Work 41 (2012) 3592-3598
DOI: 10.3233/WOR-2012-0670-3592
IOS Press

Selection of a voice for a speech signal for personalized warnings: the effect of speaker's gender and voice pitch

Sheron Machado[a], Emília Duarte[b,c], Júlia Teles[d,c], Lara Reis[b] and Francisco Rebelo[a,c*]
[a] *Ergonomics Laboratory, FMH/Technical University of Lisbon, Estrada da Costa, 1499-002 Cruz Quebrada, Dafundo, Portugal*
[b]*UNIDCOM/IADE – Institute of Arts, Design and Marketing, Av. D. Carlos I, 4, 1600-649 Lisbon, Portugal*
[c]*CIPER – Interdisciplinary Center for the Study of Human Performance, Technical University of Lisbon, Estrada da Costa, 1499-002 Cruz Quebrada, Dafundo, Portugal*
[d]*Mathematics Unit, FMH/Technical University of Lisbon, Estrada da Costa, 1499-002 Cruz Quebrada, Dafundo, Portugal*

Abstract. There is an increasing interest in multimodal technology-based warnings, namely those conveying speech-warning statements. This type of warning may be tailored to the situation as well as to the target user's characteristics. However, more information is needed on how to design these warnings in a way that ensures intelligibility, promotes compliance and reduces the potential for annoyance. In this context, this paper reports an exploratory study whose main purpose was to assist the selection of a synthesized voice for a subsequent compliance study with personalized (i.e., using the person's name) technology-based warnings using Virtual Reality. Participants were requested to listen to speech signals, gathered from a speech synthesizer and post-processed in order to change the pitch perception, and then these were evaluated by fulfilling the MOS-X questionnaire. After that, the participants ranked the voices according to their preference. The effects of the speaker's gender and voice pitch, on both ratings and ranking were assessed. The preference of the male and female listeners for a talker's voice gender was also investigated. The results show that participants mostly prefer as first choice the high-pitched female voice, which also gathered the highest overall score in the MOS-X questionnaire. No significant influence of the participants' gender was found on the assessed measures.

Keywords: Auditory warnings, speech signal, speaker's gender, voice pitch

1. Introduction

Traditionally, warnings are static and visual. However, the adoption of technology-based warnings (e.g., integrating sensors and computers) will enable delivering warning information in a more dynamic, multimodal and effective way, which can be tailored to fit individual needs and situational requirements [11, 13]. Many of the technology-based warnings will incorporate auditory warnings, which can be nonverbal sounds or speech signals. Speech warnings are those that incorporate human speech in recorded, digitized or synthesized form [3], and are the topic of interest of this study.

Due to information overload and/or visual saturation, which are common in actual working environments, the processing of warning messages by other senses, as is the case of audition, can be beneficial. The auditory modality can supplement visual warnings in advantageous ways because sound is omnidirectional and audition cannot be shut off by the receptor. These benefits are described in previous studies that found that the use of multiple sensory channels can effectively increase the amount of information processed [1], augment the speed of processing

[*]Corresponding author. E-mail: frebelo@fmh.utl.pt

[9], enhance performance while potentially decreasing workload, and promote compliance with warnings [2,12,17]. Moreover, these warnings can also be personalized (i.e., including the receptor's first name), which Wogalter et al. [14] found to be more effective than impersonalized warnings (e.g., using "Caution" instead of the name).

However, speech warning' effectiveness can be compromised, to some extent, by several factors, such as the environmental background noise, the poor quality of the sound reproduction due to technical reasons (e.g., bad sound speakers), the lack of naturalness (e.g., the so-called computer accent of synthesized speech), or subjective considerations (e.g., effect of speaker's gender) [5,7]. Regarding the speaker's gender, the literature findings fail to show strong evidences favoring either gender. However, some studies showed that male and female listeners had different preferences for the talker's voice gender [8,16]. Moreover, several studies suggest that female voices tend to produce higher urgency judgments than male voices, fact that can be attributed to the higher pitch of female voices [4,6,15].

