SUBJECTIVE EVALUATION OF THE NATURALNESS AND ACCEPTABILITY OF THREE TEXT-TO-SPEECH SYSTEMS IN FRENCH

ROBERT, Jean-Marc*, CHOINIERE, Andre*, DESCOUT, Raymond**
* Ecole Polytechnique de Montréal, Département de génie industriel, Case postale 6079, Succurcale A, Montréal, Québec, Canada H3C 3A7
** Department of Communications of Canada-Canadian Workplace Automation Research Centre, 1575 boul. Chomedey, Laval, Québec, Canada H7V 2X2

ABSTRACT

Fifty-three subjects evaluated the naturalness and acceptability of three text-to-speech systems in French which are Loquax, Télévox, and Multivox. Their evaluation bore upon 25 5-point perceptual scales. The main results indicate that the performances of the three systems are very similar to each other. They also reveal that the average evaluations of the three systems over the 25 scales are negative. Loquax: -0.14; Télévox: -0.19; Multivox: -0.38 and do not appear to differ significantly from each other. An interesting result is that the oldest system, Loquax, received the largest number of -2 ratings from the subjects, whereas Multivox obtained the largest number of +2 ratings. Some research directions are proposed for the future.

INTRODUCTION

This paper presents a subjective evaluation of the naturalness and acceptability of the voice of three text-to-speech systems in French: Loquax (1983), Télévox (1984), and Multivox (1989). Its objectives are threefold: identify the strengths and weaknesses of the voice of each system as they are perceived by French-speaking Quebecois, compare the subjective quality of the voices on different issues, and develop a tool for evaluating synthetic voices in French.

The quality of synthetic voice does matter for numerous reasons. It directly affects the user’s performance and satisfaction with the system that generates the voice. Elements that require more attention at work, or provoke hesitations, errors, or distrust towards the system increase the frustration and workload of the users and reduce their efficiency and satisfaction. The quality of voice also determines the acceptance of a service by the users and its full exploitation in the workplace. Indeed, if it is poor, the system may be rejected or sub-utilized as this happens with interactive computer systems supplied with a bad user interface. The quality of voices is also critical in domains where security is of the utmost importance (e.g., air traffic control, piloting, supervisory control of power plants, etc.); thus, it must be tested very carefully before the system is implemented. Finally, the quality of synthetic voice is of primary interest for system designers. Its evaluation allows them to clearly identify the different assets and deficiencies of their systems, measure the benefits of competitive solutions, compare their systems with other systems, and if needed, pursue the work on issues that require an improvement.

Because voice-output applications are still relatively novel, the voice quality of text-to-speech systems in French does not seem to have been evaluated so far, at least by French-speaking Quebecois. As a consequence, no tool for measuring the perception of voice quality in French yet exists, and we do not know comparative results in French for ranking synthetic voices. This study is a first step towards filling this gap. It starts from the tool developed by [1] for evaluating the quality of synthetic voices in English (see Appendix 1). The tool was translated in French and modified so as to be both more precise, cover a wider range of judgement scales, and be well-adapted to the French language (see Appendix 2). More details on the tool appear in the next paragraphs.

Furthermore, this study tackles situations where the text-to-speech systems convey guiding information over a telephone. Authors [2] have shown that the perceived quality of synthetic voices depends on situations where the voices will be used.

METHOD

Three text-to-speech systems in French were evaluated. These systems are Loquax from INRS Télécommunications (Québec, 1983), Télévox from Élan Informatique (France, 1984), and the most recent one, Multivox from the CNRT (France, 1989). Unfortunately, the system SA 101 from Infovox (Sweden, 1986) was not available for this study. Loquax is a stand-alone system, Télévox runs on a PC supplied with a speech processing interface board, and Multivox is an experimental system running on a PC system. These systems were using three different voice speeds that fall into the standard range of speed for each system.

