Semi-supervised Speech Lie Detection Algorithm Based on Multiple Features and Adaptive Threshold

Shuailong Zheng, Huawei Tao, Hang Yu, Ji Xi, Chunhua Zhu, and Xuemei Wang

***Abstract*—The limited number of labeled samples can significantly affect the accuracy of speech lie detection systems. However, semi-supervised algorithms have been shown to effectively improve detection performance under such constraints. To address this challenge, this paper proposes a semi-supervised speech lie detection algorithm that utilizes multiple features and adaptive thresholds. First, through comparative analysis, this paper identifies 312-dimensional acoustic features and 768-dimensional self-supervised features as the key elements of lie detection. Then, these acoustic features are processed using a bidirectional long short-term memory network (Bi-LSTM), which obtains comprehensive information by analyzing acoustic statistical features in both forward and backward directions. Subsequently, this paper develops a semi-supervised feature fusion module (BSFM) to extract deeper insights from basic features and achieves adaptive feature fusion through targeted training. In addition, based on a detailed analysis of the distribution characteristics of lie detection samples, this paper proposes an improved Freematch algorithm to better utilize unlabeled data. Experimental results show that the proposed algorithm outperforms most existing methods on the self-built H-Wolf dataset and the Columbia/SRI/Colorado Corpus (CSC).**

***Index Terms*—Semi-supervised, BSFM, adaptive threshold, machine learning.**

I. INTRODUCTION

LIE detection is a crucial research area in fields such as computational linguistics, psychology, and military science [1], [2]. The psychological phenomenon of lying is complex, influenced by a combination of emotions, cognition, and willpower. Traditional deception detection techniques primarily analyze physiological responses, which—though informative—require extended procedures and remain susceptible to intentional control by subjects. Recently, speech-based deception detection methods have emerged as a promising alternative, overcoming the limitations of traditional approaches by analyzing speech data and demonstrating significant research potential.

With continuous advancements in speech signal processing technology, a variety of acoustic features (e.g., MFCCs, energy, frequency) have been developed to extract key information from speech. These features effectively capture the low-level and localized characteristics of speech signals and have been widely applied in numerous speech analysis tasks. For instance, M. Graciarena et al. [3] improved system accuracy by combining prosodic and lexical features with mixed model scores based on acoustic features. Similarly, researchers at Columbia University [4] integrated multiple feature types for lie detection, achieving high accuracy on the CSC corpus. Kirchhuebel et al. [5] studied how different dialogue patterns influence deception detection by analyzing acoustic and temporal features, with a focus on emotional arousal, cognitive load, and self-control.

In recent years, the rise of deep learning has shifted the focus toward deep features, which capture higher-order and abstract characteristics of speech through deep neural networks. These features are highly sensitive to subtle changes in deceptive speech. Xie et al. [6] combined spectral features, leveraging the orthogonality and translation invariance of Hu matrices, with deep learning methods, achieving excellent classification results using deep belief networks. Liang et al. [7] employed convolutional long short-term memory (LSTM) networks to extract deep frame-level features, achieving promising recognition results on a self-constructed lie detection database. However, despite the advantages of deep features in representing data, their reliance on large labeled datasets and susceptibility to data distribution biases limit their generalizability, particularly in small-sample or unevenly distributed datasets. Insufficient labeled data remains a key challenge in advancing speech-based lie detection [8]. Semi-supervised learning has shown promise in addressing this issue. For instance, Fernandes et al. [9] achieved high-precision deception detection by analyzing three conversation recordings. They extracted cepstral/spectral energy features and then applied a hybrid approach combining the Levenberg-Marquardt optimization algorithm with an LSTM-based classification algorithm for nine different combinations of training-testing configurations. Su et al. [10] separately trained BiLSTM networks and SVM models, then fused their classification results through a decision-level scoring scheme. Fang et al. [11] proposed a hybrid model

Manuscript received December 2, 2024; revised April 21, 2025.

This work was founded in part by Innovative Funds Plan of the Henan University of Technology (2022ZKCJ13), the National Natural Science Foundation of China (No. 62001215), Open project of scientific research platform of Henan University of Technology Grain Information Processing Center (No. KFJJ2023011), Natural Science Project of Henan Provincial Department of Science and Technology, Technology research projects (No. 242102211027).

