Research on intelligent language translation system based on deep learning algorithm

Chunliu Shi

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 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.

Machine translation and computer-assisted translation are often referred to as “computer translation.” The term “machine translation” refers to translations performed by computers, while “computer-assisted translation” refers to translations performed by humans with the aid of computers (Sarria-Paja et al. 2015). 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 (Leeman et al. 2014).

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. In addition to machine translation and computer-assisted translation, more powerful artificial intelligence interactive translation will be used to accomplish many jobs.

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 artificial intelligence and human translators' connection and mutual aid. Currently, such a working mode can not only guarantee excellent translation quality, but also enable machine translation to enhance translation quality and efficiency via artificial intelligence self-learning processes. 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 (Hill et al. 2017) proved that when the subject remains awake for 24 h, 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 used the nonlinear dynamic characteristics of speech to detect speech fatigue. The literature (Haderlein et al. 2016) analyzed the relationship between relevant features and fatigue in speech recognition. The literature (Nidhyanthan et al. 2018) used three formants to recognize speech in the ill-conditioned pronunciation system. The literature (Malallah et al. 2018) 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 (Sleeper 2016) proposed a driving fatigue detection method based on multiple voice features according to the influence of human fatigue on the vocal system.

Literature (Mohan et al. 2015) reconstructs the phase space of the speech chaotic attractor and establishes a nonlinear dynamics model of the speech signal. The nonlinear characteristics of the speech under this model are retrieved in order to enhance the sufficiency and objectivity of driving tiredness detection: The speech characteristics under the conventional excitation source-filter paradigm were compared to the maximal Lyapunov exponent, approximate entropy, and fractal dimension: frequency of pitch. From many perspectives, the combination of formant and Mel frequency standard cepstrum coefficients shows the tiredness information inherent in the voice. Finally, using support vector machine technology, a multi-feature fusion classifier is created for the tiredness detection of the driver’s speech samples. Literature (Kang & Kim 2016) 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 Coefficient (MFCC), MFCC logarithmic power spectrum, speech rate, perceptual linear prediction coefficient (perceptual linear prediction, PLP), and amplitude perturbation. (Choi et al. 2015). 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 (Herbst et al. 2017). 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 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 (Muhammad Talha et al. 2021). 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. A significant variety of psychological models now indicate that temporal information is crucial in emotion perception. Changes in stress and intonation patterns, for example, are strongly linked to changes in human emotional state data. 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 (Abdel-Hamid et al. 2014; Talha et al. 2020). 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 (Kim & Stern 2016).

3 Speech processing algorithm based on deep learning

The calculation process of PESQ is shown in Fig. 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 (Noda et al. 2015):

\[
PESQ = 4.5 - 0.1 \cdot d_{sym} - 0.0309 \cdot d_{asym}
\]  

(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.

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

\[
x_{j,m} = [X_j(m - N + 1), X_j(m - N + 2), \ldots, X_j(m)]^T
\]

(2)

Among them, \(X \in \mathbb{R}^{15 \times M}\) 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, \(j \in \{1, 2, \ldots, 15\}\) is the index of 1/3 octave, and \(N = 30\) is equivalent to a frame length of 384 ms. 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 (Qian et al. 2016):

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

Among them, $\left\| \cdot \right\|_2$ 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:

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

The baseline system in this article is a deep neural network voice augmentation system based on DNN structure. The network’s input and anticipated output are the log-power spectra (LPS) feature of loud speech and the equivalent LPS feature of clean speech, respectively. The following formula may be used to describe a rigorous logarithmic power spectrum of noisy speech:

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

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 $N = \text{DFT size}/2$.

We set $n^o(t,f)$ 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:

$$y_t = [n_{t-2}, \ldots, n_{t-1}, n_t, n_{t+1}, \ldots, n_{t+2}] \quad (6)$$

We set $s(t,f)$ to represent the LPS of the clean speech corresponding to the noisy speech $n^o$. 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:

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

Among them, $K$ is the scale of the small batch, $\hat{s}_t = f(\theta, y_t)$ is the output of the network, $f(\theta)$ 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:

$$Y = [y(l-3, 1), \ldots, y(l-3, K), \ldots, y(l, 1), \ldots, y(l+2, 1), \ldots, y(l+3, K)] \quad (8)$$

Among them, $l$ 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:

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

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:

$$\text{fwSNRseg} = \frac{1}{L} \sum_{l=0}^{L-1} \frac{\sum_{k=1}^{K} W(l, k) \log \frac{x(l,k)^2}{(X(l,k) - x(l,k))^2}}{\sum_{k=1}^{K} W(l, k)} \quad (10)$$

