

The Relationship between Sound and Picture in Multimodal Oral History Discourse

Qingbei Kong

The School of International Education of Chinese Language, Beijing International Studies University, Beijing, China

kqbonly@sina.com

Keywords: Multimodal Oral History Short Video; Auditory Modes; Visual Modes; Sound-Picture Relationship

Abstract: Multimodal oral history discourse encompasses auditory, visual, and other modes, primarily consisting of language supplemented by imagery. The interplay between sound and picture comprises two categories: synchronization and counterpoint. Synchronization predominates in audiovisual materials and interview dialogues within short videos, with the framing, shooting angle, and color mode closely aligned with the video's theme and communication objectives. Sound-picture counterpoint includes correspondence, alignment, and vacant relationships, with correspondence being most prevalent.

With the development of science and technology, discourse forms in the traditional sense have been replaced by discourse composed of multiple modes, and the denotation of discourse has also changed. Multimodal oral history discourse is a discourse composed of auditory and visual modes, including oral history interview programs, oral history documentaries, feature films and short oral history videos. Taking short oral history videos in the context of new media as an example, this paper discusses the modes of multimodal oral history discourse and their relationship between sound and picture.

There are two distinct interpretations of the concepts of modality and multimodality in academic discourse: sensory theory and symbolic theory. The sensory theory conceptualizes modes as sensory organs or channels, encompassing the five human senses, as advocated by Gu Yueguo et al. On the other hand, symbolic theory views modes as symbolic forms, with language being considered as social symbols according to Halliday (1978). The term "multimodality" was introduced by the Sydney School in the 1990s to extend Halliday's social semiotic perspective on language to other sign systems and applied linguistics. This led to the development of multimodal stylistics,

multimodal social semiotics, and multimodal social interaction analysis. In this paper, we align with the sensory theory that defines modes as sensory channels. Currently, human cognition encompasses only five modes: visual modality,auditive modality,tactile modality,olfactory modality and gustatory modality;where symbol is a lower-level concept within each mode covering multiple symbols. We consider interactions involving two or more senses as constituting multimodality.

Multimodal discourse research has become an important research field in the field of discourse analysis. Huang Lihe and Zhang Delu (2019) propose that geographical semiotics and multimodal ethnography belong to the application of multimodal paradigms in other fields. Norris (2020) introduced the latest research progress in the field of multimodal discourse analysis, providing beginners and relevant researchers with research methods on multimodal, especially modal interaction, under the framework of social semiotics. Sandler (2022) pointed out that the term "multimodality" itself has ambiguity (transmission channel theory and sign theory). Based on empirical research, he believed that "multimodality" should refer to the coexistence of linguistic and gesture modes. Oittinen (2022) combines theoretical concepts with empirical results from multimodal linguistics and multimedia learning to point out the possibility of introducing appropriate forms of action or intervention into the teaching of multimodal design projects. John Flowerdew &John E Richardson (2022) proposed a study on multimodal criticism of symbolic software from three dimensions of discourse, discourse practice and social practice, and took PowerPoint as an example. Explain how to reveal the role of software in the marketization of public discourse through these three dimensions.

The above research paths, methods and fields reflect the multidisciplinary and interdisciplinary nature of multimodal discourse research. Multimodal discourse research is a multi-perspective, multi-level, interdisciplinary and cross-field comprehensive research, and scholars have made a lot of achievements. Although the academic world has different definitions and research perspectives on multimodality, the relationship between modes and how to construct new textual meaning have always been the focus of common attention of researchers.

In recent years, the research results of multimodal discourse in China have gradually increased. These studies use discourse analysis theory to discuss the aspects of discourse cohesion, multimodal meaning construction and pragmatic strategies of documentary. Zhao Haiyan and Wang Zhenhua (2022) proposed an analytical framework for the intersymbol construction of lawyer's identity to analyze how lawyers use language and gestural paralinguistics to construct identity.

The movieland has long paid attention to the relationship between visual and auditory modes. Encyclopedia of China - Film Volume (1991) divides the combination of sound and picture into sound and picture synchronization, sound and picture separation and picture counterpoint.

In Encyclopedia of China (3rd Edition) · Film and Television Volume (online edition), the combination of sound and picture is divided into sound and picture synchronization and harmony counterpoint, and sound and picture parallel and harmony opposition.

