ISSN 2616-7433 Vol. 6, Issue 7: 17-26, DOI: 10.25236/FSST.2024.060704

An Intelligent Speech Recognition Method Based on Stable Learning

Zhichao Zhou[1,2], Chaofan Hu[1,2,*]

[1]School of Mechanical and Electrical Engineering, Guilin University of Electronic Technology, 541004 Guilin, China
[2]Guangxi Key Laboratory of Manufacturing System & Advanced Manufacturing Technology, Guilin University of Electronic Technology, 541004, Guilin, China
**Corresponding author*

Abstract: *Speech is the main way of human communication, which carries both the speaker's information and the speaker's emotion. A variety of applications can harness emotion in speech to serve human needs more effectively. The deep learning algorithm is a practical solution to the classification nature of speech recognition. Various algorithms have been widely utilized for voice data and achieved remarkable performance. However, in real life, the testing data to be under a different distribution from the training data, this will cause the out-of-distribution(OOD) problem. This article proposes a new domain generalization method for speech classification based on Stable Learning (StableNet) to address the OOD problem. The StableNet can remove the connection between features through learning weights for training samples, which makes deep models learn more useful features instead of the fake connection between the discriminative features and labels. We evaluate the performance of the proposed method by conducting speech classification experiments on voice datasets. We also investigate the importance of various features on speech classification in noisy environments. The effects of proposed method on speech recognition performance are evaluated.*

Keywords: *Voice recognition; Domain Generalization; Stable Learning; Deep Learning; Signal Processing*

1. Introduction

Speech signals are the most common means of communication. Today, speech signals are used in many real life applications like speech recognition, speaker identification, gender identification and speech pre-processing for hearing and devices. With the development of multimedia communication and network communication, classification is more and more widely used in voice applications.Voice Activity Detector (VAD), used in the speech coding standard, is one of the most famous classification applications in speech processing [1-2]. Changes in speech production can be substantial and therefore have a considerable impact on the performance of speech processing applications such as recognition [3].

There have been a number of researches which focus on the variability effects of voice recognition[4]. For example, the speech recognition performance has been studied in [5-8]. Reference[9]presented a new set of speech features for the speech intelligibility detection of impaired speeches for children with Cerebral Palsy and hearing impairment. An isolated digit recognition system is developed to recognize the speeches of speech-impaired people affected with dysarthria[10]. Reference[11] presents a study exploring both conventional DNNs and deep Convolutional Neural Networks (CNN) for noise- And channel-degraded speech recognition tasks using the Aurora4 dataset. Although human listeners can understand speech under reverb conditions, suggesting that the auditory system is robust to this degradation, reverb results in a high rate of word errors for automatic speech recognition systems[12]. Normal speech is converted to difficult speech to enhance the data, and machine learning classifiers are used for classification [13]. Speech recognition is implemented using Convolutional networks [14]. Speech recognition uses a variety of modeling techniques to ensure better performance [15]. Rhythmic knowledgeof dysarthric speeches improves the system's accuracy [16] .

The Identification of Significant Speech Features. These studies show that speech variability is a challenging research problem, and traditional techniques are usually far from sufficient to improve the robustness of speech processing performance under some conditions.

ISSN 2616-7433 Vol. 6, Issue 7: 17-26, DOI: 10.25236/FSST.2024.060704

The traditional methods are too complicated and laborious to achieve because they require a lot of a priori knowledge and complex preprocessing, which limits their effectiveness and flexibility. In addition, both traditional machine learning-based and typical deep learning-based methods have a strong assumption that the data from the source domain and target domain must be acquired under the same distribution.

Therefore, we propose a new method based on stable learning and GAF to address the O.O.D problem in voice recognition field. The main contributions of this article are as follows:

(1) We propose an intelligent voice recognition method based on GAF and stable learning to deal with the poor model generalization ability caused by data distribution shift between the source and target domains in speech.

(2) In the training phase, our method separates the relevant and irrelevant features of the source domain sample data through the sample feature weighted decorrelation module and feeds the relevant features to the classifier for training, improving the model's generalization ability.

The rest of our article is organized as follows: Section 2 introduces GAF and stable learning algorithms. Section 3 discusses the structure of a generalized model for intelligent voice recognition based on GADF and stable learning. Section 4 verifies the performance of our method through defect diagnosis on two bearing datasets. Section 5 concludes the article.

