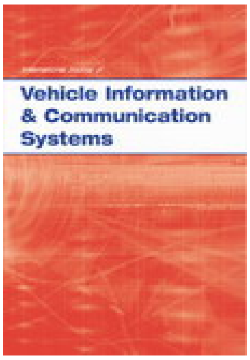


Vehicle Information & Communication Systems

International Journal of Vehicle Information and Communication Systems

ISSN online: 1741-8208 - ISSN print: 1471-0242
https://www.inderscience.com/ijvics

Research on the interactive design of electric vehicle interior based on voice sensing and visual imagery

Tao Ba, Shan Li, Ying Gao, Diyuan Tan

DOI: 10.1504/IJVICS.2023.10056749

Article History:

Received:	21 November 2022
Last revised:	10 January 2023
Accepted:	29 January 2023
Published online:	20 June 2023

Research on the interactive design of electric vehicle interior based on voice sensing and visual imagery

Tao Ba

School of Art and Design,
Zhengzhou University of Light Industry,
Zhengzhou, Henan, China
Email: bt5353968@163.com

Shan Li*

College of Landscape Architecture and Art,
Henan Agricultural University,
Zhengzhou, Henan, China
Email: shanlishan@yandex.com
*Corresponding author

Ying Gao

School of Art and Design,
Zhengzhou University of Light Industry,
Zhengzhou, Henan, China
Email: moon-gao1117@163.com

Diyuan Tan

School of Electro-mechanical Engineering,
Zhongyuan Institute of Science and Technology,
Xuchang, Henan, China
Email: tdy020926@163.com

Abstract: With the complete function of modern automobiles, in-vehicle intelligent devices are becoming more and more complex and the requirements for human-computer interaction are also increasing. The research proposes a speech recognition method that combines multi-window estimation spectral subtraction and dynamic time warping to enhance the denoising ability and speech recognition ability of in-vehicle devices. It also proposes actions based on a Gaussian hybrid segmentation algorithm and a visual image functional space segmentation algorithm. The automatic identification method and the validity of the algorithm are verified. The results show that under different input signal-to-noise ratios, the denoising capability of the method is improved by 2.45% to 31.47% over the baseline method. And the accuracy of speech recognition in the vehicle environment is 92.3% to 98.7%. It is hoped that this research can make some contributions to the upgrading of voice and visual interaction within electric vehicles.

Keywords: speech recognition; multi-window spectral subtraction; dynamic time warping; visual imagery; space segmentation.

Reference to this paper should be made as follows: Ba, T., Li, S., Gao, Y. and Tan, D. (2023) 'Research on the interactive design of electric vehicle interior based on voice sensing and visual imagery', *Int. J. Vehicle Information and Communication Systems*, Vol. 8, Nos. 1/2, pp.152–169.

Biographical notes: Tao Ba graduated from School of Art Design, Zhengzhou University of Light Industry, majoring in Industrial Design. He is currently teaching at the College of Art and Design, Zhengzhou University of Light Industry. His academic research interests include product form design and 3D printing technology application. He has published three papers and participated in two projects.

Shan Li graduated from Zhengzhou University of Light Industry from 1995 to 1999 with a Bachelor's degree in Industrial Design; graduated from Zhengzhou University of Light Industry with a Master's degree in Art and Design from 2006 to 2009. She is currently teaching at Henan Agricultural University. Her academic research direction is user experience, product design and service design. She has published 11 papers, presided over 9 projects and wrote 3 books.

Ying Gao graduated from Hubei Academy of Fine Arts from September 2000 to June 2004 with a Bachelor's degree in Industrial Design; from September 2012 to June 2015, she graduated from Hubei University of Technology with a master's degree in industrial design. She is currently teaching in the College of Art and Design, Zhengzhou University of Light Industry. Her academic research interests include product design, user experience design, and vehicle design. She has published 10 papers, participated in 10 projects, 1 textbook, 3 invention patents, and 1 utility model patent.

Diyuan Tan is currently studying Industrial Design at Zhongyuan Institute of Science and Technology. His academic research direction is user experience, interaction design and service design. He once presided over one project and published one article.

1 Introduction

Language, image and text are the three most important ways for humans to communicate with the outside world, of which language is the most important way (Urlica et al., 2022). Today, with the rise of artificial intelligence, language has also become one of the main ways of human-computer interaction. In-vehicle speech recognition has always been a key topic in the field of human-computer interaction. In the vehicle environment, human-computer interaction will be affected by various noises inevitably (Liu et al., 2021). These noises lead to large differences in the performance of speech recognition technology in practical applications and experimental environments, so speech signal enhancement and environmental noise filtering have become the main problems to be solved (Requardt et al., 2020). In addition, with the gradual promotion of electric vehicles, the market competition has become increasingly fierce. Manufacturers should not only meet the user's demand for car performance, but also meet the user's demand for automotive aesthetics. Studies had shown that excellent car exterior design and interior design could effectively improve car sales (Shanmugapriya et al., 2021). With the development of the computer industry, a large number of scholars had started to study the automatic recognition of spatial motion in combination with computer technology. At the same time, data capture technology had also been widely used in many industries and achieved outstanding results (Annamalai et al., 2020). China has a large population, and with the rapid development of the Internet users' economy and people's living standards, the number of people who own cars has surged every year and will continue to rise steadily in the future. This contains huge business opportunity. The future of electronic devices related to cars is naturally bright, and the voice control interface as an electronic device is naturally very critical. Therefore, this research has great feasibility. This research firstly combines the multi-window estimation with the commonly used spectral subtraction, adds dynamic time warping as a mathematical constraint, constructs an improved vehicle-mounted speech recognition model and proposes an analysis of visual imagery based on spatial segmentation according to the action recognition analysis methods. From the perspectives of car voice sensing and visual imagery, the voice interaction and interior design of electric vehicles were explored.

