RAPID ADAPTATION TO TARGET AND BACKGROUND TALKER VARIATION IN SPEECH-IN-SPEECH PERCEPTION

Ann R. Bradlow & Angela Cooper

Department of Linguistics, Northwestern University
abradlow@northwestern.edu; akcooper@u.northwestern.edu

ABSTRACT

The present study examined the impact of talker variation in the target or background speech streams on speech-in-speech recognition. Listeners transcribed sentences in single-talker background babble at two signal-to-noise ratios in one of three conditions: 1) variable target talker and consistent masker talker, 2) consistent target talker and variable masker talker, or 3) consistent target and masker talkers. Results showed a significant effect of signal-to-noise ratio across conditions, as well as substantial variation across target-masker pairs within conditions 1) and 2). In contrast, overall performance across all trials in all conditions was stable, suggesting that the particular energetic masking characteristics of a given target-masker talker pair within a given condition override any potential impact from talker variability in the target or in the masker across conditions. Thus, when recognizing sentence-length stimuli embedded in background speech, listeners exhibit rapid adaptation to across-trial talker variation in the target or the background.

Keywords: talker variability, speech-in-speech perception, sentence recognition

1. INTRODUCTION

Listeners must contend with substantial variability in their speech input. They are required to accommodate differences in speaking style and individual talker characteristics, including vocal tract, pitch and speaking rate changes, all of which can have consequences for the particular acoustic realization of linguistic features in the signal. Indeed, prior work has demonstrated that isolated words embedded in broadband noise were identified less accurately when the talker changed from trial-to-trial relative to when the talker remained fixed (e.g., [12] and many following). These findings point to a mechanism for perceptual adaptation to variation in talkers' voice characteristics and suggest that such a mechanism incurs a processing cost for listeners.

The challenge of input variability is further compounded by the fact that everyday conversations take place in a wide range of adverse listening conditions, including situations involving competing talkers in the background. The challenge for listeners in such situations then is to not only adapt to the idiosyncratic characteristics of the target talker, but also to locate and selectively attend to the target speech stream while tuning out other irrelevant speech streams. A variety of factors have been shown to mediate listeners' ability to effectively segregate speech streams, including the similarity between the voice characteristics of the target and background talkers (e.g., [5, 7]), spatial location of the talker and background signals (e.g., [8]), and linguistic characteristics of the target and background speech (e.g., [4, 15]).

Given that talker variability has been shown to impact speech-in-noise perception for isolated words when presented with simultaneous broadband noise (e.g., [12]), one could hypothesize that reducing talker variability in speech-in-speech contexts, either in the target or background, should enable listeners to adapt to the voice characteristics of the talker and facilitate their ability to either selectively attend to or ignore that voice. Indeed, recent work showed that long-term familiarity with a particular talker, such as a spouse of two or more decades, facilitated both tuning into and tuning out the familiar talker [11]. Listeners showed higher speech-in-speech recognition accuracy when either the target or the background talker was their long-term spouse. Moreover, other recent work has reported decrements in speech-in-speech recognition as a function of trial-to-trial changes in the language of the background [3], indicating that some aspects of background speech variation over the course of a test session can exert an influence on sentence intelligibility in speech-in-speech contexts (but see [6, 9] for limited or no influence of background talker uncertainty on keyword or sentence recognition performance).

However, prior research has not directly compared the influence of target and background talker variability on speech-in-speech perception with a consistent testing method and set of stimuli. In particular, target consistency versus variability (i.e., single versus multiple talker) effects on speech processing have generally been examined with isolated word perception in broadband noise conditions (e.g., [12]). In contrast, background variation versus consistency (i.e., masker uncertainty versus predictability) effects have necessarily focused on speech recognition in conditions with competing speech, rather than broadband noise, in the background.

