Masker design for real-time informational masking with mitigated annoyance

Yusuke Hioka[a,*], Jesin James[b], Catherine I. Watson[b]

[a] *Acoustics Research Centre, Department of Mechanical Engineering, University of Auckland, Auckland 1142 New Zealand*
[b] *Department of Electrical, Computer, and Software Engineering, University of Auckland, Auckland 1142 New Zealand*

Abstract

An alternative design of masker for real-time speech masking is proposed. Time-reversed speech, which is also adopted in the proposed design, has been commonly used for speech masking because of its efficient performance in making the target speech unintelligible using informational masking. However, previous studies revealed that a time-reversed speech causes more annoyance and cognitive distraction for the listeners. Meanwhile, although generating time-reversed speech from the target speech would help improve masking effectiveness, it would require real-time processing as the target speech would not be available beforehand. The proposed masker design utilises techniques known for minimising discontinuities in the waveform of synthesised speech in order to minimise annoyance and distraction to the listeners. The design also avoids processes that hinders real-time generation of the masker.

Results of subjective listening tests reveal that the proposed design is able to compromise the level of annoyance and the masking effect. Only marginal improvement is observed in the level of distraction. Further attempt to mitigate annoyance of the proposed masker by adding artificial reverberation did not help as it also reduces the masking effect significantly.

Keywords: Informational masking, speech intelligibility, annoyance, cognitive distraction, real-time

1. Introduction

Speech privacy in buildings is a long-standing problem in the area of building acoustics [1, 2]. Due to denser population in urban areas, the problem has been attracting more attention in various types of spaces in building such as open plan offices [2, 3]. Commonly used techniques for protecting speech privacy include installing partitions and sound absorptive materials into the building, but their installation would be costly therefore they are not suitable for frequent changes of floor plans. Speech masking is another technique that covers up the target speech (i.e. the speech that needs to be concealed) utilising the effect of auditory masking. It is realised by projecting a masker (i.e. sound that covers the target speech) into the environment from loudspeakers. Unlike other available techniques, speech masking does not require installing physical structures into the environment [4], allowing the technique to be a more cost effective option.

Choosing a right masker design is the key to the success of a speech masking system. Masker design can be classified into two categories based on the types of auditory masking induced by the masker. Traditional masker designs use *energetic masking*, which occurs when the excitation or neural response of basilar membrane on cochlea in a given frequency range caused by the target speech is less than that produced by masker [5]. Broadband stationary noises such as pink noise, HVAC (Heating, Ventilating, and Air Conditioning) systems noise [6], or some nature sound such as water based noise [7] are examples of maskers using energetic masking.

On the other hand, it has been revealed that some types of masker are able to induce *informational masking*, which hinders listeners' high level ability trying to "spotlight" particular spectro-temporal regions of sound to perceive the context in the target speech [8]. Rhebergen *et al.* [9] have found that speech-like masker is able to induce informational masking because the sound contains spectral components similar to that of the target speech. This is particularly the case when the target speech itself is used as the *seed* of the masker [10, 11]. When the masker is generated from the target speech itself, it is essential that the context in the original target speech is completely destroyed once it is in the form of masker. A commonly used speech-like

[*] Corresponding author
Email address: yusuke.hioka@ieee.org (Yusuke Hioka)

masker that addresses this issue is time-reversed speech [10, 11, 12, 13]. Time-reversing is an artificial manipulation applied to a speech which is able to convert the speech to be a meaningless sound while maintaining its "speech-like-ness" [14].

Three factors should be considered when a masker is designed. Without any doubt, maximising the masking effect should be the top priority in designing a masker. Here the effectiveness of a masker is measured by to what extent the masker is able to reduce the intelligibility of the target speech. This can be measured by either subjective listening tests or objective metrics [11]. Generally a greater masking effect is achieved when the volume of the masker is increased. Therefore masking effect has to be measured at a fixed target-to-masker ratio (TMR) when the effect of more than one masker is compared.

The second factor to be considered in masker design is minimising annoyance and distraction because a masker should not annoy listeners nor affect their cognitive performance [15, 16, 17]. There are potentially several causes of annoyance and distraction induced by a masker. Studies in cognitive research suggest the detrimental effect of variation of sound intensity to the cognitive performance, which is widely accepted as irrelevant sound effect (ISE) [18, 19]. Some previous studies [20, 21, 22] address this issue with particular focus on speech-like maskers. In terms of annoyance, it is also known that amplitude modulation (i.e. short-term variation of sound level) of a masker annoys listeners [23]. The discussion in [13, 24] particularly focused time-reversed speech, which also implies that rapid temporal variability of intensity (i.e. volume) of the masker would increase annoyance. This can also be inferred from the experimental results of the first author's previous work [25]. Overall, these studies inform that minimising the sudden temporal change of sound intensity would help reduce the annoyance and distraction caused by a masker.

The third factor is specific to sound masking using speech-like masker. Although the masking effect of a speech-like masker is improved when the masker is generated from the target speech, the target speech will not be available until the time when the masker has to be projected. Thus, masker generation has to be implemented in real-time [12]. A framework for real-time implementation of informational masking has been proposed in [26]. As shown in Figure 1, a real-time speech masking system captures the target speech using microphones, generates the masker based on the information collected from the observed target speech, and emits the masker from a loudspeaker. A series of such processes has to be realised in real-time. To this end, the amount of delay caused by the masker generation has to be minimised.

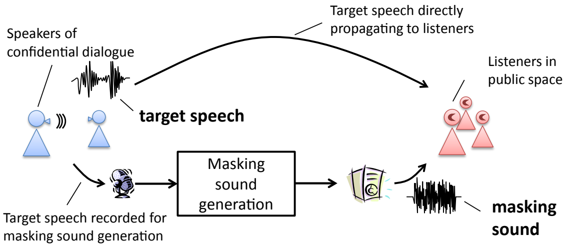

Speakers of confidential dialogue
Target speech directly propagating to listeners
Listeners in public space
target speech
Masking sound generation
masking sound
Target speech recorded for masking sound generation

Figure 1: Real-time informational masking system [27].

The aim of this study is to propose a new masker design that addresses all three factors discussed above. The authors have conducted a preliminary study which proved that minimising discontinuities between frames caused by the time-reversing contributes to mitigate annoyance [27], as the discontinuities created audible click-like sounds. This paper further investigates the performance of the proposed masker design with a comprehensive set of experimental results obtained by reconducting the subjective listening test. One significant extension from the previous report is the introduction of another masker which incorporates the ideas presented in [25] and [27] for mitigating annoyance and distraction.

The rest of this paper is organised as follows. Section 2 is devoted to elaborating the proposed masker design. Details of the subjective listening tests are presented in Section 3. Section 4 gives the experimental results and discussions. Finally the paper is concluded with some remarks in Section 5.

