Development of Mandarin speech test materials for civilian pilots in China

Mo-Sheng Hu[1,2], Jing Chen[3], Xiu-Yun Yang[2], Lei Wang[2], Wen Cao[4], Yin Bai[2], Feng-Jie Ma[2], Cai-Hong Qin[2], Shou-Qin Zhao[1], Hua Zhang[3]

[1]Department of Otolaryngology Head and Neck Surgery, Beijing Tongren Hospital, Capital Medical University, Key Laboratory of Otolaryngology Head and Neck Surgery, Ministry of Education, Beijing 100730, China;
[2]Civil Aviation Medical Assessment Institute, Civil Aviation Medicine Center, Civil Aviation Administration of China, Beijing 100123, China;
[3]Clinical Audiology Center, Beijing Institute of Otolaryngology, Beijing Tongren Hospital, Capital Medical University, Beijing 100005, China;
[4]College of Chinese Studies, Beijing Language and Culture University, Beijing 100083, China.

Air-ground radiotelephony communication provides the only way for a civilian pilot and air traffic controller to communicate during all phases of a flight. The quality of communication during the flight is crucial for aircraft safety. For this reason, it is critical that pilots have good auditory function because of the lack of visual cues during flight. Hearing loss can be endangered among civilian pilots who are routinely exposed to loud occupational noise environments such as the cockpit, and can place the aircrew and passengers at risk. Radiotelephony is a type of semi-artificial language, based on imperative sentences that are standardized, procedural, articulate, and precise clear brachylogy. Mandarin radiotelephony terminology, which is widely used in Chinese domestic air routes, has its own characteristics and comprises three parts: Mandarin Chinese terminology, English capital letters, and English abbreviations of terminology. In China, only pilots who can meet the auditory fitness for duty (AFFD)[1] standards of the Civil Aviation Administration of China (CAAC; Beijing, China) are certified as possessing sufficient hearing to ensure a safe flight. The current CAAC's AFFD test is primarily dependent on pure-tone audiometry (PTA). However, previous studies indicate that PTA may be unsuitable for deciding the ability of an individual to perform the job satisfactorily because the ability to recognize speech in steady-state noise cannot be predicted with an audiogram. Therefore, in this study, we aimed to develop a set of speech audiometry materials that is specifically designed for civilian pilots as a functional additional AFFD test for CAAC.

To be qualified as an auditory functional assessment, a suitable AFFD test must take into consideration the occupational environment, an individual's professional experience, and the auditory requirements for the job. Hence the stimuli in this study were based entirely on hearing critical tasks pertaining to radiotelephony communications. In view of the actual application situation of the Mandarin radiotelephony language in China, the content of the speech corpus should include the standard radiotelephony phraseologies required for each normal in-flight phase as well as unusual situations. Therefore, the following four classic Mandarin radiotelephony communications textbooks were finally selected: (1) International Civil Aviation Organization (ICAO) Radiotelephony Communication; (2) Radiotelephony Communication Course (second edition); (3) 900 Sentences of Pilot English Proficiency Examination of China; and (4) Guidance on Radiotelephony Communications Under Unusual/Emergency Situations. After discussing with the captain pilots and linguist in research group, we finally identified twelve categories of hearing critical tasks, as following: (1) departure (eg, pre-flight, start-up, pushback, taxi-out, and climb-out); (2) flight altitude; (3) very-high-frequency omnidirectional range and waypoints; (4) en route; (5) flight speed; (6) flight direction; (7) descent and approach; (8) transponder frequency; (9) landing; (10) after landing; (11) query normal height; and (12) non-routine condition.

Based on the above work, the development of the sentence lists needed to follow some basic principles: (1) sentences should be completely selected from speech communications with radiotelephony language; (2) intonation factors and phoneme balance should not be considered; (3) the balance of hearing critical tasks should be strictly maintained between the sentence lists; (4) excessive homogeneity and heterogeneity

Access this article online

Quick Response Code:

Website: www.cmj.org

DOI: 10.1097/CM9.0000000000000491

Correspondence to: Dr. Shou-Qin Zhao, Department of Otolaryngology Head and Neck Surgery, Beijing Tongren Hospital, Capital Medical University, Key Laboratory of Otolaryngology Head and Neck Surgery, Ministry of Education, Beijing 100730, China
E-Mail: shouqinzhao@163.com

Chinese Medical Journal 2019;132(21)
Received: 26-09-2019 **Edited by:** Yuan-Yuan Ji

should be prevented; (5) sentences should be of variable lengths, ranging from 3 to 13 Chinese characters (to avoid interference from memory factors); (6) as much as possible, monosyllabic words and spondees should be chosen as the keywords and a few trisyllabic words should be selected, as appropriate; (7) only declarative and imperative sentences should be included; and (8) no duplicate sentences should be included. As a result, 20 sentence lists were developed, each list included 20 sentences with 100 keywords [Supplementary Table 1, http://links.lww.com/CM9/A110]. Then, the recording work was conducted in an anechoic chamber in the Chinese Academy of Social Sciences in Beijing, China. An experienced sound engineer obtained voice signal acquisition by using professional recording equipment and tools. The speaker was a 47-year-old male broadcaster with more than 20 years of broadcasting experience. After digital processing, the audio files with loudness equalization were finally produced and stored for use.