Accordingly, the present study directly compared the extent to which sentence intelligibility is affected by talker variation in the target in speech-in-speech contexts with either a consistent or variable talkers in the background. Based on prior demonstrations of single versus multiple target talker benefits for word-in-noise recognition (e.g., [12], as discussed above), we might predict a significant single versus multiple target talker benefit for the present test of sentence recognition in the context of background speech. Further support for this prediction comes from prior work that has provided evidence of listener adaptation to talker consistency with sentence-length materials presented in broadband noise, albeit under quite different test conditions from the present study. Specifically, these prior studies with sentence-length materials involved either foreign-accented speech [2] or an explicit talker training paradigm [13]. In the present study we test the generalizability of this process of listener-to-talker adaptation with sentence-length materials in the more complex auditory environment of speech-in-speech rather than speech-in-noise recognition.

Alternatively, it is conceivable that with sentence-length materials, listeners would have more time to adapt to the talker than with word-length materials, thus attenuating the potentially detrimental effects of target talker variability across the test session. If this were the case, then sentence intelligibility in the context of a background talker may depend primarily on the particular characteristics of the target and background signals, including the particular target-background talker pair.

In the present study, listeners heard mono-clausal semantically felicitous sentences with simultaneously-presented single-talker background speech and were asked to transcribe the sentences in one of three conditions (Figure 1): 1) consistent target talker and variable (n=4) background talkers (Background-Variable), 2) variable (n=4) target talkers and consistent background talker (Target-Variable), and 3) consistent target talker and consistent background talker across all trials (Fixed).

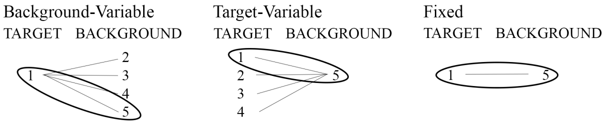

Figure 1: Experiment condition setup. Each number represents a unique talker. All conditions contained the same number of total trials. Ellipses denote critical trials to be compared from same target-background talker pair across conditions.

2. METHODS

2.1. Participants

Forty-five American English listeners, with no reported speech or hearing deficits, participated in this study (9 males, 36 females, mean$_{age}$=20 years). Listeners with knowledge of more than one language were required to have learned English first and to make use of it in 80% of their language interactions. They were randomly divided into 3 groups (n=15 in each) and received course credit for their participation.

2.2. Stimuli

The stimuli consisted of 120 Hearing-in-Noise Test (HINT) sentences [14], which are all mono-clausal, semantically felicitous, declarative sentences containing basic-level English vocabulary (e.g., “The player lost a shoe”). They were produced by 5 female American English talkers (age=18-20 years).

Sixty HINT sentences were used as target materials. For the Background-Variable and Fixed conditions, the productions of a single talker were used, and for the Target-Variable condition, the 60 sentences were divided between 4 target talkers (15 sentences per talker). The other 60 sentences were used to create background speech tracks. For the Background-Variable condition, four background tracks were created, one for each of the 4 background talkers. For Target-Variable and Fixed conditions, a single background track was created for a single background talker.

Fifteen target sentences were designated “critical” trials (n=15), such that they were always produced by the same target talker (talker 1 in Figure 1) with the same background talker (talker 5 in Figure 1) across all 3 conditions. Furthermore, the order of presentation of all sentences was fixed, such that these critical trials occurred in the same serial order position within each block in every condition. This ensured that any differences found in sentence recognition accuracy between conditions did not arise from a discrepancy in the intelligibility of the target talker, the amount of masking a particular background talker may produce, the specific sentences used or where these sentences occurred in the course of the experiment.

For each background track, the talker's sentence productions were concatenated to create a single stream of sentences (with no pauses between sentences). Target and background speech files were equated for root-mean-squared (rms) amplitude. The experiment running software (Max/MSP; Cycling '74) fixed the output level of the background tracks at 70 dB SPL and the target sentences at 60 dB SPL to produce an SNR of -10 dB in Block 1 and at 55 db SPL (SNR of -15 dB) in Block 2.

The target and background tracks were mixed online, whereby on each trial, the program randomly selected a portion of the appropriate background speech track that matched the duration of the target sentence plus an additional 300 ms. The program would delay the playing of the target sentence for 300 ms, such that the target sentence would start playing 300 ms after the onset of the background speech. This small lead time was intended as a means of facilitating listeners' ability to attend to the correct speech track.