2. Proposed masker design

The proposed masker design adopts time-reversed speech but revises its generation procedure to satisfy the requirements discussed in Section 1. This section explains the details of the proposed masker design.

2.1. Time-reversed Speech

Time-reversing is applied to a speech on a frame-by-frame basis. The speech signal is segmented into fixed length frames (e.g. 160 ms [10], 5 to 240 ms [11], and 500 ms [13]), then the signal of each frame is time-reversed, which means that the sample at the beginning of the frame is flipped to the end, and that at the end is brought to the beginning. Once the time-reversing process is applied to all frames, the position of the frames within the entire sentence is randomised. Finally these randomised frames are concatenated to generate time-reversed speech.

There are two issues that have to be addressed with regard to the time-reversed speech being used as a masker to satisfy the requirements discussed in Section 1. As already discussed, [13] reports that masking with time-reversed speech causes higher annoyance and distraction to the listeners. Further they point out such annoyance and distraction would be caused by the sudden temporal change of sound intensity. Another issue is with real-time implementation of the masking design method, which is also addressed in [12]. The original time-reversed speech requires the randomisation process, which may cause extra algorithmic delay in the masker generation. The rest of this section will be devoted to introducing a few considerations for overcoming these issues.

2.2. *Mitigating signal discontinuities*

In the field of speech signal processing, there are several fundamental methods available for mitigating the detrimental effect of discontinuities between frames. In this study, following techniques: i) frame segmentation using pitch marks, ii) frame concatenation with overlap-and-windowing are introduced to the design of masker.

2.2.1. *Variable frame size using pitch marks*

A speech signal can be categorised as voiced or unvoiced sections depending on the difference in the mechanism of the sound production. Much of the speech energy resides in the voiced section, where the sound waveform becomes nearly periodic since the vocal folds vibrate periodically during the production of a voiced sound. The fundamental frequency of this glottal pulse (also known as pitch) determines the pitch period. The pitch period can be estimated by detecting the location of the short-term energy peak of each pitch pulse in a speech signal [28].

In the proposed masker design, pitch marks in the voiced section of a speech signal are detected and utilised as a reference for the frame segmentation in the generation of time-reversed speech. Since pitch marks usually indicate the short-term energy peak in each pitch period, pitch marks themselves are not suitable for being used as the boundaries of frames since division of the speech signal at such energy peak would cause significant discontinuity of the signal. Thus, the first zero-crossing point after the energy peak of each pitch period is selected as the point of frame segmentation.

2.2.2. *Frame concatenation with Overlap-and-Windowing*

Windowing is a classical signal processing technique that is used to segment a signal into frames. Windowing is also used to minimise the discontinuity in the waveform when the signal is reconstructed by concatenating a series of frames. Typical windowing functions used for this purpose in speech signal processing are Hann and Hamming windows [29]. Concatenating windowed frames would cause more silent period in the masker especially when the edges of the selected window has a tapered shape e.g. Hann and Hamming windows, which would reduce the effect of masking. To minimise silence in the masker, adjacent frames may be slightly overlapped when they are concatenated.

2.2.3. *Adding reverberation*

Hioka *et al.* [25] suggests adding reverberation to a time-reverse speech would also mitigate annoyance caused by the masker. Since the process of adding reverberation may be deemed as applying a low-pass FIR filter to the masker, it would also have an effect to mitigate signal discontinuities. A drawback of this processing is the introduction of additional delay to the masker generation process as the length of the FIR filter coefficients for adding reverberation would typically be long, from a few hundred milliseconds up to a couple of seconds.

2.3. *No randomisation*

The original procedure for generating a time-reversed speech [10] includes randomisation of frames. For real-time implementation, this has made the algorithm require buffer that keeps the sample values of the speech that have already been observed, which would cause extra algorithmic delay in the masker generation process. In order to minimise algorithmic delay, like an attempt made by [12], the proposed masker design does not randomise the frames; meaning the frames with reversed speech are concatenated in the original order. This ensures that the amount of delay caused by generating the masker will be equivalent to the size of frames, assuming that the computational delay (i.e. time taken to run the procedure) is negligibly small.

2.4. *Algorithm for generating the proposed masker*

The overall algorithm used for generating the proposed masker is summarised below. Consider a speech signal $s(t)$ recorded from the target speaker sampled at the sampling frequency f_s, where t denotes time. This signal is used as the seed for producing the masker.

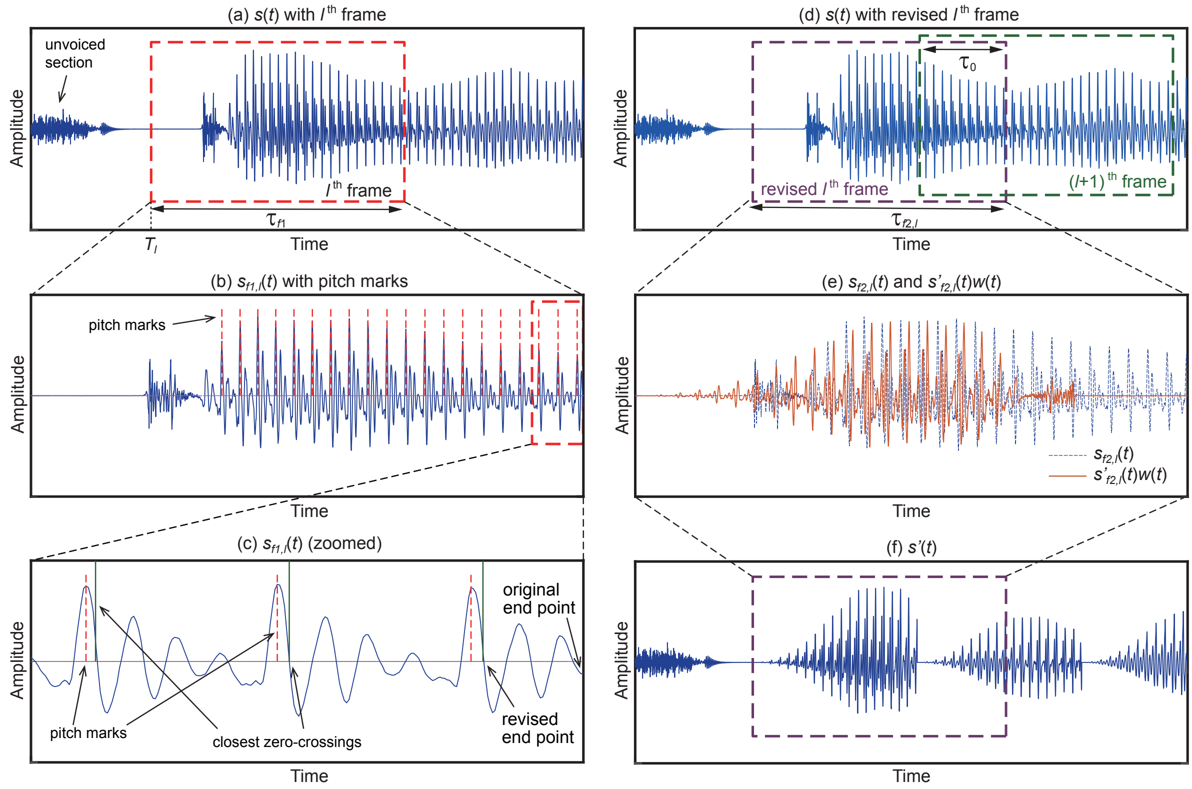


Figure 2: Procedures for generating the proposed masker.