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, $X(l, k)$ 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) The following frequency domain weighting is based on IBM’s frequency domain weighting: Frequency domain masking is a psychoacoustic paradigm for perceptual audio coding that works well. The IBM value is applied to each frequency band as a weighting factor, namely:

$$W(l,k) = \begin{cases} 1 & \text{SNR}(l,k) > 0 \\ 0 & \text{SNR}(l,k) \leq 0 \end{cases} \quad (11)$$

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: In a calm situation, ATH specifies the lowest sound energy (sound pressure level, unit is dB) of a pure tone that may be measured and heard. The relationship between energy threshold and frequency can be approximated as:

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

The frequency band weighting factor $W(l,k)$ is defined as inversely proportional to $\text{ATH}(f_q)$. The specific implementation steps are: First, according to formula (12), $\text{ATH}(f_q)$ 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 ($\text{ATH}(f_q) = \infty$), 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:

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

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 $\text{fwSNRseg}(-)$ function is used to calculate the $\text{fwSNRseg}$ 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:

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

Among them, the total number of frequency bands of training samples with index $m$ is 1, and the selection of weighting factors $W(l,k)$ 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 $\text{fwSNRseg}$ and $\text{wMSE}$ into one cost function to obtain a joint optimized cost function. It is calculated as follows:

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

$$-\text{fwSNRseg}(x_m(l,k), \hat{x}_m(l,k)) \quad (15)$$

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

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

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.

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 $\hat{\sigma}_k(m)$ 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:

$$\hat{\sigma}_k(m) = \left(1 - \tau_k(m)\right)\hat{\sigma}_k(m-1) + \tau_k(m)|X_k(m)|^2 \quad (17)$$

Among them, $\hat{\sigma}_k(m)$ 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:

\[
\alpha_k(m) = \alpha_0(1 - p_k(m))(1 - p_{\text{total}}(m)) \tag{18}
\]

Among them, \( p_k(m) \) represents the speech existence probability of each frequency band, and this parameter is calculated by the VAD algorithm. \( p_{\text{total}}(m) \) 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 Eq. (17), two parameters: a priori SNR \( \gamma_k(m) \) and a posteriori SNR \( \gamma_k(m) \) are calculated:

\[
\gamma_k(m) = \frac{|X_k(m)|^2}{\bar{X}_k(m)} \tag{19}
\]

\[
\xi_k(m) = \frac{|\hat{S}_k(m)|^2}{\bar{X}_k(m)} = \beta \xi_k(m - 1) \\
= (1 - \beta) \max\{0, \gamma_k(m - 1)\} \tag{20}
\]

Among them, the prior signal-to-noise ratio is estimated by the decision-directed approach (DDA), \( \xi_k(m) \) is the clean speech spectrum, and \( \beta \) is the recursive average weighting factor. After calculating these two parameters, the suppression gain of each frequency band can be estimated (Figs. 3, 4). This parameter is a function of \( \xi_k(m) \) and \( \gamma_k(m) \), that is:

\[
G_k(m) = g(\xi_k(m), \gamma_k(m)) \tag{21}
\]

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:

\[
H_k(m) = \frac{\xi_k(m)}{1 + \xi_k(m)} \tag{22}
\]

The input is the noisy speech amplitude spectrum feature, which includes the context window of \( M \) frames, and the study uses two modes: symmetric mode and causal mode. As a result, the symmetric and causal context input windows may be written as follows:

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

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

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

---

**Fig. 3** DNN training flowchart based on perceptually related cost function

**Fig. 4** The unified framework of traditional single-channel speech enhancement algorithms
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 Fig. 5.

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 (Fig. 6). The suppression gain estimation based on DNN is shown in Fig. 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:

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

(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:

$$\text{MSSNR} = \frac{1}{M} \sum_{m=0}^{M-1} \frac{\sum_{k=1}^{K} W_k(m) \log \frac{S_k(m)^2}{(S_k(m) - \hat{S}_k(m))^2}}{\sum_{k=1}^{K} W_k(m)}$$

(26)

Among them, $M$ and $K$, respectively, represent the total number of frames and frequency bands, and $S_k(m)$ 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.

The noisy speech is preprocessed in the enhancement step, and the amplitude spectrum characteristics are retrieved and fed into the trained DNN.

Figure 7 depicts the structure of the DNN-based suppression gain estimation speech enhancement method. The noise variance is updated, and the previous signal-to-noise
ratio and posterior signal-to-noise ratio parameter values are computed using formulae (17), (18), (19), and (20).