In this context, this paper describes an exploratory study whose main purpose consisted of selecting a synthesized voice to be used as a speech signal, as part of a personalized technological-based environmental warning, in the scope of a behavioral compliance research study that uses immersive Virtual Reality (VR).

A Text-to-Speech synthesizer generated the speech signals in a synthesis-by-rule manner. The voices differ in gender (i.e., female and male voices) and in pitch (i.e., high and low pitch). The effects of the speaker's gender and voice pitch were assessed on both subjective ratings and rankings. The preference of the male and female listeners for the talker's voice gender was also investigated. Since the combination of naturalness, intelligibility, rate of presentation, and emotion is considered as a determinant factor in the quality of the synthesized speech [7], the MOS-X (Mean Opinion Scale – Expanded) questionnaire [10] was used to assess the quality of synthesized speech. Furthermore, the MOS-X questionnaire, and its scales, allows the comparison of the ratings, assigned by the participants to the synthetic speeches, to determine the aspects that differentiates the voices.

2. Method

2.1. Participants

Participants were 20 university students (10 males and 10 females), aged between 18 and 31 (mean age = 21.65 years, SD = 2.94), who were inquired prior to the test in order to ensure that they had unimpaired hearing. They were recruited at IADE – Institute of Arts and Design, in Lisbon, Portugal.

2.2. Stimuli

The stimuli were four synthesized speech warnings, two female and two male voices, generated using a demo version of the Text-to-Speech (TTS) synthesizer Oddcast[1]. The male and female voices belong to Eusébio and Amália characters, respectively, which are the only ones that have a Portuguese pronunciation (from Portugal) available on the demo. All speech warnings were digitally stored in separate files on a laptop that served as the host computer in this study. The speech warnings were post-processed, using the GoldWave digital audio editor, V5.58, to get diverse pitch perceptions. The speech warnings tested in this study had an overall duration of 1.6 seconds and 16-bit dynamic range. The speech signal content was "Attention to the warning".

2.3. Apparatus

The four synthesized speech warnings were stored and played by the Winamp® software, v5.571, with a Toshiba® laptop model Sti Infiniti. Participants, seated in a chair in front of a desk, heard the speech sounds through headphones from Philips®, model SHP1900. The procedure took place at IADE's library, which was a quiet room.

2.4. Experimental design

The study concerns both within-subjects and between-subjects factors, so a mixed design was employed. The within-subjects factor is the type of speech signal that has four categories: female high-pitched (voice 1) and low-pitched (voice 4), male high-pitched (voice 2) and low-pitched (voice 3). The between-subjects factor is the participants' gender.

[1] http://www.oddcast.com/home/demos/tts/tts_example.php

2.5. Procedure

Participants were recruited to participate in a study, which intended to select a voice to be used in new technology-based warnings. Upon signing the inform-consent form, demographic information was gathered from the participants (e.g., age, gender, education).

The participant's task was to listen each one of the four synthetic speeches, one at time, and complete the MOS-X questionnaire for each speech. After listening to the four voices and answering the corresponding questionnaires, they had to rank the four voices by preference, taking into consideration the intended use. It was heavily emphasized that the warnings were to be heard in hazardous and/or emergency situations and that they would be activated, only in those circumstances, by proximity sensors.

The four speech signals were organized in two different sequences (see Table 1). The study participants were randomly assigned to one of these sets so that the order effect could be counterbalanced.

Table 1. Speech warnings' presentation order on each set

Set	Speech signals' presentation order			
1	Voice 1 Female high-pitched	Voice 2 Male high-pitched	Voice 3 Male low-pitched	Voice 4 Female low-pitched
2	Voice 3 Male low-pitched	Voice 4 Female low-pitched	Voice 2 Male high-pitched	Voice 1 Female high-pitched

The participants were instructed that they could replay the sounds as many times as they needed, but that they could not change the sound volume. There was no time limit to answer the MOS-X questionnaire. At the end of the procedure a brief follow-up, free-style interview was applied aiming to gather the participants' knowledge, previous experience and general opinions about the speech signals. The entire procedure lasted near to 10 minutes.

