# Research on Intelligent Language Translation System Based on Deep Learning Algorithm

**Chunliu Shi** ( scliu158@163.com )
Zhengzhou University of Aeronautics

**Research Article**

**Keywords:** Deep learning, intelligent speech, translation system, intelligent model

**Posted Date:** September 21st, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-899180/v1

# Research on Intelligent Language Translation System Based on Deep Learning Algorithm

Chunliu Shi
School of Foreign languages, Zhengzhou University of Aeronautics, Zhengzhou, Henan ,450046, China
Email: scliu158@163.com

**Abstract.** In order to improve the effect of intelligent language translation, this paper analyzes the problems of the MSE cost function used by most of the current DNN-based speech enhancement algorithms, and proposes a deep learning speech enhancement algorithm based on perception-related cost functions. Moreover, this paper embeds the suppression gain parameter estimation into the architecture of the traditional speech enhancement algorithm, and converts the relationship between the noisy speech spectrum and the enhanced speech spectrum into a simple multiplication relationship based on suppression gain combined with deep learning algorithms to construct an intelligent language translation system. Moreover, this paper evaluates the translation effect of the system, analyzes the actual results, and uses simulation tests to verify the performance of the intelligent language translation model constructed in this paper. From the experimental results, it can be seen that the intelligent language translation system based on deep learning algorithms has good results.

Keywords: Deep learning; intelligent speech; translation system; intelligent model

## 1. Introduction

With the maturity of speech recognition and processing technology, the speech data collected by the speech library is becoming more and more abundant, the speech recognition rate is getting higher and higher, and the semantic error correction and understanding ability is getting stronger. Realizing intelligent interaction through voice, and controlling smart TV through voice commands are now not just staying in the laboratory or in science fiction films, but truly becoming a reality [1].

People usually use computer translation to collectively refer to machine translation and computer-assisted translation. Machine translation mainly refers to the translation carried out by the computer, while the computer-assisted translation emphasizes the translation carried out by the human being with the assistance of the computer [2]. However, so far, there has not been a very complete machine translation system that can perfectly replace manual translation work. The reason is that language has cultural attributes in addition to its own system attributes, and the language system also has the characteristics of openness [3].

With the further development of artificial intelligence, machine translation can better meet most general translation needs. However, in the face of complex communication translation tasks that are professional, diverse, detailed and contain human emotions, machine translation is still difficult to replace human translation. In the current development of artificial intelligence, there is a trend of human-computer relations that are mirror images, embedding each other, and information each other. Therefore, currently machine translation and human translators work together. However, in this collaborative translation relationship, the previous simple "machine-assisted translation" model has been increasingly replaced by artificial intelligence interactive translation, which is an advancement in translation technology capabilities. Many tasks of machine translation and computer-assisted translation will also be completed by increasingly sophisticated artificial intelligence interactive translation.

In order to improve translation efficiency and ensure the consistency of translation content, artificial intelligence in machine translation application systems usually set system translation components, such as translation memory and term management components. Translation memory mainly refers to the equivalent corpus corresponding to the original text and the translated content of the corpus constructed by artificial intelligence in the machine translation system. When the machine translation system starts the translation work, the artificial intelligence will automatically store and compare the text materials that need to be translated in the translation corpus. During the entire translation process, the artificial intelligence scans the program code and finds that they are similar or similar. When translating the content, the artificial intelligence system will automatically match it with the content of the translation corpus. After confirming through the context and language use system, the final translation result can be quickly presented on the user interface. With the continuous improvement of artificial intelligence technology, the machine translation system that can support fuzzy matching is also constantly upgraded.The artificial intelligence system can automatically set the minimum matching value between the original text and the translation of the corpus (such as 60% or 80%). Then, the corpus in the translation memory is searched through the fuzzy matching program. Even for those sentence patterns that cannot be completely matched, the artificial intelligence can achieve similar content confirmation through machine translation, and then confirm it through the language use system or a human translator. Realize the interaction and mutual assistance between artificial intelligence and human translators. At present, such a working mode can not only ensure a high level of translation quality, but also allow machine translation to improve the quality and efficiency of translation through artificial intelligence self-learning procedures.

This paper combines deep learning algorithms to construct an intelligent language translation system, evaluates the translation effect of the system, and analyzes the effectiveness of the actual situation to provide a reference for subsequent intelligent language translation.

## 2. Related work

The voice is produced by the joint action of the vocal organs and the vocal tract. The vibration of the vocal cords produces a voice signal, and the voice causes the air to vibrate to produce a sound pressure wave. The voice signal sent by the human body contains a lot of information. Intuitive life experience and academic research conclusions show that fatigue information is implicit in human speech signals. The literature [4] proved that when the subject remains awake for 24 hours, the duration of the pause in the speech gradually increases, and the change of the fourth formant frequency of the vowel pronunciation decreases. The literature [5] used the non-linear dynamic characteristics of speech to detect speech fatigue. The literature [6] analyzed the relationship between relevant features and fatigue in speech recognition. The literature [7] used three formants to recognize speech in the ill-conditioned pronunciation system. The literature [8] extracted the fatigue feature parameters contained in speech, and proposed an effective classification of speech fatigue degree based on BP neural network. The best classification recognition rate based on BP neural network can reach 92.5%. The literature [9] proposed a driving fatigue detection method based on multiple voice features according to the influence of human fatigue on the vocal system.

