The comparison between the deletion-based methods and the mixing-based methods for audio CAPTCHA systems

Takuya Nishimoto¹, Takayuki Watanabe²

¹Graduate School of Information Science and Technology, The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
²Department of Communication, Division of Human Science, School of Arts and Sciences, Tokyo Woman’s Christian University
2-6-1 Zenpuku, Suginami-ku, Tokyo, 167-8585, Japan
nishi@hil.t.u-tokyo.ac.jp, nabe@lab.twcu.ac.jp

Abstract

Audio CAPTCHA systems, which distinguish between software agents and human beings, are especially important for persons with visual disability. The popular approach is based on mixing-based methods (MBM), which use the mixed sounds of target speech and noises. We have proposed a deletion-based method (DBM) which uses the phonemic restoration effects. Our approach can control the difficulty of tasks simply by the masking ratio. In this paper, we propose a design principle of CAPTCHA, according to which the tasks should be designed so that the large difference of performance between the machines and human beings can be provided. We also show the experimental results that support the hypotheses as follows: (1) only using MBM, the degree of task difficulty can not be controlled easily, (2) using DBM, the degree of task difficulty and safeness of CAPTCHA system can be controlled easily.

Index Terms: Security, Visual Impalement, Speech Recognition, CAPTCHA, Mixing-based Method, Deletion-based Method

1. Introduction

CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart) are popular security techniques on the Web that prevent automated programs from abusing online services.

A typical CAPTCHA is an image containing several distorted characters that appears at the Web forms [1]. Users are asked to type the distorted dirty characters to let the system know that they are human. Image-based CAPTCHAs are, however, preventing Web use of persons with visual disability [2]. Audio CAPTCHAs were created to solve this accessibility issue.

In this paper, we describe an attempt to create better audio CAPTCHA tasks in consideration of safety and usability. The difference of recognition performance between human and machine can be a criterion of safeness. Mental workload of human in listening speech was also investigated from the perspective of usability.

It is important to design the audio CAPTCHAs that the audio files are difficult to be recognized automatically by machines. In other words, it is necessary that the speech recognition performance with machines should be lower than the intelligibility of human, as shown in Figure 1.

Statistical method is the mainstream in speech recognition technology. Correct answers can be estimated from the distorted speech if there are enough learning data and the human annotations for them [3]. Many improvements in speech recognition are also proposed to cope with the noise robustness. In such situation, distorted audio files for CAPTCHAs should not be created based on trial and error, but rather be created systematically.

From another point of view, audio CAPTCHAs should not increase the mental workload of the users. The workload of listening to the tasks may increase if they are difficult to listen. Workload of memorizing many characters is also unignorable. Human auditory sensation and language cognition should be investigated to deal with the problems.

Top-down knowledge of human acts to guess the information in an incomplete stimulus. In the case of visual sensation, we can complement an image with common knowledge about the character and the vocabulary, even if part of the character image is missing, or part of the word is hidden. Similar process exists in speech understanding.

To make use of the ability of human perception, we focused on a method of speech processing that we call “deletion-based method” or DBM. There are new ways to generate the tasks in which human can use the ability of top-down knowledge, though noises and reverberations are often added to make speech difficult to listen (we call this approach “mixing-based method” or MBM). For example, we can delete some parts of an audio file on temporal axis little by little. If every 30 msec over a period of 100 msec the audio file is replaced with si-
ience, it can be considered that the 30\% of the information was deleted. If the ratio of remained sections go down, the degree of listening difficulty may increase. In this “temporal induction” method, the audio listening tasks can be controlled easily with “the ratio of information.”

Miller and Licklider [4][5] combined interrupted speech and interrupted noise maskers, i.e., speech and noise are temporally interleaved to form a continuous signal. They reported that perceived continuity of the speech signal was increased by this manipulation. This “phonemic restoration” may help the human listening, though the effect does not affect the recognition performance of machine. In other words, it serves to enlarge the performance difference of human and machine.

Examples of audio files created with deletion-based method are shown in Figure 2. Seven digit numbers are contained in a utterance. The 30\%, 50\%, and 70\% of the utterance is left and the gaps are filled with white noise, respectively.

![Figure 2: Examples of audio files created with deletion-based method (D30, D50 and D70).](image)

2. Comparison between DBM and MBM

A comparative experiment of deletion-based method (DBM) and the mixing-based method (MBM) was performed as the Exp.1. The objective is to compare the intelligibility and mental workload of DBM with those of MBM, and to examine the effect of SNR (signal-to-noise ratio) in MBM.

The task is Japanese connected digit recognition which conforms to CENSREC-4 [6]. A set of human intelligibility test consists of 75 utterances, which include three digit, four digit and five digit numbers. The balance of the occurrence of numbers is considered and the utterances were spoken by many people including men and women. Acoustic presentation was given by the headphone at the subject’s preferred reference loudness level.

The subjects were undergraduate students. They were divided into three groups (G1, G2 and G3) with different MBM task conditions. Each group consists of five persons.

The experiment consists of two parts, as shown in Table 1. At first, a set of DBN tasks with 30\% of unmasked information condition (D30) was given, followed by a set of MBM tasks with an SNR condition.

