Subjective Evaluation of Angular Displacement between Picture and Sound Directions for HDTV Sound Systems*

SETSU KOMIYAMA

NHK Science and Technical Research Laboratories, Setagaya, Tokyo 157, Japan

Several recent studies have suggested that ordinary stereophonic systems are not sufficient for HDTV use. These investigations have noted the localization error between picture and sound as a reason, but have not studied the extent to which this phenomenon causes viewer annonyance. The psychological experiment described was designed to investigate the acceptable extent of angular displacement between visual and auditory images for on-axis viewing. The results show that it is about 11° for acoustic engineers and 20° for the members of the general audience. In addition, the number of frontal channels in HDTV is discussed.

0 INTRODUCTION

Studies have been made on stereophonic systems for high-definition television (HDTV) at CBS, IRT, and NHK. All agree that ordinary stereophonic systems are not sufficient for a large HDTV screen. Torick [1] proposes a triphonic sound system which utilizes two conventional stereophonic channels and a center channel for optional use when widely spaced left and right loudspeakers are employed. NHK [2] proposes a 3–1 system consisting of three frontal channels and a rear channel. NHK's configuration is similar to the quadraphonic system that is used for motion picture sound. IRT takes a more academic position. Plenge [3] suggests that five channels are necessary in order to localize sounds in the direction of corresponding pictures. His suggestion is based on the experimental result that a discrepancy of less than 4° is not perceptible. This has been confirmed by the author as well [4].

The stability of sound localization may be indispensable to HDTV, but even in the case of movies, many products are still monophonic despite the feasibility of quadraphony. Therefore the degree to which the discrepancy annoys us should be discussed in a more practical light.

In this paper the discrepancy between sound image and video image is evaluated specifically to find the acceptable extent of discrepancy, not the just-noticeable difference (JND). To obtain practical results, the discrepancy was evaluated using human faces and voices. Moreover, the subjects, consisting of both nonexperts and experts, were encouraged to relax.

The method of successive categories was used. There is a very good reason for using this method. In general, localization of sound images does not distribute normally when associated with a visual stimulus [5]. The perceived direction depends on a number of factors, such as the subject's concentration on the visual stimulus, the compellingness of the stimulus, and the synchronization of sound to image [6]. Therefore, to ask subjects the direction of the sound image itself is not the best approach, because this question sometimes makes it difficult for the subject to concentrate on the visual stimulus. The method of successive categories is a good way to avoid this difficulty.

1 EXPERIMENT I

1.1 Experimental Procedure

The setup for this psychological experiment is shown in Fig. 1. The picture display was a 72-in (1.83-m) NTSC television projector. Pictures were projected from the front of its screen.

Ten loudspeakers (Yamaha NS-10M, flat to 20 kHz) were set on the left side of a subject. Four of the loud-

* Manuscript received 1988 May 31; revised 1988 December 5.

speakers were set just under the screen (Fig. 2), and the others were set at the same height around the subject. The loudspeakers between 0° and 45° were hidden by a curtain so that the subjects would not be aware of them, and all other loudspeakers were out of view. Assuming symmetry between left and right, no loudspeakers were set on the right side. The entire setup was in a soundproof room with a volume of 130 m^3. Though the distances to the side and rear loudspeakers were slightly shorter than to the frontal ones because of the size of the room, there was no significant difference in their loudness at the position of the subject.

Program material was a young woman sitting on a chair and reading a magazine aloud. A scene from the program and the frontal loudspeakers are shown in Fig. 2. The background behind her was a studio interior wall. Her image and voice were reproduced on the screen and by one of the loudspeakers, respectively. The image of her mouth on the screen was positioned 7° above the tweeters of the loudspeakers. The motion of her lips on the screen was synchronized with the sound.

The experimental procedure was as follows. A subject sat in front of the screen on axis and judged how much he or she was annoyed by disagreement between visual and auditory localization according to the following categories:[1]

A. Imperceptible
B. Perceptible, but not annoying
C. Slightly annoying
D. Annoying
E. Very annoying

After the subject had recorded each evaluation, he or she switched to the next presentation by pushing a button. The order of presentation was random, but identical for all subjects. The number of repetitions was five, so judgments were made by each subject 50 times (10 loudspeakers × 5 times). The subjects were allowed to spend as much time on making a judgment as they liked, and were instructed to feel at home.