Literature [10] reconstructs the phase space of the speech chaotic attractor, and establishes a nonlinear dynamics model of the speech signal. In order to improve the sufficiency and objectivity of driving fatigue detection, the non-linear features of the voice under this model are extracted: maximum Lyapunov exponent, approximate entropy and fractal dimension, and compared with the voice features under the traditional excitation source-filter model: pitch frequency The combination of formant and Mel frequency standard cepstrum coefficients reflects the fatigue information contained in the voice from different angles. Finally, a multi-feature fusion classifier is established by support vector machine technology, which is used for the fatigue recognition of the driver's voice samples. Literature [11] proposed a fatigue detection method based on speech psychoacoustics, which uses the perceptual masking process in psychoacoustics to highlight the high-sensitive fatigue frequencies, and quantifies the abnormal sounds of fatigue in speech by masking the prosodic features extracted by psychoacoustic perception. Traditional research on speech signals focuses on finding information from feature engineering, such as the short-term energy of the speech signal, short-term average zero-crossing rate, pitch frequency, formant, Mel Frequency Cepstrum Coeefficient (MFCC), MFCC logarithmic power spectrum, speech rate, perceptual linear prediction coefficient (Perceptual Linear Prediction, PLP), amplitude perturbation, etc. [12]. In terms of voice recognition fatigue, in traditional detection methods, feature engineering can be used to study the relationship between different fatigue feature parameters and fatigue by extracting features from voice samples labeled with different fatigue information. Describe the characteristics of the fatigue state, select the optimal feature describing the fatigue state by comparing different characteristics, integrate and optimize the information of different characteristics to find the optimal feature set, so as to achieve the completeness and complementarity of the fatigue information, and establish an optimal feature set. Excellent feature set [13]. Feature extraction plus classification is a typical speech emotion recognition mode. At present, a large number of researchers have studied some key features related to the emotional state of the human body. Literature [14] proposed a minimum feature set, called the Geneva minimal acoustic parameter set. It is composed of 62 features, and 88 features can be obtained through expansion. These 88 features can be used as benchmarks for future research. Combining these features with some static classifiers such as support vector machines can effectively identify the emotional state of the human body. . The traditional method of feature extraction does not pay enough attention to time information. At present, a large number of psychological models show that time information plays an important role in emotion recognition. For example, changes in stress and intonation patterns are closely related to changes in human emotional state information. One method of using time information is to find the standard deviation and mean value of the speech signal time series, and input it as an input vector to a static classifier, but this method may cause the loss of key time information, for example, after reversing the time The feature vector of the spectrogram and the original spectrogram is the same, but it does not necessarily express the same emotional state [15]. In order to overcome this shortcoming, the standard deviation, average value, and pseudo-syllable rate of voiced and unvoiced sounds are added to the input vector. However, these measures cannot make up for the missing many key time information, such as a single time feature at different times. Different patterns of change [16].

## 3. Speech processing algorithm based on deep learning

The calculation process of PESQ is shown in Figure 1. First, the clean voice signal and the enhanced voice signal need to be preprocessed, and then the voice signal is time aligned. This process includes coarse delay estimation and short sentence segmentation and alignment. Next, the loudness spectrum of clean speech and enhanced speech is obtained through the similar auditory transformation as in the calculation of Bark spectral distortion (BSD), and the symmetrical interference and asymmetrical interference terms are calculated. Finally, the PESQ score can be obtained by linearly combining the average value of symmetric interference and the average value of asymmetric interference, that is[17]:

$$PESQ = 4.5 - 0.1 \cdot d_{sym} - 0.0309 \cdot d_{asym} \quad (1)$$

Among them, $d_{sym}$ and $d_{asym}$ respectively represent the average value of symmetric interference and asymmetric interference, and these two types of interference are calculated frame by frame. Auditory masking effects are taken into account, and symmetrical interference needs to consider the absolute difference between the sound spectrum of clean speech and enhanced speech.

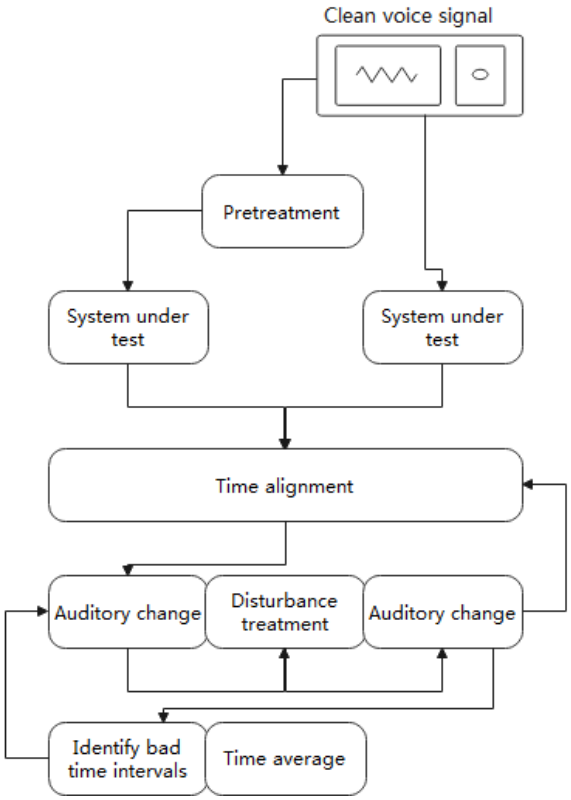


Figure 1 The calculation flow chart of PESQ measurement

The short-time envelope of the clean speech signal can be expressed as:

$$x_{j,m} = \left[ X_j(m-N+1), X_j(m-N+2), ..., X_j(m) \right]^T \quad (2)$$

Among them, $X \in R^{15*M}$ is the obtained 1/3 octave band, M is the total number of signal frames, and m is the index of the frame. At the same time, $j \in \{1,2,...,15\}$ is the index of 1/3 octave, and N=30 is equivalent to a frame length of 384ms. Similarly, $\hat{x}_{j,m}$ can be used to represent the short-time envelope of enhanced speech or noisy speech.

The time envelope of the speech disturbed by noise after normalization and clipping can be expressed as $\hat{x}_{j,m}$.

