllable range of the female synthesized speech and its prototype is more than 3.5 kHz, and the broadband spectrum of the "mei" syllable of the male synthesis speech and its prototype is more than 4 kHz.

Unless otherwise specified, the "high-frequency formants" referred to in this article are formants in the range of 3 kHz to 8 kHz. It is observed that there are two main differences in the spectral characteristics of high-frequency formants: (1) the number of formant peaks is different. In the high-frequency part, there are inconsistencies in the number of formant peaks in some of the same pronunciation groups. (2) The position of the formant is different. It is mainly manifested in the characteristics of formant frequency, direction, and morphology. Figure 10 shows the instantaneous power of the syllable "li" syllable of female synthesized speech and prototype speech, and the two curves represent the instantaneous power of the formant of the "li" syllable of female synthesized speech and its prototype speech, respectively, which can reflect the difference in formant position. F1 and F2 are very compatible with each other, but the third formant of female synthesized speech is located at 4.1kHz, and the third formant of the prototype voice is about 4.3kHz, and the formant frequency of the synthesized voice is obviously low, and the formant frequency of the synthesized voice is also low in the range of 6kHz to 7kHz.

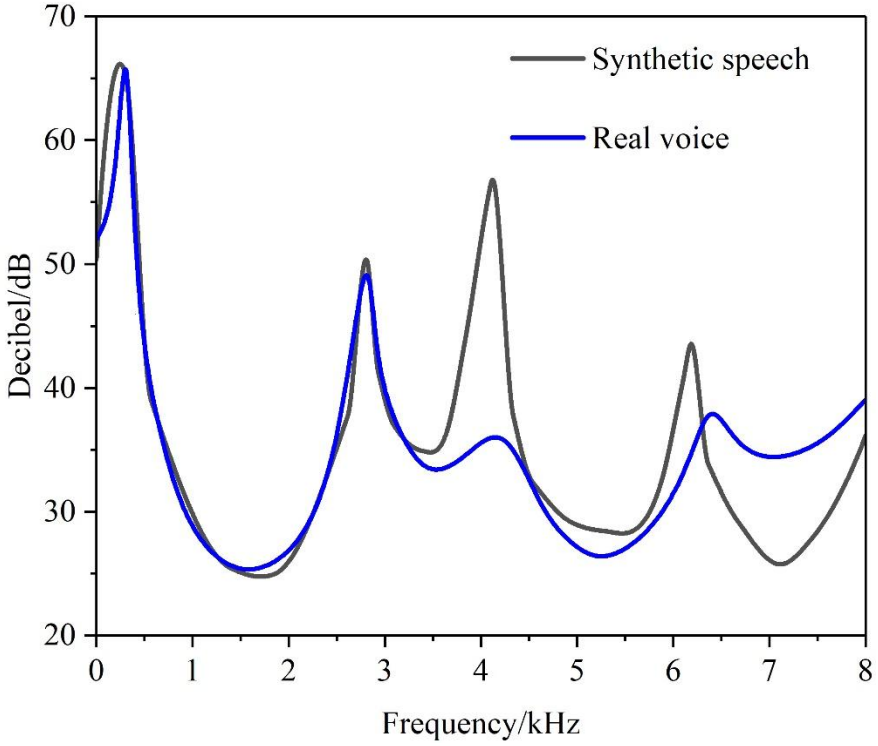


Figure 10. Female synthetic voice and prototype voice instantaneous power.

The resonance peak frequency characteristics reflect the shape of the vocal tract during the articulation process or at a certain time or moment in the articulation process. Daily identification due to recording equipment, recording quality and other issues, for more than 4kHz, especially more than 5kHz high-frequency resonance peaks less attention, but high-frequency resonance peaks is also the original vocal folds sound by the acoustic cavity resonance effect of the manifestation of the higher-frequency resonance peaks, although generally speaking, the resonance peaks of the energy tends to be weaker, but the same reflects the individual characteristics.

As far as the speech of the virtual anchor and its prototype in this study is concerned, the high-frequency formant is still quite stable. Taking the "ong" vowel of the syllable "zhong" as an example, Fig. 11 and Fig. 12 are the instantaneous power diagrams of the straight section of the finals when the syllable is pronounced multiple times by the male and female synthesized voices and their prototype voices.

In the abscissa range shown in the figure, the average frequency of the high-frequency formant of the male host is about 4.2 kHz and 7.2 kHz, respectively, and the average frequency of the high-frequency

formant frequency of the male synthesized voice is about 3.8, 5.6 and 7.6 kHz. The average frequency of the high-frequency formant of the female host's "ong" vowel is about 3.8 and 7.6 kHz, respectively, and the average frequency of the high-frequency formant frequency of her female synthesized voice is about 4.6 kHz. There is no significant difference in the low-frequency sections. In addition, the syllables such as "xi, er, shi" in the synthesized speech and its prototype speech, and the formant power diagram of the straight segment of the vowel after repeated multiple repetitions can reflect the individual stability and the difference with others.