2.3. Procedure

Listeners were instructed to transcribe English sentences presented in the presence of another background talker and were told to attend to the second, softer voice. At the end of each trial, listeners would type their response in a text box and press 'Enter' to advance to the next trial. Each sentence was only presented once. Two blocks of 30 trials each were presented, with the first block always at -10 dB SNR and the second at -15 dB SNR. The division of trials into 2 blocks at different SNRs was motivated by prior work demonstrating that the effects of certain manipulations, such as background talker language, were only revealed when listeners moved from an easier to a more adverse listening condition [15].

3. RESULTS

A strict scoring method was employed, whereby a correct sentence entailed all of the words being correctly transcribed. Homophones and apparent spelling errors were not considered incorrect; however, words with inaccurate morpheme substitutions or omissions were scored as incorrect. The primary scoring analysis was performed on only the critical trials, though an additional analysis including all of the sentences was also conducted. The scores were tabulated by SNR (-10 dB, -15 dB) and target-background condition: multiple target talkers-fixed background talker (Target-Variable), fixed target talker-multiple background talkers (Background-Variable), and fixed target and fixed background talker (Fixed).

The data were analysed using logistic linear mixed-effects regression (LMER; [1]) models, with transcription accuracy as the dependent variable. A contrast-coded fixed effect for SNR (-10 dB, -15 dB) as well as Helmert contrast-coded effects for condition (1: Target-Variable, 2: Background-Variable, 3: Fixed) were included in the model. Random intercepts for participant and item as well as random slopes for SNR by participant and Condition by item were also included.

As depicted in Figure 2, there was a significant effect of SNR (β= -0.74, SE β=0.197, $\chi^2(1)$=13.455, $p<0.001$), with listeners performing significantly worse across conditions in the harder SNR (-15 dB) relative to the easier SNR (-10 dB). However, no significant differences between conditions ($\chi^2 < 0.23$, $p>0.05$) or any condition by block interactions were found ($\chi^2 < 0.3$, $p>0.05$). This indicates that performance for identical trials from the same target-masker pair did not significantly vary as a function of the context in which they were presented. Moreover, the same analysis performed on all trials (not just critical trials) for each condition demonstrated identical findings, with a significant effect of SNR (β=0.10, SE β=0.197, $\chi^2(1)$=33.96, $p<0.001$) and no other significant effects or interactions ($\chi^2 < 0.14$, $p>0.05$). This indicates that average performance across conditions did not significantly vary as a result of target or masker talker consistency versus variability.

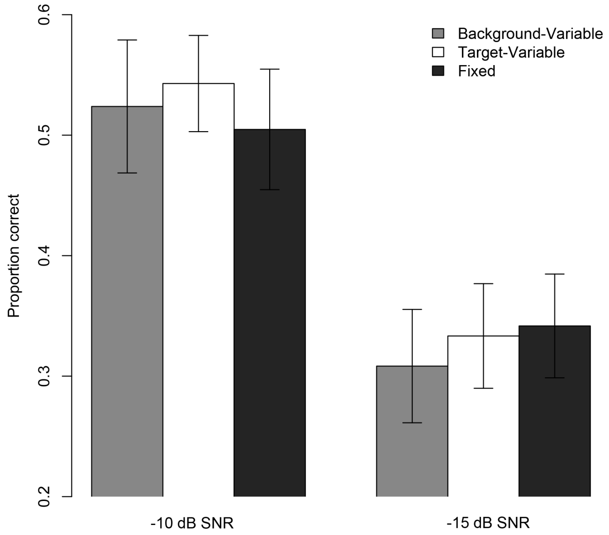

Figure 2: Mean proportion correct sentence recognition for critical trials for each condition at -10 dB SNR and -15 dB SNR (+/- 1 standard error).