Intelligibility measure: Intermediate intelligibility can be defined as the correlation coefficient between two time-domain envelopes, that is[18]:

$$d_{j,m} = \frac{\left(x_{j,m} - \mu_{x_{j,m}}\right)^T \left(x_{j,m} - \mu_{x_{j,m}}\right)}{\left\| x_{j,m} - \mu_{x_{j,m}} \right\|_2 \left\| x_{j,m} - \mu_{x_{j,m}} \right\|_2} \quad (3)$$

Among them, $\|\cdot\|_2$ represents the L2 paradigm, and $\mu(\cdot)$ represents the sample mean of the corresponding vector. Finally, STOI is obtained by averaging the intermediate intelligibility of all subbands and frames, that is:

$$STOI = \frac{1}{15M} \sum_{j,m} d_{j,m} \quad (4)$$

This paper takes the deep neural network speech enhancement system based on the mean square error (MSE) cost function as the baseline system. The framework of the baseline system is shown in Figure 2[19].

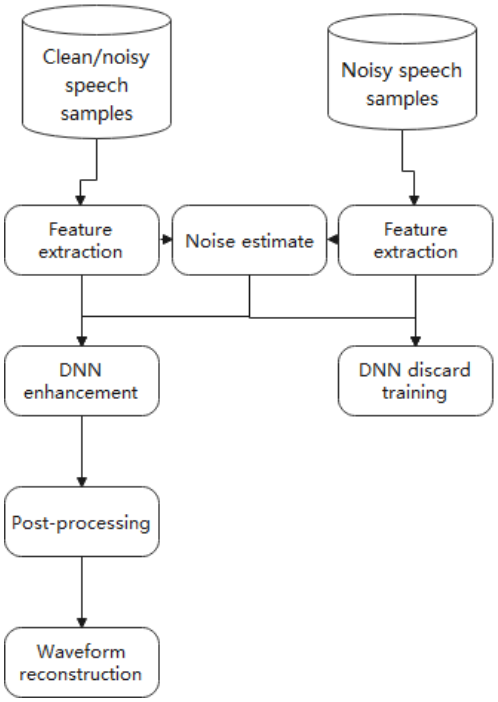

Figure 2 Speech enhancement baseline system based on DNN

The baseline DNN structure is a multilayer forward neural network. The input and expected output of the network are respectively the log-power spectra (LPS) feature of noisy speech and the corresponding LPS feature of clean speech. A strict logarithmic power spectrum of noisy speech can be expressed by the following formula, namely:

$$N(t,\mathrm{f}) = \log\left(\left|STFT\left(n^{u}\right)\right|^{2}\right) \quad (5)$$

Among them, STFT stands for short-time Fourier transform, t and f stand for time and frequency respectively, and the value range of f is from 0 to N=DFT size/2. We set nt to represent the t-th frame of N(t,f), and the frame of context extension at time t is represented by y, that is:

$$y_t = \left[n_{t-\tau},\mathrm{L}\ ,n_{t-1},n_t,n_{t+1},\mathrm{L}\ ,n_{i+\tau}\right] \quad (6)$$

We set s(t,f) to represent the LPS of the clean speech corresponding to the noisy speech $n^{*}$.The back propagation algorithm based on the MSE standard is used to train the DNN, and the small batch stochastic gradient descent algorithm is used to optimize the model parameters. The cost function used for training can be expressed by the following formula, that is:

$$MSE = \frac{1}{K}\sum_{k=1}^{K}\left\|\hat{s}_t - s_t\right\|^{2} + \lambda\left\|W\right\|_2^2 \quad (7)$$

Among them, K is the scale of the small batch, $\hat{s}_t = f\left(\theta, y_t\right)$ is the output of the network, $f\left(\theta\right)$ is the nonlinear mapping function between the input and output of the DNN, and $\theta$ is the weight W and the deviation parameter between the layers of the network. $\lambda\left\|W\right\|_2^2$ is a regularization term, and its purpose is to prevent overfitting during training.

In order to make full use of the time information of speech, we merge the amplitude spectrum features of adjacent frames into a single input feature vector. Therefore, the feature vector centered on the first frame can be constructed as:

$$Y = \left[y(l-3,1),\ldots,y(l-3,K),\ldots,y(l,1),\ldots,y(l+3,1),\ldots,y(l+3,K)\right] \quad (8)$$

Among them, 1 represents a frame, K represents the dimension of the amplitude spectrum of a frame, the number of adjacent frames on the left and right sides of the first frame is 3.The training goal of the network is the amplitude spectrum characteristics of the corresponding clean speech signal in the frequency domain, that is:

$$X = \left[x(l,1),\ldots,x(l,K)\right] \quad (9)$$

Among them, $x(l,K)$ represents the amplitude spectrum characteristic of the k-th frequency band of the first frame of the clean speech.

This paper calculates fwSNRseg in the STFT domain, and its calculation formula is as follows:

$$fwSNRseg = \frac{1}{L}\sum_{l=0}^{L-1}\frac{\sum_{k=1}^{K}W(l,k)\lg\dfrac{X(l,k)^2}{\left(X(l,k)-\hat{X}(l,k)^2\right)}}{\sum_{k=1}^{K}W(l,k)} \quad (10)$$

Among them, the total number of frames after the time domain signal is divided into frames is L, and the total number of frequency bands after STFT transformation in each frame is K. At the same time, $\hat{X}(l,k)$ represents the amplitude spectrum of the k-th

frequency band of the first frame of the clean speech signal, X(l,k) represents the amplitude spectrum of the noisy or enhanced speech in the same frequency band, and w(l,k) represents the perception-based weighting factor applied to each frequency band. This paper proposes to use Ideal Binary Mask (IBM) and Absolute Threshold of Hearing (ATH) to obtain the weight w (l,k) of each frequency band. The two weighting methods are introduced as follows:

(1) Based on IBM's frequency domain weighting: Frequency domain masking is a psychoacoustic model that can be effectively applied to perceptual audio coding. The IBM value is applied to each frequency band as a weighting factor, namely:

$$W(l,k)=\begin{cases}1\ SNR(l,k)>0\\0\ SNR(l,k)\le 0\end{cases} \tag{11}$$

Based on IBM, the idea of weighting frequency bands is: in the frequency band where speech energy dominates, the noise will be masked, so the noise is inaudible. However, in the frequency band dominated by noise energy, the speech will be masked, so the human ear cannot perceive the speech, and the $W(l,k)=0$ of these frequency bands can remove the frequency band dominated by noise energy.