S. Zheng is a master's student in Key Laboratory for Grain Information Processing and Control, Henan Key Laboratory of Grain Storage Information Intelligent Perception and Decision Making, Henan University of Technology, Zhengzhou 450001, China (e-mail: 2264300285@qq.com).

H. Tao is an associate professor in Key Laboratory for Grain Information Processing and Control, Henan Key Laboratory of Grain Storage Information Intelligent Perception and Decision Making, Henan University of Technology, Zhengzhou 450001, China (corresponding author, e-mail: thw@haut.edu.cn).

H. Yu is a master's graduate in Key Laboratory for Grain Information Processing and Control, Henan Key Laboratory of Grain Storage Information Intelligent Perception and Decision Making, Henan University of Technology, Zhengzhou 450001, China (e-mail: yuhang202303@163.com).

J. Xi is a professor of Computer Information Engineering, Changzhou Institute of Technology, Jiangsu 213000, China (e-mail: xiji@czust.edu.cn).

C. Zhu is a professor in Key Laboratory for Grain Information Processing and Control, Henan Key Laboratory of Grain Storage Information Intelligent Perception and Decision Making, Henan University of Technology, Zhengzhou 450001, China (e-mail: zhuchunhua@haut.edu.cn).

X. Wang is an associate professor in Key Laboratory for Grain Information Processing and Control, Henan Key Laboratory of Grain Storage Information Intelligent Perception and Decision Making, Henan University of Technology, Zhengzhou 450001, China (e-mail: wangxuemei @haut.edu.cn).

combining semi-supervised denoising autoencoders (DAEs) with fully supervised LSTM networks, significantly improving the accuracy of semi-supervised speech lie detection. Building on these advancements, this paper enhances the Freematch algorithm [12] by introducing an adaptive thresholding mechanism. An initial global threshold is established based on sample distribution and iteratively optimized using batch confidence. Local thresholds are generated using the exponential moving average method, and the final adaptive threshold is formed by normalizing and combining both the global and local thresholds. This iterative adjustment process enhances model performance and data utilization.

In summary, this paper proposes a semi-supervised speech lie detection algorithm that integrates multiple features and adaptive thresholds. Through a comprehensive comparative analysis of speech features, 312-dimensional acoustic features and 768-dimensional self-supervised features are selected as the basis for lie detection to ensure the comprehensive presentation of multi-dimensional speech information. This paper constructs a semi-supervised feature fusion model (BSFM) and uses the bidirectional sequence modeling capability of the Bi-LSTM network to effectively fuse and extract deep information from acoustic features and self-supervised features. In order to further improve the utilization of unlabeled data, this paper introduces an improved Freematch algorithm, which significantly improves the robustness, efficiency, and generalization ability of the model.

II. Models and methods

The algorithm framework proposed in this paper is illustrated in Fig. 1. The model consists of modules for feature extraction, feature processing and fusion, a semi-supervised adaptive threshold algorithm, and other components. It aims to fully leverage the complementary information between acoustic statistical features and pre-trained features by introducing a semi-supervised adaptive threshold strategy. This strategy is designed to improve the quality of pseudo-labels and enhance the model's classification performance. The specific details are described as follows:

A. Feature extraction module

1) Acoustic feature extraction: Selecting the appropriate set of acoustic features associated with deception is crucial for improving the recognition performance of lie detection models. Therefore, this paper selects commonly used features in speech-based lie detection from both the time and frequency domains and analyzes their roles.

Time-Domain Characteristics: In the time domain, features such as root mean square energy (RMSE) and short-time average zero-crossing rate (zero_crossing_rate) are selected. Both RMSE and the zero-crossing rate are highly sensitive to energy fluctuations and emotional variations in speech, effectively reflecting the unnatural behaviors associated with lying.

Frequency-Domain Characteristics: In the frequency domain, the selected features include Mel-frequency cepstral coefficients (MFCC), Mel spectrum, fundamental frequency (F0), spectral centroid (spectral_centroid), spectral flatness (spectral_flatness), chroma frequency (chroma_stft), spectral contrast (spectral_contrast), and magnitudes. These features, when combined with variations in frequency, pitch, noise level, and energy distribution, can effectively capture potential deceptive behavior.

