PERCEPTION OF POP-OUT VOICE IN VARIOUS CONDITIONS OF BABBLE NOISE

Mafuyu Kitahara[1], Hideki Kawahara[2], and Shigeaki Amano[3]

[1]Sophia University, [2]Wakayama University, [3]Aichi Shukutoku University
mafuyu@sophia.ac.jp, kawahara@wakayama-u.ac.jp, psy@asu.aasa.ac.jp

ABSTRACT

A perception experiment was conducted to investigate the characteristics of pop-out voice which is easy to listen to even in a noisy background. Stimuli consisted of 30 target voices (either in a high, mid, or low pop-out rank) overlapped by a forward or backward babble noise which was made by mixing normal or time-reversed short sentences in 4 to 100 voices. The authors expected that these babble noise operations would obscure temporal information of phoneme in the time-reversed babble noise, and linguistic information decreases as the number of voices of babble noise increases. Twelve listeners rated a pop-out score of the stimuli using a 5-point scale. Results indicate that the scores are similar in both forward and backward babble noises across the different number of voices in the babble noise, suggesting that the characteristics of pop-out voice are robust regardless of temporal and linguistic properties of babble noise.

Keywords: pop-out voice, babble noise, backward play-back, linguistic information

1. INTRODUCTION

There are some voices which stand out and reach the listener even in very noisy background. Amano and his colleagues have been focusing on the acoustic and perceptual properties of such voices and name those as “pop-out voice” [1-3]. It is different from so-called clear speech which concentrates on the intelligibility of linguistic information in utterances [4]. Even if the intelligibility is low, the voice itself can stand out. Pop-out voice is a concept that focuses on properties of voice, irrespective to the content of the linguistic message conveyed by that voice.

The degree of pop-out depends on various factors on the speaker’s side, such as Fo, intensity, formant frequencies, spectral envelope, speech rate and phonological structure [2]. Another important set of affecting factors is the properties of background noise. The source of the noise in natural setting may vary widely, including environmental noise, artificial noise, and speech by other people as in a cocktail party. The so-called cocktail party effect can be simulated by using the babble noise which is created by overlapping multiple voices.

The focus of the present paper is to test various conditions of babble noise in order to investigate how robust pop-out voice is. The number of voices overlapped in the generation of babble noise is an easily manipulatable property and has frequently been discussed in the literature [5-6]. For example, one past study investigated the effect of the number of voices in babble noise in detecting isolated words. Two to ten voices were overlapped for babble noise and the performance of listeners was worse in the 2-voice condition but improved when the number of voices increased [7]. In another study, the number of voices was by far wider, from 1 to 512 to find that the performance was worst when 8 voices were overlapped [8].

These studies focused on the intelligibility of target items, though. The dependent variable was the rate of correct identification of the word, syllable, or segment in most cases. In contrast, pop-out is supposed to be a property of target voice which must be independent from the intelligibility issues. To clarify the characteristics of pop-out voice, it is worth testing whether the number of voices in babble noise affect the pop-out voice in the way similar to the results obtained for intelligibility-oriented experiments.

Another issue here is the interference of language and/or linguistic content in babble noise. When the number of voices is only a few, word fragments in the noise are clearly identifiable and supposedly interfere the detection of target items. However, word fragments may not interfere if the listener does not know the language or, at least, the linguistic content is unattainable for the listener. Only a few studies have investigated the linguistically crossed speech-in-noise problem [9,10].

When the source of the babble noise came from the same language as the target item (English-English), the performance in the two-voice babble was worse than that in the different language condition (English-Mandarin) [10]. This “linguistic interference” is of particular interest for intelligibility-oriented studies by its nature. However, if the effect of linguistic interference is minimal or irrelevant for the degree of pop-out voice, this will in turn establish that pop-out voice is independent from intelligibility.