(2) Frequency domain weighting based on ATH: ATH defines the minimum sound energy (sound pressure level, unit is dB) of a pure tone that can be measured and heard in a quiet environment. The relationship between energy threshold and frequency can be approximated as:

$$ATH(f_q)=3.64\left(\frac{f_q}{1000}\right)^{-0.8}-6.5e^{-0.6\left(\frac{f_q}{1000}-3.3\right)^2}+10^{-3\left(\frac{f_q}{1000}\right)^4} \tag{12}$$

The frequency band weighting factor W(1,k) is defined as inversely proportional to $ATH(f_q)$. The specific implementation steps are: First, according to formula (12), $ATH(f_q)$ at the center frequency of each frequency band is calculated. Next, these thresholds are standardized to make their minimum value 1. Finally, by taking the reciprocal of the standardized threshold, the weighting factor W(l,k) corresponding to each frequency band can be obtained. In order to avoid the weight of the 0th frequency band being 0 $\left(ATH(f_q)=\infty\right)$, we calculate the threshold value at the 3/4 frequency range of the 0th frequency band.

The cost function proposed in this paper that should be minimized during DNN training can be expressed by the following formula, that is:

$$MSSNR=-\frac{1}{M}\sum_{m=1}^{M}fwSNRseg\left(x_m(l,k),\hat{x}_m(l,k)\right) \tag{13}$$

Among them, $x_m(l,k)$ and $\hat{x}_m(l,k)$ respectively represent the amplitude spectrum of clean speech and enhanced speech with training sample index m, and the total number of training samples is M. The fwSNRseg(·) function is used to calculate the fwSNRseg value of noisy speech or enhanced speech when clean speech is given.

DNN training should consider more frequency bands that are more important to human auditory perception. Based on the above ideas, the perceptual weighted mean square error (wMSE) cost function is proposed, that is:

$$wMSE=\frac{1}{M}\sum_{m=1}^{M}f\left[\frac{W(l,k)}{L_m}\left\|x_m(l,k)-\hat{x}_m(l,k)\right\|_2^2\right] \tag{14}$$

Among them, the total number of frequency bands of training samples with index m is l, and the selection of weighting factors w(1,k) is flexible. This paper applies perceptually-based weighting factors to each frequency band according to formulas (11) and (12).

Based on the above two perceptually related cost functions, this paper proposes to combine fwSNRseg and wMSE into one cost function to obtain a joint optimized cost function. It is calculated as follows:

$$wMSE+MSSNR=\frac{1}{M}\sum_{m=1}^{M}\left[\frac{W(l,k)}{L_m}\left\|x_m(l,k)-\hat{x}_m(l,k)\right\|_2^2-wSNRseg\left(x_m(l,k),\hat{x}_m(l,k)\right)\right] \tag{15}$$

The amplitude spectrum estimation of the enhanced speech is obtained through forward propagation, that is:

$$\hat{x}_m(l,k)=FFN(Y;\omega,b) \tag{16}$$

Among them, Y represents the input noisy speech feature, and $\omega$ and b represent the weight and deviation parameters of the DNN. Next, the corresponding clean speech amplitude spectrum is used as the training target, and the perceptual correlation cost function value between the network output and the training target is minimized through the back propagation algorithm based on gradient descent to obtain optimized weights and deviation parameters.

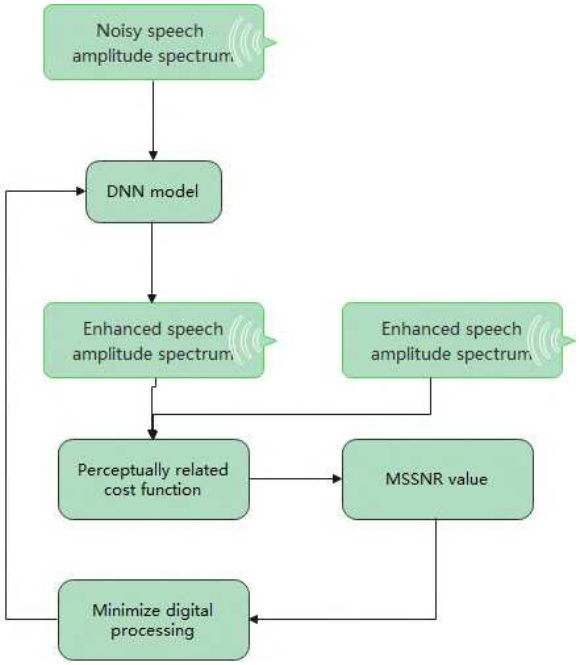


Figure 3 DNN training flowchart based on perceptually related cost function

The VAD results of each frequency band are used during the existence of the noise frame to construct and update the noise model. The noise variance $\lambda_k(m)$ of the current frame can be expressed by the recursive average formula of the noise variance of the previous frame and the noisy speech power spectrum of the current frame, that is:

$$\lambda_k(m)=\left(1-\alpha_k(m)\lambda_k(m-1)+\alpha_k(m)\left|X_k(m)\right|^2\right) \quad (17)$$

Among them, $\lambda_k(m)$ represents the noise variance of the k-th frequency band of the m-th frame, $X_k(m)$ represents the noise-containing speech spectrum, and the recursive average weighting factor $\alpha_k(m)$ is related to the speech existence probability of the current frame and frequency band. It can be expressed by the following formula:

$$\alpha_k(m)=\alpha_0\left(1-p_k(m)\right)\left(1-p_{total}(m)\right) \quad (18)$$