The specific characteristics of these features are detailed in Table I.

TABLE I
FEATURE COMBINATION

Feature name	RMSE
F0	4
spectral_centroid	3
spectral_flatness	1
MFCC	150
chroma_stft	12
melspectrogram	128
spectral_contrast	7
zero_crossing_rate	1
Magnitudes	3
Rmse	3

2) Depth feature extraction: Relying solely on acoustic features may not fully capture the complex patterns of deceptive behavior. To more effectively explore the intricate patterns and subtle changes in speech during lie detection, this study incorporates higher-level deep features alongside traditional acoustic features. Specifically, the WavLM Base+ model [13] is used to extract pre-trained features. This model consists of 12 Transformer encoder layers, 768-dimensional hidden states, and 8 attention heads, which enhance its generalization capabilities. As a result, the model successfully extracts 768-dimensional pre-trained features, denoted as X_{768}.

B. Feature processing and fusion module

The acoustic and deep features reside in different feature spaces. To enhance the fusion performance of these diverse feature types, a Feature Fusion Processing Module (BSFM) is constructed. This module aims to more accurately capture signs of deception in speech by leveraging both types of features. First, due to the significant temporal dependencies of acoustic features in expressing lies, a bidirectional Long Short-Term Memory network (BiLSTM) is employed to capture temporal changes in deceptive speech. The Gaussian Error Linear Unit (GELU) activation function is used to effectively activate the output features through a smooth probabilistic transformation. Additionally, dropout is applied after the GELU activation function to prevent overfitting. Second, the tensor stitching method is used to integrate information from different feature spaces along the feature dimension, resulting in a fused feature vector $X = [X_{BiLSTM}, X_{768}]$. This comprehensive approach ensures that the model effectively combines temporal acoustic information with deep pre-trained features, thereby enhancing the accuracy and robustness of lie detection.

C. Semi-supervised adaptive threshold algorithm

The Freematch algorithm [12] has performed well in various semi-supervised benchmarks. In this paper, we