2 Related work

The rise of in-vehicle artificial intelligence poses new challenges to the accuracy of in-vehicle speech recognition, and users expect artificial intelligence to accurately recognise voice commands in most environments. With the promotion of electric vehicles, users are no longer simply satisfied with the performance of the car. The new aesthetic concept also makes designers pay attention to the interactive design of electric vehicle interiors. In recent years, many scholars at home and abroad had proposed new improvements to vehicle speech recognition methods and interior design concepts. Scholars, such as Martinek, proposed an LMS-ICA system to reduce the background noise in the car. At the same time, they used the form of virtual instruments to mix algorithms, and used the intelligent trunking of the distributed 5G data network to exchange interference information. The overall vehicle voice system under the hybrid algorithm is integrated. The recognition rate reaches 73.03% (Martinek et al., 2022). Gao (2021) constructed a BMI data fusion framework based on a speech recognition model, and provided a smooth human-computer interaction platform based on speech, which could accept voice commands to effectively analyse BIM data. To improve the accuracy of speech recognition, Dong and Li (2020) applied a speech recognition technology to English pronunciation assessment. An improved computerised speech evaluation method was proposed. By using intonation, speed and rhythm as indicators and using multidimensional indicators as evaluation criteria, a more comprehensive and objective model of English speech evaluation was established. The results show that this method was superior to many traditional speech evaluation methods, and had better recognition performance and accuracy (Philips et al., 2021). Philips et al. (2021) proposed a cortical network identification method based on dynamic time warping combined with a support vector machine. Dynamic time warping was used as a mathematical constraint to constrain the support vector machine, which effectively improved the classification efficiency and classification performance of the support vector machine. The recognition accuracy was 74.39 to 97.56% (Philips et al., 2021). Zhang et al. (2020) proposed a noise transformation algorithm based on the reverse elimination algorithm combined with spectral subtraction, which had a better transformation effect than the traditional NMF algorithm.

Wang et al. (2020) analysed the elements of the corporate visual image recognition system, such as colour, font, auxiliary graphics, standard combination, etc. Through literature research, they follow the methods of integration, fine-tuning and supplement to discuss the basic connotation of visual image design (Wang et al., 2020). Fernandes and Espino (2021) believed that proper interior design can alleviate the problem of cognitive function degradation for elderly drivers, and the design of car interiors should focus on safety, comfort and inclusiveness. Alipaker (2020) concluded through static verification and analysis that automotive interior design will have an impact on the lean product development stage, form sub variables, standards and results of comparative problems and create basic test processes and methods. Fabian and Kupec (2021) used the software CAD and CAM to update the interior and exterior shapes of the car. He believed that the shape change of the car mainly reflects the change of the line shape, including the outer line shape and the inner line shape (Fabian and Kupec, 2021). Staniszewska et al. (2020) discussed the design of automobile interiors from the perspective of ecological environmental protection. Especially for electric vehicles, it was in line with the original intention to adopt ecological design from the supply stage (Staniszewska et al., 2020). Scholars at home and abroad had improved the methods of speech recognition or speech enhancement to a certain extent, and made innovations in the evaluation methods of speech recognition. However, from the results, the recognition accuracy needed to be improved and it was not flexible enough for the improvement of spectral subtraction and the application of dynamic time warping. Therefore, this research proposes an interactive design technology for electric vehicle interiors based on voice sensing and visual imagery, to provide a reference for the design of electric vehicle voice and visual interiors.

3 The interaction design of electric vehicle sound sensing and visual imagery

3.1 Speech recognition method based on multi-window estimated spectral subtraction and dynamic time warping

The root cause of the noise is that when the noise is estimated in the 'silent segment', the estimation of the prior signal-to-noise ratio has a large variance. It can be concluded that the strength of speech enhancement depends on the accuracy of the noise power spectrum estimation (Khan and Pierre, 2020; Nicolson and Paliwal, 2020). Since the multi-window estimation method has the characteristics of high resolution and low variance, it is combined with the common spectral subtraction method to estimate the noise signal power spectrum, and its definition is shown in the equation (1).

$$S_k^{mt}(f) = \frac{1}{L}\sum_{k=1}^{L} S_k^{mt}(f) \tag{1}$$

In equation (1), L is the number of mutually orthogonal data windows, and $S_k^{mt}(f)$ is the direct spectral function added to the first window k in the data sequence, as shown in equation (2).

$$S_k^{mt}(f) = \left| \sum_{t=1}^{N} h_k(t) s(t) \right|^2 \tag{2}$$

In equation (2), $s(t)$ is the data sequence whose length is 0. N is the data sequence $\{s(1), s(2), ..., s(N), t = 1, 2, ..., N\}$ of $h_k(t)$. k is the first data window. Since there is no corresponding inverse transform for multi-windows, the speech signal cannot be directly re-established from the windows. Instead, the spectral subtraction gain with small variance is obtained by using the characteristics of small variance of multiple windows, so as to achieve the purpose of noise elimination. The input noise file is processed into frames, the overlap rate is fixed at 50%, and the power spectrum of the framed signal is estimated by using multiple windows. Let the power spectrum estimation of the speech signal with noise be $P_y(\omega)$. And in the initial stage, there is only noise in the speech signal file, so the average value of the initial stage is taken as the noise power spectrum estimation and the spectral gain is shown in equation (3).

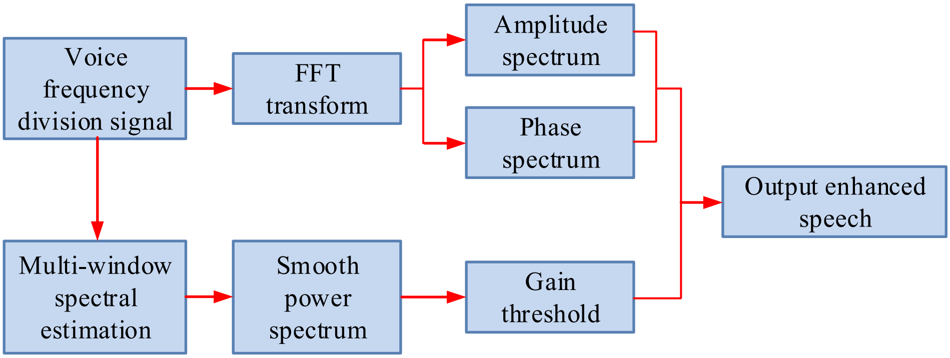
$$g(\omega) = \frac{P_y(\omega) - aP_n(\omega)}{P_y(\omega)} \tag{3}$$

In equation (3), $g(\omega)$ is the spectral gain, $P_n(\omega)$ is the noise power spectrum estimation. a is the over-reduction factor, which is used to remove the peak-value generated by the estimated variance and reduce the noise. The setting of a must be appropriate. If it is too low, the effect cannot be produced. If it is too high, the voice will be distorted. In addition, it is also necessary to set a threshold value for the spectrum gain. By assigning gain elements smaller than the threshold, the peak-to-trough variation in the spectrum estimate is moderated. The gain threshold is shown in equation (4).

$$g(\omega) = \begin{cases} g(\omega), g(\omega) > \dfrac{bP_n(\omega)}{P_y(\omega)} \\ \dfrac{bP_n(\omega)}{P_y(\omega)}, \text{others} \end{cases} \tag{4}$$

In equation (4), b is a constant. Based on previous research results, the value of this study is 0.1; $\dfrac{bP_n(\omega)}{P_y(\omega)}$ is the gain threshold (Huang et al., 2020). Assuming that the amplitude spectrum before enhancement is $|y(\omega)|$; the amplitude spectrum after enhancement is $|x(\omega)|$; and the phase spectrum is $\theta(\omega)$; the inverse FFT transform is performed. And the signal is restored to the time domain to obtain enhanced speech. The improved spectral subtraction is shown in Figure 1.

