PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Effects of length of carrier phrase on release from masking in multi-talker voice guidance

Hayato SATO[1]*, Masayuki MORIMOTO[1], Kazuhiro IIDA[2], Hiroshi SATO[3]

[1]Kobe University, Japan
[2]Chiba Institute of Technology, Japan
[3]National Institute of Advanced Industrial Science and Technology, Japan

Abstract

Multilingualization of voice guidance is necessary for international passenger facilities and world famous tourist sites. When the voice guidance is provided by public address systems, normally the same content is sequentially broadcasted in several languages. The authors are advancing a challenging project to shorten the required time by simultaneously broadcasting all of the guidances. In this paper, as a basic study, a listening test on word intelligibility of multi-talker voice guidance was performed to clarify effects of length of carrier phrase. The target stimulus consisted of two different Japanese words by two different talkers which were simultaneously reproduced. The task of the listener was to answer both of the two words. As a carrier phrase, other Japanese words spoken by the two talkers were sequentially reproduced prior to the target stimulus. The carrier phrase and the target stimulus were presented from a loudspeaker installed in front of the listener. The parameters of the test were length of the carrier phrase and combination of pitches of the two talkers' voices. As a result, the length of the carrier phrase tended to affect the word intelligibility scores. However, the slope of the score with respect to the length widely varied from positive to negative depending on the test parameters.

Keywords: Multilingualization, Voice guidance, Cocktail party effect

1 INTRODUCTION

In places full of international visitors, such as transport hubs, cosmopolitan cities, and world famous tourist sites, multilingual voice guidance is provided. When the voice guidance is provided by public address systems, normally the same content is sequentially broadcasted in several languages. However, the required time to finish the guidance for all languages is very long and the user have to wait a turn of his/her mother language.

Based on the idea that release from masking by the cocktail party effect (1) can be a solution for this problem, the authors are advancing a challenging project to shorten the required time by simultaneously broadcasting all of the guidances.

The cocktail party effect is often treated as one of the advantages of binaural listening. It is known that spatial release from masking increases the intelligibility of target speech when the target speech and the masker come from different directions. For examples, Duquesnoy (2) reported that when both the target and the masker were speech, a Speech Recognition Threshold (SRT) improvement of about 6 dB was observed by changing the directions of the target and the masker. Hawley *et al.* (3) conducted experiments using multi-talker maskers and also reported that there was an improvement in SRT of 6 to 7 dB by spatial release from masking.

On factors other than spatial release from masking that affect the segregation of multi-talker speech, there are several researches that investigated the effects of similarity between the target and the maskers. With regard to the difference in the fundamental frequency or pitch, Festen and Plomp (4) and Brungart *et al.* (5) reported that the intelligibility of the target speech was clearly improved when the gender of the target talker was different from that of the masker talker. On the linguistic difference between the target and the maskers that should

*hayato@kobe-u.ac.jp

be taken into consideration in our project, van Engen and Bradlow (6) and Calandruccio *et al*. (7) reported that the intelligibility of the target speech increased when the language of the target was different from that of the masker, in severe listening conditions (for example, when the sound pressure level of the masker was larger than that of the target). It should be noted that, as Mattys *et al*. (8) and Brouwer *et al*. (9) pointed out, which of acoustic similarity or linguistic similarity dominantly affect the intelligibility depends on the listener's proficiency on the target or the masker language.

Furthermore, it is known that the uncertainty or familiarity of the target and the masker also affects the intelligibility in multi-talker situations. Brungart and Simpson (10) compared the intelligibility when the talker or the content of masker speech were fixed and when they were random, and reported that the intelligibility scores for the fixed conditions were higher than those for the random conditions especially when the content was fixed. Nygaard and Pisoni (11) reported that the word intelligibility score was increased by training sessions to familiarize the utterance of the talker of the target word, and the training using sentences was more effective than that using words. Jonsrude *et al*. (12) reported that the intelligibility was higher when either the target or the masker was the spouse's voice than when both the target and the masker were unfamiliar voices, although the improvement was smaller when the spouse's voice was the masker than when it was the target.