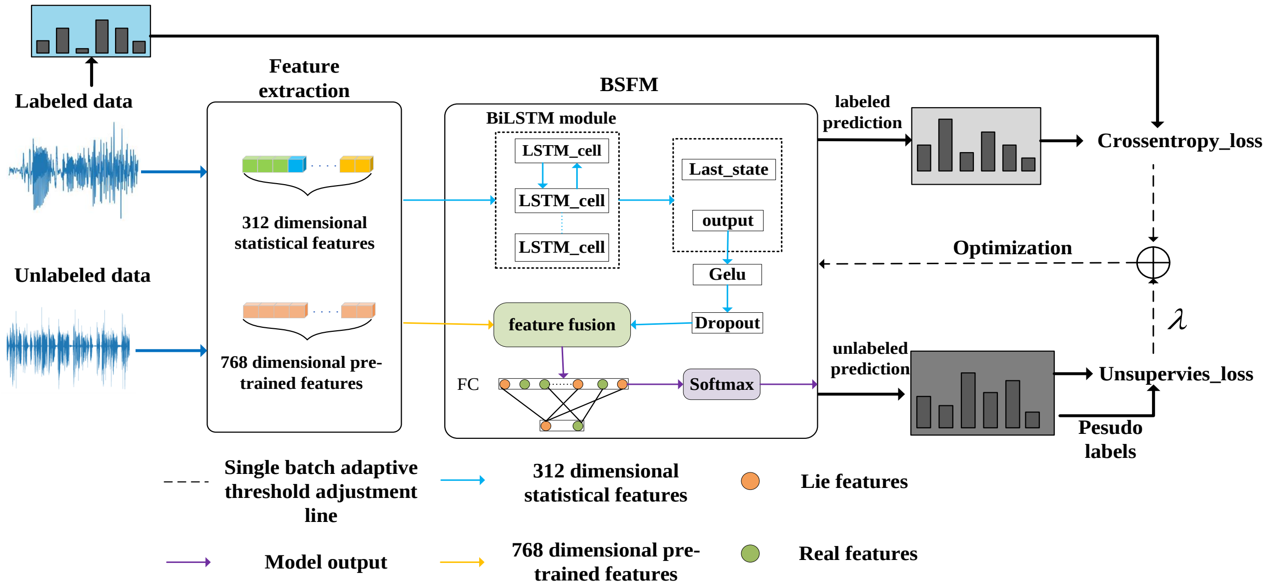

Fig. 1. Algorithm Model Framework

enhance the algorithm by introducing an adaptive threshold method (LD-SAT) specifically for speech lie detection. The improved algorithm first sets an initial global threshold based on the distribution of training set samples, which allows the model to filter out feature data with higher confidence. At the same time, the global threshold is calculated using the average maximum prediction confidence of each batch of samples. This direct and effective method helps to mitigate the impact of noise on threshold calculation. Subsequently, the threshold is gradually optimized by combining the global threshold of the previous batch of samples. This iterative process continuously improves the quality of pseudo labels. The overall algorithm flow is shown in Fig. 2. In summary, the proposed adaptive threshold method (LD-SAT) effectively improves the quality of pseudo labels and improves the robustness and accuracy of speech lie detection models.

III. Experiment and Result Analysis

A. Dataset

To verify the effectiveness of the proposed algorithm, this study used the professional CSC database, recorded by Columbia University for speech lie detection research, as well as the self-constructed H-Wolf corpus [14]. From the CSC corpus, 5,411 speech samples were selected, comprising 3,202 genuine speech samples and 2,209 deceptive speech samples. The data were split in a 9:1 ratio, resulting in 4,869 training samples and 542 testing samples. Among the training data, 600, 800, and 1,000 samples were randomly selected as labeled data, while the remaining samples were treated as unlabeled data. Additionally, 1,103 speech samples were selected from the H-Wolf corpus, which were also divided into training and testing sets in a 9:1 ratio. Within the training set, 100 and 600 samples were randomly chosen as labeled data, with the remaining samples designated as unlabeled data. These experimental setups on both the CSC and H-Wolf datasets demonstrate the robustness and applicability of the proposed algorithm in various lie detection scenarios.

B. Experimental settings and evaluation criteria

All experiments in this paper were conducted on a Windows 11 operating system using an NVIDIA RTX 3080 graphics card. The programming language used was Python, and the deep learning framework employed was PyTorch. To mitigate overfitting, the dropout rate was set to 0.9. During training, the Adam optimization algorithm was applied for up to 100 epochs, with an initial threshold of 0.7 and a learning rate of 0.00003. The number of iterations was fixed at 500. Each experiment was repeated 10 times, and the mean value was calculated to eliminate the influence of random errors. Weighted Accuracy (WA) and Unweighted Accuracy (UA) are the main evaluation metrics for recognition performance in speech lie detection, and their formulas are given in (1) and (2):

$$WA = \frac{Thenumberofsamplescorrectlydetected}{Totalsamplesize} \tag{1}$$

$$UA = \frac{TP}{TP + FN} \tag{2}$$

Where TP represents the number of positive samples predicted correctly and FN represents the number of negative samples predicted incorrectly.

C. Ablation experiment

The detailed settings for each group of ablation experiments are as follows:

1) Base: Only 312-dimensional acoustic statistical features are used and no semi-supervised features are used;
2) BSFM: pre-training features are added on the basis of Base and a semi-supervised learning algorithm is not adopted;
3) Base + SAT: a semi-supervised adaptive threshold algorithm is added on the basis of Base;
4) BSFM + SAT: The proposed algorithm.

Table II presents the ablation experimental results of the proposed algorithm on the CSC corpus. It is evident from

Algorithm 1 The training process of the adaptive optimization algorithm

Input: p_model, $label\_hist$, $time\_p$, $probs\_x\_ulb$
Output: p_model, $label\_hist$, $time\_p$

Get maximum probabilities and indices:
$max\_probs, max\_idx \leftarrow \max(probs\_x\_ulb, \mathrm{dim} = -1, \mathrm{keepdim{=}True})$
if use_quantile **then**
  $time\_p \leftarrow time\_p \cdot m + (1 - m) \cdot \mathrm{quantile}(max\_probs, 0.8)$
else
  $time\_p \leftarrow time\_p + (1 - m) \cdot \mathrm{mean}(max\_probs)$
end if
if clip_thresh **then**
  $time\_p \leftarrow \mathrm{clip}(time\_p, 0, 0.95)$
end if
Update p_model:
$p\_model \leftarrow p\_model \cdot m + (1 - m) \cdot \mathrm{mean}(probs\_x\_ulb, \mathrm{dim} = 0)$
Calculate histogram:
$hist \leftarrow \mathrm{bincount}(max\_idx.reshape(-1), \mathrm{minlength} = p\_model.shape[0]).to(p\_model.dtype)$
Update $label\_hist$:
$label\_hist \leftarrow label\_hist \cdot m + (1 - m) \cdot \left(\frac{hist}{\sum(hist)}\right)$
return p_model, $label\_hist$, $time\_p$