2.6. Measurements

The MOS-X questionnaire [10], which was translated to Portuguese, consists of 15 items, to be rated in a 7-points Likert-type scale, focusing on the following sound characteristics:

1. Listening effort (1 – Impossible even with much effort, to 7 – No effort required);
2. Comprehension problems (1 – All words hard to understand, to 7 – All words easy to understand);
3. Speech sound articulation (1 – Not at all clear, to 7 – Very clear);
4. Precision (1 – Slurred or imprecise, to 7 – Precise);
5. Voice pleasantness (1 – Very unpleasant, to 7 - Very pleasant);
6. Voice naturalness (1 – Very unnatural, to 7 – Very natural);
7. Humanlike voice (1 – Nothing like a human, to 7 - Just like a human);
8. Voice quality (1 – Significantly harsh/raspy, to 7 - Normal quality);
9. Emphasis (1 – Incorrect emphasis, to 7 – Excellent use of emphasis);
10. Rhythm (1 – Unnatural or mechanical, to 7 - Natural or rhythm);
11. Intonation (1 – Abrupt or abnormal, to 7 – Smooth or normal);
12. Trust (1 – Not at all trustworthy, to 7 – Very trustworthy);
13. Confidence (1 – Not at all confident, to 7 – Very confident);
14. Enthusiasm (1 – Not at all enthusiastic, to 7 – Very enthusiastic);
15. Persuasiveness (1 – Not at all persuasive, to 7 - Very persuasive).

The MOS-X questionnaire [10] provides the following scales: Intelligibility (average of items 1 to 4); Naturalness (average of items 5 to 8); Prosody (average of items 9 to 11); Social Impression (average of items 12 to 15) and Overall (average of all items).

Beyond the scope of the MOS-X questionnaire, participants were also requested to rank the four voices by preference (1 – most preferred to 4 – less preferred).

The dependent variables considered in this work were the participants' preference for a voice (voice ranking) and the MOS-X scales (i.e., Intelligibility; Naturalness; Prosody; Social Impression).

3. Results

The statistical analysis was performed in the IBM® SPSS® Statistics v19 and, for all of the analysis, a significance level of .05 was adopted.

3.1. Voice ranking

Since the participants ranked the four voices from 1 to 4, it was possible to determine the percentage of choices for the preferred voice. The results reveal that participants mostly choose (65%), as first option, the female high-pitched voice (voice 1). Only 20% of the participants chose the male high-pitched (voice 2) as first option, 10 % selected the female low-pitch (voice 4) and 5%, the male low-pitch (voice 3) as first option.

In spite of some constancy regarding the most preferred voice, Kendall's Concordance Coefficient reveals that there was no agreement among the 20 participants with respect to how they rank the four voices (W = .117, $X^2(3) = 7.020$, $p = .071$).

Regarding the preference of the male and female listeners for the talker's voice gender, results show that the female high-pitched voice was the first option for both females and males (the lower rank is the most preferred). However, regarding the other ranks, the preferences differ (see Figure 1). For example, the voice that emerged in second place for female and male participants was, respectively, the male high-pitched and low-pitched voices (voice 2 and 3). The least preferred voice for female listeners was the male low-pitched (voice 3) and for male listeners was the male high-pitched (voice 2).

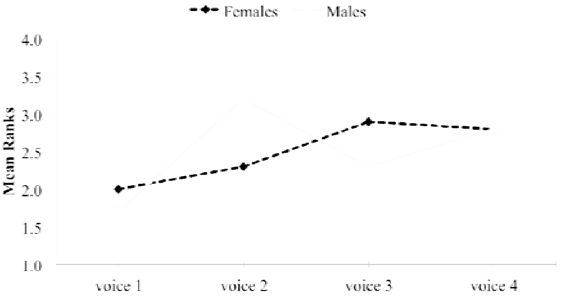

Figure 1. Voices' mean ranks by gender

In order to evaluate the gender's effect on the preference for male or female voices (regardless of the voices' pitch), the Chi-square test for homogeneity was used. The crosstabulation is presented in Table 2. The results of the Chi-square test revealed that there was no statistical significant difference between the two genders in what concerns the proportion of preferences for the female or male voices ($X^2(1) = 0.267$, $p = 1.000$).