In the present study, a method to generate unintelligible babble noise was adopted where time-reversed voices were overlapped. Due to unnatural time course of the signal, any word fragments in such backward babble noise were difficult to detect even when a very few voices were overlapped. The merit of using backward babble noise is that overall acoustic properties can be kept the same as forward babble noise created from the same set of voices: long-term spectra, mean and dispersion of Fo, mix of male and female voice, among others, are the same across the two types of noise.

2. METHOD

2.1. Participants

Twelve native Japanese speakers (eight males and four females) participated in the perception experiment. Their dialectal background varies across western and eastern regions in Japan. Their average age was 53.9 years (SD = 7.3 years). They had no significant history of hearing problems.

2.2. Pre-experiment

A pre-experiment was conducted to obtain a pop-out score of speech items for the main experiment. Speech items /gɑiʑin sɑɴ wɑ kɑmpeki ʃɯgi deɑrɯ/ (The foreigner is a perfectionist) spoken by 100 native Japanese speakers were randomly selected from the items used in [1]. The speech items were filtered with K-weighting [11], and then their root mean square (RMS) was calculated. Intensities of the items were adjusted to the same level using the RMS. After the adjustment, the items were mixed at -4dB signal-to-noise ratio with a babble noise that was generated by overlapping various sentences spoken by 10 native Japanese speakers (see [1] for precise procedure). Using a five-point scale (1: not pop-out - 5: pop-out), the 12 participants in Section 2.1 rated degree of pop-out of the 100 speech items in the babble noise. They were instructed to use the full range from 1 to 5 whenever possible, and to make intuitive decisions without too much trouble. A practice session consisting of 18 trials was given beforehand for the participants to familiarize the notion of pop-out voice [cf. 1]. The pop-out score for each item was obtained as an average of their ratings.

2.3. Stimuli

2.3.1. Target voices

Using the obtained pop-out score, target voices of 30 native Japanese speakers (15 males and 15 females) were selected from the 100 speech items in the pre-experiment to form three ranks: low (L: 1.0-2.0), middle (M: 2.5-3.5), and high (H: 4.5-5.0). Each rank contained 10 target voices spoken by five each of males and females.

2.3.2. Babble noise

The source of babble noise was the speech database having 30 Japanese sentences spoken by 14 each of male and female native Japanese speakers [12]. All sentences in the database were filtered with K-weighting [11] and their level was adjusted to the same level. A 100-second babble noise was generated by overlapping either 4, 8, 20, 40, or 100 sentences that were randomly selected from the database but had the same number of male and female speakers in each sentence group. Intensity of all the babble noises was adjusted to the same level. In addition to simple overlapping of multiple voices, reversed babble noises were prepared in this study where the source sentences were inverted in a time domain before the overlap. These backward babble noises and the usual forward babble noises were both used to test the effect of whatever linguistic information left in the babble noise on the target voice. Thus, the direction of the babble noise is one of the independent variables in our experiment.

2.3.3. Stimulus preparation

Thirty target voices (10 each for H, M, and L pop-out score ranks) were mixed at -7dB signal-to-noise ratio with the babble noises generated in the way described in the previous section. The mixing timing was such that the babble noise started 500 ms before the beginning of the target voice and ended 500 ms after the end of the target voice.

The babble noise was obtained by randomly selecting a segment from the entire 100-second babble noise so that each target voice was mixed with a different part of babble noise. Under these mixing conditions, two stimulus sets were prepared: one by using the forward babble noise and the other by using the backward babble noise.

2.4. Procedure

Stimuli stored in a personal computer were diotically presented to each participant in a random order in a quiet room at a constant intensity through