The study was approved by the Ethical Committee of the Civil Aviation Medicine Center. A total of 40 male Chinese student pilots who worked for Shenzhen Airlines and held Class I certificates of CAAC were enrolled. The average age was 23.7 years (range, 21–26 years). The mean total flight time was 229.6 h (range, 205–279 h). All the subjects had good written and oral skills both in English and Mandarin. Participants with hearing loss in both ears, or medical history of ear disorders were excluded. After conventional audiometry, we selected the relatively healthy ear as the test ear to conduct the speech test. All subjects were tested in a double-walled acoustic cabin that met the American National Standards Institute 2004 specifications for audiometric test rooms. The clinical audiometer (ie, sound pressure level [SPL]) was calibrated in line with the International Standards (IEC 645-2:1993) before administering speech audiometry. Based on the results of a small sample preliminary experiment, six stimulus intensity levels were identified, and ranged 5 to 15 decibel hearing level (dB HL) in 2 dB steps. Latin square design was applied, and the following formula was used to calculate the word recognition score (WRS): WRS = the number of correct key words/100 × 100%.

All statistical analyses were conducted using SPSS 21.0 (SPSS Inc., Chicago, IL, USA). Logistic regression analysis was performed to obtain the performance-intensity (P-I) functions and calculate the regression slopes that could be used to evaluate the sensitivity of the system and the regression intercepts for all the lists. The values of the slope and intercept for each list are in Supplementary Table 2, http://links.lww.com/CM9/A110. These values were then put into modified logistic regression equation (Equation 1) that was designed to calculate the percentage of correct performance at any specified intensity level.[2]

$$p = \frac{\exp(a + b \times i)}{1 + \exp(a + b \times i)} \times 100 \quad (1)$$

In Equation 1, "p" is the WRS, "a" is the regression intercept, "b" is the regression slope, and "i" is the intensity level (in dB HL). By putting the regression slope, intercept, and intensity level into Equation 1, the percentage of correct keyword recognition could be calculated, and the P-I functions could be obtained by statistical curve fitting.

The data of PTA and total flight hours [Supplementary Table 3, http://links.lww.com/CM9/A110], expressed as mean ± standard deviation (SD), were as follows: 10.6 ± 2.5 dB HL (range, 6.3–16.3 dB HL) and 229.6 ± 27.3 h (range, 205–279 h). For the intensity levels of 5 to 15 dB HL in 2 dB steps, the WRSs were 19.17 ± 18.68, 36.14 ± 23.32, 58.79 ± 22.77, 78.53 ± 15.07, 88.49 ± 9.28, and 93.25 ± 5.46 in ascending order, the maximum variability of the WRSs nearly 50%, and the minimum was nearly 0% and 100%. The mean threshold (50%) of P-I functions was 8.22 ± 0.35 dB HL, the mean slope at threshold was 11.34% ± 1.84% per decibel, and the mean slope of the linear region (20%–80%) was 4.50% ± 1.29% per decibel. The WRSs of the following six sentence lists (Lists 5, 7, 16–18, and 20) revealed that the non-monotonic characteristics followed the consecutive intensity levels.

By one-way analysis of variance [Supplementary Table 4, http://links.lww.com/CM9/A110], we found that all the lists were equivalent in difficulty level ($P > 0.05$). We also conducted a reliability analysis on intensity levels, scores, and test results of 100 keywords in each list. The Cronbach's α value was 0.981 (i.e., >0.80), which suggested that the 20 lists had a high internal consistency; And the validity analysis for the remaining 14 sentence lists [Figure 1] revealed that the Kaiser-Meyer-Olkin value of the test sentences was 0.905 and the Bartlett sphericity test result was $P < 0.001$, which indicated that the test materials also had good validity.

In this study, there is a good consistency between the speech reception threshold (SRT) results in quiet and PTA (8.22–10.6 dB HL). The overall mean parameters across all P-I functions of the sentence lists obtained in this study were compared with several existing materials such as the Mandarin speech test materials, which have reported that the mean SRT of its sentence lists is 23.1 dB SPL (ie, 3.1 dB HL).[3] Another recent study reported a mean SRT of the Mandarin short sentence lists as 6.3 dB HL with a 7.2% per decibel mean slope at 20% to 80% linear score region.[4] We found that the P-I functions in our study exhibited a characteristic feature of relatively high SRTs and low slopes at the linear score region which meant a relatively high difficulty and low sensitivity. The first reason is that, despite our attempt to maintain homogeneity among the participants, it was very difficult to fully avoid the floor effect caused by the relatively high PTA level ($M = 10.60$ dB HL) and some hidden hearing loss. Although our recent studies found that extended high-frequency audiometry (EHFA), which is more sensitive to inner ear injury, may be helpful for the early detection of noise-induced hearing loss in civilian pilots.[5] However, the degree of speech intelligibility deficit varies from individual to individual and unfortunately cannot be predicted effectively by PTA or EHFA. The second reason is that most student pilots had undergone flight training in English-speaking countries and only a few of them had been trained at the Civil Aviation Flight University of China (Guanghan City, Sichuan Province, China). Moreover, the 40 Chinese student pilots participating in