Subjects. Fifty-three students from the department of industrial engineering of Ecole Polytechnique de Montréal participated in the study. They were recruited through personal contacts (anads, classrooms, and were not paid for their participation).
time. They were divided into three equivalent groups of 16, 22, 15 subjects for evaluating, respectively, the systems the Télévox, Loquax, and Multivox. They were all French speaking and were unfamiliar with synthetic voices. 72% were male and 28% were female; their age varied between 20 and 29 years (\( X = 22.2 \)).

Messages. Each system generated nineteen different messages in French over the telephone in an automatic e-mail reading task. Eleven messages guided the subjects through the experiment and 8 (forming three main messages) contained familiar information regarding the students’ activities at Ecole Polytechnique. Each message lasted less than 7 seconds. All the messages prepared on each system were first recorded with a high-fidelity tape recorder (DAT Sony). These copies were then digitized with a speech processing interface board (using the 12 kHz ADPCM compression algorithm from the Texas Instrument speech board). The copies were stored in different files and were accessed at the appropriate time by a PC monitor system.

Questionnaire. Data were collected through a questionnaire in French. Its first part briefly presents the goal of the experiment. The second part collects personal data about the subjects: age, sex, mother tongue, training, previous knowledge or experience with speech technology, and personal deficiencies in audition, if any. The third part describes the procedure for using the system and completing the questionnaire. It also includes two scales made of bipolar adjective pairs (e.g., bad vs good) where the negative and the positive adjectives alternate between the two poles of the scale. This is aimed at preventing the subjects from being biased in perceiving some adjectives as positive or negative. The negative and the positive adjectives were transferred afterwards onto the proper side of the scale for data compilation and analysis (see Appendix 3). The adjective pairs are widely borrowed from the 17 scales proposed by [1]; some scales were split, some were grouped, some were deleted, some were simply adapted for French, and new scales were added (see Appendix 3 for the correspondance between our list of adjectives and that of [1]). Twenty scales have opposite poles; 3 scales do not: \#5 (quebecois vs parisian), \#7 (slow vs rapid), and \#16 (young vs old). The last two scales may correspond to a wrap-up of the evaluation. \#24 measures the subjective acceptability of the synthetic voice for presenting information, and \#25 measures the frequency of use of the system, if it was available to the subject. Each perceptual judgement was made on a 5-point scale (2:very, 1:rather, 0:neutral, -1:rather, -2:very).

Procedure. The subject could do the experiment at any time during the day and from any location where a touch-tone telephone was available. The procedure for completing the experiment was clearly described in the questionnaire that was handed to each subject. After having gone through the first three parts of the questionnaire, the subject dialed a phone number and an access code for reaching the system, listened to the series of messages from the system (he/she could ask for as many repetitions as desired), and evaluated the synthetic voice. The experiment lasted between 8 and 10 minutes.

RESULTS

Results reflect the evaluations given by subjects exposed for the very first time to synthetic voices. Figure 1 presents the results of the subjective evaluation of the three text-to-speech systems in French on 25 scales. It thus shows the performance profile of each system, and allows us to compare and classify the three systems from a user point of view. The most striking result is that the three systems have a very similar profile. Considering a difference of 0.35 or less on each 5-point scale between the two most different systems among the three, the results are very close to each other on the scales \#4 (0.18; inadequate vs adequate), \#6 (0.27; old vs warm), \#7 (0.32; efficient vs inefficient), \#14 (0.19; gross vs polite), \#15 (0.21; choppy vs continuous), \#17 (0.08; hard vs soft), \#20 (0.28; irritating vs stimulating), \#22 (0.14; hostile vs friendly), and \#25 (0.31; frequency of use: never vs always). In turn, considering a difference of 0.75 or more between the two most different systems among the three, the differences are the largest on the scales \#2 (0.76; difficult to understand vs easy to understand), \#5 (0.76; quebecois vs parisian), \#7 (1.16; slow vs rapid), \#10 (0.88; tense vs relax), and \#19 (1.06; popular vs knowledgeable).