2. Theoretical Background

2.1. Domain Generalization in voice recognition

Domain generalization aims to learn domain invariant knowledge from multiple source domains and to generalize it to the unseen target domain task, which is shown in Figure 1.

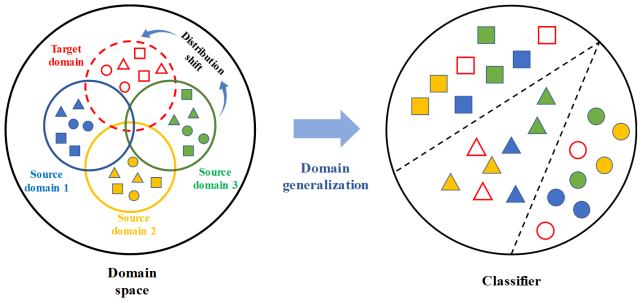


Figure 1: Definition of domain generalization.

In fault diagnosis, we supposed that there is a domain space including M source fault datasets(source domains) and one target fault dataset(target domain) denoted as $D=\{D_s \cup D_t\}$, and denoted the data set from the m th source domain with samples as $D_s^m=\left\{\left(x_s^i,y_s^i\right)\right\}\Big|_{i=1}^{n_s^m}$, where $m=1,2,3,\cdots,M$. Similarly, we denoted the target domain as $D_t=\left\{\left(x_t^i,y_t^i\right)\right\}\Big|_{i=1}^{n_t}$. It should be noted that the label space Y of each domain is the same, but the joint distributions are different($P_{XY}^i \neq P_{XY}^j, 1\leq i\neq j\leq M$)and the target domain data is inaccessible during the model training phase, which differs from traditional transfer learning and domain adaptation methods.

The ultimate goal of domain generalization is to train a predictive function h: $\mathrm{X}\rightarrow \mathrm{Y}$ on the known source domain data D_s with strong generalization ability, such that it has the minimum error on an unknown target domain data D_t :

$$\min_h \mathrm{E}_{(x,y)\in D_t}\left[\ell\left(h\left(x\right),y\right)\right] \tag{1}$$

ISSN 2616-7433 Vol. 6, Issue 7: 17-26, DOI: 10.25236/FSST.2024.060704

2.2. Gramian Angular Field (GAF)

In this study, Gramian Angular Field (GAF) was used as the method for generating images to encode one-dimensional vibration signals from bearings into two-dimensional images for model training. Reference has demonstrated through experimental verification on different datasets that encoding one-dimensional time series as images can significantly improve the accuracy of detection and classification tasks.

Given a bearing vibration signal segment $X = \{x_1, x_2, x_3, \cdots, x_n\}$ of ***n*** data points, we normalize and rescale it to [-1, 1] by the Eq(2):

$$\tilde{x}_{-1}^{i} = \frac{(x_i - max(X) + (x_i - min(X))}{max(X) - min(X)} \tag{2}$$

$$\text{or} \quad \tilde{x}_{0}^{i} = \frac{x_i - min(X)}{max(X) - min(X)} \tag{3}$$

Hence, the rescaled time series X can be represented in polar coordinates by encoding the values as the angular cosine and the time stamps as the radius with the following equation:

$$\begin{cases} \phi = \arccos\left(\tilde{x}_i\right), -1 \le \tilde{x}_i \le 1, \tilde{x}_i \in \tilde{X} \\ r = \dfrac{t_i}{N}, t_i \in \mathbb{N} \end{cases} \tag{4}$$

From eq(4), the angle is the inverse cosine of xi, t_i represents the time stamp, and ***N*** is a constant factor used to normalize the span of the polar coordinate system.

As shown in eq (5), the Gram matrix is a symmetric matrix that consists of the inner products of any k vectors in n-dimensional Euclidean space.

$$\Delta(\alpha_1, \alpha_2, \ldots, \alpha_k) = \begin{bmatrix} (\alpha_1, \alpha_1) & (\alpha_1, \alpha_2) & \ldots & (\alpha_1, \alpha_k) \\ (\alpha_2, \alpha_1) & (\alpha_2, \alpha_2) & \ldots & (\alpha_2, \alpha_k) \\ \ldots & \ldots & \ldots & \ldots \\ (\alpha_k, \alpha_1) & (\alpha_k, \alpha_2) & \ldots & (\alpha_k, \alpha_k) \end{bmatrix} \tag{5}$$

The Gram matrix measures the features and the correlations of each dimension. The diagonal elements of the multiscale matrix obtained after the inner product contain information about the different feature maps, while the off-diagonal elements contain information about the correlations between different feature maps.

