

Research on Data Augmentation and Deep Learning Algorithms for Speech Emotion Computing

Kaiyu Guo^{1,2}, Yungang Wei^{3,*}

¹ Xizang Agricultural and Animal Husbandry University, Linzhi, China

² Xizang Guoang Information Technology Co., Ltd., Lasa, China

³ Xizang University, Lasa, China

* Corresponding author: Yungang Wei

Abstract: Speech is the most natural human interaction medium, which contains both linguistic semantic information and numerous non-semantic linguistic characteristics such as emotion, attitude and mood, making it a core research object in the fields of computational linguistics and human-computer interaction. Speech emotion computing aims to explore emotional features embedded in speech signals through computer algorithms to realize automatic recognition and classification of human emotions. It is an interdisciplinary research area integrating linguistics, signal processing and artificial intelligence. At present, mainstream research on speech emotion recognition generally faces problems including insufficient annotated corpora, single sample dimensions, poor model generalization ability and low recognition accuracy under low-resource conditions. This paper systematically conducts research from two major perspectives: data augmentation and deep learning algorithms. Firstly, it sorts out the computational linguistic theoretical basis of speech emotion computing and analyzes the linguistic implications of prosodic, spectral and temporal emotional features in speech signals. Secondly, it investigates traditional data augmentation and intelligent deep learning-based augmentation methods separately, and optimizes augmentation strategies according to the linguistic characteristics of oral speech corpora. In addition, a deep learning recognition model suitable for speech emotional features is constructed to improve the feature extraction capability and classification performance of the model. Finally, extensive comparative experiments are carried out to verify the joint optimization effect of the proposed data augmentation algorithms and deep learning models. Experimental results demonstrate that the improved multi-dimensional data augmentation method can effectively enrich speech corpora and address the shortage of low-resource annotated data. Combined with the optimized deep learning model, the proposed method significantly improves the accuracy and stability of speech emotion recognition, providing technical support for natural language interaction scenarios such as intelligent customer service, human-computer dialogue and emotion monitoring.

Keywords: Speech Emotion Computing, Data Augmentation, Deep Learning, Computational Linguistics, Feature Extraction, Emotion Recognition.

1. Introduction

From the perspective of computational linguistics, the scarcity of high-quality annotated speech emotion corpora constitutes a critical bottleneck restricting performance improvements of recognition models. Manual annotation of speech corpora is costly and time-consuming, while existing public corpora suffer from limited sample sizes, monotonous scenarios, and insufficient data covering dialects and niche application scenarios. Consequently, deep learning models are prone to overfitting during training, leading to weak generalization performance in complex real-world environments. Furthermore, conventional speech emotion recognition systems mostly rely on manually designed shallow features, which fail to effectively capture implicit deep linguistic emotional features within speech signals, resulting in low feature utilization and unsatisfactory recognition accuracy^[1]. In view of the above challenges, this paper carries out research on data augmentation and deep learning algorithms for speech emotion computing. Data augmentation technology is adopted to compensate for the lack of corpus resources, while deep learning algorithms enable automatic extraction of deep emotional features. This research bears important theoretical value and engineering practical significance for refining the research system of speech interaction in computational linguistics and promoting

the industrial application of affective intelligent human-computer interaction technologies^[2].

2. Relevant Theoretical and Technical Foundations

2.1. Computational Linguistic Basis of Speech Emotion Computing

From a computational linguistic perspective, human oral language conveys information across four layers: semantic, syntactic, pragmatic and emotional layers. The emotional layer refers to implicit emotional expressions in linguistic communication, which are intuitively reflected through acoustic properties of speech including prosody, timbre and rhythm. The core objective of speech emotion computing is to establish mapping relationships among "acoustic features – linguistic emotional features – human emotion categories", and achieve precise judgment of speakers' emotions via quantitative computerized analysis of diverse speech signal characteristics^[3].

2.2. Core Feature System of Speech Emotion

Speech signals are typical time-series data containing three categories of key emotional features: temporal features, frequency-domain features and prosodic features, which collectively form the fundamental feature system for speech

emotion computing and recognition. The specific connotations of each feature and their correlations with emotions are elaborated as follows: Temporal features reflect intuitive time-varying properties of speech signals, including short-time energy, zero-crossing rate, speech speed and utterance duration. Frequency-domain features represent core deep characteristics of speech signals, mainly comprising Mel-Frequency Cepstral Coefficients (MFCC), spectral entropy and spectral contrast^[4].