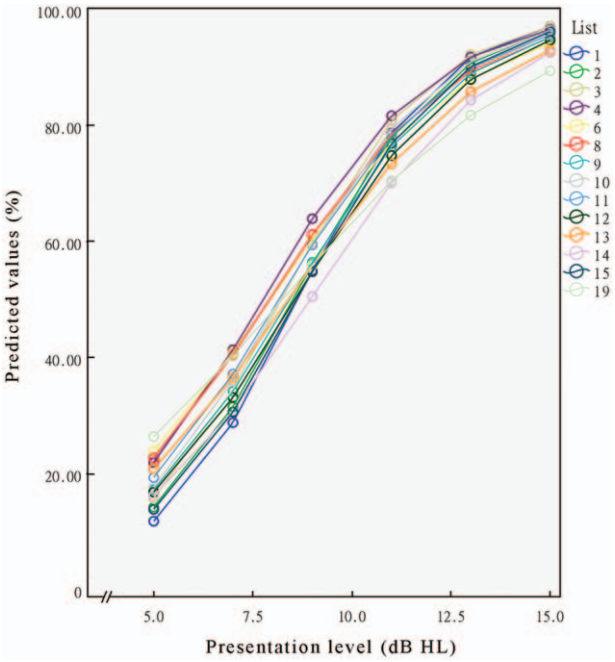

Figure 1: The cluster of P-I function fitting curves of the 14 sentence lists for the 40 pilots. dB HL: Decibel hearing level; P-I: Performance-intensity.

this study had normal hearing and slightly over 200 h of flight experience, but none of them had experience in piloting the commercial jet. Hence, the pilots who were trained abroad may not have a good knowledge of the radiotelephony language spoken in Mandarin.

As reported in this manuscript, the SDs for intensity levels that produced approximately 50% correct word recognition was over ±20% (58.79 ± 22.77). This means the individual scores varied widely over approximately 90%, which was nearly the entire range of possible WRSs. Since all the subjects were audiometrically normal, these differences may be because of their fluency and familiarity with the Mandarin radiotelephony language that is directly related to flight experience. The decision to prevent an experienced pilot with hearing loss from flying should be reconsidered. Thus, in view of the fact that flight experience can compensate for hearing loss to some extent. It is concluded that the novel material is in line with the requirements for a suitable functional AFFD test. Further studies are needed to conduct the listening tasks (i.e., the 14-sentence lists) in real-world noise environments to establish the auditory pass/fail criteria for CAAC.

Acknowledgements

The authors would like to thank Professor Ai-Jun Li and Hong-Li Liang from the Chinese Academy of Social Sciences for their assistance in providing digital recordings. The authors thank Captain Ji Li from Shenzhen Airlines, Captain Shi-You Wang from China United Airlines, First Officer Tie-Chuan Hu from Chengdu Airlines, and Gao-Yuan Fan from Beijing Capital Airlines for their advice and discussion with regard to radiotelephony communications. The authors thank Dr. Wen-Fang Wu from the School of Biomedical Engineering of Capital Medical University for statistical guidance.

Funding

This study was supported by a grant from the Safety Capacity Building Project of Civil Aviation (No. FSDSA0038).

Conflicts of interest

None.

References

1. Tufts JB, Vasil KA, Briggs S. Auditory fitness for duty: a review. J Am Acad Audiol 2009;20:539–557. doi: 10.3766/jaaa.20.9.3.
2. Nissen SL, Harris RW, Jennings LJ, Eggett DL, Buck H. Psychometrically equivalent Mandarin bisyllabic speech discrimination materials spoken by male and female talkers. Int J Audiol 2005;44:379–390. doi: 10.1080/14992020500147615.
3. Zhang H, Wang S, Chen J, Lin S, Wang L, Guo L. Performance-intensity function of Mandarin monosyllable and sentence materials for normal-hearing subjects (in Chinese). J Clin Otolaryngol Head Neck Surg 2008;22:1–4. doi: 10.3969/j.issn.1001-1781.2008.01.001.
4. Zhao DY, Li XL. P-I curves for mono-and disyllable words lists and short sentence lists (in Chinese). Chin J Otol 2015;4:608–612. doi: 10.3969/j.issn.1001-1781.2008.01.001.
5. Ma FJ, Gong SS, Liu S, Hu MS, Qin CH, Bai Y. Extended high-frequency audiometry (9-20 kHz) in civilian pilots. Aerosp Med Hum Perform 2018;89:593–600. doi: 10.3357/AMHP.5029.2018.

How to cite this article: Hu MS, Chen J, Yang XY, Wang L, Cao W, Bai Y, Ma FJ, Qin CH, Zhao SQ, Zhang H. Development of Mandarin speech test materials for civilian pilots in China. Chin Med J 2019;132:2638–2640. doi: 10.1097/CM9.0000000000000491