Table 2. Crosstabulation table for participant's gender by preference for male and female voices

	Voice preference		
Participants' Gender	Female voice	Male voice	*Total*
Feminine	7 (70%)	3 (30%)	10
Masculine	8 (80%)	2 (20%)	10
Total	14	6	20

3.2. MOS-X ratings

To assess the quality of the synthesized speeches under evaluation and to gain further knowledge about what aspects differ the MOS-X questionnaire was used.

Figure 2 depicts the median ratings for the items of the MOS-X and in Figure 3 we can compare the four voices mean ratings for the five MOS-X scales.

Both female voices (1 and 4) presented the higher median ratings for all items, and were never below the central point (the value 4) of the Likert-type scale. The median ratings for voice 1 were overcome only once, by voice 4, on the item 14 (Enthusiasm), which was one of the worst rated items on all voices. The Intelligibility items (1 to 4) attained the higher median ratings for all voices, in opposition to the Naturalness, Prosody, and Social Impression items that presented lower median ratings. The voice presenting more items that were poorly rated was voice 3 (male low-pitched).

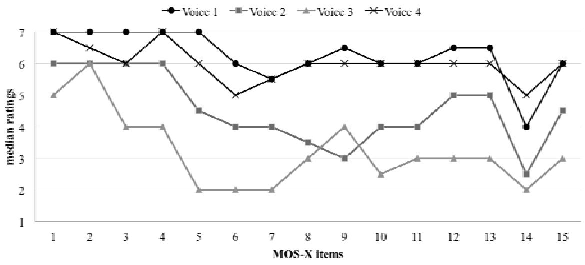

Figure 2. Voices' median ratings for the MOS-X questionnaire

When comparing the mean ratings attained by the four voices, regarding the MOS-X scales (see Figure 3), results show that both female voices 1 and 4 (voices with the highest overall mean rating) attained very close mean ratings in all scales, and only in Social Impression scale the voice 4 attained higher mean than voice 1. The highest mean ratings occurred for Intelligibility, followed by Prosody scales in both female voices, and the lowest for Social Impression for voice 1 and Naturalness for voice 4.

The male voices attained intermediate mean ratings for all scales. Voice 3 (male low-pitched) was always the worst rated, strongly penalized in what regards to Naturalness, Prosody and Social Impression.

Table 3. Friedman tests and multiple comparisons results for MOS-X scales

	Friedman tests		Multiple comparisons					
MOS-X Scales	$X^2(3)$	*p*	Voice 1 vs. 2	Voice 1 vs. 3	Voice 1 vs. 4	Voice 2 vs. 3	Voice 2 vs. 4	Voice 3 vs. 4
Intelligibility	31.451	< .001	.061	< .001	1.000	.397	.042	< .001
Naturalness	41.313	< .001	.013	< .001	1.000	.086	.086	< .001
Prosody	32.037	< .001	.009	< .001	1.000	.035	.035	< .001
Social impression	28.665	< .001	.141	.002	1.000	.004	.004	< .001
Overall	38.606	< .001	.042	< .001	1.000	.120	.013	< .001

Table 4. Mann-Whitney tests results regarding gender effect on MOS-X scales

	Voice 1			Voice 2			Voice 3			Voice 4		
	U	W	*p*	U	W	*p*	U	W	*p*	U	W	*p*
Intelligibility	49.5	104.5	.967	33.0	88.0	.210	25.0	80.0	.058	46.5	101.5	.790
Naturalness	48.5	103.5	.925	39.0	94.0	.430	39.5	94.5	.443	41.0	96.0	.518
Prosody	49.0	104.0	.959	42.0	97.0	.571	34.0	89.0	.231	38.5	93.5	.400
Social impression	30.0	85.0	.136	47.0	102.0	.839	40.5	95.5	.491	33.5	88.5	.222
Overall	41.5	96.5	.541	42.0	97.0	.565	37.5	92.5	.362	38.0	93.0	.382