Among them, $p_k(m)$ represents the speech existence probability of each frequency band, and this parameter is calculated by the VAD algorithm. $p_{total}(m)$ represents the overall speech existence probability of the m-th frame, and this parameter is obtained by averaging $p_k(m)$ of all frequency bands. Next, according to the noise spectrum variance given by equation (17), two parameters: a priori SNR $\xi_k(m)$ and a posteriori SNR $\gamma_k(m)$ are calculated:

$$\gamma_k(m)=\frac{\left|X_k(m)\right|^2}{\lambda_k(m)} \quad (19)$$

$$\xi_k(m)=\frac{\left|\hat{S}_k(m)\right|^2}{\lambda_k(m)}=\beta\xi_k(m-1)=(1\text{-}\beta)\max\left\{0,\gamma_k(m-1)\right\} \quad (20)$$

Among them, the prior signal-to-noise ratio is estimated by the decision-directed approach (DDA), $\hat{S}_k(m)$ is the clean speech spectrum, and β is the recursive average weighting factor. After calculating these two parameters, the suppression gain of each frequency band can be estimated. This parameter is a function of $\xi_k(m)$ and $\gamma_k(m)$, that is:

$$G_k(m)=g\left(\xi_k(m),\gamma_k(m)\right) \quad (21)$$

Finally, by applying the suppression gain to the noisy speech spectrum, an enhanced speech spectrum can be obtained.

Different suppression criteria g(.) are based on different statistical assumptions and optimization criteria. For example, for the Gaussian distribution of speech and noise signals, the optimal suppression criterion in the sense of MMSE is Wiener filtering, that is:

$$H_k(m)=\frac{\xi_k(m)}{1+\xi_k(m)} \quad (22)$$

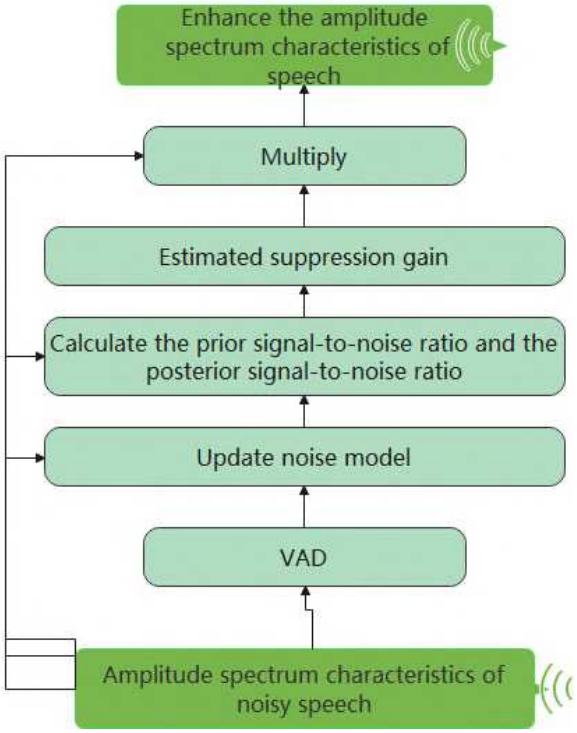


Figure 4 The unified framework of traditional single-channel speech enhancement algorithms

This paper adopts two modes: symmetric mode and causal mode, and the input is the noisy speech amplitude spectrum feature including the context window of M frames. Therefore, the symmetric context and causal context input windows can be expressed as:

$$X = \left[ x_{m-\frac{M-1}{2}}, \ldots, x_m, \ldots, x_{m+\frac{M-1}{2}} \right] \tag{23}$$

$$X = \left[ x_{m-\frac{M-1}{2}}, \ldots, x_m \right] \tag{24}$$

Among them, $x_m$ represents the amplitude spectrum eigenvectors of different frequency bands on the time frame m. The amplitude spectrum feature of the past frame is used as the context input, and the goal is to restore the clean speech amplitude spectrum feature vector of the last frame of the context window, as shown in Figure 5.

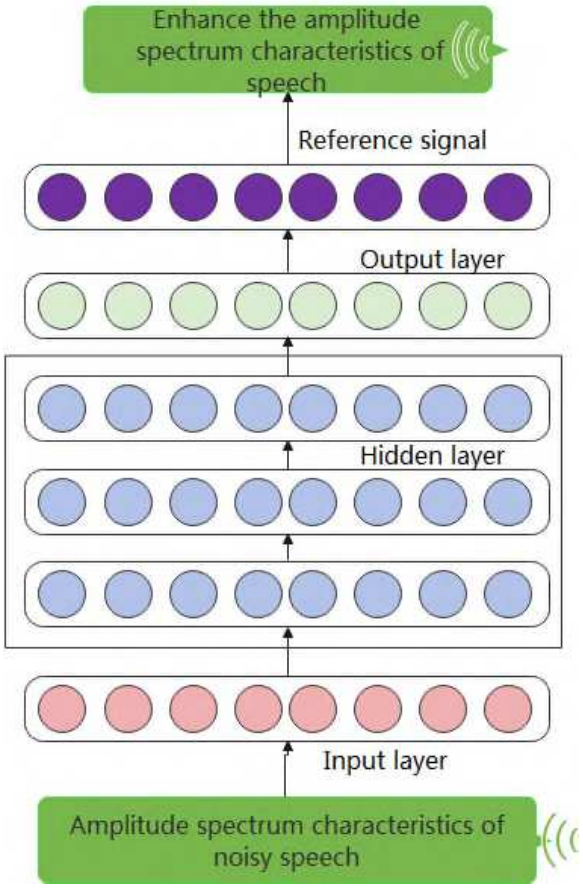


Figure 5 End-to-end DNN regression speech enhancement architecture

One way to minimize speech distortion is to replace the direct estimation of the amplitude spectrum characteristics of the clean speech with the estimation of the suppression gain of each frequency band for the output of the DNN. The suppression gain estimation based on DNN is shown in Figure 7. Among them, in the training process, the input of the DNN is also the amplitude spectrum feature $x_k(m)$ of the noisy speech, and the target output of the DNN is set as:

$$G_k(m) = \frac{S_k(m)}{X_k(m)} \tag{25}$$

Among them, $S_k(m)$ is the amplitude spectrum characteristic of the clean speech signal. This paper enhances the noisy speech signal frame by frame. Therefore, the cost function is defined as the following formula:

$$MSSNR = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{k=1}^{K} W_k(m)\lg\frac{S_k(m)^2}{\left(S_k(m)-\hat{S}_k(m)^2\right)}}{\sum_{k=1}^{K} W_k(m)} \tag{26}$$