- **Step 1:** The generation of the masker is conducted by frame-by-frame basis starting from the beginning of the speech signal. A duration τ_{f1} is extracted as the l-th frame denoted by $s_{f1,l}(t) = s(t)$ ($T_l \leq t \leq T_l + \tau_{f1}$) (Figure 2(a)), where l is the frame index (i.e. $l = 1$ for the first frame) and T_l is the starting point of the l-th frame. The extracted signal is examined if it predominantly covers the voiced or unvoiced sections of the speech signal. If it is determined as an unvoiced section, the signal is used as it is without further processing and the algorithm moves to the next frame.

  Voiced/Unvoiced signal detection procedure: The voiced or unvoiced detection is based on 3 features of the speech signal: (i) zero crossing rate, (ii) magnitude sum function, and (iii) pitch. The (i) zero crossing rate is the number of times the speech signal crosses the zero amplitude per unit time. This is calculated for every frame l of the speech signal based on the formula: $\mathrm{ZC}_l = 0.5\sum_{n=1}^{f_s\tau_{f1}-1}\left|sgn(s_{f1,l}\left(\frac{n+1}{f_s}\right) - sgn(s_{f1,l}\left(\frac{n}{f_s}\right))\right|$, where n is the sample index of the signal. The zero crossing rate provides good distinction between the voiced and unvoiced sections of the speech signal as number of the zero crossings is lower for the voiced sections compared to the unvoiced sections [30]. In unvoiced sections presence of noise-like signal causes more number of zero crossings, while for the voiced sections the pitch frequency component has lower number of zero crossings.

  The (ii) magnitude sum function (MSF) is calculated for every frame l of the speech signal based on the formula: $\mathrm{MSF}_l = \sum_{n=1}^{f_s\tau_{f1}}\left|sgn\left(s_{f1,l}(\frac{n}{f_s})\right)\right|$. The voiced sections of speech have higher energy compared to the unvoiced sections which are quantified by the MSF. The (iii) pitch is calculated using autocorrelation of the speech frame [31].

  In all the three methods, a threshold is heuristically set based on the overall feature value of the speech signal. If the feature value of a particular speech frame is higher than its corresponding threshold, the frame is marked as voiced. Otherwise, the frame is marked as unvoiced. Ultimately only frames whose results from all the three metrics are marked as voiced are classified as voiced frames, otherwise they are determined as unvoiced frames. Details of these voiced or unvoiced detection can be found in [30, 32].

- **Step 2:** If the frame is a predominantly voiced section (like the l-th frame $s_{f1,l}(t)$ in Figure 2(a)), then pitch marks (Section 2.2.1) within the frame are detected. These pitch marks refer to the peak of every time period of the voiced section of the speech (red dashed lines in Figure 2(b)).

  Pitch marks detection procedure: The pitch marks are detected based on the short term energy spectrum [33]. For this an approximate pitch contour is calculated based on the energy peaks, and the pitch marks are placed at the peaks of the short term energy function.

- **Step 3:** Once the pitch marks are identified, then the zero-crossing point closest to each pitch mark is searched (green solid lines in Figure 2(c)).

- **Step 4:** For frame segmentation using pitch marks, the start and end points of the frame are replaced by the zero crossings closest to the first and last pitch marks in the frame (Figure 2(c)). The new speech frame with the revised start and end points is given as $s_{f2,l}(t)$ $(0 \leq t \leq \tau_{f2,l})$, where $\tau_{f2,l}$ is the length of the new frame (Figure 2(d)). When the start or end points are already in the silence of the speech, e.g. the start point of the l-th frame, it is left unchanged.

- **Step 5:** The signal in the revised frame, $s_{f2,l}(t)$, is time-reversed by flipping it across the time axis providing $s'_{f2,l}(t) = s_{f2,l}(\tau_{f2,l} - t)$. A windowing function $w(t)$ is multiplied to the time-reversed speech frame providing $s'_{f2,l}(t)w(t)$ (Figure 2(e)).

- **Step 6:** Consider the next frame, which starts from τ_o before the end of the previous frame where τ_o is the length of the frame overlap (Figure 2(d)). Repeat Steps 1 to 5, and the resultant windowed frame is concatenated with the previous frame where overlapped part is replaced by the signal of the new frame (Figure 2(f)).

- **Step 7:** Steps 1 to 6 are repeated until the end of the entire speech signal $s(t)$. If the last frame is shorter than τ_{f1}, it is time-reversed, flipped and concatenated with the previous frame by overlapping. Finally, a time-reversed speech with frame concatenation with overlap-and-windowing (OLaW) $s'(t)$ is provided (Figure 2(f)).

- **Step 8 (additional):** Artificial reverberation may be added to the generated masker $s'(t)$ providing $s'_{\text{rev}}(t) = h(t) * s'(t)$, where $*$ denotes the convolution and $h(t)$ is the impulse response of the artificial reverberation [25].

3. Subjective listening test

Performance of the proposed masker design was verified and compared with other types of masker design by subjective listening tests. Details of the tests are presented in this section.

3.1. Stimuli

Six different types of maskers listed below were used as stimuli in the subjective listening test.

1. Pink noise (*Pink*).
2. Babble noise [34] (*Babble*)
3. Time-reversed speech with randomly re-ordered frames [10] (*T-rev*).
4. *T-rev* with artificial reverberation [25] (*T-rev + Reverb*).
5. Proposed design (Step 1 to 7) (*OLaW*)
6. Proposed design with artificial reverberation (Step 1 to 8) (*OLaW + Reverb*)

Of those six types, *Pink* and *Babble* are the conventional types of masker that have been commonly used for speech masking [35, 36]. Since generation of these masker do not require the target speech as a *seed*, their properties are also independent from that of the target speech. *Pink* is a typical masker that purely relies on energetic masking whereas *Babble* may induce partial effect of informational masking as it consists of speech too, but does not originate from the target speech. On the other hand, the masker of remaining types are all generated from the target speech, which induce informational masking. The two types proposed by previous studies, *T-rev* [10] and *T-rev + Reverb* [25], both involve randomisation of the time-reversed frames so that these masker designs involves more computational overheads for real-time masking systems. The difference between *T-rev* and *T-rev + Reverb* is that the latter adds artificial reverberation to the masker generated by *T-rev* which has proven to be able to reduce annoyance caused by the masker [25]. *OLaW* is the proposed masker using Hann window [29] for $w(t)$ in Step 5 stated in Section 2.4. Finally *OLaW + Reverb* is the masker generated by adding the same artificial reverberation used to generate *T-rev + Reverb* to *OLaW*. For the amount of artificial reverberation added for *T-rev + Reverb* and *OLaW +*

Table 1: Comparison of masker used for the experiments

Stimulus	Masking effect (E: energetic, I: Informational)	Overhead for real-time implementation*
Pink	E	N/A
Babble	E+I (partial)	N/A
T-rev	E+I	+++
T-rev + Reverb	E+I	++++
OLaW	E+I	+
OLaW + Reverb	E+I	++

*More number of "+" indicates higher amount of computational overhead.