After transforming the rescaled time series into the polar coordinate system, we can easily leverage the angular perspective by considering the trigonometric difference between each point to identify the temporal correlation within different time intervals, which allows us to capture the temporal dynamics of the time series in a more intuitive and interpretable way.

2.3. Stable Learning

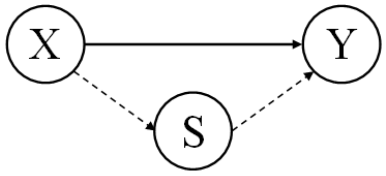


Figure 2:Causality with false associations: X represents related features(cause), S represents unrelated features, and Y represents category labels(effect).

As shown in Figure 2, stable learning is based on causal reasoning. When training a model, the real labels are affected by the relevant features X. However, if there are many irrelevant features S, the model

ISSN 2616-7433 Vol. 6, Issue 7: 17-26, DOI: 10.25236/FSST.2024.060704

will predict the labels based on both X and S. This will create spurious correlations that reduce the model's generalization ability.

Stable learning assumes that the distribution shift is caused by spurious correlations between irrelevant features and labels, which are in turn caused by subtle associations between relevant and irrelevant features.

Stable learning aims to address the OOD problem by removing the dependencies between features through learning weights for training samples. This helps deep models focus more on the true relationship between discriminative features and labels.

To eliminate the dependency between any pair of features in the sample representation space and improve the model's generalization ability, we introduce a hypothesis test statistic to measure two random variables.

Consider a measurable, positive definite kernel k_A on the domain of random variable A and the corresponding RKHS is denoted by $\mathcal{H}_{\mathcal{A}}$.If k_B and $\mathcal{H}_{\mathcal{B}}$ are similarly defined, the cross-covariance operator from $\sum AB$ to $\mathcal{H}_{\mathcal{B}}$ is $\mathcal{H}_{\mathcal{A}}$ as follows:

$$< h_A, \sum ABh_B >= \mathbb{E}_{AB}\left[h_A(A)h_B(B)\right] - \mathbb{E}_A[h_A(A)]\mathbb{E}_B[h_B(B)] \tag{6}$$

for all $h_A \in \mathcal{H}_{\mathcal{A}}$ and $h_B \in \mathcal{H}_{\mathcal{B}}$.

HSIC is a criterion for measuring the independence between features, which determines whether two features are independent based on the size of the cross-covariate factor of the sample. When the Hilbert-Schmidt norm $\sum AB$ is 0, it can indicate that the two features of AB are unrelated. However, calculating the size of using HSIC requires a lot of time and resources, so HSIC is not suitable for deep learning. The Frobenius norm corresponds to the Hilbert-Schmidt norm in Euclidean space, and the amount of calculation is less than the Hilbert-Schmidt norm, so the Frobenius norm can be used as the independence test statistic.

Let the partial cross-covariance matrix be:

$$\hat{\Sigma}_{AB} = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(\mathbf{u}(A_i) - \frac{1}{n}\sum_{j=1}^{n}\mathbf{u}(A_j)\right)^T \cdot \left(\mathbf{v}(B_i) - \frac{1}{n}\sum_{j=1}^{n}\mathbf{v}(B_j)\right)\right], \tag{7}$$

here

$$\begin{aligned}\mathbf{u}(A) &= \left(u_1(A), u_2(A), \ldots u_{n_A}(A)\right), u_j(A) \in \mathcal{H}_{\text{RFF}}, \forall j,\\ \mathbf{v}(B) &= \left(v_1(B), v_2(B), \ldots v_{n_B}(B)\right), v_j(B) \in \mathcal{H}_{\text{RFF}}, \forall j.\end{aligned} \tag{8}$$

We sample nA and nB functions separately from $\mathcal{H}_{\text{RFF}}$, as shown in eq(9), which represents a function space with the following form of random Fourier features:

$$\mathcal{H}_{\text{RFF}} = \left\{h : x \to \sqrt{2}\cos(\omega x + \phi) \middle| \omega \sim N(0,1), \phi \sim \text{Uniform}(0, 2\pi)\right\} \tag{9}$$

i.e. ω is sampled from the standard normal distribution, and φ is sampled from the uniform distribution.

Then, the independence test statistic I_{AB} is defined as the Frobenius norm of the partial covariance matrix, i.e.,

$$I_{AB} = \left\| \hat{\sum} A, B, w \right\|_F^2 \tag{10}$$

As I_{AB} decreases to zero, the two variables A and B tend to become independent. Therefore, I_{AB} can effectively measure the independence between random variables.