Fig. 2. Flow chart of semi-supervised adaptive threshold algorithm

the table that the Base model achieves accuracies of 63.03%, 64.01%, and 63.75% when the number of labeled samples is 600, 800, and 1,000, respectively. This indicates that the 312-dimensional acoustic statistical features possess strong characterization capabilities. When the number of labeled samples is 600, 800, and 1,000, the BSFM model achieves accuracies of 63.03%, 65.00%, and 64.28%, respectively, representing improvements of 0.00%, 0.99%, and 0.53% over the Base model. This demonstrates that adding pre-trained features effectively enhances the model's feature representation ability. The Base + SAT algorithm introduces an adaptive thresholding mechanism to fully utilize the characteristics and structural information of unsupervised samples, thereby extracting high-quality pseudo-labels. With 600, 800, and 1,000 labeled samples, the model achieves accuracies of 62.85%, 65.35%, and 66.07%, respectively, which are increases of 0.18%, 1.34%, and 2.32% over the Base model. The BSFM + SAT algorithm proposed in this paper achieves accuracies of 63.93%, 65.71%, and 66.96% for 600, 800, and 1,000 labeled samples, respectively, outperforming all other configurations. Analysis of the ablation experiment results reveals that the BSFM feature fusion model not only retains important characterization information from the acoustic statistical features but also fully leverages the generic representations from the pre-trained features. Additionally, the adaptive threshold method enhances the utilization of unlabeled data, thereby improving the accuracy of lie detection. Fig. 3 illustrates that the proposed algorithm combines the advantages of feature fusion and the adaptive threshold method, resulting in improved classification accuracy and stability. Table III displays the experimental results of the proposed algorithm and each ablation task on the H-Wolf corpus. It is clear that the WA and UA metrics of the proposed algorithm outperform those of the other ablation tasks, further validating the effectiveness of the proposed approach.

TABLE II
DETECTION ACCURACY (%) OF EACH GROUP OF ABLATION EXPERIMENTS ON CSC CORPUS

Database	Model	Labels number		
		600	800	1000
CSC	Base	63.03	64.01	63.75
	BSFM	63.03	65.00	64.28
	Base + SAT	62.85	65.35	66.07
	BSFM + SAT	**63.93**	**65.71**	**66.96**

TABLE III
DETECTION ACCURACY (%) FOR EACH GROUP OF ABLATION EXPERIMENTS ON THE H-WOLF CORPUS

Database	Model	Labels number			
		100	100	600	600
		WA	UA	WA	UA
CSC	Base	59.77	56.98	64.39	66.78
	BSFM	61.43	60.78	65.75	66.78
	Base + SAT	60	58.29	66.06	68.24
	BSFM + SAT	**62.57**	**63.60**	**68.26**	**69.78**

D. Comparative experiment

To further verify the detection performance of the proposed algorithm, it is compared with other feature sets, semi-supervised algorithms, and speech lie detection