Reverb, 0 dB was chosen for the direct-to-reverberation ratio (DRR) used in [25]. The length of the filter for adding the artificial reverberation was 0.275 s. Table 1 summarises the properties of each type of masker used in the tests.

The audio file "Speech Babble" in Noise Data database of Signal Processing Information Base (SPIB) [34] was used for *Babble*. On the other hand, all other maskers (10 for each masker type) were generated from 40 speech sentences randomly selected from the corpus of the Harvard sentences [37]. The Harvard sentences consist of phonetically balanced sentences that contain specific phonemes in the same frequency of occurrence seen in the English language. 150 ms was selected as the frame size of time reversing (τ_{f1} for the proposed method) by following [13]. The length of the frame overlap (τ_0) for the proposed method (*OLaW* and *OLaW+Reverb*) was set to 50 ms. All maskers used in the listening tests were normalised by their power in order to keep the TMR at the listener's position consistent. Based on the finding from a preliminary experiment, the TMR value at the participant's seat was set at −3 dB because it provided a reasonable degree of masking effect without causing significant ceiling or flooring effects.

Except the Annoyance Test where only the masker is evaluated, both the target speech and masker were played at the same time by saving the signals as a stereo file, with one track containing the masker and another track containing the target speech. This allows both tracks to be perfectly synchronised and played simultaneously from different loudspeakers. Although the listening test was conducted using an off-line system (i.e. all masker was generated prior to the listening test), in order to replicate the real-time implementation, the masker was delayed by 160 ms compared to the target speech by taking into account the algorithmic delay of 150 ms (equal to the frame size) for the masker generation plus 10 ms figuring in the computational delay. The sampling rate of the audio files was 16 kHz.

3.2. Testing environment

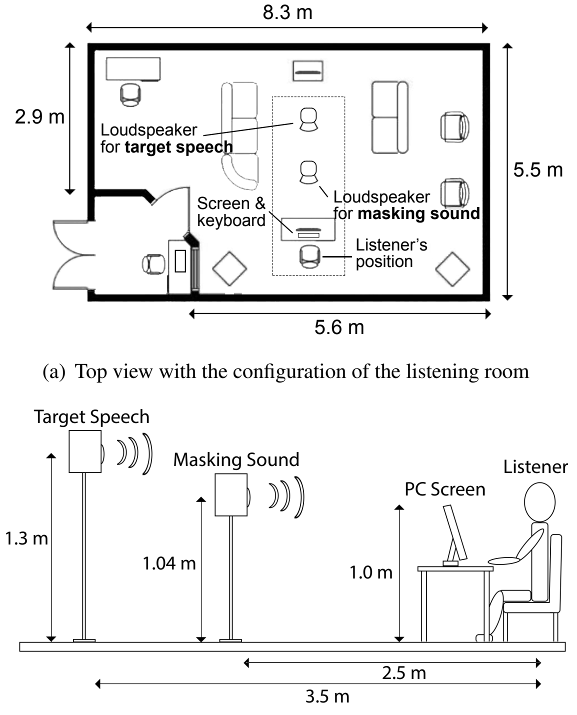
The listening tests were conducted in the listening room of the Acoustics Laboratory at the University of Auckland, New Zealand, the general layout of which is shown in Figure 3(a). The listening room was sound proofed and had an acoustically separated area for the examiner and computer to be located, with the aim to isolate the participants from any potential noise sources. The level of ambient noise in the listening room was 18 dB(A) and the reverberation time was approximately 0.3 s at the frequencies between 100 and 1 kHz.

Two loudspeakers were placed in the middle of the listening room. As shown in Figure 3(b), the front loudspeaker, which acted as the masker source, was placed 2.5 m away from the participant's seat. The back loudspeaker, which was utilised to project the target speech, was placed at a distance of 3.5 m away from the participant's seat. The height (from the floor to the centre of the cone) of the front loudspeaker was 1.30 m which is kept lower than that of the back loudspeaker being 1.04 m, and both were placed in direct sight from the participant's seat to ensure that the participant was able to perceive the sound waves directly propagating from the loudspeakers. A computer screen, a wireless keyboard and mouse placed in front of the participant's seat were used by participants to enter their responses through a Graphical User Interface (GUI).

The volume of the loudspeakers were calibrated to ensure the TMR at the participant's seat was set at −3 dB. The volume of the back loudspeaker was initially adjusted to project a normalised speech file at the level of 58 dB(A) at 1 m away from the loudspeaker, i.e. front loudspeaker's position. The decision to use 58 dB(A) was made since the standard average sound level of a typical conversational speech at normal volume ranges between 55 and 60 dB(A) at the distance of 1 m. The sound level at the participant's seat was then measured which was 52 dB(A). To make sure the TMR was kept at −3 dB, the volume of the front loudspeaker was hence calibrated to have the sound level of 55 dB(A) at the participant's seat.

3.3. Participants

The participants were widely recruited from staff and students at tertiary institutes around Auckland region. Participants were given a voucher worth 10 NZD when they agree to participate the test. A group of 20 participants (5 females and 15 males; 17 in age group 18 – 29,



(a) Top view with the configuration of the listening room

(b) Side view

Figure 3: Position of loudspeakers and participants.

one in 30 – 39, two in 60+) were tested. Based on their self-reporting, none of them had any hearing disabilities, and all participants were native speakers of English or bilingual speakers who speak English as their first language. All participants sat the entire test, each taking approximately one hour.

3.4. *Testing Design*

The main design criteria to be considered in this study was three-fold apart from applicability to real-time implementation which has already been addressed in the design procedure; namely the intelligibility of the target speech, the annoyance caused by the projection of the masker and its distraction. In order to measure the performance of the proposed masker design in terms of these three criteria, three tests were designed and conducted. The test was approved by the The University of Auckland Human Participants Ethics Committee (015933).