As described above, many factors need to be considered in order to optimize multilingual simultaneous voice guidance. Considering the above-mentioned past studies, a multi-loudspeaker system will be the most effective to multilingual simultaneous voice guidance, because it can utilize spatial release of masking which is assumed to have the largest effect on the intelligibility in multi-talker situations. However, considering that the guidance is provided to a number of unspecified listeners scattered in the space, it is desirable to be able to keep a high intelligibility as much as possible even when the guidance is provided from a single loudspeaker because the degree of spatial release of masking depends on the position and orientation of the listener. Therefore, in the present study, a simple system using a single loudspeaker was investigated as a basic study.

The purpose of the present study is to investigate methods to improve the intelligibility of multilingual simultaneous voice guidance using a single loudspeaker system. Specifically, a method of providing the target talker's speech as the carrier phrase just before the target speech was investigated. The past studies (11) (12) showed that relatively long-term learning of target speech features increased the intelligibility, however, it is unclear whether the increase can be obtained by listening the target taker's speech just before the target. The length of the carrier phrase and the gender of the talker, which were expected from the past studies to affect the intelligibility, were used as the parameters of the listening test.

In the case of using multi-language speech materials as test stimuli, it is necessary to design complicated experimental conditions in order to investigate the effect of the listener's proficiency on the target/masker language and the effect of the combination among languages. Here, to simplify the purpose of the study, only the condition considered the most severe that both the target and the masker language are the listener's mother language was tested. In this case, since the listener can not distinguish which speech stimulus is the target or the masker, the listener was asked to answer both of them.

2 METHOD

2.1 Test words

240 Japanese words were used as test words. Test words were selected from the FW03 word list (13) to be most familiar to young adults. Each test word has 4 syllables. Test words were spoken by two female talkers (*fhi* and *fto*) and a male talker (*mya*), and recorded in an anechoic room. The fundamental frequencies averaged over 10 sampled words were 262 Hz, 253 Hz, and 145 Hz for the talkers of *fhi*, *fto*, and *mya*, respectively.

2.2 Stimuli and test conditions

Each stimulus was consisted of a carrier phrase (or sound) and two target words as shown in Fig. 1. The carrier phrase was consisted of several words sequentially spoken by two different talkers. The target stimulus

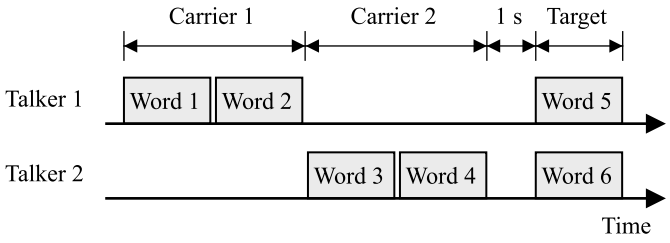

Figure 1. Example of temporal pattern of stimulus (carrier phrase: 2 talkers × 2 words).

was two different words simultaneously spoken by the two different talkers who were the same as those for the carrier phrase. In order to analyze effects of pre-listening the voices of talkers for the target stimulus, a steady-state speech-like noise was used as a carrier sound. The spectrum of noise was determined using Eq. 1.

$$S_{\mathrm{N}}(f) = \frac{S_{fhi}(f)}{4} + \frac{S_{fto}(f)}{4} + \frac{S_{mya}(f)}{2}, \tag{1}$$

where, $S_{\mathrm{N}}(f)$, $S_{fhi}(f)$, $S_{fto}(f)$, and $S_{mya}(f)$ are long-time average spectra for the carrier noise, the test words spoken by *fhi*, *fto*, and *mya* for the frequency of f, respectively. The length of the noise was set to the average length of 4 test words (3.3 s). The interval between the carrier phrase (sound) and the target stimulus was 1 s. The equivalent continuous A-weighted sound pressure level for each word and that for the noise were set to be 65 dB.