Among them, M and K respectively represent the total number of frames and frequency bands, and $\hat{S}_k(m)$ is the amplitude spectrum feature of enhanced speech, which is obtained by multiplying the output of DNN and the amplitude spectrum of noisy speech. $W_k(m)$ represents a weighting factor based on human auditory perception applied to each frequency band.

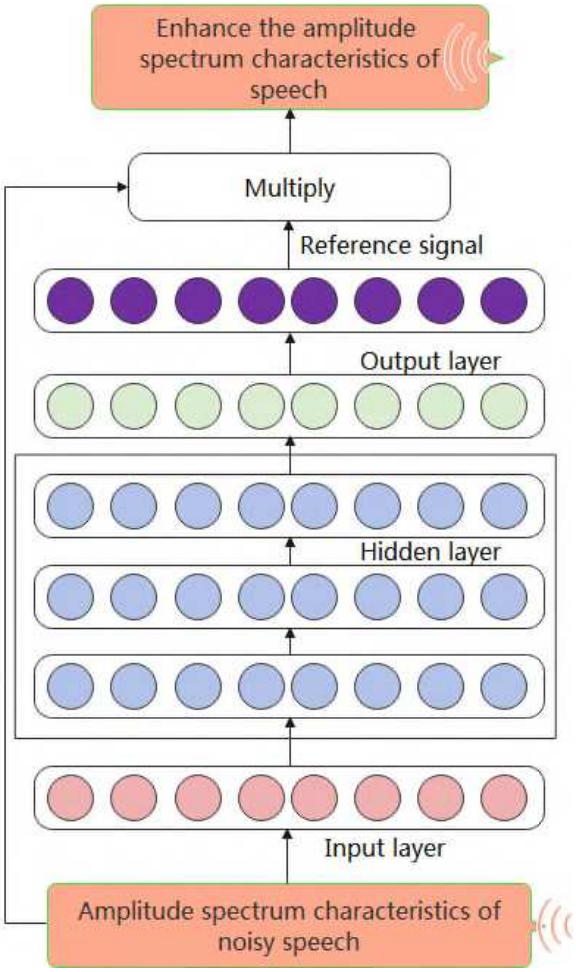


Figure 6 Speech enhancement architecture of suppression gain estimation based on DNN

In the enhancement stage, the noisy speech is preprocessed and the amplitude spectrum features are extracted and input into the trained DNN.

The structure of the DNN-based suppression gain estimation speech enhancement algorithm structure is shown in Figure 7. According to formulas (17), (18), (19), and (20), the noise variance is updated, and the prior signal-to-noise ratio and posterior signal-to-noise ratio parameter values are calculated.

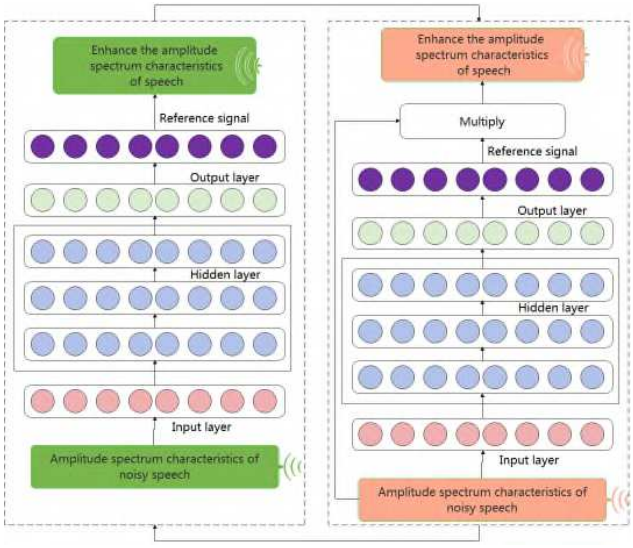


Figure 7 Structured DNN-based suppression gain estimation speech enhancement architecture

The target output of the network is set to the speech presence probability (SPP) of each frequency band. SPP is obtained by the VAD algorithm based on the statistical model. If $H_0^k$ represents the state where the voice in the k-th frequency band exists, $H_1^k$ represents the state where the voice does not exist, and $Y_k$ represents the noisy voice spectrum, the SPP of each frequency band can be expressed as:

$$P\left(H_1^k \mid Y_k\right) = \frac{1-q_k}{1-q_k+q_k\left(1+\xi_k^{'}\right)\exp\left(-v_k^{'}\right)} \quad (27)$$

$$\xi_k^{'} = \frac{\xi_k}{1-q_k}, v_k^{'} = \frac{\xi_k^{'}}{\xi_k^{'}+1}\gamma_k \tag{28}$$

Among them, there is $q_k = P\left(H_0^k\right)$. $\xi_k$ and $\gamma_k$ respectively represent the a priori SNR and the posterior SNR of each frequency band.