3.4.1. *Intelligibility Measurement Test*

The first test measured how well the intelligibility of the target speech was reduced by the masker. This was realised by projecting a target speech and masker to a participant simultaneously. Participants were required to transcribe the sentence of the target speech through the GUI provided. A total of 60 Harvard sentences consisting of 5 to 10 words were used for testing the six different types of masker, namely 10 different Harvard sentences were used for each type of masker. Additional 3 sentences were projected without a masker to test the participants' proficiency in English. The whole testing process was controlled by the GUI, which was used to play the stimuli as well as collecting participants' responses. Participants were given up to 30 seconds to enter the sentence they heard. They were able to move on to the next stimulus by pressing a button on the GUI if their response was completed before 30 seconds had elapsed, otherwise the GUI automatically proceeded to the next stimulus. The order of the type of masker projected was randomised but the same order and same sentences were utilised for all participants to ensure a uniform test environment.

Once the data was collected, the scoring of the test was conducted based on number of words that were correctly transcribed by the participants. Each correct word was given a score of 1, and incorrect words were given no score. If a similar phoneme coverage word was obtained, a half mark was given. Percentage of points gained out of the total number of words in each sentence was calculated, which was used as the evaluation metric. A lower percentage indicates that the masker was more effective in terms of reducing the intelligibility of the target speech.

3.4.2. *Distraction Measurement Test*

The second test evaluated the degree of cognitive distraction caused by the masker when a participant was performing a task. A cognitive memory test similar to that used in [13] was designed, whereby the participant was subjected to a target speech and masker, and was asked to memorise numbers/words presented on the GUI. The test consisted of two parts. In the first half of the test, participants memorised a sequence of 9 randomly selected integers ranging from 0 to 9 that appeared on the GUI one after the other. The participants were exposed to the target speech and masker while seeing the numbers/words. They were given 30 s to type the answers after memorising them. There was a retention period of 3 s before participant was able to enter their answer. The numbers were presented in their numeric formats, not in words.

In the second half of the test, the participants memorised a sequence of 7 words that were randomly ordered and presented to the GUI one after the other. The words consisted of monosyllable and disyllable words that are most commonly used in English. Examples of the words included *right*, *table*, *water*, *think*, *know*, *money*, and *run*, which had no relation to each other. The task was to memorise the words in the correct order, while again being exposed to a target speech and masker.

The choice of using 9 numbers and 7 words was made based on some preliminary experiments conducted before the test. It was observed that using more numbers made it too difficult to memorise correctly, which would make it difficult to judge the effect of the masker as a cause of distraction. On the other hand, remembering few numbers (< 5) was very easy, again causing a significant ceiling effect. Thus 9 was selected as the optimum number for memorising numbers. Likewise, 7 was given as the optimum number for the test using words.

Each type of masker was tested by 6 memory questions (3 for numbers and another 3 for words), in total 36 memory questions were given for each participant. Same as the intelligibility measurement test, the type of masker presented was randomised but the same stimuli in the same order were utilised for all the participants. The scoring was done in a similar manner to that in the intelligibility test; namely a correct number or word was given 1 mark each and the percentage out of the total numbers/words (9 or 7) was calculated. Unlike the intelligibility test, a higher score here indicates a better masker, causing less distraction.

3.4.3. Annoyance Measurement Test

Finally, an assessment from the participants regarding the level of annoyance of each masker was collected. The difference of this test from the distraction measurement test is that the participants rated the masker subjectively rather than performing a cognitive task to measure the distraction objectively. In order to isolate the annoyance caused by the target speech, only the masker was projected to the participants in this test.

A Visual Analogue Scale (VAS) ranging from 0 to 10, 0 being most annoying while 10 being least annoying, was used to evaluate the level of annoyance. In total 10 sets of stimuli were tested, each set consisting of *Pink*, *Babble*, and four other types of maskers generated from the same Harvard sentence. Participants rated the maskers in each set where they were allowed to compare the maskers. Unlike other tests, participants were allowed to listen to the maskers as many times as they liked. The raw score of the VAS was used as a measure of annoyance.

4. Results and Discussion

4.1. Statistical Analysis

For each test, the results collected through the experiments were analysed statistically. Wilcoxon signed rank test was utilised to compare the differences of the median scores across different types of maskers. Matlab was used for every statistical analysis conducted. Statistical significance was tested with the significance level $\alpha = 0.05$ (5%) after applying Bonferroni correction [38].

4.2. Intelligibility Measurement Test

The box plot in Figure 4 shows the distribution of the percentage of correct answer across all (i.e. $N = 20$) participants collected from the intelligibility measurement test. The distribution of each type of masker consists of 200 (= 20 participants ×10 sentences) samples. The average and standard deviation of the scores for the tests with no masker were 99.44% and 3.19%, respectively, suggesting all participants had full proficiency in English.

Figure 5 shows the z-scores of the Wilcoxon signed rank test. It can be said that there were statistically significant differences between the median scores of every masker except the combinations of *Pink* vs *Babble*, *Pink* vs *OLaW+Reverb*, *Babble* vs *OLaW+Reverb*, and *T-rev* vs *T-rev+Reverb*.

Among the types of masker evaluated, *T-rev* and *T-rev+Reverb* showed the highest effectiveness (i.e. lower median percentage of correct answer) and the result of statistical test suggests there was no significant difference between the two masker types. This result agrees with the observation reported in the previous studies [25, 27]. On the other hand, *Pink* and *Babble* were the least effective masker (i.e. highest median percentage of correct answer). The fact that there was no significant difference between the two masker types suggests the effect of informational masking induced by *Babble* may have been marginal making it rely on only energetic masking. The broad distribution of the scores seen in *Pink* and *Babble* was mainly caused by the difference of sentences rather than by the difference of individuals, which implies further tests with different TMRs might be worthwhile to gain deeper insights in future studies. In the box plot, the median score of *OLaW* can be seen in-between these two groups of existing maskers. The statistical test result also shows that *OLaW* is more effective than *Pink* or *Babble* but less effective than *T-rev* or *T-rev+Reverb*. Finally, adding artificial reverberation to the proposed masker seems to have an detrimental effect to the performance of

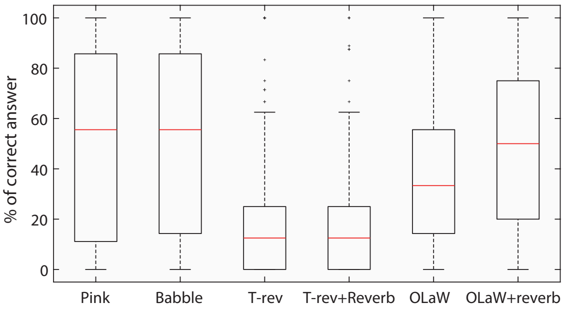
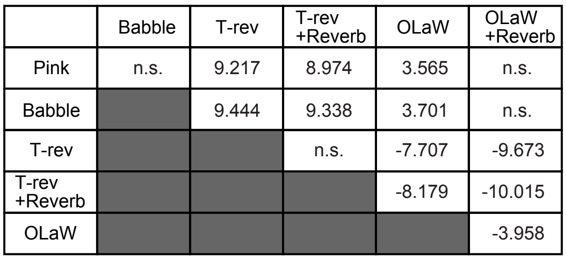


Figure 4: Boxplot showing the distribution of intelligibility scores. The red lines in the boxes show the median of overall scores while the bottom and top edges of the boxes denote the first and third quartiles of the distribution, respectively. The whiskers show the 1.5 interquartile range.



