PROTECTING SPEECH PRIVACY FROM NATIVE/NON-NATIVE LISTENERS - EFFECT OF MASKER TYPE

Hinako Masuda[1], Yusuke Hioka[2], Jesin James[3], and Catherine I. Watson[3]

[1]Faculty of Science and Technology, Seikei University

[2]Acoustics Research Centre, Department of Mechanical Engineering, University of Auckland

[3]Electrical, Computer, and Software Engineering Department, University of Auckland
h-masuda@st.seikei.ac.jp, yusuke.hioka@ieee.org, c.watson@auckland.ac.nz

ABSTRACT

Sound masking is a technique used for protecting speech confidentiality, which is realised by adding maskers to cover target speech. Numerous research have demonstrated the native and non-native differences in perceiving masked speech, but few have compared the effect of masker types from the perspective of effectiveness of speech privacy. This paper reports on the result of an English word identification task by native and non-native listeners where five types of maskers are implemented. Natives, non-natives residing in English-speaking country, and non-natives residing in Japan were tasked to write down sentences when masking sound was present. Results showed that while the three groups of listeners' baseline performance differed significantly, no significant difference was observed between two groups of non-native listeners in the presence of maskers. Additionally, five types of maskers impacted native and non-native listeners differently, which provides novel evidence on future implementation of maskers to increase speech privacy.

Keywords: word identification, sound masking, speech privacy, noise, non-native

1. INTRODUCTION

We have all experienced difficulty in listening to speech in noisy environments. Sound masking is a technique that makes use of this difficulty to keep conversations confidential. It is realised by projecting additional interfering sound, the masker, that *covers* the contents in the target speech. The degrees of difficulty differs according to the maskers and listeners - for example, if the target speech is spoken in a non-native language, the challenge caused by sound masking would be even greater. Much research has looked into the effect of noise for non-native perception, using data from various language groups and processing levels such as segments (consonants/vowels), words and sentences [3, 4, 14, 15].

This non-native perception challenge has been a point of interest to linguists and phoneticians for several decades, but it is only recently that we are looking deeper into how various types of maskers (e.g. noise and reverberation) affects non-native speech perception [2, 11]. As stated above, many researchers have looked into the effect of non-native speech perception, but the types of maskers used in the experiments were mostly limited to, for example, babble noise, pink noise, stationary noise, and reverberation, and we have come to understand that certain types of maskers are more damaging to speech perception (conversely, more effective in masking sounds). For example, competing speech maskers are less detrimental than babble noise [10]. Less known is the effectiveness of more complex maskers in attempting to mask speech from various listeners (e.g. native and non-native).

Various designs of masker have been studied, which can be categorised into two different classes. Conventional maskers such as pink noise and HVAC (heating, ventilation, and air conditioning) systems noise [16] rely on the *energetic* masking, which occurs when the excitation or neural response in a given frequency range, due to the target speech, is less than that produced by the background noise [7]. On the other hand, some more recent studies proposed using speech-like signals for the masker [6, 8, 9]. The idea is stimulated by *informational* masking which hinders listeners' ability trying to "spotlight" particular spectro-temporal region of sound to perceive the context [12].

This paper reports the results of an English word identification task assigned to three groups of listeners varying in nativeness under conditions where sound masking is generated using different types of masker, both energetic and informational. The study addresses the questions: i) What type of masker would be more effective in masking speech? and ii) Are the effects different between native and non-

native listeners depending on the types of masker?

2. EXPERIMENT

A subjective listening test was conducted to investigate the effect of sound masking on word identification by native and non-native listeners.

2.1. Stimuli

Five types of maskers were used in the experiment: 1) Pink noise (*Pink*) [16], 2) Babble noise (*Babble*) [13], 3) Time-reversed speech with randomly re-ordered frames (*T-rev*) [8], 4) *T-rev* with artificial reverberation (*T-rev + Reverb*) [6], and 5) Time-reversed speech using overlap-and-add with Hanning window (*OLAW*) [5]. Of those five types, *Pink* and *Babble* are classified as energetic masking whereas the other three types are classified as informational masking since these types are generated from a speech signal. The two types proposed by previous studies, *T-rev* [8] and *T-rev + Reverb* [6], both involve randomisation of the time-reversed frames. The difference between *T-rev* and *T-rev + Reverb* is that the latter adds artificial reverberation to the masker generated by *T-rev*. *OLAW* is generated similarly to *T-rev* but does not involve the frame randomisation process. The time-reversed frames are concatenated using the overlap-and-add technique with windows in order to minimise discontinuities in the signal [5].