Table1 shows test conditions used in the present study. Test parameters were the length of carrier sound (2, 4, and 6 words), the type of carrier sound (speech and noise), and the combination of talkers (*fhi–fto* and *fhi–mya*).

Table 1. Test conditions

Condition	Carrier 1	Carrier 2	Target
1	*fhi*: 1 word	*fto*: 1 word	*fhi* + *fto*
2	*fhi*: 2 words	*fto*: 2 words	*fhi* + *fto*
3	*fhi*: 3 words	*fto*: 3 words	*fhi* + *fto*
4	noise (4 words length)		*fhi* + *fto*
5	*fhi*: 1 word	*mya*: 1 word	*fhi* + *mya*
6	*fhi*: 2 words	*mya*: 2 words	*fhi* + *mya*
7	*fhi*: 3 words	*mya*: 3 words	*fhi* + *mya*
8	noise (4 words length)		*fhi* + *mya*

2.3 Participants

16 young adults participated in the listening test. They were university students in their twenties. The results of pure-tone audiometry show that all participants had normal hearing sensitivity.

2.4 Procedures

160 test words were used as the target words, and the rest 80 test words were used to construct the carrier phrase. Each participant listened to 10 stimuli per condition, in other words, a total of 80 stimuli. Each target test word was presented to the same participant only once, while the words for the carrier phrase were presented twice. The combinations of the test conditions and the target words were different for each participant. All of

the target words were presented once for each condition after 16 participants finished the listening test. The listening test was divided into 5 sessions to listen to 16 stimuli. The stimuli were presented in random order.

The stimulus was presented from a loudspeaker (Eclipse, TD508II) at a distance of 1.5 m in front of the participant in an anechoic room. Each participant was asked to take dictation of both of the target words as they listened using katakana characters (Japanese phonograms). Before the listening test, each participant listened to 8 stimuli which consisted of words other than the 240 test words as an exercise.

3 RESULTS AND DISCUSSION

Word intelligibility scores for each conditions were obtained from the collective results of the listening test for all participants. This means that the number of samples for each condition was 160. The word intelligibility score is the percentage of the target words written down correctly.

3.1 Similar pitch conditions

Figure 2 shows the results for the conditions 1–4, where the target words were spoken by *fhi* and *fto*, in other words, the two target words had a similar pitch. The x-axis of Fig. 2 represents the carrier phrase/sound for each condition. For example, the label "4 words" corresponds to the condition 2 where the carrier phrase consists of 2 words by *fhi* and 2 words by *fto*. Table 2 shows p-values obtained by Chi-squared tests between the scores.

The scores of *fto* were higher than those of *fhi* in all conditions. The p-values of Chi-squared test between the scores of *fhi* and *fto* for the conditions 3 and 4 were 0.058 and 0.089 respectively, and they were almost statistically significant. Considering that there was a difference in the scores even under the condition 4 where the noise was presented instead of the carrier phrase, the interference between the voices of *fhi* and *fto* seemed to be asymmetry regardless of whether or not the listeners had listened to the voices in advance. In order to clarify the cause of this difference, a detailed analysis of speech stimuli is required, but it is beyond the scope of the present study and not be done here.

The score tended to increase as the length of the carrier phrase increased. In the case of *fhi*, the increase was only 3 percent points, and the p-values did not show statistical significance. In the case of *fto*, the increase was slightly large at 8 percent points, and the p-value between the conditions 1 and 3 (between the carrier phrase of 2 words and that of 6 words) was 0.08. The effect size (Cohen's h) for the comparison between the conditions 1 and 3 was 0.198, and this means that one can expect that the length of the carrier phrase causes a *small* difference of the scores. However, the difference can not be detected as a statistically significant difference because of the small number of samples in the present study. The difference in the effect of the carrier phrase length between *fhi* and *fto* can be explained by the temporal position of *fto*'s voice in the carrier phrase: While the *fto* carrier phrase was always adjacent to the target words, the *fhi* carrier phrase was more widely separated from the target words under longer carrier phrase conditions. It is likely that the listeners could pay attention to the voice of *fto* quickly by using the fresh image of the voice in their working memory, and this caused relatively large amount of informational masking release.