Figure 1 Improved speech enhancement algorithm flow



The main flow of the improved spectral subtraction is: (1) Make speech signal framing. (2) Perform FFT transformation, and calculate the amplitude spectrum and phase spectrum. (3) Multi window spectrum estimation: noise power spectrum and noisy speech power spectrum are estimated by smoothing power spectrum, and then spectrum gain is calculated. (4) Finally, combine the second step and the third step, the enhanced voice is obtained. It should be noted that since residual noise or possible signal distortion will affect the speech signal, a gain function $G_i(k)$ needs to be set, as shown in equation (5).

$$G_i(k) = \frac{|P_{yi}(k)| - a|P_{di}(k)|}{|P_{yi}(k)|}, k = 1, 2, ..., N-1 \tag{5}$$

In equation (5), i is the serial number of the frame, $P_{di}(k)$ is the power spectrum estimation of noise and $P_{yi}(k)$ is the power spectrum estimation with noise, as shown in equation (6).

$$P_{yi}(k) = P_{si}(k) + P_{di}(k), k = 0, 1, ..., N-1 \tag{6}$$

In equation (6), $P_{si}(k)$ is the power spectrum estimation of the pure speech signal. For the setting of the over-reduction factor, $a \in [1, 7]$ is usually taken. Set the gain compensation threshold as T, when the gain coefficient is smaller than this threshold, a new gain coefficient needs to be taken to achieve a better denoising effect, as shown in equation (7).

$$T = \frac{bP_{di}(k)}{P_{yi}(k)}, i = 1, 2, \cdots \tag{7}$$

Since the human ear is not sensitive to the phase of speech, when reconstructing the speech signal, it is necessary to obtain the phase directly from the original speech signal with noise. The enhanced speech signal $S(n)$ is shown in equation (8).

$$S(n) = IFFT\left[|S_i(k)| \cdot \exp\left(i \arg Y_i(k)\right)\right] \tag{8}$$

In equation (8), $IFFT$ is the inverse fast Fourier transform, n is the time sample point and $\arg\left(Y_i\left(k\right)\right)$ is the phase spectrum of the noisy speech signal. In reality, the speech signal has great randomness, and the time consumed by the same person to speak the same sentence at different times is also different. Therefore, it is necessary to time-regulate the speech input. Let the feature vector sequence of the reference template be $a_1, a_2, \ldots, a_M$, and let the input speech feature vector be $b_1, b_2, \ldots, b_N$, which needs to be satisfied $M \neq N$. The purpose of dynamic time-warping is to find the time-warping function. Let the time axis of the input template be n, and let the time axis of non-linear mapping to the reference template be m, then the function w is shown in equation (9).

$$D = \min_{w(n)} \sum_{n=1}^{N} d\left[n, w\left(n\right)\right] \tag{9}$$

In equation (9), $d\left[n, w\left(n\right)\right]$ is the distance between the n frame input vector and the m is the frame reference vector. D is the distance measured between the two templates under the optimal time warping. Dynamic warping is an optimisation algorithm whose principle is to transform an N-stage decision-making process into N single-stage processes to simplify complex calculations. Its boundary conditions and continuity conditions are shown in equation (10).

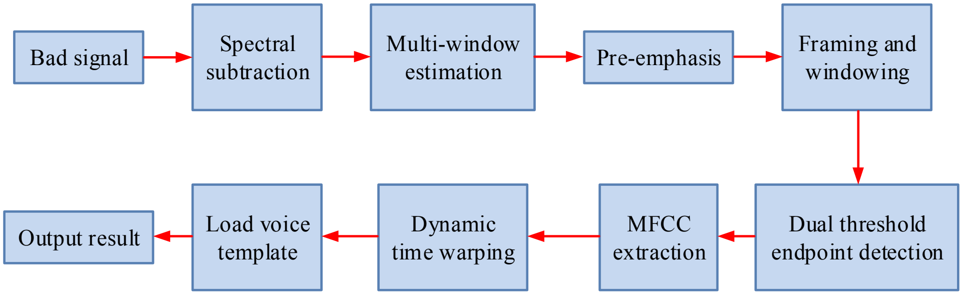
$$\begin{aligned} & w\left(1\right) = 1, w\left(N\right) = M \\ & w\left(n+1\right) - w\left(n\right) = \begin{cases} 0,1,2 \; w\left(n\right) \neq w\left(n-1\right) \\ \;\; 1,2 \; w\left(n\right) = w\left(n-1\right) \end{cases} \end{aligned} \tag{10}$$

In equation (10), the boundary condition and the continuous condition are the constraints of the regularisation function $w\left(n\right)$ in the actual problem. Let $d\left[n, m\right] = d\left[n, w\left(n\right)\right]$ be the distance between the frame vector b_n and the sum a_m. And the minimum cumulative distance is shown in equation (11).

$$D\left(n, m\right) = \min_{w(j)} \sum_{j=1}^{n} d\left[j, w\left(j\right)\right] \tag{11}$$

According to the general principle of dynamic time warping, consider starting from the last stage of the whole process and so on, and advance to the starting point one by one. The composition of the entire vehicle voice recognition system is shown in Figure 2.

Figure 2 The flow of the vehicle voice recognition system



In Figure 2, the system is mainly divided into two parts, speech enhancement, and speech recognition. First of all, it is necessary to enhance the vehicle's voice signal. The method used is the multi-window estimation spectrum subtraction. After enhancement, the recognition stage is entered. At this time, dynamic time warping and common voice templates need to be introduced. So far, the construction of the vehicle voice interactive system is completed.