2.3. Core Technologies of Speech Data Augmentation

Combined with the characteristics of speech emotional features, this paper integrates traditional data augmentation and intelligent deep learning-based augmentation technologies, which fall into two primary categories: Traditional speech data augmentation technologies mainly include time-domain transformation, frequency-domain transformation and noise superimposition. Intelligent deep learning-based augmentation relies on Generative Adversarial Networks (GAN). Through adversarial training between generators and discriminators, GANs autonomously learn the feature distribution rules of real speech emotional samples to generate high-quality novel speech emotional samples^[5].

2.4. Theoretical Foundations of Core Deep Learning Models

Speech emotional signals are time-series data characterized by contextual relevance and dynamic variations. Conventional static deep learning models cannot effectively capture temporal dependencies. This paper focuses on improved Recurrent Neural Networks (namely Long Short-Term Memory, LSTM) integrated with Convolutional Neural Networks (CNN) for research.

3. Speech Data Preprocessing and Multi-Dimensional Data Augmentation

3.1. Dataset Construction and Preprocessing

Experimental data in this paper are derived from the CASIA Chinese Emotional Speech Database, an authoritative standard dataset for Chinese speech emotion computing developed by the Institute of Automation, Chinese Academy of Sciences, which perfectly matches the research context of oral emotion analysis in computational linguistics. The database contains pure Chinese oral speech recorded by four professional announcers (2 males and 2 females), covering six basic emotional labels: neutral, happy, angry, sad, surprise and fear. All utterances adopt standardized daily general sentences to avoid semantic interference with emotion recognition and ensure single experimental variables. The original dataset contains 4,800 speech samples with a sampling rate of 16kHz and single-channel recording. After removing invalid samples with distortion, excessive silence or ambiguous emotional features, 3,200 pure base samples are screened out with roughly balanced distribution across all emotion categories. The sample distribution of the original dataset is shown in Table 1.

To conform to real human-computer interaction scenarios, 1,000 manually collected and annotated daily chat speech samples are supplemented to enrich oral scene data. The constructed base dataset combines standard broadcast speech

and authentic daily oral speech, which enhances the generalization capability of the model. Raw speech data suffer from noise interference, inconsistent duration and redundant invalid segments, thus unified preprocessing operations are required.

Table 1. Sample Distribution of Original Dataset

Emotion Category	English Label	Raw Sample Count	Sample Proportion	Speech Duration Range
Neutral	Neutral	535	16.7%	1.2s-3.5s
Happy	Happy	542	16.9%	1.0s-3.2s
Angry	Angry	538	16.8%	0.9s-3.0s
Sad	Sad	527	16.5%	1.5s-3.8s
Surprise	Surprise	531	16.6%	1.0s-2.9s
Fear	Fear	527	16.5%	1.3s-3.6s

3.2. Single-Dimension Data Augmentation Experiments and Defect Analysis

To verify the limitations of individual data augmentation methods, single-dimension augmentation experiments are conducted using four independent approaches: speed perturbation, pitch shifting, noise addition and GAN generation. Each method expands the total sample size to 6,400, and the effects are evaluated via sample feature analysis and preliminary experiments.

Experimental observations reveal prominent drawbacks of standalone traditional augmentation methods: Speed perturbation only modifies speech speed features, generating new samples with high similarity to original data and single feature dimensions, which fails to substantially enrich data diversity and still carries risks of model overfitting. Excessive pitch shifting during pitch augmentation causes emotional label drift; originally neutral speech acquires sharp timbre characteristics, leading to label confusion. Noise addition only improves the anti-interference capability of models without generating samples with novel emotional features, exerting limited effects on supplementing low-resource emotion categories.

Table 2. Detailed Sample Increment of Integrated Augmentation

Augmentation Method	Initial Sample Count	New Samples Generated	Total Samples After Augmentation	Valid Sample Retention Rate
Time-Domain Speed & Pitch Perturbation	3200	3200	6400	96.2%
Noise Superimposition Augmentation	6400	3200	9600	95.8%
GAN Intelligent Generation Augmentation	9600	3200	12800	93.5%

Given the multiple drawbacks of single-dimension augmentation, this paper proposes a multi-dimensional integrated data augmentation scheme combining traditional and intelligent augmentation techniques, balancing sample diversity, feature authenticity and label accuracy. Detailed statistics of sample increments in each augmentation stage are

recorded to ensure traceability and reproducibility of experimental data. The overall sample growth details are listed in Table 2.