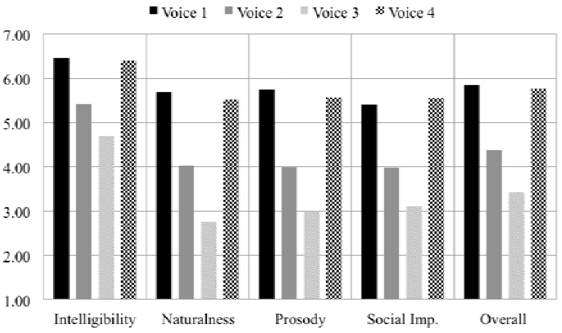

Figure 3. Voices' means ratings for the MOS-X scales

To determine if the ratings differed significantly according to the voice being evaluated, the Friedman's two-way ANOVA by ranks tests were used, for each one of the five MOS-X scales. The results show significant statistical differences between at least two voices ($p < .05$) for each of the scales (see Table 3).

Post-hoc multiple comparisons of mean ranks were conducted, through the Bonferroni-Dunn test, for all the scales in order to identify the voices that differ. As shown in Table 3, significant differences were found between the male and female low-pitched voices (3 vs. 4), and between the female high-pitched and male low-pitched voices (1 vs. 3), for all MOS-X scales. In what concerns the female and the male high-pitched voices (1 vs. 2), significant differences were achieved only for Naturalness, Prosody and Overall scales. With respect to male high-pitched and female low-pitched voices (2 vs. 4) significant differences were attained for all scales with the exception of Naturalness. Regarding both high and low-pitched male voices (2 vs. 3) differences were found only for Prosody and Social impression. Relating to the remaining comparison, for both high and low-pitched female voices (1 vs. 4), no statistical significant differences were found.

The Mann-Whitney tests (see Table 4) show that gender did not have a significant effect on any of the scales of the MOS-X questionnaire ($p > .05$ for all scales and for all voices).

3.3. Follow-up interviews

When asked, in the follow-up interview, about their preferences, the participants that preferred female voices stated that they were used to hearing the same type of voice in public transportation and/or computer software.

4. Discussion and conclusions

The goal of this exploratory study was to gain knowledge that can be used in the selection of a synthesized voice for a subsequent VR-based compliance study with personalized warnings. The approach taken involved the ranking of four speech warnings, varying regarding the speaker's gender and voice pitch, and its rating, through the MOS-X questionnaire. The male and female preference for a voice's gender was also investigated.

This study used only two female and two male voices, at two different pitches, and involved a reduced sample, so the results should be seen as preliminary and cannot be generalized.

The summary results of voices evaluation in what concerns the ranking and the rating of the MOS-X scales are presented in Table 5. The results suggest that participants, when asked to rank the four voices, preferred the female high-pitched voice (voice 1). In second place emerged the male low-pitched voice (voice 3). However, the differences on the mean ranks amongst the three voices worst positioned are negligible, suggesting that there were different opinions in how to rank them. The female high-pitched voice was the first choice for both females and males. When ignoring the voices' pitch, no significant gender effect was verified regarding the talker's voice gender.

The assessment of the MOS-X scales revealed how the voices differ and gave an overall mean rating for each voice. The voice with higher overall mean was the female high-pitched voice, similar to what was found with the rankings. However, the second highest valued voice was the female low-pitched, which was the less preferred in the ranking classification (see Table 5).

Table 5. Summary results of voices evaluation according to the rankings and the rating of the MOS-X scale

Position	Mean ranks*	MOS-X Overall
1º	Voice 1 (1.85)	Voice 1 (5.84)
2º	Voice 3 (2.60)	Voice 4 (5.77)
3º	Voice 2 (2.75)	Voice 2 (4.38)
4º	Voice 4 (2.80)	Voice 3 (3.42)

* Lower value is the most preferred

These results suggest that the preference could rely on other factors than those assessed by the MOS-X questionnaire, such as familiarity with the female voices, in auditory warnings presented in daily routines. The participants on the follow-up interview highlighted this aspect. Additionally, the fact that the decision regarding the preferences for alternative voices was made in a hierarchical manner and, therefore, no ties were allowed, can also be on the root of these apparently contradictory results.