Forty-four persons served as subjects. They consisted of 14 acoustic research engineers (experts) and 30 nonexperts, including 11 women. The sound pressure level was set at about 65 dB(A) maximum, or virtually equal to the loudness of the natural voice of the model. The lights in the room were turned off to shut out visual stimuli other than the video image.

1.2 Experimental Results

All the results are shown in Table 1. The values represent the percentage of each response to each stimulus. It is clear that the larger the discrepancies become, the lower the scores, except for the loudspeaker at 180°. As sounds coming from behind are sometimes perceived as coming from in front, it is not unusual that the scores for 180° are not minima.

The same data are illustrated in Figs. 3 and 4. These figures show that the experts judged the discrepancies more strictly than the nonexperts. The nonexperts responded with a much higher ratio of categories B and C: "perceptible, but not annoying" and "slightly annonying" than did the experts. For the nonexperts, perceptible discrepancies are often acceptable.

14.3% of the experts and 14% of the nonexperts judged 0° localization not to be category A, which might be because of the 7° discrepancy between the picture and the front loudspeaker in the median plane.

Since categories A to E represent only orders and not values, the averages and the standard deviations of the evaluations cannot be calculated directly from Table 1. If we were to assume that the categories do actually represent equal psychological intervals, the frequency distributions of the stimuli are obviously often not normal. In other words, these categories are given proper values (category values) by assuming the normal distribution of data [7]. These calculated values for the three subject groups (experts, nonexperts, and total) are shown in Table 2, where each category E is forced to an origin and each standard deviation is a

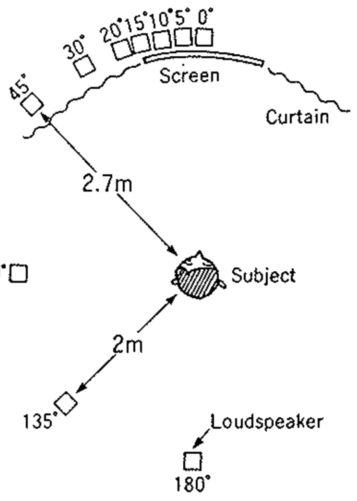

Fig. 1. Setup for experiment I.

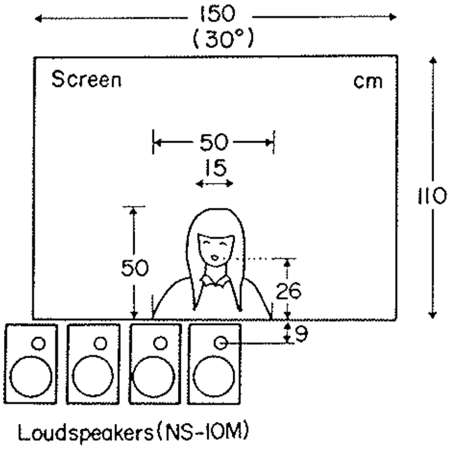

Fig. 2. Placement of picture on television screen and frontal loudspeakers (1° = 5 cm).

[1] CCIR Rec. 500, "Impairment Scale of Study Group 11."

Table 1. Percentage of responses to each stimulus.

	Acoustic research engineers					Nonexperts				
Discrepancy	A	B	C	D	E	A	B	C	D	E
0°	85.7	14.3	0	0	0	86	13.3	0.7	0	0
5°	70	24.3	1.4	4.3	0	78	18.7	3.3	0	0
10°	15.7	44.3	24.3	14.3	1.4	48	42.7	8.7	0.7	0
15°	4.3	24.3	35.7	20	15.7	20.7	54.7	21.3	3.3	0
20°	1.4	15.7	34.3	25.7	22.9	5.3	52	34	8.7	0
30°	0	5.7	30	20	44.3	3.3	24	50	19.3	3.3
45°	0	0	11.4	22.9	65.7	0	6	32.7	42.7	18.7
90°	0	0	0	7.1	92.9	0	0	10.7	28	61.3
135°	0	0	0	10	90	0.7	0	5.3	26	68
180°	1.4	0	0	21.4	77.1	2	18.7	14	32.7	32.7

unit for each group. In calculating these values, we excluded the data for 135° and 180°, because they had confusion of front and back.