3.2 Visual image analysis method based on spatial segmentation

BVH 3D data file is adopted, and the reading mode is a line. The data module is divided into the hierarchical structure of the target bone and the rotation of the bone joint action from top to bottom, and the format is ASCII data coding (Nambiar et al., 2022). The data includes two parts. One is by using json sentences to store the information on the joint points of the human body model. The root node is the origin of the object coordinate system, and the coordinates are expressed as (0, 0, 0). Other nodes are used as extensions of the root node. Therefore, the action trajectory of the node can be represented by the stored displacement offset relative to the previous node. The root node has six data volumes, and the other nodes have only *X*, *Y* and *Z* variables. The second is the storage of the entire sequence data. It is started with MOTION, following by the variable frames. The value is the frame number of the action sequence, the third line represents the collection interval of the two frames of data. Next is the representation of the action sequence. With 66 data per line, the data is stored as Euler angles compared to the previous node. Since the stored data is based on the action trajectory of the previous node, the original data is pre-processed first and then the identification and analysis are performed.

The pre-processing of data needs to convert the 3D information of the original data to the coordinates of the world coordinate axis through the rotation matrix. Among them, 66 columns of data represent the coordinates of 21 joint points of the human body. Each node uses three-dimensional information for storage. The extra 3 columns of data refer to the coordinates of the centre point of the human body. The coordinates of any node are shown in the equation (12).

$$\begin{pmatrix} x_c \\ y_c \\ z_c \end{pmatrix} = R_r \cdot \left(R_{r-i} \cdot \left(\cdots \left(R_2 \cdot \left(R_1 \cdot \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} \right) \right) \right) \right) \tag{12}$$

The specific process includes three steps. The first is to read the data according to the behaviour standard and judge the motion data block part. Second, the data is stored in row units and separated by spaces. The third is to transform the original data into the world coordinate axis through coordinate transformation. At the same time, the data is stored in row units (Silpachai, 2020). During the action, the velocity of the joint will appear at a minimum value when the movement is transformed. By using a series of minimum values to segment the segment, the action image can be recognised. For the upper limbs, the study uses Gaussian Mixture Segmentation (GMM) for data segmentation, and its probability density function is expressed as shown in equation (13).

$$p(x) = \sum_{k=1}^{K} p(k) p\left(x \middle| k\right) = \sum_{k}^{K} \pi_k N\left(x \middle| \mu_k, \Sigma_k\right) \tag{13}$$

In equation (13), the coefficient π_k represents the probability of Gaussian distribution, the value is greater than or equal to 0. And the probability density function of Gaussian distribution is $\left(x|\mu_k, \Sigma_k\right)$, then equation (14) is obtained.

$$\left(x|\mu_k, \Sigma_k\right) = \frac{1}{\sqrt{2\pi \Sigma_k}} \exp\left(-\frac{\left(x-\mu_k\right)^2}{2\Sigma_k}\right) \tag{14}$$

In equation (14), Σ_k and μ_k are the variance and mean of the Gaussian distribution, respectively. In the process of using GMM model, it is needed to set the number of *K*, and calculate the number of wave crests that each random speed threshold passes through according to the speed curve, regardless of positive and negative directions. The main direction is calculated according to the normal of the plane. The specific process is divided into two steps. One is to calculate the intersection vector sum $\overline{n}_2$ of the three extracted nodes 1, 2 and 3. The coordinates $\overline{n}_1$ of the three nodes are $\left(x_1, y_1, z_1\right)$, $\left(x_2, y_2, z_2\right)$ and $\left(x_3, y_3, z_3\right)$.The two calculated equations are:

$$\overline{n}_1 = \left(x_1 - x_2, y_1 - y_2, z_1 - z_2\right) \tag{15}$$

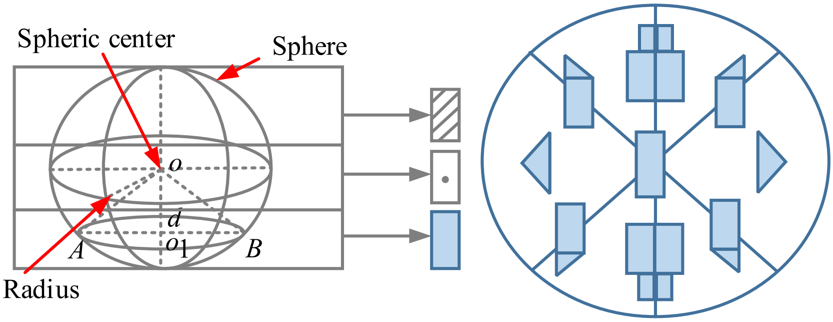
$$\overline{n}_2 = \left(x_3 - x_2, y_3 - y_2, z_3 - z_2\right) \tag{16}$$

According to the right-hand rule, the calculation equation of the main direction is:

$$\overline{n} = \overline{n}_1 * \overline{n}_2 \tag{17}$$

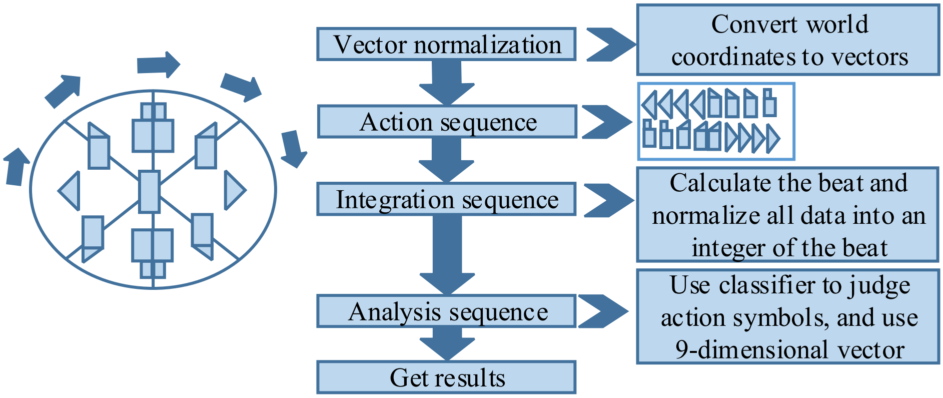
From equations (15) to (17), the vector direction is the *z*-axis direction in the object coordinate system; the centre of the three nodes is the centre of the coordinate axis; the line parallel to the node is the *X*-axis. The straight line perpendicular to the plane *Y*-axis. Then we construct the right-hand coordinate, and each joint point obtains the object coordinate through coordinate transformation. It should divide the space according to the rules of the sphere, and get 27 subspaces. The horizontal and vertical directions are 9 and 3 spaces, respectively. Then we record according to the spatial position of the joint points. The schematic diagram is shown in Figure 3.