It should be noted that the scores of *fto* in the condition 4 was higher than that in the condition 1, although the p-value between them was 0.231. This implies that information included in the carrier phrase may degrade the scores in some cases.

3.2 Different pitch conditions

Figure 3 shows the results for the conditions 5–8, where the target words were spoken by *fhi* and *mya*, in other words, the two target words had a difference in pitch. Table 3 shows p-values obtained by Chi-squared tests between the scores.

The differences between scores of *fhi* and *mya* were not significant in all conditions. The scores ranged from 81% to 88%, and they were higher than the similar pitch conditions shown in Fig. 2. Table 4 shows p-values

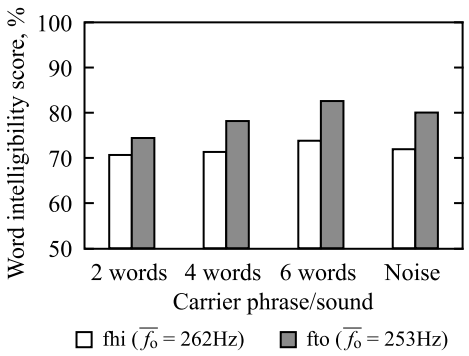

Figure 2. Word intelligibility scores for each target stimulus in the conditions 1–4.

Table 2. The results (p-value) of Chi-squared tests for conditions 1–4.

(a) Between talkers

Condition			
1	2	3	4
0.453	0.157	0.058	0.089

(b) Between conditions

Talker	Pair of conditions					
	1–2	1–3	1–4	2–3	2–4	3-4
fhi	0.902	0.533	0.805	0.617	0.901	0.706
fto	0.431	0.077	0.231	0.325	0.680	0.567

obtained by Chi-squared tests between the scores of *fhi* for each carrier phrase/sound. Except the case that the carrier phrase was 6 words, the scores of *fhi* in the different pitch conditions were significantly higher than those in the similar pitch conditions (p<0.05). This result can be explained by informational masking release due to the difference in pitch between competing voices.

The score tended to decrease as the length of the career phrase increased, although the p-values did not show statistical significance. This tendency was rather apparent for *fhi*, and the p-value between the conditions 5 and 7 was 0.124. It is curious that this tendency was opposite to the similar pitch conditions. As mentioned in the previous section, this result implies that the carrier phrase used in this study had the useful effect while it also had the detrimental effect on the scores. A possible explanation for these results is as follows. The carrier phrase was a sequence of irrelevant words and its content changed with each presentation. Although the listeners were not asked to memorize the carrier phrase, this feature might have increased the listener's working memory load and consequently decreased the scores. In the different pitch conditions, the listeners could easily segregate the two target words so that the useful effect of the carrier phrase was disappeared while the detrimental effect remained.

Table 3. The results (p-value) of Chi-squared tests for conditions 5–8.

(a) Between talkers

Condition			
5	6	7	8
0.741	0.652	0.459	0.636

(b) Between conditions

Talker	Pair of conditions					
	5–6	5–7	5–8	6–7	6–8	7–8
fhi	0.210	0.124	0.741	0.772	0.356	0.225
mya	0.636	0.636	0.636	1.000	1.000	1.000

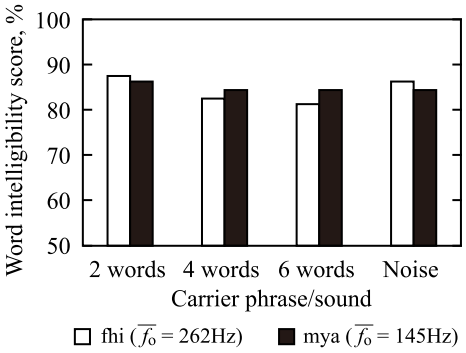

Figure 3. Word intelligibility scores for each target stimulus in the conditions 5–8.

Table 4. The results (p-value) of Chi-squared tests between the scores of *fhi* in the similar pitch conditions and those in the different pitch conditions for each carrier phrase/sound.