Specific scheme details are as follows: Aiming at the defects of single-dimension traditional augmentation, this paper puts forward a multi-dimensional integrated data augmentation framework combining traditional and intelligent augmentation, prioritizing sample diversity, feature authenticity and label accuracy. The framework consists of three core components:

- 1) Hierarchical traditional augmentation: differentiated augmentation strategies customized according to acoustic feature characteristics of speech under distinct emotion categories;
- 2) Supplementary intelligent augmentation via GAN;
- 3) Data filtering and category balance optimization.

4. Construction of Deep Learning Speech Emotion Recognition Model

4.1. Overall Architecture Design

This paper constructs a CNN-LSTM hybrid deep learning model to accommodate both local features and temporal dependencies of speech emotional signals. The model comprises five layers: input layer, feature extraction layer, temporal memory layer, fully connected layer and output layer. It can effectively capture shallow acoustic features and deep linguistic emotional features embedded in speech signals, with an overall structure compatible with speech emotion computing tasks.

4.2. Structural Design and Performance Optimization of Each Module

4.2.1. Input Layer

The input layer receives preprocessed and augmented speech samples, extracts multi-dimensional emotional features including MFCC, prosodic features and short-time energy, and generates standardized feature matrices with unified dimensions as model inputs to guarantee data normalization while retaining complete feature information.

4.2.2. Feature Extraction Layer

The feature extraction module is built with multi-layer Convolutional Neural Networks, including 3 convolutional layers and 2 pooling layers. Convolutional layers adopt convolution kernels of varying sizes to extract local deep emotional features of speech spectra and prosody, capturing subtle discrepancies in emotional characteristics. Max pooling is applied in pooling layers to reduce feature dimensionality, filter key valid features and eliminate redundant interfering features, lowering computational overhead and improving feature extraction efficiency.

4.2.3. Temporal Memory Layer

Double-layer LSTM networks form the temporal feature memory module, which takes local features extracted by CNN as input to further explore contextual correlations within time-series speech signals and memorize dynamic emotional variation patterns throughout utterances. This compensates for the incapability of standalone CNNs to model temporal dependencies, conforming to linguistic rules of continuous emotional changes in oral speech.

4.2.4. Fully Connected Layer and Output Layer

The temporal feature outputs of LSTM are flattened and processed through two fully connected layers for feature

fusion and dimension transformation, integrating local spatial features and global temporal features. The output layer adopts the Softmax activation function corresponding to six basic emotion categories to realize multi-class classification of speech emotions.

5. Experimental Design and Result Analysis

5.1. Experimental Environment and Evaluation Metrics

All experiments are implemented based on Python with the PyTorch deep learning framework, and CUDA acceleration is deployed to boost training efficiency. Hardware configurations include an Intel i7 CPU, 16GB RAM and an RTX3060 GPU. To comprehensively evaluate the performance of the proposed algorithm and model, accuracy, precision, recall and F1-score are adopted as core evaluation metrics. Overfitting risks and convergence speed during model training are also analyzed to comprehensively assess the effectiveness of the optimized data augmentation algorithm and hybrid model.

5.2. Comparative Experiment Setup

Three groups of comparative experiments are designed to highlight the superiority of the proposed scheme: Experiment 1: Single LSTM model trained on raw original dataset for emotion recognition; Experiment 2: CNN-LSTM hybrid model trained on dataset expanded via single-dimension traditional augmentation; Experiment 3: CNN-LSTM hybrid model trained on dataset processed by the proposed multi-dimensional integrated augmentation scheme. Comparisons among results of the three groups separately quantify performance gains brought by optimized data augmentation algorithms and the advantages of the hybrid model architecture.

5.3. Experimental Results and Data Analysis

Quantitative core metrics of the three experimental groups are summarized in Table 3, including overall accuracy, average precision, average recall and average F1-score, which intuitively demonstrate performance gaps among different schemes.

Table 3. Core Performance Metrics

Experiment Group	Experimental Scheme	Accuracy (%)	Precision (%)	Recall (%)	F1-Score
Group 1	Raw Dataset + Standalone LSTM	78.32	77.56	76.89	0.75
Group 2	Single-Dimension Traditional Augmentation + CNN-LSTM	85.67	84.92	84.35	0.83
Group 3	Multi-Dimensional Integrated Augmentation + CNN-LSTM (Proposed)	92.15	91.86	91.53	0.91

For fine-grained analysis, individual recognition metrics of each emotion category under the proposed optimal scheme

are calculated to reflect recognition performance for subtle emotional distinctions, as shown in Table 4.