	Babble	T-rev	T-rev +Reverb	OLaW	OLaW +Reverb
Pink	n.s.	9.217	8.974	3.565	n.s.
Babble		9.444	9.338	3.701	n.s.
T-rev			n.s.	-7.707	-9.673
T-rev +Reverb				-8.179	-10.015
OLaW					-3.958

Figure 5: z-scores of the Wilcoxon signed rank test for intelligibility scores. Values show the z-scores of the pairs of maskers the null hypothesis of which is rejected by 5% significance level after applying Bonferroni correction.

the masker given that *OLaW+Reverb* performed significantly worse than *OLaW* whereas there was no significant differences from *Pink* and *Babble*.

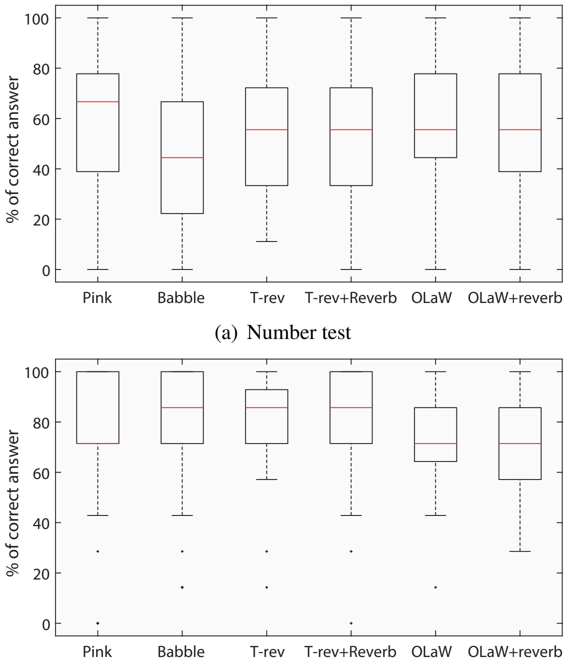
Overall the results indicate that the proposed method is not as effective as the original time-reversed speech (*T-rev*) nor its variant (*T-rev+Reverb*) but still maintained the advantage of informational masking over the conventional speech masking relying predominantly on energetic masking (*pink* and *Babble*). The performance degradation from the original time-reversed speech may be explained by a few reasons. One would be the absence of the frame randomisation in the generation process. A previous study has revealed that the intelligibility of the contents in a time-reversed speech (without randomisation) does not degrade when the frame size is shorter than 50 ms and it still maintains 50% intelligibility at 130 ms [14]. This means a time-reversed speech without randomisation would sound more like a normal speech resulting in making the effectiveness of the time-reversed speech similar to that of *Babble*. Although the frame size used for generating *OLaW* was a bit longer 150 ms, similar phenomenon might have happened to the proposed masker. For more insights, further experiments that investigate the effect of randomisation process with different frame length need to be conducted. Another possible cause would be removing discontinuities in a waveform which generate pulses that would have added extra masking effect.

4.3. Distraction Measurement Test

The box plots in Figure 6 show the distribution of scores across 20 participants collected from the distraction measurement test. The scores are represented by the percentage of correct answers, which incorporates the results obtained from each memory test, i.e. 9 numbers or 7 words, into one set of data by converting the number of correct answer to percentage. The distribution of each type of masker consists of 60 (= 20 speakers ×3 word sets/number sets) samples. Figure 7 shows the matrix table of the z-scores of the Wilcoxon signed rank test. Since there was no combination with significant difference using the same criterion, no z-scores table is required for the results of the word test.

Overall, different trend can be seen between the number and word tests. In the number test (Figure 6(a)), *Pink* marked the highest median score (i.e. less distraction) whereas *Babble* performed the worst (i.e. more distraction), leaving all other types of masker in a similar range (around 55%). The statistical test result shows there were only two combinations that show a significant difference, i.e. *Babble* vs *T-rev* and *Babble* vs *OLaW*, *Babble* being more distractive than others. This may be the result of participants having been more distracted because of the high intelligibility of contents in *Babble*.

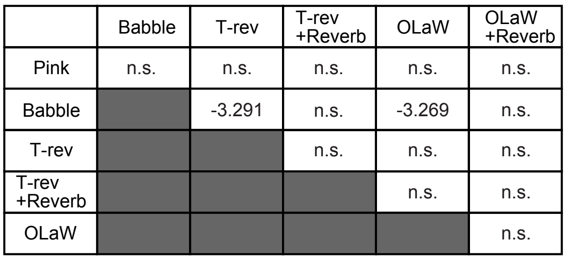
On the other hand, no difference was observed from the word test suggesting the experiment was not well designed to measure the degree of distraction. A possible reason for seeing such results would be the fact that the experiment tested only at a single value of the TMR fixed at −3 dB. As briefly mentioned in Section 1, distraction could be caused not only by the masker but also the target speech. The experiment would have measured the distraction caused only by the masker for some masker such as *T-rev* and *T-rev+Reverb* given that their median intelligibility scores were very low. On the other hand, the scores for maskers identified as less effective, e.g. *Pink* and *Babble*, would have included distraction caused both by the masker and target speech. Further study should be conducted to isolate the causes of distraction using different TMR. In addition, it is also recommended to design a more effective testing method in order to accurately measure the degree of distraction caused by masker.



(a) Number test

(b) Word test

Figure 6: Boxplot showing the distribution of percentage of correct answer provide by the memory test. The red lines in the boxes show the median of overall scores while the bottom and top edges of the boxes denote the first and third quartiles of the distribution, respectively. The whiskers show the 1.5 interquartile range.



	Babble	T-rev	T-rev +Reverb	OLaW	OLaW +Reverb
Pink	n.s.	n.s.	n.s.	n.s.	n.s.
Babble	▨	-3.291	n.s.	-3.269	n.s.
T-rev	▨	▨	n.s.	n.s.	n.s.
T-rev +Reverb	▨	▨	▨	n.s.	n.s.
OLaW	▨	▨	▨	▨	n.s.

Figure 7: z-scores of the Wilcoxon signed rank test distraction scores for the number test. Values show the z-scores of the pairs of maskers the null hypothesis of which is rejected by 5% significance level after applying Bonferroni correction.