## 4. Intelligent language translation system based on deep learning algorithms

The translation system is mainly about data preprocessing, word vector pre-training, and translation model training and testing. First of all, data preprocessing needs to further process the data into a machine-recognizable format, and the result of word segmentation after the corpus is also converted into a vector that can be recognized by the model. The technology involved in the training of the translation model and the internal network structure are very complex. To improve the overall performance of the translation model, it is necessary to have a thorough understanding of the internal structure and to be able to propose improvements and then perform experimental verification. The other is the test link, which not only needs to evaluate the BLEU index of the model's translation performance, but also needs to carry out specific examples to test the accuracy of the translation results. The overall architecture diagram of the machine translation system is shown in Figure 8:

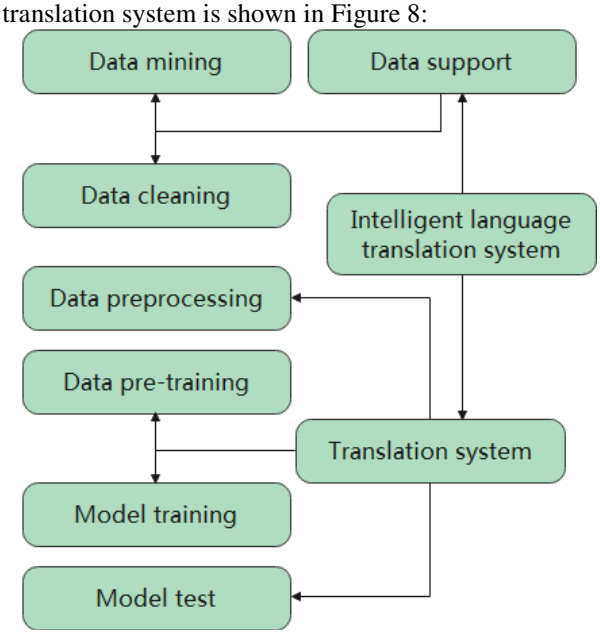


Figure 8 Intelligent language translation model

After constructing the above system, this paper evaluates the effect of intelligent language translation on the system, and conducts simulation experiments. First of all, this paper verifies the effect of the algorithm constructed in this paper on speech recognition, and obtains the following results.

Table 1 Statistical table of speech recognition results

| NO | Speech Recognition | NO | Speech Recognition |
|---|---|---|---|
| 1 | 92.2 | 15 | 92.9 |
| 2 | 93.2 | 16 | 93.4 |
| 3 | 90.9 | 17 | 89.3 |
| 4 | 91.6 | 18 | 92.8 |
| 5 | 93.8 | 19 | 91.4 |
| 6 | 91.1 | 20 | 89.8 |
| 7 | 90.2 | 21 | 93.5 |
| 8 | 91.4 | 22 | 92.5 |
| 9 | 92.8 | 23 | 90.5 |
| 10 | 92.3 | 24 | 93.1 |
| 11 | 92.4 | 25 | 91.7 |
| 12 | 92.8 | 26 | 93.0 |
| 13 | 93.5 | 27 | 90.1 |
| 14 | 93.2 | | |

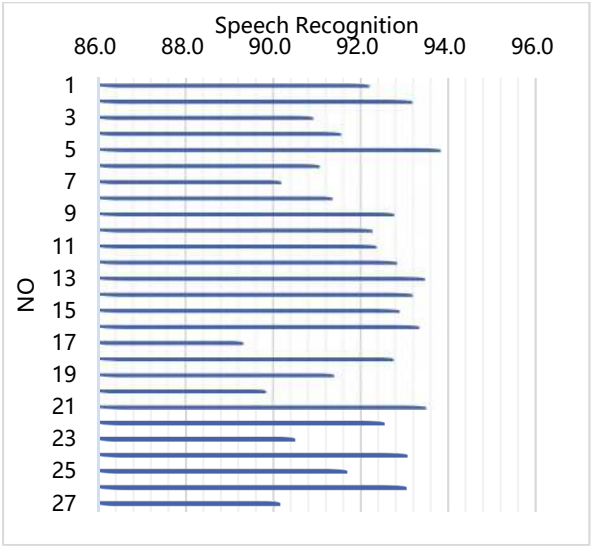


Figure 9 Statistical diagram of the speech recognition results of the deep learning algorithm

It can be seen from the above research that the intelligent language translation system based on the deep learning algorithm proposed in this paper has a good speech recognition effect, and the intelligent language translation effect is carried out on this basis. The test results are shown in Table 2 and Figure 10.

Table 2 Statistical table of the practical effect of the system

| NO | Translation effect | NO | Translation effect |
|---|---|---|---|
| 1 | 84.5 | 15 | 94.2 |
| 2 | 91.3 | 16 | 88.7 |
| 3 | 92.2 | 17 | 90.9 |
| 4 | 93.0 | 18 | 91.8 |
| 5 | 85.7 | 19 | 90.0 |
| 6 | 89.3 | 20 | 91.7 |
| 7 | 81.6 | 21 | 87.4 |
| 8 | 91.4 | 22 | 88.4 |
| 9 | 87.9 | 23 | 92.9 |
| 10 | 90.1 | 24 | 93.5 |
| 11 | 94.6 | 25 | 88.6 |
| 12 | 88.2 | 26 | 81.1 |
| 13 | 84.6 | 27 | 81.7 |
| 14 | 84.6 | | |

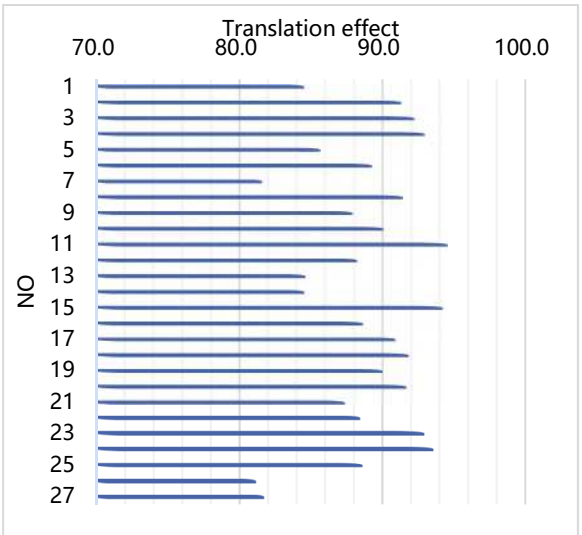


Figure 2 The performance of the intelligent speech translation system

From the above research, it can be seen that the intelligent speech translation system based on deep learning constructed in this paper has good practical performance.