ISSN 2616-7433 Vol. 6, Issue 7: 17-26, DOI: 10.25236/FSST.2024.060704

3. Proposed method

In this section, Ds represents labeled source domain data, and Dt represents unlabeled target domain data. It is worth noting that the joint distribution $P(XY)$ of each domain is different, but the Y is the same. The purpose of the proposed method is to obtain predictive labels for the target domain by studying domain invariant features across domains. The main assumption for our domain generalization task is DGD. In this section, the main introduction is the framework of the model.

The proposed deep stable learning-based model for bearing fault diagnosis is illustrated in Figure 3. In summary, the architecture of the voice recognition model mainly consists of three parts, which are data transformation, sample weighting, and voice recognition module respectively.

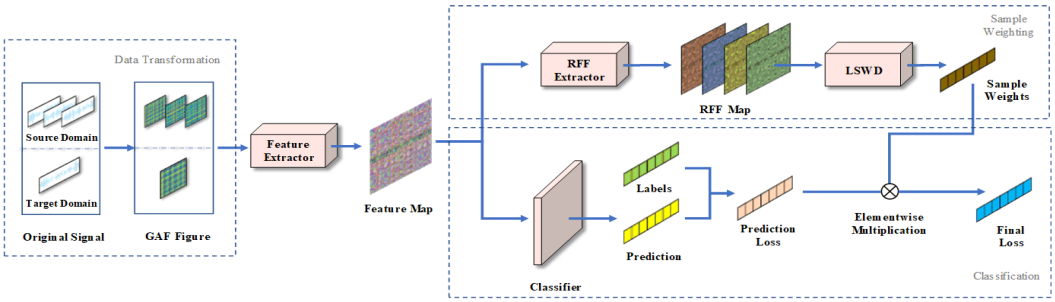


Figure 3:Proposed framework.

3.1. Data transformation

In the data transformation stage, this article used GADF to convert one-dimensional vibration signals into two-dimensional images. Specifically, the following steps are taken: first, the one-dimensional vibration signal is divided into signal fragments of the same length.

Where fs denotes the sampling frequency, rs denotes the bearing rotation speed and n denoted the number of data points within a complete rotation cycle.

Secondly, the dimension of the signal fragment is transformed from 1x to 3x*x images through the GADF. Specific details of the GADF have been provided in Chapter 2 and will not be repeated here.

3.2. Sample weighting

In the sample weighting stage, stable learning eliminates dependencies between features in the representation space through sample weighting, and measures general independence through RFF.

We use $w \in \mathbb{R}_{+}^{n}$ to represent sample weights and $\sum_{i=1}^{n} w_i = n$. After weighing, the partial cross-covariance matrix of random variables $\boldsymbol{A}$ and $\boldsymbol{B}$ in equation 3 can be calculated as follows:

$$\hat{\Sigma}_{AB;\mathbf{w}} = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(w_i\mathbf{u}(A_i) - \frac{1}{n}\sum_{j=1}^{n} w_j\mathbf{u}(A_j)\right)^{T} \cdot \left(w_i\mathbf{v}(B_i) - \frac{1}{n}\sum_{i=1}^{n} w_j\mathbf{v}(B_j)\right)\right] \tag{11}$$

Here $\boldsymbol{u}$ and $\boldsymbol{v}$ are the RFF mapping functions explained in eq (8).

3.3. Voice recognition

After data preprocessing and dataset partitioning, input the source domain data into the model for model training. In the training phase, Resnet 34 is used as the feature extractor, and the network parameters as shown in the table. The softmax regression function is used as the classifier, and SGD is used as the optimization function. It should be noted that during the training phase, only the source domain data can be used for model training and validation, while the target domain data cannot be accessed.

ISSN 2616-7433 Vol. 6, Issue 7: 17-26, DOI: 10.25236/FSST.2024.060704

The extracted features have two paths: one is for learning the feature weights of the samples, and the other is for predicting with the classifier.