Figure 3 Schematic diagram of action space identification



Action identification symbols are determined according to the spatial rules, and the vertical and horizontal determinations are carried out, respectively. The vertical determination adopts the foot node J_{y1} and J_{y2}. They refer to the coordinates of the first node of an action and the last vertical of the foot. When the interpolation is less than 10, it refers to the low position, greater than 20 is the high position and between 10 and 20 is the middle position. The horizontal direction is determined by action horizontal space cutting rule, and reach the goal of determining the angle between the *x*- and *z*-axes. The spatial segmentation algorithm proposed in this study is shown in Figure 4. The first step is to normalise the world coordinate vector, the second step is to identify each frame of score data as a dense series of action symbols according to the spatial law, and the third step is to use the beat and other judgment criteria to integrate the action sequence symbols, and finally. The classification is used to optimise the results.

Figure 4 Flow of space segmentation algorithm


Vector normalization
Convert world coordinates to vectors
Action sequence
Integration sequence
Calculate the beat and normalize all data into an integer of the beat
Analysis sequence
Use classifier to judge action symbols, and use 9-dimensional vector
Get results

With the help of the vector angle in the vertical direction, first determine the action symbol of the action vector $\overline{x}, \overline{y}, \overline{z}$ of each frame, $\overline{d}_1$ refers to the three-dimensional coordinates of the object coordinate system. A plane can be formed between any two. The projection vectors of the vector $\overline{d}_1$ on the three planes are $\overline{d}_x$, $\overline{d}_y$ and $\overline{d}_z$. The calculation equation is:

$$\theta_{x,y,z} = \arccos\left(\frac{\overline{x}, \overline{y}, \overline{z} \bullet \overline{d}_{x,y,z}}{\left|\overline{x}, \overline{y}, \overline{z}\right|\left|\overline{d}_{x,y,z}\right|}\right) \tag{18}$$

The action symbols in the horizontal direction can be determined by calculating the angle between the action vector and the *z*-axis and the *x*-axis. The action symbols in the vertical direction can be determined by calculating the angle between the action vector and the *y*-axis. And the dense action sequences with the basic characteristics of the action can be obtained. The beat method is used to integrate the action sequence, determine the minimum beat, obtain the high-frequency signal using the wavelet change function and calculate the action components. The minimum beat is expressed as 1, and the action sequence is represented by different numbers such as 1, 2, 3. The array with the sub array length of J is classified for statistics, and finally the symbol sequence with the length of J is obtained, and the adjacent sequences with the same symbol are normalised. The action sequence obtained after the above three steps is still not the final sequence. When performing action recording, only four space symbols are needed: in-situ low, front-middle, in-situ high and rear-middle. At the same time, it is necessary to introduce two classifications when performing action recording. The controller judges whether the sign is preserved through the positive and negative samples. The research uses a Support Vector Machine (SVM) to separate the positive and negative samples through the hyperplane, and the objective function expression is shown in the equation (19).

$$\arg\ \max\left\{\min\left(lab\left(w^{T}x+b\right)\right)\cdot\frac{1}{\|w\|}\right\} \tag{19}$$

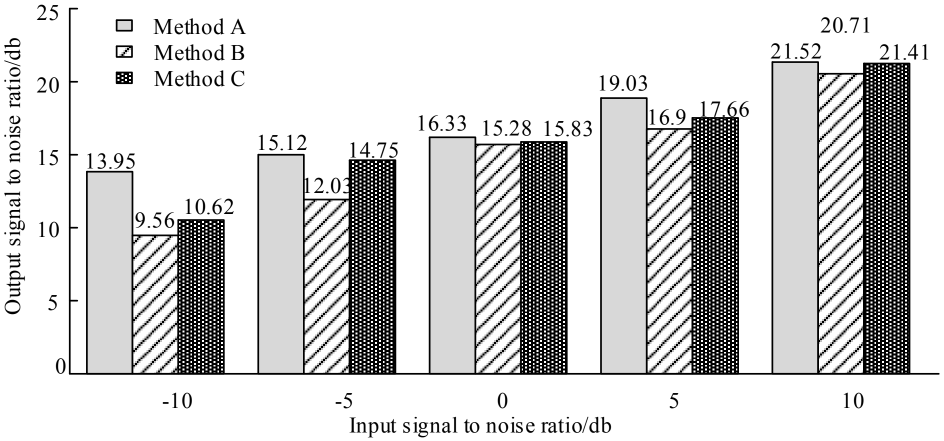
The sample category is represented by lab, the hyperplane equation is represented by $w^{T}x+b$ and the result is obtained after continuous iteration. Feature selection directly determines the accuracy of the SVM classifier. Since the minimum beat lengths of different actions are inconsistent, 9-dimensional velocity data is selected as the feature of the classifier, which is obtained by the three data components of the data to be detected in three directions. Finally, the action symbols are input into the classifier to get the final result.

4 Analysis of the results of denoising and speech recognition ablation and visual imagery methods

To verify the denoising effect of the improved multi-window estimation spectral subtraction method, a denoising experiment was carried out. The datasets of ‘531 hours of vehicle noise collected by a microphone and mobile phone’ and ‘245 hours of voice collected by a mandarin phone in a vehicle environment’ developed by Datatang were used as a test set. The data set covers almost all in-car environments such as different weather, different operating environments, window opening or closing, air conditioning opening or closing, music opening or closing and different speeds. The layer is used as the test set of the Apollo autonomous driving software developed by Baidu. Tests were carried out under the conditions of signal-to-noise ratios of 0 db, –5 db, –10 db, 5 db and 10 db, respectively. The comparison methods are spectral subtraction, multi-window spectral subtraction without dynamic time warping, and traditional speech recognition methods, which are recorded as method A, method B and method C. The results are shown in Figure 4.

In Figure 4, under five different input SNRs, the output SNR of method A is higher than method B and method C. When the input signal-to-noise ratio is 0db, the output signal-to-noise ratios of the three methods are close, and the denoising effect of method A is only 6.43% and 3.06% higher than method B and method C. At 5db, the improvement effect of method A is obvious, which is 31.47%, 23.87%, 20.44% and 2.45%, respectively. It can be seen that the dynamic time warping module does not significantly improve the denoising effect, while the multi window estimation module significantly improves the spectral subtraction. To further verify the effect of method A, the speech recognition effects of different methods are tested. It is Considered that the fundamental frequency of male vocalisation is between 60 Hz and 200 Hz, and the peak is around 125 Hz. While the fundamental frequency of female vocalisation is between 150 Hz and 500 Hz, and the peak is around 300 Hz. There is a big difference between men and women (Huang et al., 2020). To ensure the validity of the experiment, a set of pronunciation templates are designed for males and females. In the laboratory environment, 50 commonly used car words such as switch window, wiper, week, play, close, next song, heating and cooling were recorded. For the convenience of calculation, 100 subjects with healthy vocal cords, 50 men and 50 women, were selected. All of them could speak standard Mandarin, and some of them had unobvious accents. These subjects should imitate the actual environment as much as possible, and the input signal-to-noise ratio during the test is still –10 db, –5 db, 0 db, 5 db, 10 db.