The definition of sound and picture synchronization is basically the same in both versions. According to the definitions of separation of sound and painting, counterpoint of sound and painting, parallel of sound and painting, and separation of sound and painting, the separation of sound and painting in the 1991 edition of the Encyclopedia of China is similar to the parallel of sound and painting in the online edition, and the counterpoint of sound and painting in the 1991 edition of the Encyclopedia of China is similar to the opposition of sound and painting in the online edition.

The relationship between sound and picture has also been discussed in film and television circles. Wei Lu (2020) analyzed the effect of the relationship between sound and painting on the creation of artistic conception in films. This paper takes short oral history videos as the corpus, combines the theory of multimodal discourse analysis and the theory of sound and picture relations in domestic film and television circles, and explores the sound and picture structural relations of multimodal

oral history discourse.

1. Corpus source and annotation

The corpus of this paper is multimodal, including target corpus and reference corpus.

The target corpus is oral history short video. Oral history short video is a short video of oral history subject that is less than 10 minutes in length and broadcast through new media clients such as the Internet, public accounts and we-media, and watched by users through mobile terminals such as mobile phones and tablet computers. Target language materials are selected from major websites, clients and public accounts, involving politics, economy, people's livelihood and other fields, mainly including "Family Oral History", "1911 Revolution", "Oral Beihang", "Tianjian oral history" and other oral history video materials of different periods and different themes. By random sampling, we randomly selected 50 video samples from the above websites, clients and public accounts, and built a multi-modal corpus of "Oral History Short videos" as the target corpus, with a total duration of about 210 minutes and a total word count of about 45,000 words of manually transcribed video text.

Reference material selected from Li Ziqi series of short videos. We randomly selected 30 short video samples from the "Li Ziqi Gu Xiang Gu Shi" series of Tencent video website, and built a multimodal corpus of "Li Ziqi Traditional Craft Short Video" as a reference corpus, with a total length of about 68 minutes and a total of 649 words of manually translated video text. Although the oral history short video and Li Ziqi's traditional craft short video belong to the same category of narrative discourse, the oral history short video takes the language symbol in the auditory mode as the main narrative means, while the traditional craft short video takes the picture symbol in the visual mode as the main narrative means. Therefore, we choose the short video series of Plum Seven as the reference material. By comparing them in terms of mode selection and sound and picture relationship, we reveal the characteristics of the modal relationship and sound and picture relationship in multimodal oral history discourse.

Oral history short video is a multimodal discourse constructed by both visual and audio modes. The audio modes of oral history short video include spoken language (hereinafter referred to as language), music and sound. Among them, vocal language includes explanatory discourse, interviewee discourse and interviewer discourse. The visual modes of oral history short video include pictures, charts, labels, captions and video images. Among them, the image screen includes the interviewer screen, the interviewee screen, the scene reproduction screen, the location screen, the interlude screen, the opening and the end of the film. As shown in Figure 1:

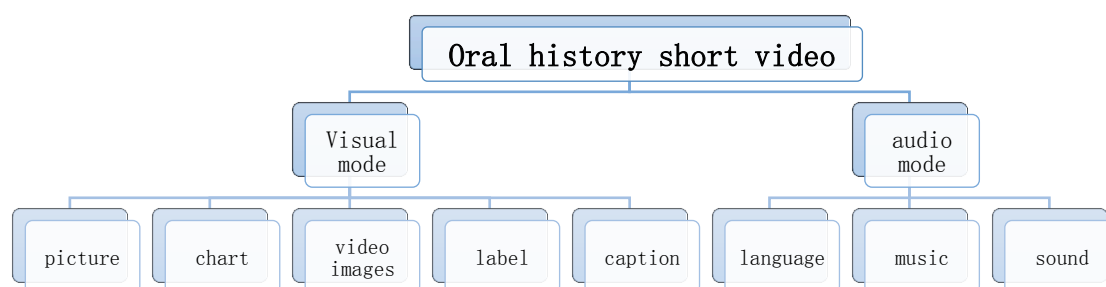


Figure 1 Audiovisual modes and symbols of oral history short video

We use Elan6.2 to build our own annotation template, marking 5 categories of audio-visual symbols, including language, picture, chart, logo and screen, as shown in Figure 2 (music, audio

and subtitle symbols are not involved for the time being) :