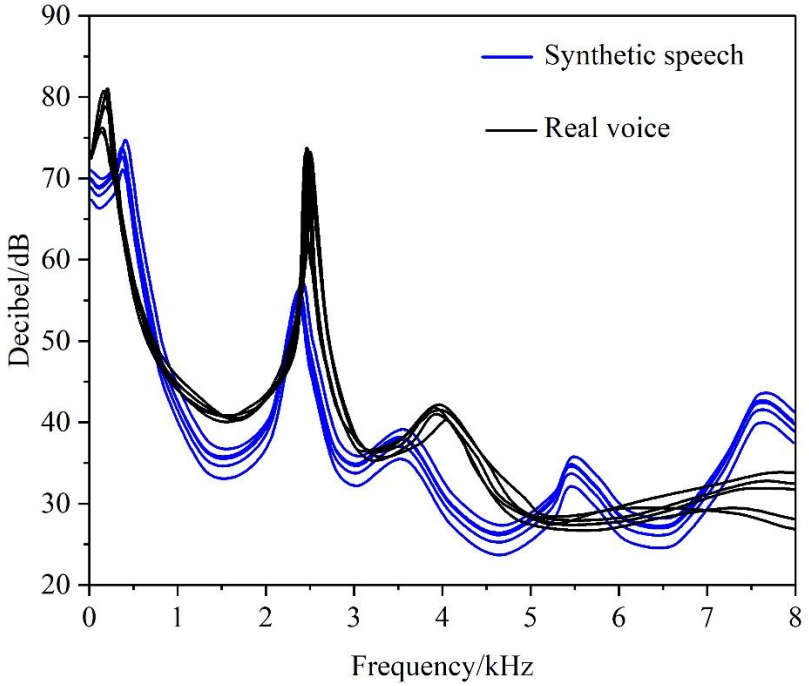


Figure 11. Male synthetic voice and original speech ong" instantaneous power.

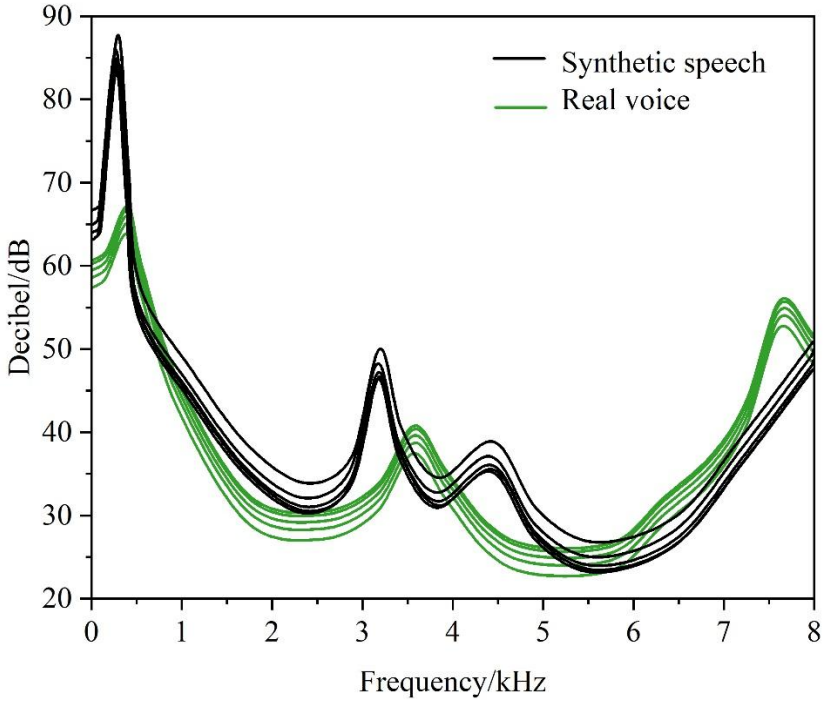


Figure 12. Female synthetic voice and original speech ong" instantaneous power.

5. Conclusion

In this paper, a deep learning based vocal synthesis and meta-learning vocal verification system is investigated to achieve deep forgery defense for live e-commerce of agricultural products.

The acoustic features of different people and the acoustic features of the same person are compared separately. The similarity of speech spectrograms described by the same speech of the same person is found to be extremely high. The similarity of the speech spectrograms described by the same speech of different people is low and the difference is high, which confirms the stability of the acoustic pattern and the uniqueness of the resonance peaks. At the same time, the feasibility of this paper's model to accurately recognize the voiceprint features of anchors enhances the credibility of this paper's system.

From the results of the spectrogram analysis, the synthesized speech lacks emotion and naturalness in auditory perception compared with the prototype speech, and is prone to sentence break errors and other situations. The differences between the high-frequency resonance peaks of the synthesized speech and the prototype speech are mainly manifested in the high-frequency resonance peaks above 4 kHz, and some syllables also have large differences from 3 kHz to 4 kHz. In this paper, the deep forgery defense system for live agricultural products e-commerce can effectively distinguish between real language and synthetic language, reflecting the importance of voiceprint verification in deep forgery detection, and providing an effective means of deep forgery defense for live agricultural products e-commerce.