Though mean evaluation and standard deviation for each stimulus can be obtained by using the values in Tables 1 and 2, some inaccuracies may be caused by the truncation of the data in Table 1. We therefore used the regressions of distances between stimulus and category means as obtained from each discrepancy on the category values shown in Table 2. The resulting regressions are illustrated in Figs. 5 and 6, where the abscissa indicates the scale of category values and the ordinate indicates the distances from stimulus to category means, measured in standard deviations. The fact that the regressions are linear on the scale of category values indicates that the scales shown in Table 2 were appropriate.

The mean evaluation for each discrepancy can be found on the position where the regression line crosses the abscissa. (The distance between stimulus and category mean is zero at that point.) The standard deviation corresponds to the reciprocal of the slope. The resulting means and standard deviations are graphed in Figs. 7 and 8. The ordinates represent evaluation values for the stimuli by assuming category E to be 0 and each standard deviation to be a unit.

The author believes that the acceptable limit of the discrepancy is the sound direction corresponding to the boundary between categories B and C in Figs. 7 and 8. This was about 11° for the experts and about 20° for the nonexperts.

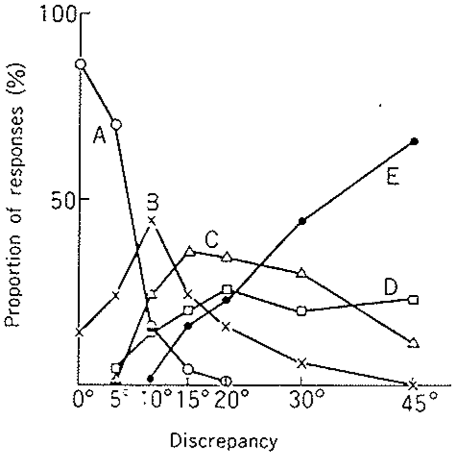

Fig. 3. Proportion of responses to each discrepancy (experts).

Table 2. Category mean values measured in standard deviation units. (Category E is forced to 0.)

	A	B	C	D	E
Expert	4.171	2.818	1.886	1.128	0
Nonexpert	4.89	3.442	2.272	1.184	0
Total	4.22	2.838	1.788	0.983	0

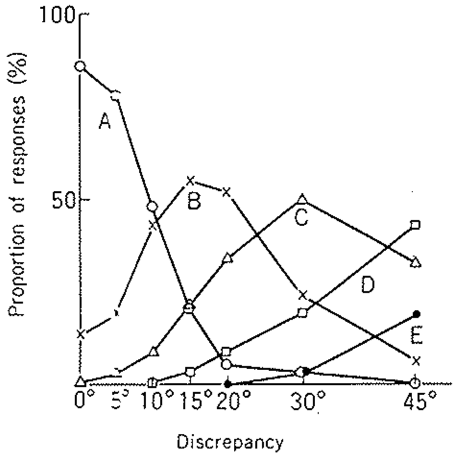

Fig. 4. Proportion of responses to each discrepancy (nonexperts).

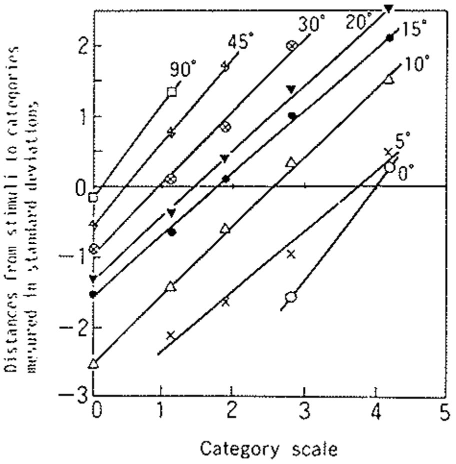

Fig. 5. Regression of distances from stimuli to categories on the common category scale (experts).

2 EXPERIMENT II

2.1 Experimental Procedure

In this experiment, an actual television program was used in order to determine whether the results of experiment I would apply to general programs as well. The program material was a pop song by a female singer. The audio track had been recorded in a live performance with a multichannel recorder and mixed down to 3–1 quadraphony. In the center channel, her vocal was the only signal, while the orchestra and reverberation effects were recorded in the left and right channels. The rear channel included the ambient sounds, but they were not reproduced in this experiment. The image of the singer was always on the screen, but sometimes varied in placement.