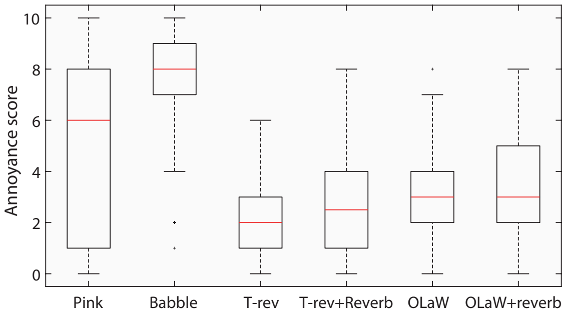


Figure 8: Boxplot showing the distribution of annoyance scores. The red lines in the boxes show the median of overall scores while the bottom and top edges of the boxes denote the first and third quartiles of the distribution, respectively. The whiskers show the 1.5 interquartile range.

4.4. Annoyance Measurement Test

For the annoyance measurement test, the VAS scores collected from each participant were averaged across maskers generated from ten different sentences. The distribution of the scores are provided in Figure 8, with the scores out of 10 being the least annoying sound based on the participants' opinions. The distribution of each type of masker consists of 200 (= 20 participants ×10 sentences) samples. Figure 9 shows the matrix table of the z-scores of the Wilcoxon signed rank test. It can be seen that there were statistically significant differences between the median scores of every masker except the combinations of *T-rev+Reverb* vs *OLaW*.

According to Figure 8 *Babble* received the highest median score meaning participants rated it as the least annoying masker. This is followed by *Pink* but it has much broader distribution suggesting participants' preference on *Pink* are more divided compared to *Babble*. On the other hand, all masker using the target speech as their seed received much lower scores. Such order of annoyance score could be explained by the familiarity of the sound [39] because both *Pink* and *Babble* can be easily seen in our day-to-day life whereas most participants would have never been exposed to time-reversed speech before. Of those maskers based on time-reversed speech, *T-rev* was perceived as the most annoying masker with lower scores. *T-rev+Reverb* and *OLaW*, which were on par, received slightly better scores and a even higher score was given to *OLaW+Reverb*. The order of preference for the existing maskers agree with the results reported in [25].

Overall these observations suggest that the proposed masker has some advantage over the original time-reversed speech (*T-rev*) to mitigate annoyance. The extent of improvement in annoyance is comparable to but is not more than that of adding an artificial reverberation (*T-rev+Reverb*). However, it is worthwhile to emphasise that the proposed masker is suitable for real-time implementation which is not the case for *T-rev+Reverb*. Further mitigation of annoyance is also expected by adding artificial reverberation to the proposed masker.

4.5. Overall discussion

Looking over the results of the three measurement tests allows us to draw a few further findings. By comparing the results of the intelligibility and annoyance

	Babble	T-rev	T-rev +Reverb	OLaW	OLaW +Reverb
Pink	-8.064	8.577	7.221	6.035	3.974
Babble		12.274	12.249	11.990	12.030
T-rev			-4.137	-5.670	-8.349
T-rev +Reverb				n.s.	-5.979
OLaW					-4.326

Figure 9: z-scores of the Wilcoxon signed rank test for annoyance scores. Values show the z-score of the pairs of maskers the null hypothesis of which is rejected by 5% significance level after applying Bonferroni correction.

measurement tests, there is a clear trade-off between the masking effect and annoyance; namely a more effective masker causes more annoyance, and vice versa. This trend agrees with the findings in similar previous studies [13, 25, 27]. Among such trade-off, *OLaW* is located between the two extremes allowing it to stand as a compromise between masking effect and annoyance. Although *T-rev+Reverb* has to be called an even better compromise given that its annoyance was lower while its effectiveness was as good as *T-rev*, it has to be reminded that *T-rev+Reverb* is not suitable for real-time implementation. *OLaW+Reverb* may be another option for real-time implementation, however, adding reverberation to *OLaW* does not seem to help improve its overall performance as it degrades the effectiveness as a masker while the improvement in annoyance is marginal. Thus, overall *OLaW* has proven itself to be the best compromise between effectiveness and annoyance for a real-time speech masking system.

In terms of distraction, on the other hand, *OLaW* has some advantage over *Babble*, causing less distraction while achieving better masking effect. However, this has to be studied further since it has not been clear whether the masker or the target speech are responsible for distraction.

5. Conclusions

This study has proposed a new design of masker for real-time informational masking. The proposed design aimed at compromising the trade-off between annoyance and distraction to the listeners while maintaining the effect of masking the target speech. In addition, the new masker design allows a system to generate a masker in real-time, i.e. the masker is instantly generated from the target speech recorded, in order to make the most use of the advantage of informational masking. Following a hypothesis that the discontinuities between frames that occur in the process of generating a time-reversed speech would be responsible for the annoyance and distraction, fundamental techniques used in speech processing are applied to the time-reversed speech to generate the proposed masker.

The performance of the proposed masker was evaluated by three different types of subjective listening tests. The results indicated that the proposed masker compromises the trade-off between masking effect and annoyance while enabling real-time implementation of a speech masking system. Adding artificial reverberation to the proposed masker did not help improve the overall performance. Compared to annoyance, only limited improvement has been observed in mitigating distraction, due to the fact that distraction could be caused both by the the target speech and masker.

Further experiments have to be conducted to isolate the cause of distraction which may be achieved by testing the masker at different TMR. Comparison with other commonly used masker types such as music or natural sounds would also be another interesting topic for future studies.

Acknowledgements

The authors would like to thank all anonymous participants of the subjective listening tests conducted for this study. This work was supported by Faculty Research Development Fund at the University of Auckland.

References

[1] W. J. Cavanaugh, W. R. Farrell, P. W. Hirtle, B. G. Watters, Speech privacy in buildings, The Journal of the Acoustical Society of America 34 (4) (1962) 475–492.

[2] E. Sundstrom, R. K. Herbert, D. W. Brown, Privacy and communication in an open-plan office: A case study, Environment and Behavior 14 (3) (1982) 379–392.

[3] J. Bradley, Acoustical design for open-plan offices, Construction Technology Update 63 (2004) 1–6.

[4] R. Chanaud, Progress in sound masking, Acoustics Today 3 (2007) 21–26.

[5] B. W. Y. Hornsby, T. A. Ricketts, E. E. Johnson, The effects of speech and speechlike maskers on unaided and aided speech recognition in persons with hearing loss, Journal of the American Academy of Audiology 17 (6) (2006) 432–447.

[6] K. Ueno, H. Lee, S. Sakamoto, A. Ito, M. Fujiwara, Y. Shimizu, Experimental study on applicability of sound masking system in medical examination room, in: Acoustics08 Paris07, 2008, pp. 2374–2378.

[7] V. Hongisto, J. Varjo, D. Oliva, A. Haapakangas, E. Benway, Perception of water-based masking soundslong-term experiment in an open-plan office, Frontiers in psychology 8 (2017) 1177.