A and ***B*** represent a pair of random eigenvectors of a sample, the cross-covariance matrix of $\hat{\Sigma}_{AB;\boldsymbol{w}}$ the sample can be obtained through. The Frobenius normal form I_{AB} of the cross-covariance matrix $\hat{\Sigma}_{AB;\boldsymbol{w}}$ is taken as the metric of $\boldsymbol{w}$. The loss function of the weight learning is shown in eq(12):

$$J\left(w,\lambda\right)=\frac{1}{\lambda}\left\|\hat{\sum}AB,w\right\|_F^2+\left\|w\right\|_2^2 \tag{12}$$

Optimize $\boldsymbol{w}$ through the SGD optimizer, as shown in eq(16).

$$\begin{aligned} V_{t+1} &= m*V_t+l_w*\nabla_w J\left(w,\lambda\right) \\ w_{t+1} &= w-V_{t+1} \end{aligned} \tag{13}$$

where $\boldsymbol{m}$ is the momentum, l_w is the learning rate of the weight learning. Initially, $w_0=(1,1,1,\cdots,1)^T$. $\boldsymbol{w}$ is trained with 30 epochs for each sample. l_w and $\boldsymbol{m}$ are set up to 0.6 and 0.9 in weight learning stage respectively.

The softmax regression function is shown in eq(14)

$$f\left(x_i\right)=softmax\left(x_i\right)=\frac{e^{x_i}}{\sum_{i=0}^{C}e^{x_i}} \tag{14}$$

The loss function of the classifier is the cross entropy loss function, as shown in eq(15):

$$J\left(\theta,b\right)=-\frac{1}{M}\sum_{i=1}^{M}y_i\log f\left(z(\theta,b)\right) \tag{15}$$

Where $z(\theta,b)$ represents the output of the feature exactor, M denotes the number of classes. After calculating the sample prediction loss of the classifier, it is multiplied by the corresponding element of the feature weight to obtain the final prediction loss.

4. Experimental study

In this section, the method is evaluated on experimental data obtained from public speech datasets to demonstrate the feasibility and efficiency of the proposed method.

4.1. TALSER voice emotion datasets

Voice emotion data set for good future teacher class audio, a total of 4541 audio, a total time of 12.5 hours. The recordings were made in a quiet indoor environment with only one speaker per audio message. The tagging includes Pleasure and Arousal. Each audio clip has A P value and A value in the range of arousal [-3, 3]. The higher the value, the higher the pleasure or arousal. Data size :12.5 hours; Sampling rate :16KHz; Sampling bit sound:16bit; Recording equipment:ordinary microphone; Speaker:42, Male: 18, female: 24; Data format: Voice:wav mono; Annotation result:txt; Audio length :10s; Accuracy :96%.

Table 1: Domain division of TALSER datasets.

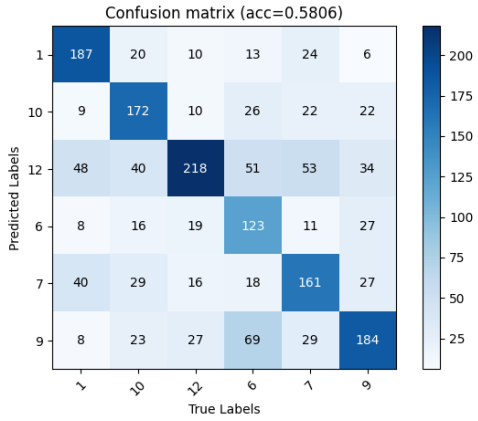
Person	Source Domain	Target Domain
1	1, 2, 3, 4, 5	6
2	2, 3, 4, 5, 6	1
3	1, 3, 4, 5, 6	2
4	1, 2, 4, 5, 6	3
5	1, 2, 3, 5, 6	4
6	1, 2, 3, 4, 6	5

We treat different persons in the TALSER dataset as different domains. When partitioning the dataset, we use one person as the target domain and the other three persons as the source domains. For example,

ISSN 2616-7433 Vol. 6, Issue 7: 17-26, DOI: 10.25236/FSST.2024.060704

1, 2, 3, 4, 5, 6 means six people. Five of themare source domains and one is the target domain. Table 1 shows the complete dataset partitioning method.

Each rotation cycle contains 400 data points, so we collect one sample every 400 data points. In the balanced sample experiment, we collect 300 samples of each speech type from each source domain, ensuring equal sample size across domains. In the imbalanced sample experiment, we reduce the number of samples from each domain by 20 from top to bottom. Experimental results and analysis are shown below.