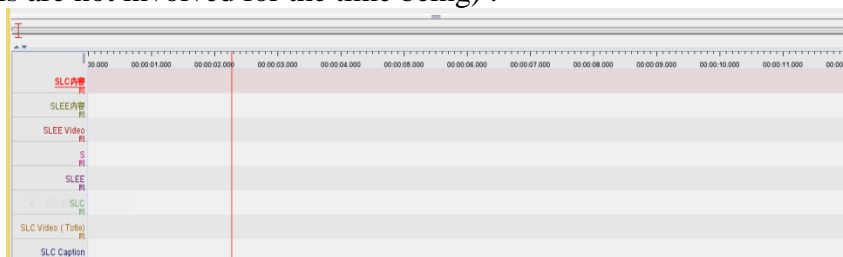


FIG.2 Schematic diagram of Elan6.2 marking audiovisual symbols

We built a target corpus (referred to as target database, labeled MBK) and a reference corpus (referred to as contrast database, labeled DBK), and used a combination of machine annotation and manual annotation to label and extract the modes and modal relationships of the corpus.

When conducting multi-modal data statistics, we used the multi-file processing function in Elan6.2 to conduct hierarchical statistics on the files in the self-built multi-modal corpus, and the results were shown in Figure 3:

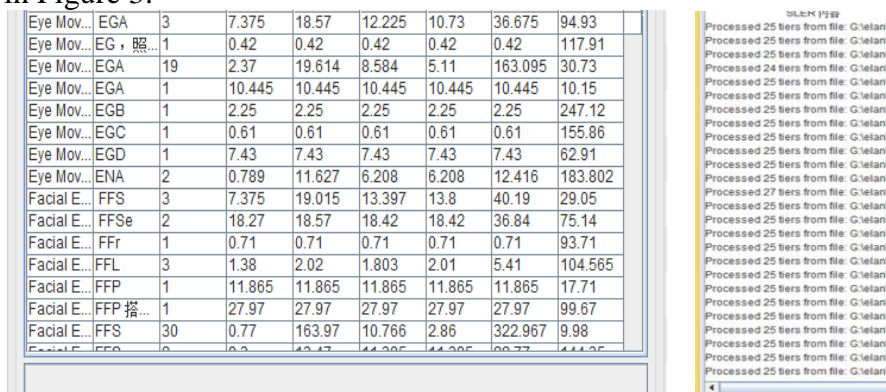


Figure 3. Statistical diagram of multimodal file annotation (excerpt)

Due to the differences in sample size and duration of a single sample between the target corpus and the reference corpus, the statistics and analysis of the original modal counts could not accurately reflect the characteristics and rules of the multimodal sound-picture relationship in short oral history videos. Therefore, we standardized the frequency of the modes/symbols in the samples to obtain their standardized frequencies. The frequency obtained based on the unified benchmark is the standardized frequency, and the calculation formula is as follows:

Modal/symbol normalization frequency = (original count of modes/symbols/text duration) * text specification

This formula is applicable to the standardized frequency calculation of the use frequency of audiovisual symbols such as pictures, pictures, signs, languages, etc. The original count refers to the actual number of times that a certain audiovisual symbol appears in a discourse, and the discourse quantity is 100 minutes.

The data analysis in this paper is carried out on the basis of standardized frequency. In the specific operation process, we use AntConc3.5.8, SPSS24.0 and other software tools to organize and analyze the data, describe the corresponding statistical measures as comprehensively as possible, and analyze and interpret the statistical results on this basis.

The statistical methods used in this paper include sampling, descriptive statistics and non-parametric test. Since the data in this paper do not conform to the normal distribution, we use

two independent samples Mann-Whitney U test to examines whether there are significant differences in modal distribution and sound and picture combination between oral history short video discourse and Li Ziqi's traditional craft short video discourse. The significance level P of the test was 0.05.

2.The sound and picture of multimodal oral history discourse

The "sound" in the audio-visual relationship is equivalent to the auditory mode in the theory of multimodal discourse analysis, and the "picture" is equivalent to the visual mode. Therefore, the research on the audio-visual mode relationship in the multimodal oral history discourse is essentially the study of its audio-visual mode relationship.