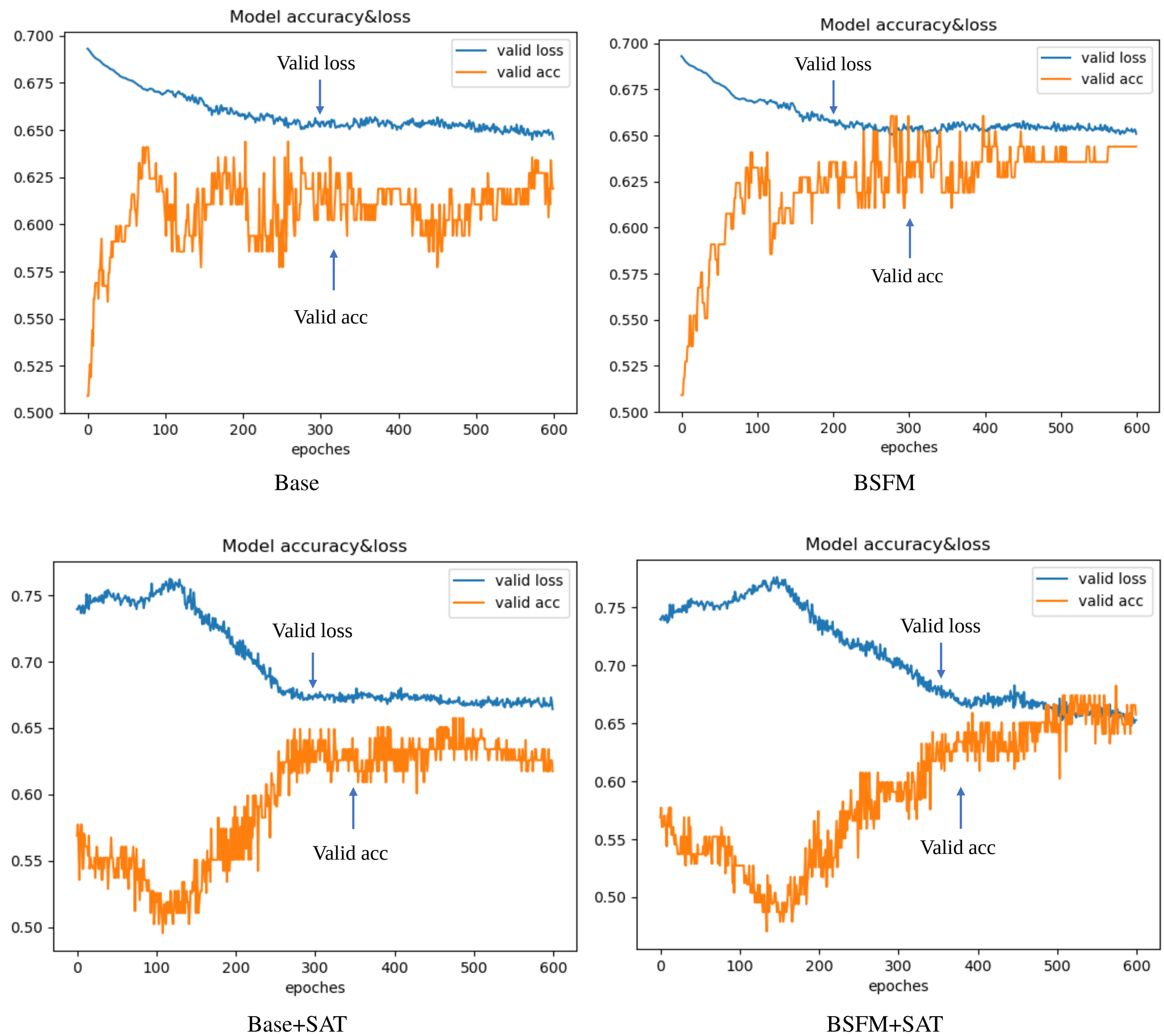

Fig. 3. The change trend of accuracy and loss of each group of ablation tasks on the H-wolf corpus

algorithms. The specific settings for each group of algorithms are as follows:

1) 384D + SAT: Combines 384-dimensional acoustic statistical features [15] with a semi-supervised adaptive thresholding algorithm. The 384-dimensional acoustic statistical features are among the most commonly used in lie detection.
2) mixmatch: Combines 312-dimensional acoustic statistical features with the classical MixMatch algorithm [16];
3) 312D + SAT: Combines 312-dimensional acoustic statistical features with the semi-supervised adaptive thresholding algorithm;
4) AE+MT+CR: A hybrid network model based on a self-encoding network and a mean-teacher network, combined with a consistency regularization method;
5) fixmatch: Adds the classic FixMatch algorithm to the feature fusion model BSFM [17].