Two loudspeakers were added to the setup that had been used in experiment I, at directions of ±30° in order to reproduce the sounds of the orchestra. The loudspeakers for the vocal were switched on in random order by the subjects. The sound pressure level was about 80 dB(A).[2] Judgments were made in the same way as in experiment I, except that the subjects were instructed to make them while the lips of the singer could be seen. Seven research engineers and five nonexperts served as subjects.

2.2 Experimental Results

The method of processing the data was the same as used in experiment I, except for the use of the numerical values shown in the bottom line of Table 2 as category values. The results are illustrated in Fig. 9. The solid line shows the results of this experiment, while the broken line shows the results of experiment I using the same subjects. There is no significant difference between the two lines, except for the 0° direction. The difference at the 0° direction is not unusual, however, because the image of the singer did not remain at the center of the screen.

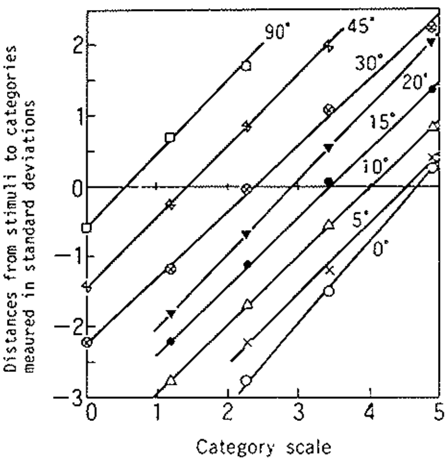

Fig. 6. Regression of distances from stimuli to categories on the common category scale (Nonexperts).

3 DISCUSSION AND CONCLUSION

When watching HDTV from the optimum viewing

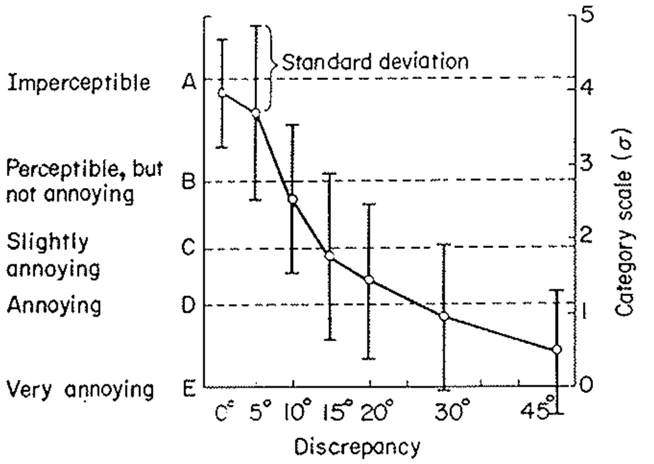

Fig. 7. Evaluation of discrepancy between video and sound images in horizontal plane (experts).

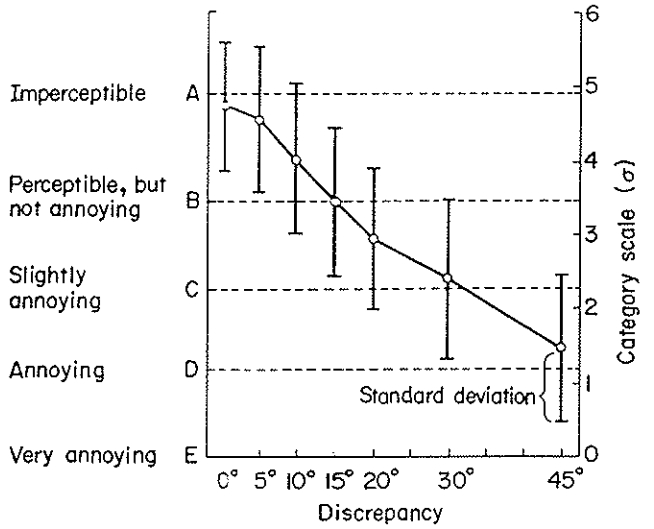

Fig. 8. Evaluation of discrepancy between video and sound images in horizontal plane (Nonexperts).