All types of maskers for informational masking (i.e. 3 to 5) were generated from speech sentences randomly selected from the corpus of the Harvard sentences [1] spoken by a male speaker with British accent. The same speech was used as the target speech assuming the maskers were generated from the target speech recorded beforehand. 150 ms was selected as the size of the frames and the length of the frame overlap for *OLAW* was set to 50 ms. All maskers were normalised by their power in order to keep the target-masker ratio (TMR) at the listener's position consistent. The sampling rate of the audio files was 16 kHz. 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. To replicate the delay caused by generating maskers from the recording of target speech, the masker was played 160 ms later compared to the target speech.

Following the settings used in the previous studies, e.g. [6, 8, 9], the TMR value was set at -3 dB because it provided a reasonable degree of masking effect without causing ceiling or flooring effects.

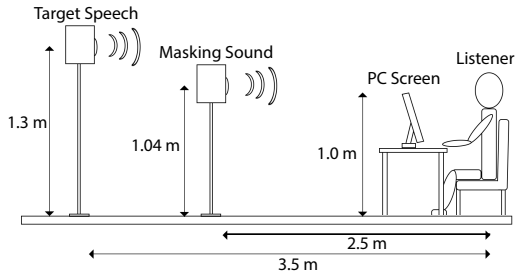

Figure 1: Experimental setup for listening task.

2.2. Testing environment

The listening tests were conducted in the listening room of the Acoustics Laboratory at the University of Auckland, NZ (New Zealand) and the semi-anechoic room at Seikei University, Japan. Both sites were sound proofed and had an acoustically separated area for the examiner and computer to be located thus isolating the participants from any potential noise sources.

Two loudspeakers were placed in the middle of the rooms, as shown in Figure 1. The front loudspeaker (masker source) was placed 2.5 m away from the participant's seat. The back loudspeaker (target speech) was placed at a distance of 3.5 m away from the participant's seat. The front loudspeaker was kept lower than the back loudspeaker, 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 to enter their responses through a Graphical User Interface (GUI).

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 sound level of the normalised speech at the participant's seat 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.

2.3. Participants

A total of 55 listeners participated in the experiment. Based on their self-reporting, none of the participants had any hearing disabilities. Twenty native speakers of English took part in the experiment as the L1 Group (17 participants in age group 18-29, one in 30-39, two in 60+) while another twenty non-native speakers took part in the experiment as the L2a Group (16 in 18-29, four in 30-39). The par-

ticipants in the L1 and L2a Groups were recruited in NZ universities and the participants in the L2a Group had all been residing in NZ at the time of experiment, and their length of residence in English-speaking countries varied from 1.5 to 9 years (min: 1 year, max: 9 years, mean: 3 years, median: 2 years). The native languages of the L2a Group varied (e.g. Mandarin, Cantonese, Arabic, Tamil, Hindi). The participants in the L2a Group had no difficulty communicating in English. Their English proficiency level was measured to be advanced, with eleven participants reporting their IELTS scores to be between 6.5-8.5, median 7).

Fifteen native speakers of Japanese took part as the L2b Group (age group: 18-29). The participants in the L2b Group were recruited in a university in Japan. All L2b Group participants learned English as their second language and most participants in this group had no experience of living outside of Japan. Their English proficiency level was measured to be between lower intermediate and advanced (TOEIC scores 420-915, median 650). Three participants had experience of living abroad (7 years, 1 year and 1.5 years); however, their proficiency level was intermediate (TOEIC scores 650, 600 and 590). Generally speaking, the L2a Group had more English input compared to the L2b Group. All participants were paid upon completion of the experiment. The procedure for the data collection of L1 and L2a Groups was approved by Auckland University Human Participants Ethics Committee, and the procedure for the data collection of L2b Group was approved by Seikei University Ethics Committee.

2.4. Procedure

The subjective listening test measured how intelligibility of a target speech is affected when the masker is also projected to a participant. Overall, 50 Harvard sentences consisting of 7 to 9 words were used in the test, i.e. 10 different Harvard sentences were used for each type of masker stated in Section 2.1. In addition to the 50 sentences an extra 5 sentences were used without a masker to test the participants' baseline performance.