Tables IV and V show that the recognition performance of the 312-dimensional acoustic statistical features combined with the SAT method (312D + SAT) surpasses that of the 384-dimensional acoustic statistical features with SAT (384D + SAT) and the MixMatch algorithm overall. This indicates that the 312-dimensional acoustic statistical features are better suited for lie detection tasks compared to the 384-dimensional features. Furthermore, compared to the MixMatch algorithm, the adaptive threshold method (SAT) demonstrates significant improvements in enhancing the model's classification performance. The proposed algorithm also outperforms AE+MT+CR and fixMatch across all tests. This underscores that integrating new acoustic statistical features with pre-trained features allows the model to learn more effective feature representations, thereby improving its classification performance. Additionally, the adaptive threshold algorithm effectively adjusts the threshold range and filters out higher-quality pseudo-labels, further enhancing the model's classification accuracy. Overall, these results highlight the advantages of combining advanced acoustic statistical features with pre-trained features and using an adaptive thresholding strategy to improve lie detection performance.

TABLE IV
DETECTION ACCURACY (%) OF EACH ALGORITHM ON CSC CORPUS

Database	Model	Labels number		
		600	800	1000
CSC	384D+ SAT	62.32	63.57	61.96
	Mixmatch	63.14	64.44	65
	312D+ SAT	62.85	65.35	66.07
	AE+MT+CR	62.12	62.68	63.06
	Fixmatch	62.67	**65.89**	65.71
	Proposed	**63.93**	65.71	**66.96**

TABLE V
DETECTION ACCURACY (%) OF EACH ALGORITHM ON H-WOLF CORPUS

Database	Model	Labels number			
		100	100	600	600
		WA	UA	WA	UA
CSC	384D+ SAT	62.27	63.56	63.25	68.02
	Mixmatch	61.43	60.78	65.75	66.78
	312D+ SAT	60	58.29	66.06	68.24
	AE+MT+CR	59.61	52.53	66.34	64.82
	Fixmatch	61.59	61.42	67.42	68.60
	Proposed	**62.57**	**63.60**	**68.26**	**69.78**

E. Confusion Matrix

The model performance is systematically evaluated through confusion matrix. In the experiment on CSC dataset (Fig. 4), as the sample size increases, the recognition performance of the model for real and deceptive sentences shows different performance: when the number of labeled samples increases from 600 to 800, the recognition accuracy of real sentences increases from 64% to 67%, while the recognition rate of deceptive sentences remains stable at 60%; when the sample size is expanded to 1000, the recognition rate of real sentences remains at 67%, and the recognition rate of deceptive sentences remains at 60%. It is worth noting that under all experimental conditions of CSC dataset, the recognition accuracy of the model for deceptive sentences remains stable above the threshold of 60%.

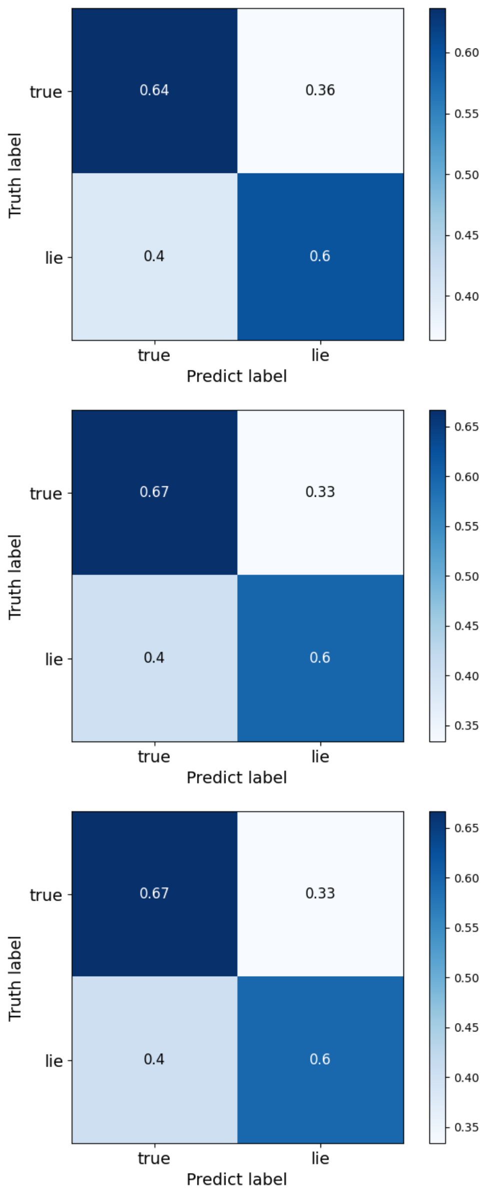

Fig. 4. Confusion matrix of the model when the number of CSC database labels is different