When one examines the scales where a difference of at least 0.5 exists between a system and its two rivals, one discovers that the Télévox voice is the most rapid (\#7: 0.75 vs -0.41 and -0.27), the most tense (\#10: -0.81 vs -0.23 and -0.09), the most boring (\#12: 0.38 vs -0.91 and -0.93), and the least foreign (\#18: 0.50 vs -1.05 and -1.07). The Loquax voice is the least knowledgeable (\#19: 0.32 vs 1.38 and 0.93), and the most unacceptable (\#24: -0.45 vs 0.06 and 0.40). Finally, the Multivox voice is the easiest to understand (\#2: 0.13 vs -0.63 and -0.64).

The three largest differences bear upon the scale \#7 between Télévox & Loquax (1.16), Televox & Multivox (1.02), and the scale \#19 between Loquax & Télévox (1.06), and the scale \#24, between Loquax & Multirox (0.95).

Another important result from Figure 1 is that the evaluation points on the 25 scales are mostly negative. Actually, the average evaluations of the three systems over the 25 scales taken altogether are -0.14 (standard deviation = 0.924) for Multivox, -0.39 (S.D. = 0.882) for Télévox, and -0.39 (S.D. = 0.867) for Loquax (which is the oldest system). With such averages and important standard deviations, there should not be significant differences between the average evaluations of the three systems.
Quality of synthetic voices in French
The current study seems to be the first to evaluate the naturalness and acceptability of synthetic voices in French by French-speaking Quebecois. So, let’s combine the results of the three systems on each scale to get a portrait of the advancement of synthetic voice in French from a user viewpoint. The main characteristics of synthetic voices in French are the following: the voice is cold (#6: -0.89), boring (#13: -0.75), choppy (#15: -0.72), foreign (#18: -0.89), and artificial (#23: -1.15). The main positive characteristics are that the voice is polite (#14; 0.96) and knowledgeable (#19; 0.81). On the average, subjects requested 2.06 message repetitions (Télécyox: 3.22; Loquax: 1.23; Multivox: 1.85) for one or several of the 3 main messages that could be repeated during the experiment.

Strong ratings
Figure 2 presents for each system the frequency of +2 and -2 ratings on each scale. The interest of such results is to highlight the systems and the scales that generate the strongest ratings from the subjects. These extreme ratings carry more meaningfulness in system design. So it appears that the Loquax system generated exclusive -2 ratings (i.e., without +2 rating on the same scale) on 16 scales (#1, 3, 6, 7, 9, 10, 12, 13, 15, 16, 17, 18, 19, 21, 22, 23). Compared to 9 scales for Télécyox (#4, 6, 9, 10, 12, 15, 20, 21, 25), and to 8 scales for Multivox (#6, 12, 13, 15, 17, 18, 20, 23). As a corollary, Loquax generated exclusive +2 ratings on 3 scales only (#5, 14, 22) as compared to 5 scales for Télécyox (#3, 7, 14, 16, 19), and 8 scales for Multivox (#5, 8, 11, 14, 19, 21, 22, 24). These results confirm those presented in Figure 1.

Each system also generated several +2 and -2 ratings on the same scale, which reflects important within-a-group differences towards the same issue of a system. Loguax received such ratings on the scales #4, 13, 14, 16, 18, 19, 20, 23. Multivox, on the scales #3, 9, 10, and 25, and Télécyox, on the scales #2, 5, 8, 17, 23, and 24.

Finally, one or several systems received more than 20% of -2 ratings on 9 scales: #3, 6, 9, 12, 13, 15, 16, 20, 23. The scales #6 (cold vs warm), #15 (choppy vs continuous) and #23 (artificial vs natural) had the highest percentages of such ratings for the three systems. Only the scales #5 (quebecois vs parisian), #7 (slow vs rapid), #14 (gross vs polite), and #15 (choppy vs continuous) deserved more than 20% of +2 ratings for one or several systems; the scale #14 obtained the highest percentages for the three systems.