Funding

This research was supported by the Jiangxi Agricultural University Academic Affairs Office School level Teaching Reform Project: Research on the "Restricted Competition Teaching Method" of the "Management Communication" Course from the Perspective of Learning and Practice Competitions (Project No. 2024B2Z17). Social Science Planning Project of Nanchang City, Jiangxi Province: Research on Entrepreneurial Obstacles and Countermeasures of New Vocational Farmers in the Main Grain Production Areas of Jiangxi Province under the Digital Economy (Project No. YJ202316).

References

1. Huang, X. (2023). Optimization of Marketing Strategy for “E-Commerce Live Streaming+ Agricultural Products” in the New Media Era. American Journal of Industrial and Business Management, 13(10), 1094-1103.
2. Dong, X., Zhao, H., & Li, T. (2022). The role of live-streaming e-commerce on consumers’ purchasing intention regarding green agricultural products. Sustainability, 14(7), 4374.
3. Zhou, G. J. (2022). Analysis and Research on the New Model of E-Commerce Poverty Alleviation: E-Commerce Live Broadcast of Agricultural Products. Journal of Accounting, Finance & Management Strategy, 17(2), 33-59.
4. Cao, J. (2024). Innovative Approaches and Value of Live Streaming for Agricultural Assistance—A Case Study of Douyin Platform. International Journal of Frontiers in Sociology, 6(6).
5. Sivathanu, B., Pillai, R., & Metri, B. (2023). Customers' online shopping intention by watching AI-based deepfake advertisements. International Journal of Retail & Distribution Management, 51(1), 124-145.
6. Fan, Y., Xie, M., Wu, P., & Yang, G. (2022, June). Real-time deepfake system for live streaming. In Proceedings of the 2022 International Conference on Multimedia Retrieval (pp. 202-205).
7. Ajimah, N., Ezukwoke, N., Dialoke, I., Odaba, A., & Iloanusi, O. (2020). Overview of voice biometric systems: voice person identification and challenges. TECHISD Proceedings, 2020, 57-62.
8. Harrison, O., Reed-Jones, J., Morrison, K., Robinson, C., & Jones, K. (2023). The Effect of Noise Reduction Upon Voiceprint Integrity. ICISNA 23 Proceedings Book, 211-213.
9. Shafik, A., Sedik, A., Abd El-Rahiem, B., El-Rabaie, E. S. M., El Banby, G. M., Abd El-Samie, F. E., ... & Iliyasu, A. M. (2021). Speaker identification based on Radon transform and CNNs in the presence of different types of interference for Robotic Applications. Applied Acoustics, 177, 107665.
10. Gao, Y., Singh, R., & Raj, B. (2018, April). Voice impersonation using generative adversarial networks. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 2506-2510). IEEE.
11. Gu, Y., Zhao, X., Yi, X., & Xiao, J. (2022, November). Voice conversion using learnable similarity-guided masked autoencoder. In International Workshop on Digital watermarking (pp. 53-67). Cham: Springer Nature Switzerland.
12. Lin, Y. S., Chen, H. Y., Huang, M. L., & Hsieh, T. Y. (2024). Data Augmentation for Voiceprint Recognition Using Generative Adversarial Networks. Algorithms, 17(12), 583.
13. Sun, W. Z., Wang, J. S., Zheng, B. W., & Li, Z. F. (2021). A novel convolutional neural network voiceprint recognition method based on improved pooling method and dropout idea. IAENG International Journal of Computer Science, 48(1), 202-212.
14. Wang, J. X. (2021). Meta-learning in natural and artificial intelligence. Current Opinion in Behavioral Sciences, 38, 90-95.

15. Paudyal, R., & Pyakurel, D. (2024). Enhancing the Efficiency of Deep Learning Models for Handwritten Text Recognition by Utilizing Meta-learning Optimization Techniques. Journal of Advanced College of Engineering and Management, 9, 1-13.
16. Pattanaik Rakesh Kumar & Mohanty Mihir Narayan. (2022). Nonlinear system identification for speech model using linear predictive coefficients based radial basis function. Journal of Information and Optimization Sciences,43(5),1139-1150.
17. Meryam Telmem,Naouar Laaidi & Hassan Satori. (2025). The impact of MFCC, spectrogram, and Mel-Spectrogram on deep learning models for Amazigh speech recognition system. International Journal of Speech Technology,28(1),1-14.
18. Sai Chen,Hong Cui Wang,Jia Jia,Ye Teng An & Jian Wu Dang. (2013). Comparison of Mel Frequency Ceptrum Coefficient and Perceptual Linear Predictive in Perceptual Measurement of Chinese Initials. Applied Mechanics and Materials,2700(411-414),291-297.
19. Zhiguo Wang,Zheng Wang,Kai Sun,Shi Chen,Yongfeng Zheng,Xiuyang Fang & Zhenbing Cai. (2025). Fretting fatigue damage and crack propagation of shot-peening dovetail joints assisted with the U-Net model. International Journal of Fatigue,199,109074-109074.