The study of the relationship between sound and picture in multimodal oral history texts needs to be based on the central content expressed by sound (mainly language) and picture and their combined relationship.

2.1Sound in multimodal oral history discourses

The sounds of multimodal oral history texts are auditory modes, including language signs, music signs and sound signs.

Language symbols refer to the discourse of both explanation and interview. In general, multimodal oral history discourse adopts the mode of "explanatory discourse + interviewee discourse", which belongs to interpretive oral history discourse.

Music symbol refers to a kind of art form of sound flow, composed of music symbol and song symbol, this paper will not make a detailed distinction. Acoustic symbols refer to sounds other than language and music in auditory modes, simulating or reproducing sounds in real life. (Zhao Yuming and Wang Fushun, 1999) In multimodal oral history discourse, language symbols are used to narrate and convey new information to achieve the purpose of discourse communication and assume the main communicative function, while music and audio symbols are used to render atmosphere, set off context, deepen the theme, and connect the lens or the conversation wheel. Therefore, in the auditory mode, the language symbols are in the foreground, and the music and sound symbols are in the background. Only linguistic symbols are discussed.

2.2 Pictures in multimodal oral history discourse

The images of multimodal oral history texts are visual modes, including pictures, charts, images, captions, logos and other symbols.

We conducted statistics and analysis on the modal intensity of various visual symbols in the multimodal oral history discourse, and compared it with the short video of Li Ziqi's traditional craft, as shown in FIG. 4:

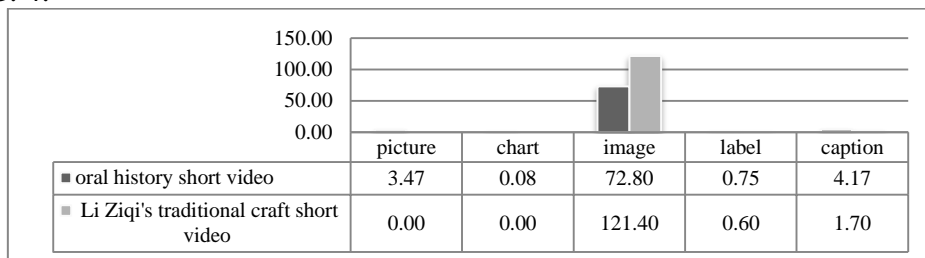


Figure 4. Visual modal symbol intensity comparison between two types of short video discourse

As can be seen from Figure 4, there are differences in the types and modal intensity of visual symbols in these two types of short videos. In the multimodal oral history discourse, there are five kinds of visual symbols, among which the modal intensity of image symbol is the highest, that of chart symbol is the lowest, and the modal intensity of subtitle symbol, picture symbol and identification symbol is between image symbol and chart symbol. In the short video referring to the traditional craft of Liziqi, there are no pictures and charts, only image symbols, subtitle symbols and identification symbols on the screen, and the modal intensity of image symbols is much higher than that of subtitle symbols and identification symbols.

The sounds and pictures in the multimodal oral history discourse are not independent, but combined to tell historical events to the audience. As mentioned above, the auditory mode in the theory of multimodal discourse analysis is equivalent to "sound" in the theory of sound and picture relations, and the visual mode is equivalent to "picture". The combination of audiovisual modes in multimodal oral history texts is the combination of sound and picture. Starting from the most basic combination of sound and picture, this paper makes a statistical analysis of the relationship between sound and picture in multimodal oral history discourse with the help of ELAN, SPSS and other software.

3.The combination of sound and picture in multimodal oral history discourse

Although sounds and pictures in multimodal oral history texts belong to different symbolic systems in nature, they both have the function of transmitting information. The former propagates through auditory modes, while the latter propagates through visual modes.

However, sound and picture in multimodal oral history texts play different roles in conveying information. In multimodal oral history texts, information is mainly transmitted by sound, and language symbols carry the main narrative function. The information of traditional craft short video is mainly transmitted by the picture, the picture bears the main narrative function, and the language symbols and music and sound symbols in the auditory mode are the auxiliary and supplementary to the picture. Therefore, the multimodal oral history discourse is based on language and supplemented by pictures. Sound and picture are basically master-slave relationships.