The experiment was controlled by the GUI, which was used to play the stimuli as well as collecting participants' responses. Participants were required to transcribe the sentence of the target speech through the GUI provided on a computer screen. 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 s had elapsed, otherwise the GUI automatically proceeded to the next stimulus. The order of presented masker type was randomised; however the same sentences in the same order were utilised for all participants to ensure a uniform test environment. Each participant spent approximately 20 to 25 minutes for the whole process.

The marking system of the listening test was binary - each response either got a full (1) or no mark (0). Responses were given a full mark 1) when the word was spelled accurately or 2) when the word contained a non-critical spelling mistake (e.g. 'it's' written as 'its'). Partially incorrect responses (e.g. 'beauty' written as 'beautiful') received no mark.

3. RESULTS AND DISCUSSION

Figure 2 shows the baseline performance of the three listener groups (sentences without a masker). The language effect is observed, as L1 Group had near-perfect scores while scores gradually decreased from L2a to L2b, which corresponds to the amount of English input they have received (i.e. L2a has had more English input than L2b). A one-way Analysis of Variance showed a significant main effect of language ($F(2,162) = 112.1$, $p < .001$), and post hoc analysis using the Bonferroni correction revealed significant differences in the language pairs L1 and L2a ($p < .001$), L1 and L2b ($p < .001$), and L2a and L2b ($p < .001$). The significant difference between L1/L2a and L1/L2b indicate that, to no surprise, the ability to perceive unmasked speech varies depending on whether the target sounds are in one's native or non-native language, as documented in previous research, e.g. [10]. Moreover, the statically significant difference between L2a and L2b groups provide evidence that the amount of input one received in the non-native (target) language is reflected on how well they can perceive speech in an unmasked condition.

Figure 3 shows the accuracy rates by the three listener groups merged across all types of maskers (overall) and that of each masker. The ranking of the effectiveness of masking is similar among the three listener groups, with *T-rev* and *T-rev + Reverb* located at the higher end of the effectiveness spectrum (i.e. lower accuracy) and *Babble* located at the lower end (i.e. higher accuracy). Analysis of variance showed a significant interaction between language and masker type ($F(8,208) = 23.66$, $p < .001$), and significant main effects of language ($F(2,52) = 33.41$, $p < .001$) and masker type ($F(4,208) = 94.90$, $p < .001$). In terms of the language effect, post hoc analysis using the Tukey Kramer test showed significant differences between

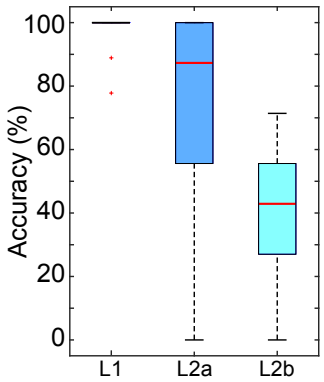

Figure 2: Accuracy rates in condition without masker (%).

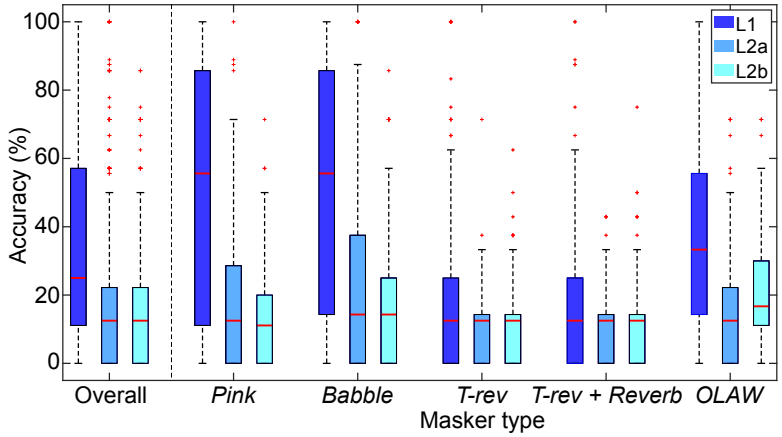

Figure 3: Accuracy rates in conditions with maskers (%).

L1 and L2a ($p < .01$), L1 and L2b ($p < .01$) and a tendency in significance between L2a and L2b ($p = .07$). The asymmetry in the results of L2a/L2b groups in unmasked and masked speech is one of the key findings of this experiment. The amount of input one receives in the non-native language is reflected in the accuracy of unmasked speech task but not necessarily so in masked speech task. In other words, the more input one receives in a language, the higher the performance in unmasked speech task, as indicated in the ranking L1>L2a>L2b. On the contrary, whether the target speech is in one's native or non-native language affects performance in the case of masked speech task, as indicated in the ranking L1>L2a, L2b.