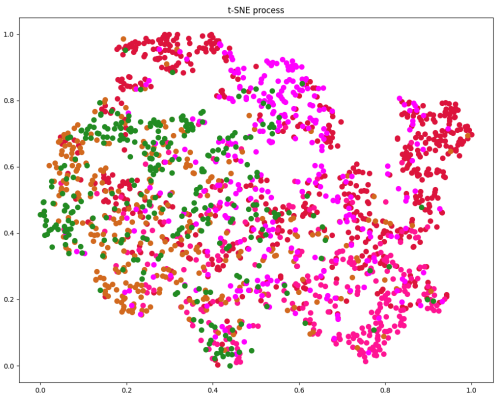


Figure 4:confusion matrix of the proposed method. *Figure 5:Visualization maps of the proposed method.*

As shown in Figure 4, confusion matrix of the proposed method is shown. We use t-SNE technology to visualize the output features of the last fully connected layer of the proposed method in four generalization experiments. The results are shown in Figure 5. The figure shows that most of the training and testing data are clustered together, and different clusters are well separated. This indicates that the proposed method has a strong feature learning ability and an excellent domain invariant feature extraction ability, which corresponds to a average accuracy of 58.06% . Therefore, the proposed diagnostic method has a strong domain generalization ability.

To verify the feasibility and superiority of the proposed method, comparison experiments are conducted.

The proper hyperparameters of methods under comparison are determined from previous studies and experimental requirements to achieve satisfactory performance.

As shown in Figure 6, confusion matrix of the compared method is shown. We use t-SNE technology to visualize the output features of the last fully connected layer of the proposed method in four generalization experiments. The results are shown in Figure 7. The figure shows that most of the training and testing data are clustered together, and different clusters are well separated. This indicates that the proposed method has a strong feature learning ability and an excellent domain invariant feature extraction ability, which corresponds to a average accuracy of 46.5% . Therefore, the proposed diagnostic method has a poor domain generalization ability.

ISSN 2616-7433 Vol. 6, Issue 7: 17-26, DOI: 10.25236/FSST.2024.060704

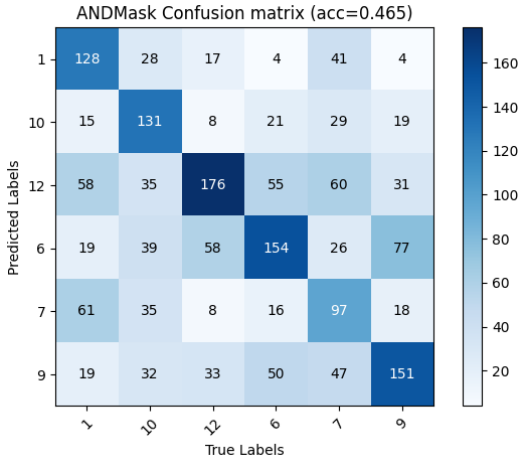


Figure 6: confusion matrix of the compared method.

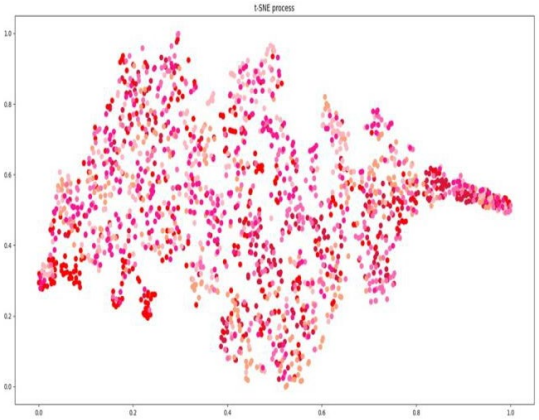


Figure 7: Visualization maps of the compared method.

4.2. TALASR Speech Recognition dataset

Speech recognition data set for future online course teachers teaching audio, covering Chinese, mathematics two subjects. A total of 80 speakers, each audio only one speaker. The annotated data contains the subject and speaker numbers. Data size :100 hours; Sampling rate :16KHz; Sampling bit sound :16bit; Recording equipment: ordinary microphone; Speaker:80 persons; Recording time: April to May, 2018; Data format: Voice: wav mono; Annotation result: txt; Audio length:1 ~ 60s;

We treat different persons in the TALASR dataset as different domains. When partitioning the dataset, we use one person as the target domain and the other three persons as the source domains. For example, 1, 2, 3, 4, 5, 6 means six people. Five of themare source domains and one is the target domain. Table 2 shows the complete dataset partitioning method.

Table 2:Domain division of TALASR datasets.