Drawing on the research results of the relationship between sound and picture in the field of film and television, we divide the relationship between sound and picture in multimodal oral history texts into two categories: the relationship between sound and picture synchronization and the relationship between sound and picture counterpoint. Sound and picture synchronization refers to a sound editing and processing method in which sound and picture content appear and disappear simultaneously in multimodal oral history discourse. The relationship between sound and picture synchronization generally appears in audiovisual materials of interviews or interjected broadcasts. Sound and picture counterpoint means that sound and picture are independent and look at each other, explaining the connotation of the same thing from different aspects. This kind of sound and picture relationship strengthens the internal connection between sound and picture image, and is more appealing. Acoustographic counterpoint can be divided into three sub-categories: correspondence relation, alignment relation and vacancy relation. Among them, correspondence is divided into single correspondence and multiple correspondence.

We have marked and counted the types and quantities of audio-picture relationships in multimodal oral history texts and Li Ziqi's short video of traditional craft, and the specific results are shown in Table 1:

Table 1 Numerical statistics of the phonographic relationship between two types of multimodal discourse

Sound-picture relationship		Proportion of relational subclasses		The proportion of the sound-picture relationship	
		Multimodal oral historical discourse	Li Ziqi traditional craft short video	Multimodal oral historical discourse	Li Ziqi traditional craft short video
Sound and picture synchronization	language and picture synchronization	94.85%	1.42%	60.81%	30.02%
	sound and picture synchronization	5.15%	98.58%		
Sound and picture counterpoint	the corresponding relation	54.83%	100%	21.49%	69.98%
	the alignment relation	40.36%		15.82%	
	The vacancy relation	4.81%		1.88%	
total		100%	100%	100%	100%

As can be seen from Table 1, among the phonographic relationships in multimodal oral history texts, the phonographic synchronization relationship accounts for 60.81% and the phonographic counterpoint relationship accounts for 39.19%. Synchronization relationship is divided into language and picture synchronization and sound and picture synchronization, mainly language and picture synchronization relationship; The counterpoint relationship can be divided into three categories: correspondence relationship, column relationship and vacancy relationship. The corresponding relationship has the highest proportion, followed by column relationship and vacancy relationship has the lowest proportion.

In the short video of Liziqi's traditional craft, the proportion of sound and picture synchronization is about 30.02%, and the proportion of sound and picture counterpoint is about 69.98%. However, most of the synchronization relations here are the synchronization between the picture and the sound, the synchronization between the picture and the language accounts for only 1.42%, and the counterpoint relationship is completely the relationship between the picture and the music, which has nothing to do with the language.


3.1 Sound and picture synchronization

The sound and picture synchronization relationship of multimodal oral history discourse generally appears in audiovisual materials of interviews or episodic broadcasting. The sound and picture synchronization relationship of Li Qiqi's traditional short videos (we only discuss the synchronization relationship between language and picture) appears in the pictures of Li Qiqi's own words and images.

3.1.1 The sound and picture synchronization of the inserted data

In the multimodal oral history discourse, in order to increase the concretization and vividness, the director often inserts some audiovisual materials related to the theme, which are embedded in the multimodal oral history discourse as a whole. The language and the picture in the audiovisual

data are both in and out.


sound	picture
<p>Yes, socialist democracy has been trampled on the most violently, how can the people not raise their eyebrows and take out their swords? When the Gang of Four wanted to push our country into the abyss of feudal fascist suffering, how could the people not raise their eyebrows and take out their swords?</p>	

The above example is selected from the oral history of "Central New Film Group", "Remembering the New film 60 years · Commemorating the beloved Premier Zhou". The interviewee Chen Jinchu recalled the past when he shot "Ten Miles Long Street to send the Premier". In the interview, the audio and video archive "Raised eyebrows and sword out of the sheath" was inserted. The picture shows people gathering in Tiananmen Square to mourn the death of Premier Zhou, and the voice is a voiceover from the film "Raised eyebrows and swords out of their sheaths." When the picture disappears, so does the voice-over. The two are combined in a way of sound and painting synchronization, and the lack of either side will affect the audience's understanding.

3.1.2 The sound and picture synchronization of both sides of the interview

In the multimodal oral history discourse, the relationship between interview language and picture is also the synchronization of sound and picture.