While all groups had difficulty identifying words accurately in the stimuli with maskers, salient differences between L1 and L2a/L2b Groups were observed in *Babble*, *Pink* and *OLAW* maskers. Conversely, *T-rev* and *T-rev + Reverb* conditions were difficult for all groups of listeners, regardless of their native language. Indeed, post hoc analysis using Tukey Kramer test showed significant differences ($p < .01$) between L1 and L2a/L2b in *Babble*, *OLAW*, and *Pink*, but not in *T-rev* and *T-rev + Reverb*. No significant differences were found between L2a and L2b in any of the five maskers.

A plausible explanation for the different masker effects observed in the five types of maskers is the processing methods used to generate them. *T-rev* and *T-rev + Reverb*, the two maskers that had adverse effects on both native and non-native participants, were generated by randomised reordered frames. No significant differences were observed in the accuracy rates of the two maskers in the three participant groups. The *OLAW*, on the other hand, is similar to T-Rev, but instead generated with a different processing (overlap-and-add and Hanning window). The common point in *Babble*, *Pink* and *OLAW* is that they have not been generated using randomised frames.

In addition, an interesting factor that is worth further investigation is the relationship between the signal processing utilised in generating the maskers and annoyance caused by the maskers, as annoyance is an important factor that needs to be considered in creating a stress-free masker. Previous research [5] has demonstrated that, at least in the case of native participants, they may be related.

To summarise, three finding can be drawn from Figure 3. Firstly, maskers generated with randomised frames may be a universally effective masker. Secondly, the *OLAW* maintains effectiveness to mask speech, but only to non-native listeners. Thirdly, maskers that are not generated with randomised frames are less effective as maskers to native listeners.

4. CONCLUSION

A word identification task was assigned to three participant groups differing in nativeness: native listeners (L1), non-native listeners living in an English-speaking country (New Zealand) (L2a), and non-native listeners living in non-English-speaking country (Japan) (L2b). Five types of maskers were implemented to the stimuli to explore their effectiveness in masking target sounds. Results indicated that maskers generated with a certain processing (i.e. randomised frames) is effective to all listeners, while maskers generated without this processing was only effective to non-native listeners. This result provides novel evidence and indicates direction towards effective masking in occasions where speech privacy is needed.

5. REFERENCES

[1] 1969. IEEE recommended practices for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics* 7(3), 225–246.

[2] Brouwer, S., Van Engen, K. J., Calandruccio, L., Bradlow, A. R. 2012. Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *The Journal of the Acoustical Society of America* 131(2), 1449–1464.

[3] Cutler, A., Garcia Lecumberri, M. L., Cooke, M. 2008. Consonant identification in noise by native and non-native listeners: Effects of local context. *The Journal of the Acoustical Society of America* 124(2), 1264–1268.

[4] Florentine, M. 1985. Non-native listeners' perception of american- english in noise. *Internoise 85*.

[5] Hioka, Y., James, J., Watson, C. August 2017. A design of comfortable masking sound for real-time informational masking. *Internoise 2017*.

[6] Hioka, Y., Tang, J., Wan, J. Dec 2016. Effect of adding artificial reverberation to speech-like masking sound. *Applied Acoustics* 114, 171–178.

[7] Hornsby, B. W. Y., Ricketts, T. A., Johnson, E. E. 2006. 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), 432–447.

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

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

[10] Lecumberri, M. L. G., Cooke, M. 2006. Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America* 119(4), 2445–2454.

[11] Lecumberri, M. L. G., Cooke, M., Cutler, A. 2010. Non-native speech perception in adverse conditions: A review. *Speech Communication* 52(11), 864 – 886.

[12] Leek, M. R., Brown, M. E., Dorman, M. F. May 1991. Informational masking and auditory attention. *Perception & Psychophysics* 50(3), 205–214.

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

[14] Mayo, L. H., Florentine, M., Buus, S. 1997. Age of second-language acquisition and perception of speech in noise. *Journal of Speech, Language, and Hearing Research* 40(3), 686–693.

[15] Meador, D., Flege, J. E., Mackay, I. R. A. 2000. Factors affecting the recognition of words in a second language. *Bilingualism: Language and Cognition* 3(1), 55–67.

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