Table 4. Category-Wise Emotion Recognition Metrics

Emotion Category	Accuracy (%)	Precision (%)	Recall (%)	F1-Score
Neutral	95.32	94.87	95.11	0.95
Happy	93.68	93.25	93.47	0.93
Angry	94.21	93.96	94.02	0.94
Sad	89.75	89.32	89.56	0.89
Surprise	91.36	91.05	91.22	0.91
Fear	88.92	88.64	88.79	0.88

Experimental data indicate that the scheme combining standalone LSTM with raw data achieves poor overall performance, with drastically reduced recognition accuracy for low-resource emotions such as sadness and fear. Single-dimension data augmentation only yields marginal performance improvements and exhibits limited capability to distinguish highly similar subtle emotions. The proposed framework integrating multi-dimensional integrated augmentation and CNN-LSTM hybrid model achieves an overall recognition accuracy exceeding 92%, with comprehensive optimization of metrics across all emotion categories. Notably, it substantially enhances recognition performance for low-resource and acoustically similar emotions, delivering outstanding generalization and robustness.

Experimental results verify that data augmentation algorithms fundamentally resolve the shortage of annotated speech emotion corpora and provide sufficient high-quality training data for deep learning models. The CNN-LSTM hybrid model effectively extracts local deep emotional features and global temporal features of speech, consistent with the emotional expression rules of Chinese oral speech. The combination of the two technologies generates synergistic optimization effects, achieving significant improvements in recognition accuracy and model stability compared with conventional solutions, and exhibiting stronger adaptability for speech emotion computing tasks under complex real-world environments.

5.4. Analysis of Experimental Limitations

Analysis of experimental results reveals several aspects requiring further optimization in this research:

1) Distinction accuracy for subtly differentiated emotions (sadness vs. fear, neutral vs. mild happiness) needs improvement, as these emotions share highly similar acoustic characteristics and pose challenges for feature discrimination;

2) Edge samples generated during augmentation possess ambiguous feature boundaries, interfering with model classification performance;

3) Recognition accuracy slightly declines under extremely noisy complex environments, leaving room for anti-interference performance optimization.

6. Conclusion

From the perspective of computational linguistics and in response to practical application demands of speech emotion computing, this paper systematically investigates data augmentation and deep learning algorithms targeting prevailing industrial challenges including insufficient annotated corpora, monotonous sample dimensions, low recognition accuracy and poor model generalization. Firstly, it sorts out the linguistic theoretical foundation of speech emotion computing and analyzes core dimensions and corresponding mapping rules of Chinese speech emotional features. Secondly, it compares the advantages and disadvantages of various speech data augmentation technologies and designs a multi-dimensional integrated augmentation framework combining traditional and intelligent augmentation methods, which effectively expands corpus volume and enriches sample diversity to address the scarcity of low-resource speech emotion data. Furthermore, a CNN-LSTM hybrid deep learning model is constructed, combining the strengths of Convolutional Neural Networks in capturing local features and Long Short-Term Memory networks in modeling temporal dependencies to accurately extract deep linguistic emotional features of speech. Extensive comparative experiments validate the effectiveness of the proposed scheme. The optimized framework delivers remarkably higher recognition accuracy, improved model stability and stronger generalization capability, providing technical references for research fields including natural language interaction and affective intelligent computing.

References

- [1] Kragel, P. A., Reddan, M. C., Labar, K. S., et al. (2019). Emotion schemas are embedded in the human visual system. *Science Advances*, 5(7), 1–15.
- [2] Ankit, K., Yu, Z. W., Hillary, S., et al. (2021). A preliminary exploration of virtual reality-based visual and touch sensory processing assessment for adolescents with autism spectrum disorder. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29(1), 619–628.
- [3] Liu, T. J., Li, F. L., & Jiang, Y. (2017). Cortical dynamic causality network for auditory-motor tasks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(8), 1092–1099.
- [4] Davis, K., Biddulph, R., & Balashek, S. (1952). Automatic recognition of spoken digits. *Journal of the Acoustical Society of America*, 24(6), 637–644.
- [5] Jelinek, F., Bahl, L., & Mercer, R. (1975). Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 21(3), 250–256.