Person	Source Domain	Target Domain
1	1, 2, 3, 4, 5	6
2	2, 3, 4, 5, 6	1
3	1, 3, 4, 5, 6	2
4	1, 2, 4, 5, 6	3
5	1, 2, 3, 5, 6	4
6	1, 2, 3, 4, 6	5

According to eq(12), each rotation cycle contains 400 data points, so we collect one sample every 400 data points. In the balanced sample experiment, we collect 300 samples of each speech type from each source domain, ensuring equal sample size across domains. In the imbalanced sample experiment, we reduce the number of samples from each domain by 20 from top to bottom. Experimental results and analysis are shown below.

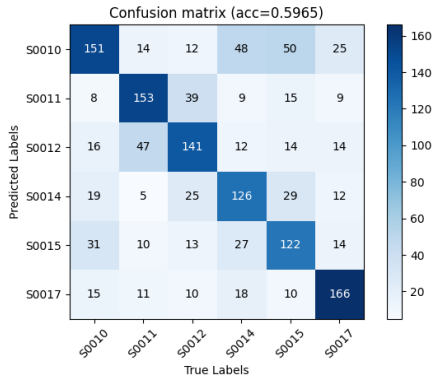


Figure 8: Confusion matrix of the proposed method.

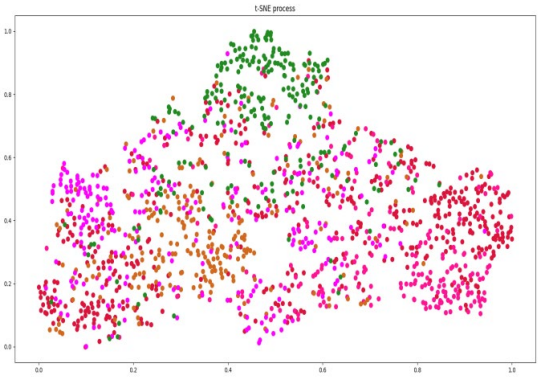


Figure 9:Visualization maps of the proposed method.

ISSN 2616-7433 Vol. 6, Issue 7: 17-26, DOI: 10.25236/FSST.2024.060704

As shown in Figure 8, confusion matrix of the proposed method is shown. We use t-SNE technology to visualize the output features of the last fully connected layer of the proposed method in four generalization experiments. The results are shown in Figure 9. The figure shows that most of the training and testing data are clustered together, and different clusters are well separated. This indicates that the proposed method has a strong feature learning ability and an excellent domain invariant feature extraction ability, which corresponds to a average accuracy of 59.65%.

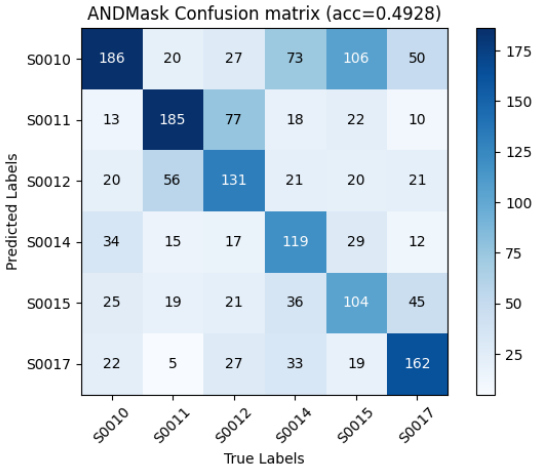


Figure 10:Confusion matrix of the compared method.

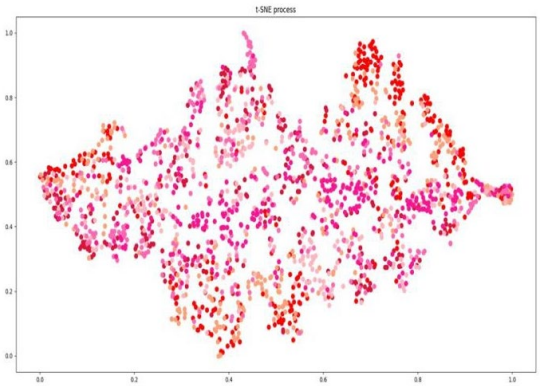


Figure 11:Visualization maps of the compared method.

To verify the feasibility and superiority of the proposed method, comparison experiments are conducted.

The proper hyperparameters of methods under comparison are determined from previous studies and experimental requirements to achieve satisfactory performance.