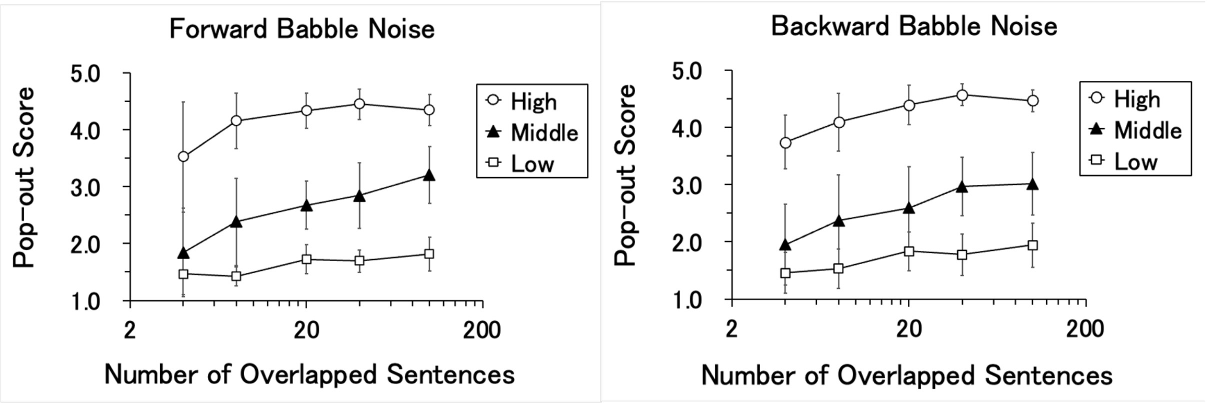

Figure 1: Average pop-out score as a function of the number of sentences (log scale) in the babble noise and pop-out rank of the target voice. Left: forward babble noise. Right: backward babble noise.

circumaural headphones (SONY, MDR-7506) with an audio interface (Roland, Rubix24). The participants rated the degree of pop-out using a five-point scale (1: not pop-out - 5: pop-out). They rated the set of forward babble noise (150 trials) and then the set of backward babble noise (150 trials).

3. RESULTS

Figure 1 depicts average ratings by the participants as a function of the number of sentences. The left and the right panel show the forward and backward babble noise condition, respectively. Three lines in each panel correspond to the rank of the pop-out score of target voices, H, M, and L. Statistical analyses using linear mixed effects models were applied with rating score as the dependent variable, and the number of sentences in the babble noise, the pop-out score rank, and the direction of the babble noise as the independent variables. Intercepts of participants and target voices were random factors.

Table 1: LME results for the number of sentences in the babble (n_babble), the direction of the babble (dir_babble: backward as the baseline), and the rank of pop-out score (rank_L/M: H as the baseline)

	Estimate	*t value*	*Pr(>|t|)*
intercept	3.962	27.161	<.001
n_babble	.006	5.573	<.001
dir_babble	-.014	-.191	.848
rank_L	-2.349	-14.481	<.001
rank_M	-1.695	-10.450	<.001
n_babble:dir_babble	-.003	-.214	.830
n_babble:rank_L	-.001	-.876	.381
n_babble:rank_M	.003	2.125	.033
dir_babble:rankL	-.064	-.590	.555
dir_babble:rankM	-3.672	-.341	.733

The statistical model indicates that the main effects of the number of sentences and the pop-out score rank were significant (p < .001), but the main effects of the direction was not significant (n.s.). The interaction between the number of sentences and M-ranked voices was significant (p < .05). These results suggest that when the number of sentences in the babble noise increase, the pop-out score increases for the M-ranked target voices. However, for the H- and L-ranked voices, the effect of the number of voices was not confirmed. Moreover, the direction of the babble noise did not change the overall picture.

4. DISCUSSION AND CONCLUSION

The results of the experiment have several implications on the nature of the pop-out voice. First, the linguistic information left in the babble noise had negligible effect on the recognition of the target voice. The fact that forward and backward babble noise conditions did not significantly make a difference in the performance of listeners imply that the backward noise whose linguistic information was destroyed by inverting the source also had a similar effectiveness to the forward babble noise in disturbing the target voice. This, in turn, suggests that pop-out voice is independent from the notion of intelligibility. It is a property of voice irrespective to the linguistic content.

Second, the fewer the number of sentences in the babble noise, the more disturbing effect the babble noise had on the target voice. As shown in Figure 1, the average pop-out score becomes larger as the number of sentences increases, which was statistically evident for the M-ranked target voices.