IV. CONCLUSION

To further optimize the multi-feature semi-supervised speech lie detection algorithm, this paper proposes a semi-supervised lie detection algorithm based on multiple features and adaptive thresholds. The aim is to explore new features with strong representation ability that can be used in speech lie detection research, enhancing the understanding of speech signals. Meanwhile, an adaptive threshold adjustment algorithm is introduced to better filter high-quality features, thereby improving the model's classification performance. Experimental results on the CSC dataset confirm the effectiveness of the proposed algorithm for speech lie detection tasks.

REFERENCES

[1] Y. Zhi and W. Hong, "Research and analysis of speech lie detection technology," Guangdong Public Security Technology, vol. 29, pp. 48–50, 2021.

[2] Z. Li, L. Ruiyu, X. Yue, and Z. Dongze, "Research status and prospects of speech lie detection technology," Data Collection and Processing, vol. 32, pp. 246–257, 2017.

[3] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg and S. Kajarekar, "Combining Prosodic Lexical and Cepstral Systems for Deceptive Speech Detection," 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, France, 2006, pp. I-I.

[4] Enos F, Shriberg E, Graciarena M, et al. Detecting deception using critical segments[C]//Interspeech. 2007: 2281-2284.

[5] Khawaja, M. A., Chen, F., & Marcus, N. (2014). Measuring Cognitive Load Using Linguistic Features: Implications for Usability Evaluation and Adaptive Interaction Design. International Journal of Human–Computer Interaction, 30(5), 343–368.

[6] XIE Yue, LIANG Ruiyu, BAO Yongqiang, ZHAO Li, WANG Shijia. Deception detection with spectral features based on deep belief network[J]. ACTA ACUSTICA, 2019, 44(2): 214-220.

[7] Y. Xie, R. Liang, H. Tao, Y. Zhu and L. Zhao, "Convolutional Bidirectional Long Short-Term Memory for Deception Detection With Acoustic Features," in IEEE Access, vol. 6, pp. 76527-76534, 2018.

[8] Hirschberg J, Benus S, Brenier J M, et al. Distinguishing deceptive from non-deceptive speech[C]//Interspeech. 2005, 2: 4.

[9] S. V. Fernandes and M. S. Ullah, "Use of Machine Learning for Deception Detection From Spectral and Cepstral Features of Speech Signals," in IEEE Access, vol. 9, pp. 78925-78935, 2021.

[10] Su, B. -H., Yeh, S. -L., Ko, M. -Y., Chen, H. -Y., Zhong, S. -C., Li, J. -L., Lee, C. -C. (2018) Self-Assessed Affect Recognition Using Fusion of Attentional BLSTM and Static Acoustic Features. Proc. Interspeech 2018, 536-540.

[11] Yuanbo FANG, Hongliang FU, Huawei TAO, Ruiyu LIANG, Li ZHAO, "A Novel Hybrid Network Model Based on Attentional Multi-Feature Fusion for Deception Detection" in IEICE TRANSACTIONS on Fundamentals, vol. E104-A, no. 3, pp. 622-626, March 2021.

[12] Wang Y, Chen H, Heng Q, et al. Freematch: Self-adaptive thresholding for semi-supervised learning[J]. arXiv preprint arXiv:2205. 07246, 2022.

[13] S. Chen et al., ”WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing,” in IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1505-1518, Oct. 2022.

[14] Hung, H.; Chittaranjan, G. The idiap wolf corpus: Exploring group behaviour in a competitive role-playing game. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 879–882.

[15] Schuller B, Steidl S, Batliner A. The interspeech 2009 emotion challenge[J]. 2009.

[16] Berthelot D, Carlini N, Goodfellow I, et al. Mixmatch: A holistic approach to semi-supervised learning[J]. Advances in neural information processing systems, 2019, 32.

[17] Sohn K, Berthelot D, Carlini N, et al. Fixmatch: Simplifying semi-supervised learning with consistency and confidence[J]. Advances in neural information processing systems, 2020, 33: 596-608.