The Intelligibility aspect was the one that was better rated for all voices, while the worst rated ranged from Social Impression for voices 1 and 2 (high-pitched), Naturalness for voices 4 and 3 (low-pitched), and Prosody also for voice 2 (see Figure 3). Further analysis would be required to understand the degree to which the speeches' aspects, measured by MOS-X, are correlated to the participants' preferences.

A limitation of this evaluation is that speech signals were evaluated as being isolated from background noise. Under normal circumstances, surrounding sounds can mask auditory warnings. Further studies should also address the message content, perceived urgency and reaction time. Additionally, other speech parameters (e.g., rate and rhythm) should be investigated.

The MOS-X seemed appropriate to evaluate the speech warnings, regardless of the shortness of the speech could have affected the participants' ability to evaluate the speeches.

Although emphasizing that the presented results cannot be generalized, this study provided data-based arguments favoring the adoption of the high-pitched synthesized female voice to be used on the speech signals in the subsequent VR-based compliance study.

Acknowledgements

This study was developed in the scope of a research project (PTDC/PSI-PCO/100148/2008) funded by the Portuguese Science Foundation (FCT).

References

[1] C. Wickens, et al., An introduction to human factors engineering, 2nd ed., Pearson Prentice Hall, Upper Saddle River, NJ, 2004.

[2] E. Duarte, et al., Behavioral compliance in Virtual Reality: Effects of warning type, in: Advances in Cognitive Ergonomics. Advances in Human Factors and Ergonomics Series, D. B. Kaber and G. Boy, eds., CRC Press/Taylor & Francis, Boca Raton, FL, 2010, pp. 812-821.

[3] E. Haas, and J. Edworthy, An introduction to auditory warnings and alarms, in: Handbook of Warnings, M. S. Wogalter, ed., Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 2006, pp. 189-198.

[4] E. Hellier, et al., The perceived urgency of speech warnings: Semantics versus acoustics. Human Factors, 44 (2002), 1-17.

[5] J. Edworthy, and E. Hellier, Complex nonverbal auditory signals and speech warnings, in: Handbook of Warnings, M. S. Wogalter, ed., Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 2006, pp. 199-220.

[6] J. Edworthy, et al., Acoustic, semantic and phonetic influences in spoken warning signal words. Applied Cognitive Psychology, 17 (2003), 915-933.

[7] J. M. Noyes, E. Hellier, and J. Edworthy, Speech warnings: a review. Theoretical Issues in Ergonomics Science, 7 (2006), 551-571.

[8] J. Wilding, and S. Cook, Sex differences and individual consistency in voice identification. Perceptual and Motor Skills, 91 (2000), 535-538.

[9] K. Miller, Channel interaction and the redundant-targets effect in bimodal divided attention. Journal of Experimental Psychology: Human Perception and Performance, 17 (1991), 160-169.

[10] M. D. Polkosky, and J. R. Lewis, Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X. International Journal of Speech Technology, 6 (2003), 161-182.

[11] M. S. Wogalter, and C.B. Mayhorn, The future of risk communication: Technology-based warning systems, in: Handbook of Warnings, M. S. Wogalter, ed., Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 2006, pp. 783-794.

[12] M. S. Wogalter, and S. L. Young, Behavioural compliance to voice and print warnings. Ergonomics, 34 (1991), 79-89.

[13] M. S. Wogalter, and V.C. Conzola, Using technology to facilitate the design and delivery of warnings. International Journal of Systems Science, 33 (2002), 461-466.

[14] M. S. Wogalter, et al., Personalization of warning signs – The role of perceived relevance on behavioral compliance. International Journal of Industrial Ergonomics, 14 (1994), 233-242.

[15] R. S. Brazegar, and M. S. Wogalter, Intended carefulness for voiced warning signal words, in: Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting, 1998, pp. 1068-1072.

[16] T. W. Whipple, and M. K. McManamon, Implications of using male and female voices in commercials: An exploratory study. Journal of Advertising, 31 (2002), 79-91.

[17] V. C. Conzola, and M. S. Wogalter, Using voice and print directives and warnings to supplement product manual instructions. International Journal of Industrial Ergonomics, 23 (1999), 549-556.