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 \( H_k^0 \) represents the state where the voice in the \( k \)-th frequency band exists, \( H_k^1 \) 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:

\[
P(H_k^1|Y_k) = \frac{1 - q_k}{1 - q_k + q_k(1 + \xi'_k) \exp(-\gamma'_k)}
\]

\[
\xi'_k = \frac{\tilde{\xi}_k}{1 - q_k}, \quad \gamma'_k = \frac{\tilde{\gamma}_k}{\tilde{\xi}_k + 1}
\]

Among them, there is \( q_k = P(H_k^0) \). \( \xi_k \) 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. To begin, data preprocessing must transform the data into a machine-readable format, and the output of word segmentation after the corpus must also be translated into a vector that the model can identify (Table 1). 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 must not only assess the model’s translation performance using the BLEU index, but also conduct particular instances to ensure that the translation results are accurate. The overall architecture diagram of the machine translation system is shown in Fig. 8:

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 (Fig. 9).
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 Fig. 10.

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. Using a large-scale corpus for training without distinction and screening will not only increase the complexity of training and system overhead, but it will also result in mistranslation of vocabulary, terminology, and sentences across 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.

Author contributions CS is responsible for planning, applying the methodology, deduced the results, writing and editing the manuscript.

Funding No funds received.

Declarations

Conflict of interests The authors declare that they have no conflict of interest.

References

Abdel-Hamid O, Mohamed A-R, Jiang H, Deng L, Penn G, Yu D (2014) Convolutional neural networks for speech recognition.

IEEE/ACM Trans Audio Speech Lang Process 22(10):1533–1545

Choi H-N, Byun S-W, Lee S-P (2015) Discriminative feature vector selection for emotion classification based on speech. Trans Korean Inst Electr Eng 64(9):1363–1368

Haderlein T, Döllinger M, Matoušek V, Nöth E (2016) Objective voice and speech analysis of persons with chronic hoarseness by prosodic analysis of speech samples. Logop Phoniatr Vocol 41(3):106–116

Herbst CT, Hartegard S, Zanger-Borch D, Lindestad PÅ (2017) Freddie Mercury—acoustic analysis of speaking fundamental frequency, vibrato, and subharmonics. Logop Phoniatr Vocol 42(1):29–38

Hill AK, Cárdenas RA, Wheatley JR, Welling LL, Burriss RP, Claes P, Shriver MD (2017) Are there vocal cues to human developmental stability? Relationships between facial fluctuating asymmetry and voice attractiveness. Evol Hum Behav 38(2):249–258

Kang TG, Kim NS (2016) DNN-based voice activity detection with multi-task learning. IEICE Trans Inf Syst 99(2):550–553

Kim C, Stern RM (2016) Power-normalized cepstral coefficients (PNCC) for robust speech recognition. IEEE/ACM Trans Audio Speech Lang Process 24(7):1315–1329

Leeman A, Mixdorff H, O’Reilly M, Kolly M-J, Dellwo V (2014) Speaker-individuality in Fujisaki model f0 features: implications for forensic voice comparison. Int J Speech Lang Law 21(2):343–370

Li J, Deng L, Gong Y, Haeb-Umbach R (2014) An overview of noise-robust automatic speech recognition. IEEE/ACM Trans Audio Speech Lang Process 22(4):745–777

Malallah FL, KNYMG S, Abdulameer SD (2018) Vision-based control by hand-directional gestures converting to voice. Int J Sci Technol Res 7(7):185–190

Mohan G, Hamilton K, Grasberger A, Lammert AC, Waterman J (2015) Realtime voice activity and pitch modulation for laryngectomy transducers using head and facial gestures. J Acoust Soc Am 137(4):2302–2302

Nidhyananthan SS, Muthugeetha K, Vallimayil V (2018) Human recognition using voice print in labview. Int J Appl Eng Res 13(10):8126–8130

Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T (2015) Audio-visual speech recognition using deep learning. Appl Intell 42(4):722–737

Qian Y, Bi M, Tan T, Yu K (2016) Very deep convolutional neural networks for noise robust speech recognition. IEEE/ACM Trans Audio Speech Lang Process 24(12):2263–2276

Sarria-Paja M, Senoussaoui M, Falk TH (2015) The effects of whispered speech on state-of-the-art voice based biometrics systems. In: Paper Presented at the 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE)

Sleeper M (2016) Contact effects on voice-onset time in Patagonian Welsh. J Acoust Soc Am 140(4):3111–3111

Talha M, Sohail M, Tariq R, Ahmad MT (2021) Impact of oil prices, energy consumption and economic growth on the inflation rate in Malaysia. Cuad Econ 44(124):26–32

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.