Figure 5 Comparison of noise removal effect of different methods



In Table 1, there is a slight difference in the recognition accuracy of male and female templates, and the overall trend is similar. The recognition accuracy of method A is higher than method B and method C under the five signal-to-noise ratios. Since the signal-to-noise ratio of vehicle-mounted noise is usually between 5 db and 10 db in practice, it can be considered that the recognition accuracy of method A ranges from 92.3 to 98.7%.

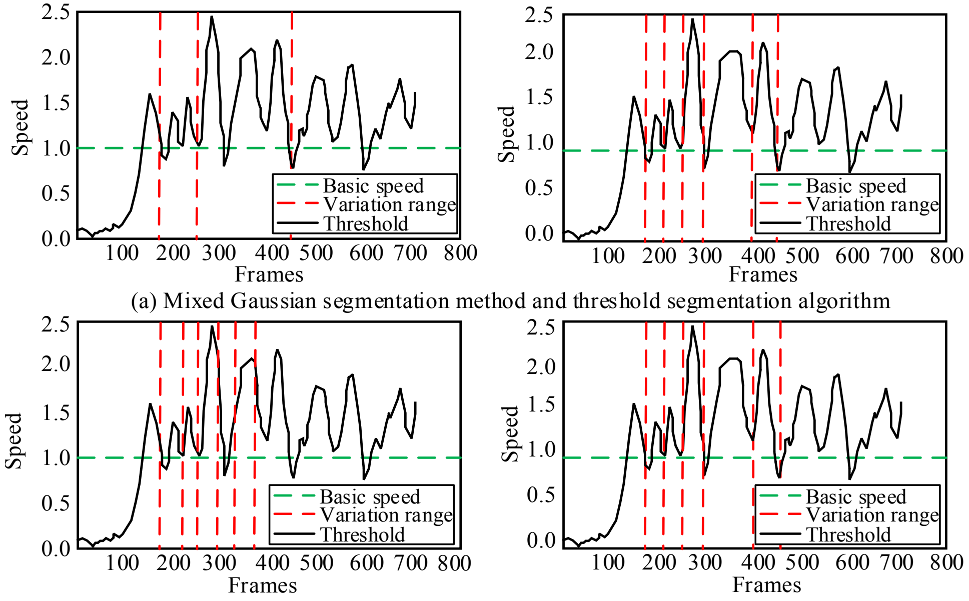
Table 1 Speech recognition accuracy of different methods

Signal-to-noise ratio		*–10 db*	*–5 db*	*0 db*	*5 db*	*10 db*
Male template recognition accuracy	Method A	79.90%	84.60%	90.10%	94.50%	98.70%
	Method B	71.4%	79.70%	85.10%	89.50%	95.20%
	Method C	71.30%	78.60%	85.40%	90.10%	96.80%
Female template recognition accuracy	Method A	77.50%	82.10%	88.60%	92.30%	97.50%
	Method B	71.20%	77.90%	82.50%	85.30%	91.90%
	Method C	72.80%	79.10%	84.70%	89.10%	94.60%

The study further compares the three algorithms of threshold segmentation, beat segmentation and mixture Gaussian segmentation. Figure 6(a) shows the results of the mixture Gaussian segmentation method and the threshold segmentation algorithm.

Figure 6(b) shows the result of mixture Gaussian segmentation method and beat segmentation. As can be seen from Figure 6(a), although the calculation speed is relatively fast, many motion clips above the threshold have not been segmented and the error of manually setting the threshold has increased. The mixed Gaussian segmentation proposed on the basis of the threshold can accurately segment the motion data. On the whole, the latter is obviously better than the former. In Figure 6(b), the computer action tempo is not fixed, so the actions obtained by the tempo segmentation method are not accurate. And there will be some wrong action symbols and wrong symbol lengths. The results are better than the former.

Figure 6 Comparison of speed threshold, beat and mixed Gaussian segmentation algorithm