Acceptance of the voice
The three systems have a par1sian accent (#5; 0.56, 0.14, 0.93) according to French-speaking Quebecois subjects. This accent does not seem to be embarrassing for the perception of voice quality since Multivox is the best system according to the subjects while it has the most par1sian accent (#5; 0.93). The par1sian accent is not a barrier for naturalness and acceptability of synthetic voices heard by Quebecois. Notice also that the subjects had some difficulty in identifying the voice accent; three subjects perceived an african accent in Multivox, and one subject perceived the same accent in Loquax.

CONCLUSION
The three synthetic voices evaluated in this study were quite intelligible to the subjects, however they call for significant improvements in order to be more natural and acceptable. Thus, researchers should work on improving voice quality before concentrating on the regional accents of synthetic voices.

ACKNOWLEDGEMENTS
This study is part of a joint collaboration between Ecole Polytechnique de Montréal and Communications Canada - Canadian Workplace Automation Research Centre.

The authors thank Gilles Bergeren for helping to analyse the data and construct the figures.

REFERENCES

APPENDIX 1
Adjective pairs of Nusbaum et al.(1984)
1. Interesting/Boring
2. Coarse/Polished
3. Choppy/Continuous
4. Young/Old
5. Hard/Easy
6. Comfortable/Frustrating
7. Gentle/Harsh
8. Confusing/Clear
9. Annoying/Pleasant
10. Smooth/Rough
11. Foreign/American
12. Knowledgeable/Uneducated
13. Stimulating/Tiring
14. Distracting/Improves concentration
15. Halting/Fluent
16. Friendly/Hostile
17. Grating/Melodious
APPENDIX 2

Adjective pairs used in this study

1. Mauvaise/Bonne
2. Difficile à comprendre/Facile à comprendre
3. Déplaisante/Plaisante
4. Inadéquate/Adéquate
5. Québécoise/Parisiennne
6. Froide/Chaleureuse
7. Lente/Rapide
8. Inefficace/Efficace
9. Désagréable/Agréable
10. Tendue/Détendue
11. Insatisfaisante/Satisfaisante
12. Passive/Vivante
13. Ennuyante/Intéressante
14. Froide/Chaleureuse
15. Saccadée/Continue
16. Vieille/Jeune
17. Dure/Douce
18. Étrangère/Familière
19. Populaire/Instruite
20. Irritante/Stimulante
21. Nuit à la concentration/Aide à la concentration
22. Hostile/Amicale
23. Artificielle/Naturelle
24. Présentation d’information: Inacceptable/Acceptable
25. Utilisation: Jamais/Toujours

APPENDIX 3

Relation between the adjective pairs of Nusbaum et al. and the adjective pairs used in this study

<table>
<thead>
<tr>
<th>This study</th>
<th>NUSBAUM et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td># 3</td>
<td>translation of # 9</td>
</tr>
<tr>
<td>13</td>
<td>&quot; &quot; 1</td>
</tr>
<tr>
<td>16</td>
<td>&quot; &quot; 4</td>
</tr>
<tr>
<td>19</td>
<td>&quot; &quot; 12</td>
</tr>
<tr>
<td>20</td>
<td>&quot; &quot; 13</td>
</tr>
<tr>
<td>21</td>
<td>&quot; &quot; 14</td>
</tr>
<tr>
<td>22</td>
<td>&quot; &quot; 16</td>
</tr>
<tr>
<td># 5</td>
<td>modification of # 11</td>
</tr>
<tr>
<td>8</td>
<td>&quot; &quot; 8</td>
</tr>
<tr>
<td># 10, 11</td>
<td>split from # 6</td>
</tr>
<tr>
<td># 17</td>
<td>grouping of # 5, 7, 10</td>
</tr>
<tr>
<td>15</td>
<td>&quot; &quot; 3, 15, 17</td>
</tr>
<tr>
<td></td>
<td>deletion # 2</td>
</tr>
</tbody>
</table>

Addition: #1, 2, 3, 4, 6, 7, 9, 12, 14, 18, 23, 24, 25