Carrier phrase/sound			
2 words	4 words	6 words	Noise
<0.001	0.017	0.108	0.002

3.3 Toward applying the carrier phrase to multi-talker broadcast

The results obtained in the present study suggested that the carrier phrase may have both useful and detrimental effects on the intelligibility of the target multi-talker speech. The longer the carrier phrase, the stronger the useful effect, meanwhile, however, the interval from the carrier phrase to the target speech also affected. Therefore, there is room for optimization as to how to arrange each talker's voice in the time axis. With regard to the detrimental effect, it is necessary not to increase the working memory load of the listener as much as possible. For example, using the fixed carrier phrase without critical information every time to decrease uncertainty may be a reasonable solution. Although the increase in intelligibility by the carrier phrase in the present study was small, if it was the result that the useful and detrimental effects were offset, it can be expected that the intelligibility will further increase by optimization.

4 CONCLUSIONS

As a basic study toward realizing multilingual simultaneous voice guidance, the effect of carrier phrase on multi-talker speech was investigated using a word intelligibility test. The target stimulus consisted of two different Japanese words by two different talkers which were simultaneously reproduced. The task of the listener was to answer both of the two words. The carrier phrase consisted of the two talkers' voices. The parameters of the test were length of the carrier phrase and combination of pitch of the talker's voice. The results suggested the following conclusions.

(1) The length of the carrier phrase tended to affect the word intelligibility scores. However, the slope of the score with respect to the length varied from positive to negative depending on the test parameters.

(2) From (1), information contained in the carrier phrase may have both useful and detrimental effects on word intelligibility scores. In order to optimize the carrier phrase so that only useful effects can be obtained, further studies are required focusing on each parameter of the carrier phrase such as length, difference in pitch between talkers' voices, order of each talker's voice, uncertainty of contents.

ACKNOWLEDGEMENTS

The authors express their gratitude to the listeners. This work was partially supported by JSPS KAKENHI Grant Number 18H01597. The authors would like to thank Sakino Kataoka for her practical support in this work.

REFERENCES

1. Cherry EC. Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Am. 1953;25(5):975–9.

2. Duquesnoy AJ. Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons. J Acoust Soc Am. 1983;74(3):739–43.

3. Hawley ML, Litovsky RY, Culling JF. The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. J Acoust Soc Am. 2004;115(2):833–43.

4. Festen JM, Plomp R. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. J Acoust Soc Am. 1990;88(4):1725–36.

5. Brungart DS, Simpson BD, Ericson MA, Scott KR. Informational and energetic masking effects in the perception of multiple simultaneous talkers. J Acoust Soc Am. 1990;110(5):2527–38.

6. Van Engen KJ, Bradlow AR. Sentence recognition in native- and foreign-language multi-talker background noise. J Acoust Soc Am. 2007;121(1):519–26.

7. Calandruccio L, Dhar S, Bradlow AR. Speech-on-speech masking with variable access to the linguistic content of the masker speech. J Acoust Soc Am. 2010;128(2):860–9.

8. Mattys SL, Carroll LM, Li CKW, Chan SLY. Effects of energetic and informational masking on speech segmentation by native and non-native speakers. Speech Commun. 2010;52:887–99.

9. Brouwer S, Van Engen KJ, Calandruccio L, Bradlow AR. Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. J Acoust Soc Am. 2012;131(2):1449–64.

10. Brungart DS, and Simpson BD. Within-ear and across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty. J Acoust Soc Am. 2004;115(1):301–310.

11. Nygaard LC, Pisoni DB. Talker-specific learning in speech perception. Percept Psychophys. 1998;60(3):355-76.

12. Johnsrude IS, Mackey A, Hakyemez H, Alexander E, Trang HP, Carlyon RP. Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. Psychol Sci. 2013;24(10):1995–2004.

13. Amano S, Sakamoto S, Kondo T, Suzuki S. Development of familiarity-controlled word lists 2003 (FW03) to assess spoken word intelligibility in Japanese. Speech Commun. 2009;51:76–82.