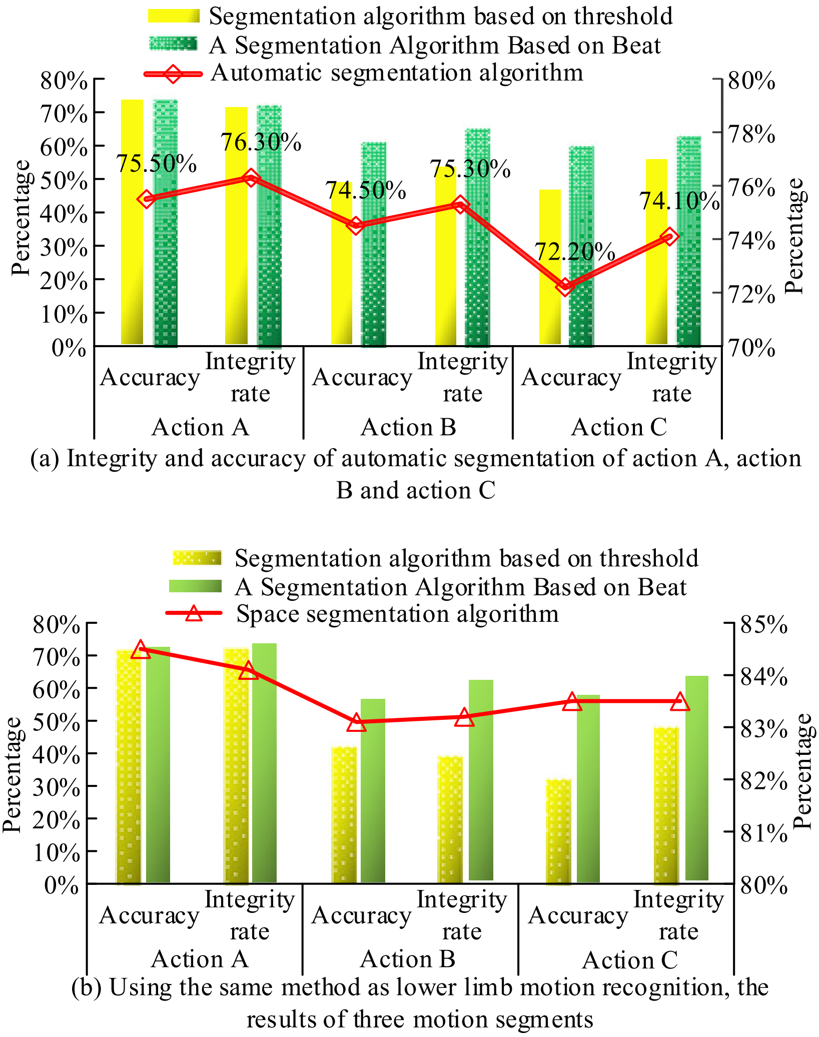


(a) Mixed Gaussian segmentation method and threshold segmentation algorithm

(b) Mixed gaussian segmentation method and beat segmentation algorithm

This experiment collects 48 action clips, each action is within 1 minute, a total of 30,000 motion data, and each action corresponds to 48 action symbols. Figure 7(a) shows the automatic segmentation of action A, action B and action C. By comparing the completeness and accuracy of the three algorithms, it is found that the segmentation accuracy and completeness are all around 75% for simple actions, and the difference is not particularly large. In the data segmentation of actual fast and slowly mixed action clips, the completeness and accuracy of the mixed Gaussian segmentation are better than the other two segmentation algorithms. In terms of action B symbol generation, the accuracy of the Gaussian mixture segmentation algorithm is about 60% higher than that of the other two algorithms. Using the same method as lower limb action recognition, the results of the three action segments are shown in Figure 7(b). In the data segmentation of action segments with mixed speed and slowness, especially the action C segment, the integrity and accuracy of spatial segmentation are obviously due to the other two segmentation algorithms. The accuracy of spatial segmentation is about 2 times higher than that of the threshold segmentation algorithm. The beat segmentation algorithm is about 60% higher.

Figure 7 Accuracy and integrity statistics of three segmentation algorithms



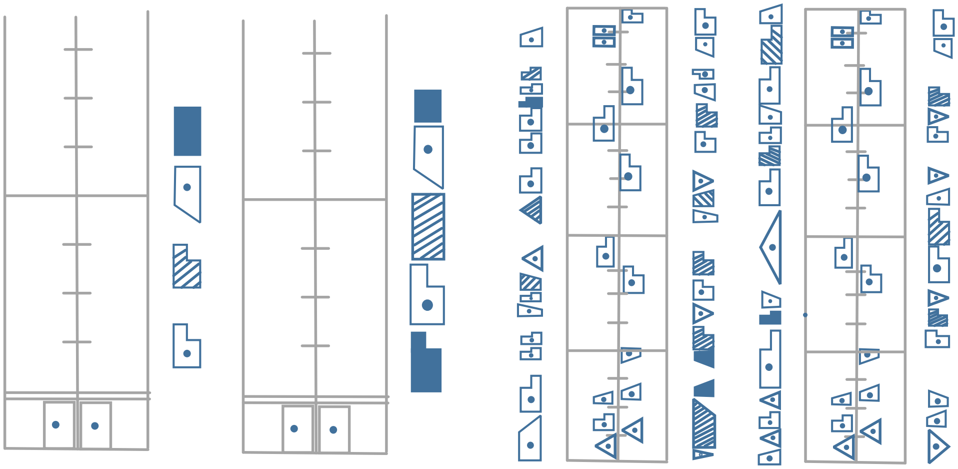
(a) Integrity and accuracy of automatic segmentation of action A, action B and action C

(b) Using the same method as lower limb motion recognition, the results of three motion segments

The images obtained by the spatial segmentation method are four frames in the action sequence: low in situ, middle in front, high in situ and middle behind. Figure 7 illustrates that the algorithm effectively avoids the complex recording of intermediate processes and incorporates them into methods such as cadence, thresholding, classifiers and spatial classification. The regularity and accuracy of the final action recognition is ensured. Figures 8(a) and 8(b) represent the results of arm motion based on the threshold segmentation method and spatial segmentation method, respectively. It can be seen from

Figure 8(a) that the threshold segmentation method obtains satisfactory results in terms of completeness and accuracy of action symbols, while the spatial segmentation method has very good results in both completeness and accuracy. *Effect*: It can be seen from Figure 8(b) that the accuracy and completeness of the spatial segmentation method are better than the beat segmentation method in the action symbols, especially in the detail part, the spatial segmentation method has a better effect.

Figure 8 Arm motion based on threshold segmentation and space segmentation

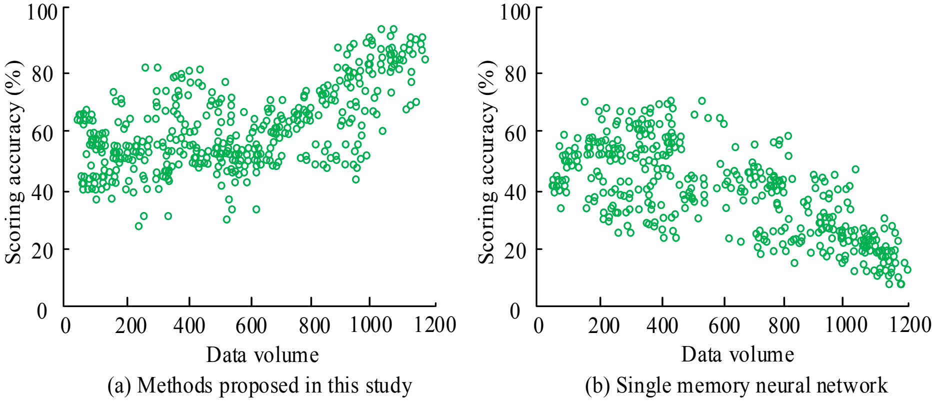


(a) Threshold segmentation method

(b) Space division method

To better analyse the application effect of the method proposed in this study, the data set of a large automobile interior design company is used as the experimental data. The optimisation algorithm proposed in this study is compared with the single method of the model. The results are shown in Figure 9. As can be seen from Figure 9(a), the algorithm proposed in this study can achieve better scoring results. And its scoring accuracy shows an upward trend with the increase of sample data, from 40.12 to 91.20% accuracy. The accuracy of its data samples is based on the upper left of the overall area, and there is a certain positive correlation between the correct scoring data and the overall sample. However, the single method in Figure 9(b) is a simple memory neural network, but the accuracy rate of sample data scoring decreases with the increase of sample data volume, from 57.74 to 14.17%. The sample values are generally located in the lower left of the overall area, reflecting a certain negative correlation. It is difficult for a single algorithm to effectively identify the degree of fit between actions when faced with a large amount of sample data, resulting in a large discrepancy between the scoring results. Therefore, the optimisation algorithm proposed in this study can effectively reduce the recognition error in interior applications, and has good application practice.