As shown in Figure 10, confusion matrix of the compared method is shown. We use t-SNE technology to visualize the output features of the last fully connected layer of the proposed method in four generalization experiments. The results are shown in Figure 11. The figure shows that most of the training and testing data are clustered together, and different clusters are well separated. This indicates that the proposed method has a strong feature learning ability and an excellent domain invariant feature extraction ability, which corresponds to a average accuracy of 49.28% . Therefore, the proposed diagnostic method has a poor domain generalization ability.

5. Conclusion

In this paper we have introduced a intelligent voice recognition method based on stable learning. We carefully designed the stable learning networks so that they can classify speech accurately. The classification experiment shows that proposed method best in the difference in speech. Meanwhile, our proposed method significantly outperforms the conventional speech recognition methods for mixed speech. This approach improved the accuracy of a speech classification task compared to traditional method. In future work, we will improve the performance of proposed approach in more complex condition. Further improvements require more stable functionality to achieve the full potential of training classifications.

Acknowledgments

The financial sponsorship from the project of Natural Science Foundation of Guangxi (Grant no. 2021GXNSFBA075050), It's also sponsored by Guangxi Key Laboratory of Manufacturing System & Advanced Manufacturing Technology, School of Mechanical and Electrical Engineering, Guilin University of Electronic Technology (Grant no. 20-065-40-004Z).

ISSN 2616-7433 Vol. 6, Issue 7: 17-26, DOI: 10.25236/FSST.2024.060704

References

[1] ITU-T G.7299 "A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70". 1996, International Telecommunication Union.
[2] ETSI GSM 06.32 "Full rate speech; VAD for full rate speech traffic channel", 1998, European Telecommunication Standards Institute.
[3] J.H.L. Hansen, S. Bou-Ghazale, "Robust speech recognition training via duration and spectral-based stress token generation", IEEE Rans. Speech Audio Proc., (3):415-421, 1995.
[4] J.H.L. Hansen, S. Bou-Ghazale, "Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database", EUROSPEECH-97, pp. 1743-1746.
[5] J.H.L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect", IEEE Rans. Speech Audio Proc., vol. 2, pp. 598-614, Oct. 1994.
[6] B.A. Hanson, T. Applebauni, "Robust speaker-independent word recognition using instantaneous, dynamic and acceleration features: Experiments with Lombard and noisy speech", ICASSP-90, pp. 857-860.
[7] J.C. Junqua, "The Lombard reflex and its role on hunian listeners and autoniatic speech recognizers", J. Acous. Soc. Am., vol. 93, pp. 510-524, Jan. 1993.
[8] R. Ruiz et al., "Time- and spectruni-related variabilities in stressed speech under laboratory and real conditions", Speech Comm., vol. 20, pp. 111-129, 1996.
[9] Rosdi F, Mustafa M B, Salim S, et al. Automatic Speech Intelligibility Detection for Speakers with Speech Impairments: The Identification of Significant Speech Features[J]. Sains Malaysiana, 2019, 48(12):2737-2747.
[10] Revathi, A,Nagakrishnan, R, Sasikaladevi,N. Comparative analysis of Dysarthric speech recognition: multiple features and robust templates[J]. Multimedia Tools and Applications, 2022, 22(81),31245-31259.
[11] Mitra V , Wang W , Franco H , et al. Evaluating robust features on Deep Neural Networks for speech recognition in noisy and channel mismatched conditions. 2014.
[12] Mitra V , Wen W , Franco H . Deep convolutional nets and robust features for reverberation-robust speech recognition[C] 2014 IEEE Spoken Language Technology Workshop (SLT).
[13] Jiao Y, Tu M, Berisha V, Liss J (2018) Simulating Dysarthric Speech for Training Data Augmentation in Clinical Speech Applications. 2018 IEEE international conference on acoustics, speech, and signal processing (ICASSP), Calgary, pp 6009–6013.
[14] Takashima Y, Nakashima T, Takiguchi T, Ariki Y (2015) Feature extraction using pre-trained convolutive bottleneck nets for dysarthric speech recognition. 2015 23rd European Signal Processing Conference(EUSIPCO), Nice, pp 1411–1415.
[15] España-Bonet C, Fonollosa JA (2016) Automatic speech recognition with deep neuralnetworks for impaired speech. In: International Conference on Advances in Speech and Language Technologies forIberian Languages. Springer, Cham, pp 97–107.
[16] Sloane S, Dahmani H, Amami R et al (2012) Using speech rhythm knowledge to improve dysarthric speech recognition. Int J Speech Technol 15:57–64.