As for the H-ranked target voices, there seems to be a ceiling effect across babble noise conditions, while the L-ranked target voices were identified as close to "not pop-out" across the board.

There are remaining issues to be addressed for the future study of pop-out voice. First, the effect of other noises than babble noise needs to be explored. A body of research about "speech-in-noise" has already explored a variety of different noises, such as vocoded speech, speech-shaped noise, and environmental noise, but it is an intelligibility-oriented study which did not explore the characteristics of "voice-in-noise".

Second, the linguistic information of the target voice itself needs further investigation. There is an on-going study of pop-out voice perception by listeners who have no experience in the language used for the target voice [13], which suggests that pop-out voice is independent from language. That is, English monolingual listeners can detect pop-out voice spoken in Japanese. Then, what characteristics of the target voice did the listeners pick up even though they did not know the language? Paralinguistic and prosodic information in the voice may be relevant for this issue, which remains largely unexplored.

Third, the procedure of presenting forward babble noise first, and backward babble noise second might have a confounding effect on our conclusions in which the direction of the babble noise do not matter. Counter-balancing the material is necessary in the future experiments.

Application of pop-out voice is another possible area of future study. Voices that pop out in noisy environment is in itself useful for public announcement. Disaster and other emergency situations need such voice that addresses people who need to evacuate immediately.

5. ACKNOWLEDGMENT

This research was supported by JSPS Kakenhi 20H00291.

6. REFERENCES

[1] Amano, S., Kawahara, H., Banno, H., Maki, K., Yamakawa, H. 2022. Evaluation experiment of pop-out voices in babble noise. *Proc. of the Acoustical Society of Japan, Spring 2021 meeting*.

[2] Amano, S., Kawahara, H., Banno, H., Maki, K., Yamakawa, H. 2021. Acoustic features of pop-out voice in babble noise. *Acoustic Science and Technology*. 43, 105-112 (in Japanese).

[3] Kitahara, M., Tajima, K., Yoneyama, K., Kitamura, T., Kawahara, H., Amano, S. 2022. Detection of pop-out voice at various SN ratios. *Proc. of the Acoustical Society of Japan, Fall 2022 meeting* (in Japanese).

[4] Uchanski, R. M. 2005. Clear speech. In: D. B. Pisoni and R. E. Remez (eds), *The Handbook of Speech Perception*. Blackwell, 207-235.

[5] Miller, G. A. 1947. The masking of speech. *Psychol. Bull.* 44, 105–129.

[6] Bronkhorst, A. W. 2000. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions, *Acust. Acta Acust.* 86, 117–128.

[7] Freyman, R. L., Balakrishnan, U., Helfer, K. S. 2004. Effect of number of masking talkers and auditory priming on informational masking in speech recognition, *J. Acoust. Soc. Am.* 115, 2246–2256. doi:10.1121/1.1689343

[8] Simpson, S. A., Cooke, M. 2005. Consonant identification in N-talker babble is a nonmonotonic function of N, *J. Acoust. Soc. Am.* 118, 2775–2778. doi:10.1121/1.2062650

[9] Van Engen K. J. and Bradlow A. R. 2007. Sentence recognition in native- and foreign-language multi-talker background noise, *J Acoust. Soc Am.* 121, 519–526. doi: 10.1121/1.2400666.

[10] Van Engen, K. J. 2012. Speech-in-speech recognition: a training study. *Lang. Cogn. Processes* 27, 1089–1107. doi: 10.1080/01690965.2012.654644

[11] *Recommendation ITU-R BS.*1770-4. 2015.

[12] Atake, Y., Irino, T., Kawahara, H., Lu, J., Nakamura, S., Shikano, K. 2000. Robust estimation of fundamental frequency using instantaneous frequencies of harmonic components. *IEICE Trans. Inf. Syste.* (Jpn Ed.), J83-DII, 2077-2086 (in Japanese).

[13] Amano, S., Yamakawa, K., Kawahara, H. 2022. Effects of native language on detection of pop-out voice. *Proc. of the Acoustical Society of Japan, Fall 2022 meeting* (in Japanese).