The MBM task condition consists of M0, Mm10 and Mm20, which correspond to SNR of 0dB, -10dB and -20dB, respectively.

To make the disturbing signals, utterances of Japanese sentence were fragmented as short periods (250msec), shuffled and combined. The same-gender disturbance voices as the target utterances were selected. The start of the disturbing signal is one second before than that of the target utterance, and the end of the disturbing signal is one second after than that of the target utterance.

NASA-TLX [7] was used to evaluate mental workload. In this method, the trial subjects answer the amount of workload by rating the following subscales: Mental, Physical, and Temporal Demands, Frustration, Effort, and Performance. The weights of the subscales are derived for each participant by placing an order how the 6 dimensions are related to their personal definition of workload.

According to the ANOVA (analysis of variance) results as shown in Table 2, the difference of the intelligibility among the groups was marginally significant, although the conditions of DBN were same among the groups. The intelligibility of MBM also showed significant difference among the conditions of M0, Mm10 and Mm20 ($p < 0.05$). The multiple comparison test (LSD and Tukey HSD), however, indicated only the significance between M0 and Mm10.

![Table 2: The ANOVA results of intelligibility (%) (Exp.1).](table)

Concerning the mental workload, the effect of the MBM task difference was investigated. To cancel the individual difference, subtraction of DBM (D30) score from MBM (M0, Mm10 and Mm20) score was performed. The results shown in Table 3 suggested that the WWL of M0 was lower than that of D30. According to the ANOVA, however, significant difference was not observed.

The result of every subject was shown in Figure 3 and Figure 4. The hypothesis that the SNR condition of MBM gives significant difference for intelligibility or mental workload was not supported.

![Table 3: The average and SD of WWL difference (%) (Exp.1).](table)

HTK (Hidden Markov Model toolkit) 1 was used for the ASR experiment. Statistical acoustic models for ASR were trained with 8440 utterances that were processed using MBM. This means the assumption that the correct answers for all audio

1http://htk.eng.cam.ac.uk/
Table 1: The experimental setup and the average (SD) of intelligibility (%) (Exp.1). The V1 and V2 correspond to the vocabularies.

<table>
<thead>
<tr>
<th>Group</th>
<th># Persons</th>
<th>Trial 1</th>
<th>Trial 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>5</td>
<td>V1-D30</td>
<td>82.9 (7.0)</td>
</tr>
<tr>
<td>G2</td>
<td>5</td>
<td>V1-D30</td>
<td>69.1 (9.1)</td>
</tr>
<tr>
<td>G3</td>
<td>5</td>
<td>V1-D30</td>
<td>74.7 (7.6)</td>
</tr>
</tbody>
</table>

Figure 3: Intelligibility of every trial of every subject (Exp.1).

Figure 4: Difference of workload of every subject (Exp.1).

files were given to break this CAPTCHA system. The evaluation of ASR was performed using 1001 utterances that were not included in training data. The trained HMM consists of 18 states with 20 mixtures of Gaussian for each number.

D30, D50 and D70 indicate the conditions that the unmasked information ratios are 30%, 50% and 70%, respectively.

The sentence recognition performance of the tasks using HTK is shown in Table 6. The ASR performances were 40.5%, 79.0% and 89.4% for D30, D50 and D70, respectively. The average and SD of intelligibility by human is shown in Table 7. The average (SD) of normalized WWL is shown in Table 8.

Table 5: The experimental setup (Exp.2). The V1, V2 and V3 correspond to the vocabularies.

<table>
<thead>
<tr>
<th>Group</th>
<th># Persons</th>
<th>Trial 1</th>
<th>Trial 2</th>
<th>Trial 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>5</td>
<td>V1-D70</td>
<td>V2-D50</td>
<td>V3-D30</td>
</tr>
<tr>
<td>G2</td>
<td>6</td>
<td>V1-D30</td>
<td>V2-D70</td>
<td>V3-D50</td>
</tr>
<tr>
<td>G3</td>
<td>6</td>
<td>V1-D50</td>
<td>V2-D30</td>
<td>V3-D70</td>
</tr>
</tbody>
</table>

Table 6: The sentence recognition performance of Exp.2 tasks using HTK (%).

<table>
<thead>
<tr>
<th>Condition</th>
<th>D30</th>
<th>D50</th>
<th>D70</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>40.5</td>
<td>79.0</td>
<td>89.4</td>
</tr>
</tbody>
</table>

Table 7: The human intelligibility (%) (Exp.2).

<table>
<thead>
<tr>
<th>Condition</th>
<th>D30</th>
<th>D50</th>
<th>D70</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average (SD)</td>
<td>84.2 (6.5)</td>
<td>97.4 (2.7)</td>
<td>99.0 (2.0)</td>
</tr>
</tbody>
</table>

As the result of ANOVA shown in Table 9, the human intelligibility showed significant difference among the conditions. According to the multiple comparison test (LSD), difference of intelligibility was significant between D30-D50 and between D30-D70.