In our self-built multimodal oral history discourse corpus, there are very few short video samples containing interviewer's pictures. Among the 50 samples, there were only 2 samples containing interviewer images, with a discourse document rate of 4% and a discourse mean of 1.15. The interviewer has few pictures and few words. In videos containing images of the interviewer, the interviewer usually asks questions to the interviewee. The camera usually cuts to the interviewee immediately after the question is asked. For example:

sound	picture
<p>When did Yang Yi start to learn crosstalk?</p>	

The above examples are selected from the public number "Family Oral Documentary" "iCare minute to minute: Father Yang Shaohua does not let his son talk crosstalk". In the picture, the interviewer is seated, smiling, facing the camera side, hands folded and naturally drooping, and

leaning slightly forward to ask the interviewee questions. The interviewer's language is synchronized with the picture. The interviewer asked, "When did Yang Yi start to learn crosstalk?" The camera immediately cuts to the interviewees Yang Shaohua, Yang Yi and his son. At the same time, the interviewer's words disappear with the picture. The interviewer's voice is clear and standard, and even if you do not look at the interviewer's picture, it will not affect the audience's understanding.


In our self-built multimodal oral history discourse corpus, the text document rate of short video samples containing interviewees' images is 100%, and the average text is 42.1. The duration and frequency of interviewees' images are very high. It can be seen that the interviewee picture is the necessary content of multimodal oral history discourse.

Generally speaking, the interviewees in the picture do not have large body movements. We refer to the proposal of Wang Lifei and Wen Yan (2008). The respondents' body language is divided into five categories: Hand Movement, Head Movement, Facial Expression, Eye Movement and Posture. In multimodal oral history texts, more than 78.46% of respondents used posture to tell stories, accompanied by Hand Point, Head Move, Smile, Frown and other gestures. Respondents tend to use more hand movements and head movements, but less facial expressions and eye movements, as shown in Figure5:

类别	出现次数	所占时间	所占时间	所占时间	所占时间	所占时间	所占时间
Hand Move...	359	0.23	12.38	1.751	1.39	628.71	13.505
Head Move...	160	0.19	9.954	1.157	0.832	185.078	11.272
Facial Expre...	59	0.37	163.97	9.659	3.32	569.869	9.98
Posture	214	1.41	163.97	15.702	11.31	3360.159	8.85
Eye Movem...	29	0.42	19.614	8.046	6.48	233.341	10.15

Figure 5. Statistical chart of body language annotation of respondents

On the whole, compared with the pictures that show human behavior or large-scale actions, the interviewees' pictures are relatively static, mostly like pictures, which only play an auxiliary role in information transmission. For example:


sound	picture
Lian Lijuan: "I checked with him, I said that it is still like that, it is a fibroma, our opinion is surgical removal." In the end, she opted for surgery. I had surgery and he stood behind me, and it turned out to be a fibroid. He said you Chinese doctors are great."	

The above example is selected from the oral history of old experts of Peking Union Medical College Hospital. The interviewee is Professor Lian Lijuan from the Department of Obstetrics and Gynecology. The interviewee is recalling her diagnosis and surgery for Princess Monique. The interviewee in the picture takes a seated position, and when recalling the French doctor standing behind her, the left arm naturally raises and points back, echoing the words "he stands behind me and looks".

In addition to the body language, the framing, lighting, color selection mode and the position of the interviewees' pictures are synchronized with their words, which reflects positive interactive significance.

In the short video of Liziqi's traditional craft in the reference corpus, there are no two sides of the interview corresponding to the multimodal oral history discourse, and there are few language

symbols. These language symbols are mostly Li Qiqi's own words, forming a synchronous relationship with Li Qiqi in the picture, which is different from the audio-picture synchronization relationship, which accounts for only 1.42% of Li Qiqi's traditional craft short videos.

sound	picture
Li Ziqi: "There are so many flowers outside today!"	

The above example is selected from the "Rose flower cake" in the reference corpus. Li Ziqi went out to collect rose flower cake for making. When she came home, she exclaimed, "There are many flowers outside today." In the picture, Li Ziqi is positive and side to the camera, the vision, talking to herself, walking and talking, and the sound and painting are synchronized.

3.2 Sound and picture counterpoint

In multimodal oral history texts, picture symbols such as pictures, charts, scene representations, and location symbols have a sound and picture counterpoint relationship with their corresponding languages. Picture and sound are independent and interrelated, which can be either superficial or deep. However, in the short video of Li Ziqi's traditional craft, we have not found the counterpoint between language and picture.