[8] M. Leek, M. Brown, M. Dorman, Informational masking and auditory attention, Perception & Psychophysics 50 (3) (1991) 205–214.

[9] K. S. Rhebergen, N. J. Versfeld, W. A. Dreschler, Release from informational masking by time reversal of native and non-native interfering speech, The Journal of the Acoustical Society of America 118 (3) (2005) 1274–1277.

[10] A. Ito, A. Miki, Y. Shimizu, K. Ueno, H. Lee, S. Sakamoto, Oral information masking considering room environmental condition part1: Synthesis of maskers and examination on their masking efficiency part1 synthesis of maskers and examination of their masking efficiency, in: Proceedings of Inter-Noise 2007, 2007.

[11] Y. Wang, P. Leistner, P. Li, Objective evaluation of speech intelligibility for speech masked by time reversed masker processed from the target speech and the effects of its parameters, Acta Acustica united with Acustica 98 (5) (2012) 820–826.

[12] T. Arai, Masking speech with its time-reversed signal, Acoustic Science & Technology 31 (2) (2010) 188–190.

[13] B. Jing, A. Liebl, P. Leistner, J. Yang, Sound masking performance of time-reversed masker processed from the target speech, Acta Acustica United with Acustica 98 (4) (2012) 135–141.

[14] K. Saberi, D. Perrott, Cognitive restoration of reversed speech, Nature 398 (6730) (1999) 760.

[15] A. Haapakangas, E. Kankkunen, V. Hongisto, P. Virjonen, D. Oliva, E. Keskinen, Effects of five speech masking sounds on performance and acoustic satisfaction. implications for open-plan offices, Acta Acustica united with Acustica 97 (4) (2011) 641–655.

[16] V. Hongisto, D. Oliva, L. Rekola, Subjective and objective rating of spectrally different pseudorandom noisesimplications for speech masking design, The Journal of the Acoustical Society of America 137 (3) (2015) 1344–1355.

[17] L. Brocolini, E. Parizet, P. Chevret, Effect of masking noise on cognitive performance and annoyance in open plan offices, Applied Acoustics 114 (2016) 44 – 55.

[18] C. P. Beaman, D. M. Jones, Role of serial order in the irrelevant speech effect: Tests of the changing-state hypothesis., Journal of Experimental Psychology: Learning, Memory, and Cognition 23 (2) (1997) 459.

[19] S. J. Schlittmeier, T. Weißgerber, S. Kerber, H. Fastl, J. Hellbrück, Algorithmic modeling of the irrelevant sound effect (ise) by the hearing sensation fluctuation strength, Attention, Perception, & Psychophysics 74 (1) (2012) 194–203.

[20] S. Schlittmeier, T. Weissgerber, S. Kerber, H. Fastl, J. Hellbrck, An algorithm modelling the irrelevant sound effect (ISE), in: Proc. Acoustics 08, Paris, France, 2008, pp. 1197–1201.

[21] M. Keus van de Poll, J. Carlsson, J. E. Marsh, R. Ljung, J. Odelius, S. J. Schlittmeier, G. Sundin, P. Sörqvist, Unmasking the effects of masking on performance: The potential of multiple-voice masking in the office environment, The Journal of the Acoustical Society of America 138 (2) (2015) 807–816.

[22] A. Liebl, A. Assfalg, S. J. Schlittmeier, The effects of speech intelligibility and temporalspectral variability on performance and annoyance ratings, Applied Acoustics 110 (2016) 170 – 175.

[23] B. Schäffer, R. Pieren, S. J. Schlittmeier, M. Brink, Effects of different spectral shapes and amplitude modulation of broadband noise on annoyance reactions in a controlled listening experiment, International journal of environmental research and public health 15 (5) (2018) 1029.

[24] T. Renz, P. Leistner, A. Liebl, Auditory distraction by speech: Comparison of fluctuating and steady speech-like masking sounds, The Journal of the Acoustical Society of America 144 (2) (2018) EL83–EL88.

[25] Y. Hioka, J. Tang, J. Wan, Effect of adding artificial reverberation to speech-like masking sound, Applied Acoustics 114 (Dec 2016) 171–178. doi:10.1016/j.apacoust.2016.07.014.

[26] T. Sannohe, T. Arai, K. Yasu, A method of creating speech masker using a database in a sound masking system, Journal of the Acoustical Society of Japan 71 (8) (2015) 382–389.

[27] Y. Hioka, J. James, C. Watson, A design of comfortable masking sound for real-time informational masking, INTER-NOISE and NOISE-CON Congress and Conference Proceedings 255 (6) (2017) 1449–1457.

[28] L. A, A short guide to pitch-marking in the festival speech synthesis system and recommendations for improvements, Local Language Speech Technology Initiative.

[29] P. Taylor, Text-to-Speech Synthesis, 1st Edition, Cambridge University Press, New York, NY, USA, 2009.

[30] L. Rabiner, R. Schafer, Digital Processing of Speech Signals, Prentice-Hall signal processing series, Prentice-Hall, 1978.

[31] J. Makhoul, Linear prediction: A tutorial review, Proceedings of the IEEE 63 (4) (1975) 561–580. doi:10.1109/PROC.1975.9792.

[32] W. C. Chu, Speech Coding Algorithms: Foundation and Evolution of Standardized Coders, 1st Edition, John Wiley & Sons, Inc., New York, NY, USA, 2003.

[33] T. T. Swee, S. H. S. Salleh, M. R. Jamaludin, Speech pitch detection using short-time energy, in: International Conference on Computer and Communication Engineering (ICCCE'10), 2010, pp. 1–6.

[34] Signal Processing Information Base, Speech babble, http://spib.linse.ufsc.br/noise.html.

[35] T. Saeki, T. Tamesue, S. Yamaguchi, K. Sunada, Selection of meaningless steady noise for masking of speech, Applied Acoustics 65 (2) (2004) 203 – 210.

[36] H. D. Lewis, V. A. Benignus, K. E. Muller, C. M. Malott, C. N. Barton, Babble and random-noise masking of speech in high and low context cue conditions, Journal of Speech, Language, and Hearing Research 31 (1) (1988) 108–114.

[37] IEEE recommended practices for speech quality measurements, IEEE Transactions on Audio and Electroacoustics 7 (3) (1969) 225–246.

[38] F. Curtin, P. Schulz, Multiple correlations and bonferronis correction, Biological Psychiatry 44 (8) (1998) 775 – 777. doi:http://dx.doi.org/10.1016/S0006-3223(98)00043-2.

[39] W. Ellermeier, A. Zeitler, H. Fastl, Predicting annoyance judgments from psychoacoustic metrics: identificable versus neutralized sounds, in: Proc. of 33rd Intern. Congress on Noise Control Engineering INTER-NOISE 2004, Prague, Czech Republic, 2004.