Concerning mental workload, WWL scores were normalized within every subject so that their average and standard deviation became 50 and 10 respectively at first, followed by the comparison among the conditions. As the ANOVA results shown in Table 9, significant difference of normalized-WWL between the groups was observed. According to the multiple comparison test (LSD), difference of normalized-WWL was significant between D30-D50 and D30-D70.

3. Comparison between Human and Machine with DBM

An experiment that compares the intelligibility and mental workload between the conditions of DBM was performed. The automatic speech recognition (ASR) performance was also examined. Experimental setup, shown in Table 5, is similar to that of previous section, though this experiment is to compare between the conditions in each subject. The subject was seventeen undergraduate students.

HTK was used for the ASR experiment. Statistical acoustic models for ASR were trained with 8440 utterances that were processed using DBM. This means the assumption that the correct answers for all audio files were given to break this CAPTCHA system. The evaluation of ASR was performed using 1001 utterances that were not included in training data. The trained HMM consists of 18 states with 20 mixtures of Gaussian for each number.

D30, D50 and D70 indicate the conditions that the unmasked information ratios are 30%, 50% and 70%, respectively.

The sentence recognition performance of the tasks using HTK is shown in Table 6. The ASR performances were 40.5%, 79.0% and 89.4% for D30, D50 and D70, respectively. The average and SD of intelligibility by human is shown in Table 7. The average (SD) of normalized WWL is shown in Table 8.

Table 8: The average (SD) of normalized WWL (Exp.2).

<table>
<thead>
<tr>
<th>Condition</th>
<th>D30</th>
<th>D50</th>
<th>D70</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>50 (10)</td>
<td>50 (10)</td>
<td>50 (10)</td>
</tr>
</tbody>
</table>
Table 8: The average (SD) of normalized WWL (Exp.2).

<table>
<thead>
<tr>
<th>Condition</th>
<th>D30</th>
<th>D50</th>
<th>D70</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>60.6</td>
<td>45.2</td>
<td>44.2</td>
</tr>
<tr>
<td>(SD)</td>
<td>5.5</td>
<td>7.5</td>
<td>6.9</td>
</tr>
</tbody>
</table>

Table 9: The ANOVA results of intelligibility (%) and normalized WWL (Exp.2).

<table>
<thead>
<tr>
<th>Group</th>
<th>Intelligibility</th>
<th>N-WWL</th>
</tr>
</thead>
<tbody>
<tr>
<td>D30-D50</td>
<td>&lt; *</td>
<td>&gt; *</td>
</tr>
<tr>
<td>D30-D70</td>
<td>&lt; *</td>
<td>&gt; *</td>
</tr>
<tr>
<td>D50-D70</td>
<td>ns</td>
<td>ns</td>
</tr>
</tbody>
</table>

Intelligibility: \( F = 66.95 (p < 0.01) \)
Normalized WWL: \( F = 20.21 (p < 0.01) \)

4. Discussions

The differences of human intelligibility and the machine recognition performance of Exp.1 and Exp.2 are shown in Table 10 and Table 11, respectively. Concerning the Exp.1, the subtraction of ASR performance from human intelligibility was 43.7 point, 18.4 point and 9.6 point for D30, D50 and D70, respectively.

Concerning the Exp.2, the subtraction of ASR performance from human intelligibility was 43.7 point, 18.4 point and 9.6 point for D30, D50 and D70, respectively. Although the ASR method was not optimized enough, the results supported the effectiveness of DBM (Figure 5).

The performance differences of D30 (DBM) and Mm10 (MBM) are close (43.7 point and 44.8 point), and their WWL are also very close, i.e. the average difference is 0.7, according to the results of Exp.1/G2. It indicates that these conditions can be the benchmarks for the purpose of comparison between the MBM and DBM.

The results also suggest that we need further investigation of the experimental setup. Using DBM, conditions of lower information levels, such as 20% or 10% should be examined. Using MBM, the number of subjects may be insufficient, because the intelligibility of Mm10 has large deviation and its average is lower than that of Mm20.

Table 10: The difference of human intelligibility and the machine recognition performance (MBM in Exp.1, point).

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mm20</th>
<th>Mm10</th>
<th>M0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>78.1</td>
<td>44.8</td>
<td>24.7</td>
</tr>
</tbody>
</table>

5. Conclusions

Two experiments of mixing-based method (MBM) and deletion-based method (DBM) for audio CAPTCHAs were presented. Our results supported the hypothesis as follows: (1) only using MBM, the degree of task difficulty can not be controlled easily, (2) using DBM, the degree of task difficulty and safeness of CAPTCHA system can be controlled easily.

Future works include the utilization of state-of-art ASR techniques and investigations of effective combinations of MBM and DBM. It is also interesting to verify how the "phonemic restoration" effect increases the intelligibility under the conditions of DBM. The effect of word familiarity and the language dependency problem should also be investigated.

6. Acknowledgments

We would like to thank Kenta Nishiki, Ryo Tanemura, Jonathan Le Roux, Nobutaka Ono, Shigeki Sagayama, Chihiro Fukuoka and Hitomi Matsumura for their support and assistance in this research.

7. References