In the counterpoint relation of sound and picture, according to the correlation of the center of sound and picture, we further divide the counterpoint relation of sound and picture into three sub-categories: correspondence relation, countercolumn relation and vacancy relation.


3.2.1 The corresponding relation of sound and picture counterpoint

In the counterpoint relationship between sound and picture, when the language center and the picture center are cross-referenced and corresponding, they tell the same historical figure or the same event together, there is a corresponding relationship between language and picture. In the corresponding relationship, the language and the picture are simply and directly combined together, showing a one-to-one or one-to-many reference relationship, and the content is generally simple and concentrated, easy for the audience to understand and remember. The correspondence between language and picture is one of the most common sound and picture relationships in multimodal oral history texts, accounting for about 21.49% of the overall sound and picture structure relationship and 54.83% of the sound and picture counterpoint relationship in short video. Correspondence includes single correspondence and multiple correspondence.

(1)Single correspondence

When there is only one language center and only one picture center, and the two form a correspondence, when the same historical figure or different aspects of the same thing are told and displayed together, such a sound and picture correspondence is a single correspondence. In the single corresponding relationship, the picture focuses on showing historical figures and things from a concrete point of view; Language focuses on telling historical events from an abstract point of view, and language is directly related to the picture. The language is mostly explanatory words, and


there are also a small number of interviewees' words. What appears on the screen is pictures, charts and pictures of objects. For example:

sound	picture
Under the mushroom cloud, thousands of armed police who rushed to the scene were stopped outside.	

In the above example, the explanatory discourse center is "thousands of public security armed police who rushed to the scene were stopped outside". The picture shows the picture of public security armed police who were stopped by trucks to help the scene, and the language center corresponds to the picture center.

(2) Multiple correspondence

When the language has multiple centers, and the picture also has multiple centers, and the picture center corresponds to the language center one by one, and the two jointly tell different aspects of multiple historical figures and things, such a sound and picture correspondence is a multiple correspondence. The language is mostly explanatory words, and there is a small amount of interviewees' words. The pictures appear on the screen are pictures of people, objects or scenes. For example:


sound	picture
Following Sun Yat-sen's instructions, Wang Jingwei, Dai Jitao and eleven others made a decision on the Northern political situation that the provisional executive government could only be a temporary de facto government...	

In the above example, there are two parallel centers in the explanatory discourse, one is "Wang Jingwei" and the other is "Dai Jitao". The pictures of Wang and Tao appear on the picture, and the language center corresponds to the picture center one by one, forming multiple correspondence.

3.2.2 The alignment of sound and picture

In the sound and picture counterpoint relationship, when the picture center and the language center are not simple one-to-one correspondence, and the two tell different but related content, the language and the picture are in a line relationship. In the multimodal oral history discourse, the relationship between sound and picture can provide the maximum information to the audience in a limited time, which is conducive to mobilizing the audience's thinking, deepening understanding and emphasizing the key points. The language in the relationship of sound and picture is mostly explanatory discourse, and the interviewee's discourse is little. The picture can be a statement of action, or it can be a static sentence. The center of the picture is mainly the figurative characters, objects, environment and behavior of the characters, while the language center is the summary,


comment and explanation of the center of the picture. In the multimodal oral history discourse, this kind of sound and picture relationship exists in large numbers. As one of the most common sound and picture relationships, the proportion of the overall sound and picture structure relationship in short video is about 15.82%, and the proportion of the sound and picture counterpoint relationship is about 40.36%. For example:

sound	picture
<p>During the war, a total of more than 110 comrades from the Central News Documentary Film Studio participated in the battlefield shooting, and they and other journalists together with their blood and lives recorded that a thrilling battle and a moving story.</p>	

The language center is in the anti-American aid during the new film factory sent more than one hundred war photographers, the center of the picture is a new film photographer and others group pictures, the photographer and the explanation of the discourse in the "more than 110 comrades" is the relationship between the local and the whole, the language is a summary of the picture.