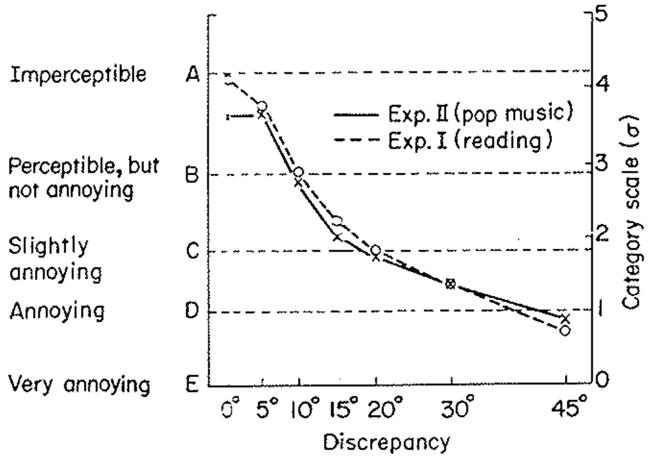

Fig. 9. Evaluation of discrepancy for actual program.

[2] CCIR Rec. 562-2.

distance, the viewing angle of the screen is about 30°. With a center channel, the discrepancies between the directions of picture and sound can be limited to less than 15° in any scene. As the acceptable extent of discrepancy in this experiment was found to be 20° for nonexperts and 11° for experts, this improvement may be satisfactory for the majority of the television audience, but not for a sensitive audience, such as acoustic engineers.

If the discrepancy is held to less than 11°, even an expert audience should be satisfied. Acoustic engineers, however, are an exceptional audience, as they are prone to pay attention to the localization and quality of sound first of all. Actually, some of the acoustic engineers in this experiment suggested that they would not have been annoyed by the discrepancy if they had not been asked to judge it. The author therefore believes that a center channel is sufficient to improve the stability of sound localization, at least in the horizontal plane.

4 ACKNOWLEDGMENT

The author wishes to thank Dr. K. Ohgushi and Dr. E. Miyasaka for the opportunity to study this problem, and Dr. K. Nakabayashi for his pioneering work.

5 REFERENCES

[1] E. Torick, "A Triphonic Sound System for Television Broadcasting," *Soc. Mot. Pic. Telev. Eng. J.*, vol. 92, pp. 843–848 (1983 Aug.).

[2] K. Ohgushi, S. Komiyama, K. Kurozumi, A. Morita, J. Ujihara, and K. Tsujimoto, "Subjective Evaluation of Multi-Channel Stereophony for HDTV," *IEEE Trans. Broadcast.*, vol. BC-33, pp. 197–202 (1987 Dec.).

[3] G. Plenge, "Sound Design and Sound Transmission in a Future High-Density Television (HDTV) System," presented at the 79th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 33, p. 1006 (1985 Dec.), preprint 2306.

[4] S. Komiyama, "Visual Factors in Sound Localization for HDTV" (in Japanese), *J. Acoust. Soc. Jpn.*, vol. 43, pp. 664–669 (1987).

[5] H. L. Pick, Jr., and D. H. Warren, "Sensory Conflict in Judgments of Spatial Direction," *Perception & Psychophys.*, vol. 6, no. 4, pp. 203–205 (1969).

[6] R. B. Welch and D. H. Warren, "Immediate Perceptual Response to Intersensory Discrepancy," *Psychol. Bull.*, vol. 88, no. 3, pp. 638–667 (1980).

[7] J. P. Guilford, *Psychometric Methods* (McGraw-Hill, New York, 1954).

THE AUTHOR

Setsu Komiyama was born in Nagano City, Japan, in 1953. He studied electrical engineering at Tokyo University, where he received an M.S.E.E. degree in 1977. He then joined NHK (Japan Broadcasting Corporation) and worked as an engineer at Shizuoka Broadcasting Station for three years. In 1980 he joined NHK Science and Technical Research Laboratories and has worked in research on visual–auditory interaction for HDTV sound systems. In 1984 he developed a synchronizer system for audio tape recorders having no time-code track. He is a member of the AES and the Acoustical Society of Japan.