Figure 9 Comparison of interior application accuracy scores under different algorithms



(a) Methods proposed in this study

(b) Single memory neural network

5 Conclusions

To address the vehicle noise and visual recognition problems, this study proposes a design solution for electric vehicle interior interaction based on speech induction and visual images. The speech recognition method combining multi window estimated spectral subtraction and a speech recognition method combining multi-window estimated spectral subtraction and dynamic time warping is adopted, and a dynamic time warping and the automatic action recognition method based on Gaussian mixture segmentation algorithm and space segmentation algorithm is adopted. The results show that under different input SNR, the denoising ability of the method is 2.45 to 31.47% higher than that of the benchmark method. The accuracy of speech recognition in-vehicle environment is 92.3 to 98.7%. The computer action beat is not fixed, which leads to the inaccurate action obtained by the beat segmentation method. And some wrong action symbols and wrong symbol lengths will appear. However, the method proposed in this study has no limitation of action rules and is superior to other algorithm models in the results. The integrity and accuracy of action B and action C are improved by 200% compared with the speed threshold segmentation method and about 60% compared with the beat segmentation method.

The accuracy and integrity of the action symbols generated by the space segmentation method are superior to the speed threshold and beat segmentation methods. The integrity and accuracy of action B and action C are improved by 200% compared with the speed threshold segmentation method. The integrity and accuracy of action B and action C are improved about 60% compared with the beat segmentation method. The algorithm proposed in this study can achieve good scoring results, and its scoring accuracy rate shows an upward trend with the increase of sample data volume, from 40.12 to 91.20%. However, there are still some deficiencies in this study. The experiment did not consider whether the improved speech recognition method can correctly recognise the voice when the user of the electric vehicle is a dialect. The future research direction can focus on breaking through the dialect problem, which will be conducive to the development of electric vehicle interior design.

References

Alipaker, F. (2020) 'The 'static' and 'dynamic' design verification stages of the lean development process: automotive industry', *World Engineering and Technology*, Vol. 8, No. 1, pp.74–91.

Annamalai, S., Nagarajan, B. and Kumar, H.V. et al. (2020) 'Process and analysis with demographic methodological refinement of bus body industry', *Materials Today: Proceedings*, Vol. 33, No. 7, pp.3549–3557.

Dong, J. and Li, S. (2020) 'English speech recognition and multidimensional pronunciation evaluation', *Frontier of Education Research*, Vol. 10, No. 3, pp.184–188.

Fabian, M. and Kupec, F. (2021) 'Use of 3D parametric models in the automotive component design process', *Advances in Science and Technology – Research Journal*, Vol. 15, No. 1, pp.255–264.

Fernandes, S. and Espino, Y.C. (2021) 'Neuropsychological aspects of aging and driving for inclusive automotive interior design', *Journal of Transportation Technologies*, Vol. 11, No. 3, pp.390–403.

Gao, Z. (2021) 'Intelligent building BIM fusion data analysis framework based on speech recognition and sustainable computing', *International Journal of Networking and Virtual Organisations*, Vol. 25, No. 1, pp.83–101.

Huang, Z., Lin, J. and Yang, H. et al. (2020) 'An algorithm based on text position correction and encoder-decoder network for text recognition in the scene image of visual sensors', *Sensors*, Vol. 20, No. 10, pp.2942–2956.

Khan, M.A. and Pierre, J.W. (2020) 'Separable estimation of ambient noise spectrum in synchrophasor measurements in the presence of forced oscillations', *IEEE Transactions on Power Systems*, Vol. 35, No. 1, pp.415–423.

Liu, Y., He, F. and Wen, J. et al. (2021) 'Visual analytics of large-scale e-government text data via simplified word cloud', *Data Science and Infometrics*, Vol. 2, No. 1, pp.29–51.

Martinek, R., Baros, J. and Jaros, R. et al. (2022) 'Hybrid in-vehicle background noise reduction for robust speech recognition: the possibilities of next generation 5G data networks', *Computers, Materials and Continua*, No. 6, pp.4659–4676.

Nambiar, A., Rubel, T. and Mccaull, J. et al. (2022) 'Dropping diversity of products of large US firms: models and measures', *PLOS ONE*, Vol. 17, No. 3, pp.1–22.

Nicolson, A. and Paliwal, K.K. (2020) 'Spectral distortion level resulting in a just-noticeable difference between an a priori signal-to-noise ratio estimate and its instantaneous case', *The Journal of the Acoustical Society of America*, Vol. 148, No. 4, pp.1879–1889.

Philips, R.T., Torrisi, S.J. and Gorka, A.X. et al. (2021) 'Dynamic time warping identifies functionally distinct fMRI resting state cortical networks specific to VTA and SNC: a proof of concept', *Cerebral Cortex*, Vol. 32, No. 6, pp.1142–1151.

Requardt, A.F., Ihme, K. and Wilbrink, M. et al. (2020) 'Towards affect-aware vehicles for increasing safety and comfort: recognising driver emotions from audio recordings in a realistic driving study', *IET Intelligent Transport Systems*, Vol. 14, No. 6, pp.1265–1277.

Shanmugapriya, P., Rajakani, V. and Parthasarathy, P. et al. (2021) 'Enhancing the noise immunity in speech signal by using combined filtering technique', *Annals of the Romanian Society for Cell Biology*, Vol. 25, No. 1, pp.5330–5340.

Silpachai, A. (2020) 'Prosodic structural and tonal contextual modulation of voice onset time and consonant-induced fundamental frequency in the three-way laryngeal contrast in Thai', *The Journal of the Acoustical Society of America*, Vol. 148, No. 4, pp.2725–2725.

Staniszewska, E., Klimecka-Tatar, D. and Obrecht, M. (2020) 'Eco-design processes in the automotive industry', *Production Engineering Archives*, Vol. 26, No. 4, pp.131–137.

Urlica, A., Kamberi, L. and Boguslawska-Tafelska, M. (2022) 'Communication and language learning in virtual environments through an eco-semiotic lens', *Book chapters-LUMEN Proceedings*, Vol. 17, No. 19, pp.182–187.

Wang, W., Lv, J. and Dai, Z. (2020) 'Exploration of visual image design in sports colleges and universities: taking the visual image design of Wuhan Sports Institute as an example', *Technium Social Sciences Journal*, Vol. 14, No. 1, pp.117–126.

Zhang, S., Jian, Z. and Sun, M. et al. (2020) 'Noise-robust voice conversion based on joint dictionary optimization', *Journal of Acoustics*, Vol. 39, No. 2, pp.259–272.