3.2.3 Vacancy relation in sound and picture counterpoint

In the sound and picture counterpoint relation, when the language center and the picture center are irrelevant or they do not match, the language and the picture are empty. In the vacancy relation, there is no direct relationship between the language center and the picture center. The language in the vacancy relation of sound and picture is mostly interpretive discourse. Most of the language centers are historical figures, events or abstract relationships, properties and states that are difficult to express directly in pictures. The pictures are mostly static sentences, and the narrative sentences are few. The center of the picture is mostly figurative characters, objects and environments, which are used as metaphors. In the multimodal oral history discourse, the proportion of the sound and picture relationship in the overall sound and picture structure relationship of short video is about 1.88%, and the proportion of the sound and picture synchronization relationship is about 4.80%. For example:

sound	picture
<p>He (Zhang Yuanji) begged the government to buy these books, while raising money everywhere." And the great Qing Dynasty is already in turmoil, where is the mind to protect the book?</p>	

The above example is from the National Photo Album: The First Good Thing. The center of the explanatory discourse is the political turmoil in the late Qing Dynasty, and the center of the picture

is the dangerous bridge on the river. There is no direct connection between the language center and the picture center. Because it is difficult for the picture to directly represent the abstract state of wind and rain in the late Qing Dynasty, the floating boat and bridge are used to metaphor the current situation.

4. Conclusion

On the basis of self-built multimodal corpus, this paper takes the theory of multimodal discourse analysis as the theoretical basis, combined with the theory of sound and picture relationship, and takes short oral history videos as the corpus to analyze the sound and picture relationship of multimodal oral history discourse.

The phonographic relationship between picture and language in multimodal oral history discourse includes two categories: phonographic synchronization and phonographic counterpoint. The counterpoint relationship of sound and picture includes three sub-categories: correspondence relationship, countercolumn relationship and vacancy relationship.

Short oral history video is a very important class of multimodal oral history discourse, and the discussion of the relationship between sound and picture is the discussion of the relationship between audiovisual modes. At present, the research on the relationship between sound and picture in multimodal oral history is still in the exploratory stage, and there are still many problems to be further studied.

References

- [1] Bateman, J. *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents* [M]. New York: Palgrave-Macmillan, 2008.
- [2] Flewitt, R. *Bringing ethnography to a multimodal investigation of early literacy in a digital age* [J]. *Qualitative Research*, 2011(3): 293-310.
- [3] Gunther Kress and Theo van Leeuwen, *Reading Images: The Grammar of Visual Design* (2nd ed) [M]. London: Routledge, 2006.
- [4] Halliday M.A.K. & Hasan. *Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective* [M]. Deakin: Deakin University Press, 1985b.
- [5] Holsanova, J. *Reception of multimodality: Applying eye tracking methodology in multimodal research* [A]. In Jewitt, C. (ed). *The Routledge Handbook of Multimodal Analysis* [C]. London: Routledge, 2014: 287-298.
- [6] Jewitt, C. *The Routledge Handbook of Multimodal Analysis* [M]. London: Routledge, 2014.
- [7] Jewitt, C., Jeff Bezemer, Kay O'Halloran. *Introducing Multimodality* [M]. Routledge, 2016.
- [8] John Flowerdew and John E. Richardson. *Routledge Handbook of Critical Discourse Studies* [M]. Routledge, 2022.
- [9] Kress, G. & T. Van Leeuwen. *Multimodal Discourse: The Modes and Media of Contemporary Communication* [M]. London: Arnold, 2001.
- [10] Kress, G. *Multimodality: A Social Semiotic Approach to Contemporary Communication* [M]. London: Routledge, 2010.
- [11] Longacre, Robert E. *The grammar of discourse* [M]. Plenum Press, 1983.
- [12] M. O'Toole, *The Language of Displayed Art* [M]. London: Routledge, 1994/2011.
- [13] Norris, S. *Analyzing Multimodal Interaction: A Methodological Framework* [M]. London: Routledge, 2004.
- [14] Norris Sigrid. *Multimodal Theory and Methodology: For the Analysis of (Inter)action and Identity* [M]. Routledge, 2020.

- [15] Scollon, R. & S. Scollon. *Discourses in Place: Language in the Material World* [M]. London: Routledge, 2003.
- [16] Tuire Oittinen. *Multimodal and collaborative practices in the organization of word searches in lingua franca military meetings* [J]. *Journal of Pragmatics*. 2022: 41-55.
- [17] Wendy Sandler. *Redefining Multimodality* [J]. *Frontiers in Communication*. 2022.