## 5. Conclusion

Machine translation often faces the problem of inconsistency between the training data and the test set sentence to be translated. It is especially obvious in the field of scientific and technological intelligence. There are big differences between different types of scientific and technological literature. If a large-scale corpus without distinction and screening is used for training, it will not only increase the complexity of training and system overhead, but also cause mistranslation of vocabulary, terminology and sentences in different fields, resulting in poor translation performance and high translation costs. In particular, the training corpus of neural machine translation is too large, and the number of unregistered words increases due to the limitation of the vocabulary. Therefore, the self-adaptation problem in the field of machine translation has always been a problem to be solved.

This paper combines deep learning to construct an intelligent speech translation system, and studies a deep learning speech enhancement algorithm based on DNN. By analyzing the existing problems of the MSE cost function used by most of the current DNN-based speech enhancement algorithms, this paper proposes a deep learning speech enhancement algorithm based on the perception-related cost functions, and the reliability of the model in this paper is verified through experimental research.

## Declarations

Funding-No funds received.
Conflicts of Interests-Not applicable.
Availability of data and Material-Not applicable.
Code Availability-Not applicable.
Authors' Contributions- Chunliu Shi is responsible for planning, applying the methodology, deduced the results, writing and editing the manuscript.

## References

[1]. Ngoc Q. K. Duong, HienThanh Duong. A Review of Audio Features and Statistical Models Exploited for Voice Pattern Design[J], computer science, 2015, 03(2):36-39.
[2]. Sarria-Paja M , Senoussaoui M , Falk T H . The effects of whispered speech on state-of-the-art voice based biometrics systems[J], Canadian Conference on Electrical and Computer Engineering, 2015, 2015(1):1254-1259.
[3]. Leeman A , Mixdorff H , O'Reilly M , et al. Speaker-individuality in Fujisaki model f0 features: Implications for forensic voice comparison[J], International Journal of Speech Language and the Law, 2015, 21(2):343-370.
[4]. Hill A K , Rodrigo A. Cárdenas, Wheatley J R , et al. Are there vocal cues to human developmental stability? Relationships between facial fluctuating asymmetry and voice attractiveness[J], Evolution & Human Behavior, 2017, 38(2):249-258.
[5]. Marcin Woźniak, Dawid Połap. Voice recognition through the use of Gabor transform and heuristic algorithm[J], Nephron Clinical Practice, 2017, 63(2):159-164.
[6]. Haderlein T , Michael Döllinger, Václav Matoušek, et al. Objective voice and speech analysis of persons with chronic hoarseness by prosodic analysis of speech samples[J], Logopedics Phoniatrics Vocology, 2015, 41(3):106-116.
[7]. Nidhyananthan S S , Muthugeetha K , Vallimayil V . Human Recognition using Voice Print in LabVIEW[J], International Journal of Applied Engineering Research, 2018, 13(10):8126-8130.
[8]. Malallah F L , Saeed K N Y M G , Abdulameer S D , et al. Vision-Based Control By Hand-Directional Gestures Converting To Voice[J], International Journal of Scientific & Technology Research, 2018, 7(7):185-190.
[9]. Morgan Sleeper. Contact effects on voice-onset time in Patagonian Welsh[J], acoustical society of america journal, 2016, 140(4):3111-3111.
[10]. Mohan G , Hamilton K , Grasberger A , et al. Realtime voice activity and pitch modulation for laryngectomy transducers using head and facial gestures[J], Journal of the Acoustical Society of America, 2015, 137(4):2302-2302.
[11]. Kang T G , Kim N S . DNN-Based Voice Activity Detection with Multi-Task Learning[J], Ieice Transactions on Information & Systems, 2016, E99.D(2):550-553.
[12]. Choi, HaNa, Byun, SungWoo, Lee, SeokPil. Discriminative Feature Vector Selection for Emotion Classification Based on Speech[J], Transactions of the Korean Institute of Electrical Engineers, 2015, 64(9):1363-1368.
[13]. Herbst C T , Hertegard S , Zangger-Borch D , et al. Freddie Mercury—acoustic analysis of speaking fundamental frequency, vibrato, and subharmonics[J], Logopedics Phoniatrics Vocology, 2016, 42(1):1-10.
[14]. Al-Tamimi J . Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: Implications for formal representations[J], Laboratory Phonology, 2017, 8(1):1-40.
[15]. Abdel-Hamid O, Mohamed A, Jiang H, et al. Convolutional neural networks for speech recognition[J], IEEE/ACM Transactions on audio, speech, and language processing, 2014, 22(10): 1533-1545.
[16]. Kim C, Stern R M. Power-normalized cepstral coefficients (PNCC) for robust speech recognition[J], IEEE/ACM Transactions on audio, speech, and language processing, 2016, 24(7): 1315-1329.
[17]. Noda K, Yamaguchi Y, Nakadai K, et al. Audio-visual speech recognition using deep learning[J], Applied Intelligence, 2015, 42(4): 722-737.
[18]. Qian Y, Bi M, Tan T, et al. Very deep convolutional neural networks for noise robust speech recognition[J], IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(12): 2263-2276.
[19]. Li J, Deng L, Gong Y, et al. An overview of noise-robust automatic speech recognition[J], IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(4): 745-777.

Shi Chunliu is an associate professor working on foreign language teaching in School of Foreign Languages in Zhengzhou University of Aeronautics, China. She completed her BA study from Xi'an International Studies University and MA study from Beijing Foreign Studies University. Her Research interests include tourism English, tourism translation, cross-culture communication, and computer-aided translation. More than 10 paper and 4 books published.