In contrast to this resistance to variation in speech-in-speech recognition from target or background talker variation across conditions, there were considerable differences in sentence intelligibility across conditions and SNRs as a function of particular target and background talker combinations (Table 1). For example, in the Background-Variable condition, performance on one pair (target talker 1 with background talker 4) averaged 43% correct while performance on the same target talker with a different background talker (target talker 1 with background talker 2) averaged 54% correct, a difference of 11 percentage points. Similarly, in the Target-Variable condition, performance on one pair (target talker 2 with background talker 5) averaged 32% correct while performance on a different target talker with the same background talker (target talker 4 with background talker 5) averaged 49% correct, a difference of 17 percentage points. These differences highlight the range in intelligibility that can arise by changing the vocal characteristics of the target or background talker. That is, the stability across conditions (shown in Figure 2) stands in stark contrast to the variability (shown in Table 1) across the various target-background talker combinations within the Target-Variable and Background-Variable conditions.

Table 1: Mean proportion correct sentence recognition (standard deviation in parentheses) for each unique target-background talker pair (*Fixed accuracy only for critical trials). All trials: mean proportion correct by condition across talker pairs.

Condition	Talker Pair	Accuracy	All trials
Fixed*	1 - 5	0.42 (0.2)	0.47 (0.1)
Background -Variable	1 - 5	0.41 (0.2)	0.47 (0.1)
	1 - 4	0.43 (0.2)	
	1 - 3	0.52 (0.1)	
	1 - 2	0.54 (0.2)	
Target-Variable	1 - 5	0.43 (0.1)	0.43 (0.1)
	2 - 5	0.32 (0.2)	
	3 - 5	0.47 (0.2)	
	4 - 5	0.49 (0.1)	

4. DISCUSSION & CONCLUSIONS

The present work revealed highly stable speech-in-speech recognition accuracy across conditions that varied with respect to talker variability versus consistency in either the target or the background. While a significant effect of SNR was found, no significant differences in sentence recognition were found between any of the conditions, where the sentences to be identified and the target-background talker combination in which they were produced remained the same across conditions—only the context in which they were presented differed. This pattern of results indicates that with sentence-length stimuli, listeners exhibit rapid adaptation to talker and background talker variation.

One possible explanation for listeners' resistance to talker variation in both the target and background in this study is that listeners were provided with sentence-length stimuli, each containing 4-7 words in well-formed sentences with appropriate prosody, which may have provided listeners with sufficient time and linguistic information to adapt to the talker. Prior research demonstrating performance decrements as a product of target talker variability has primarily been restricted to recognition of monosyllabic words (e.g., [12]), which provides a much shorter time span over which to adapt to talker characteristics. Although, as noted in the introduction, some prior work has provided evidence of listener adaptation to talker consistency with sentence-length materials presented in broadband noise (e.g., [2] and [13]), although these studies involved foreign-accented speech and an explicit talker training paradigm, respectively.

Another possible explanation for the discrepancy between the present study and previous work showing a detrimental effect of target talker variability is that prior research utilized much more extensive variability, with 10 to 15 different talkers [10, 12] as compared to the 5 talkers used in this experiment. It is conceivable that performance decrements would emerge with more extensive talker variability in the target and/or in the background, though [9] did employ 10 different background talkers (with a single fixed target talker) and found no effect of fixing the background talker.

The relative stability of sentence recognition scores of the same target-background pair across different presentation contexts suggests that the energetic masking of the particular background talker on the particular target talker was the overriding influence on intelligibility. The findings of the current study indicate that listeners were proficient at accommodating the variability of a relatively small number of talkers in either the background or target speech. This suggests that in more natural communicative contexts, with a more restricted set of speakers (it is perhaps less common to engage in conversation with 15 different people) and sentence-length stimuli providing greater opportunity to adapt to talker variation, the prevailing influence on one's ability to tune into the appropriate speech stream is the talker and linguistic content characteristics of the target and of the background.

5. REFERENCES

[1] Baayen, R., Davidson, D., Bates, D. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390-412.

[2] Bradlow, A., Bent, T. 2008. Perceptual adaptation to non-native speech. *Cognition,* 106, 707-729.

[3] Brouwer, S., Bradlow, A. 2014. Contextual variability during speech-in-speech recognition. *J. Acoust. Soc. Am*. 136, EL26-EL32.

[4] Brouwer, S., Van Engen, K., Calandruccio, L., Bradlow, A. 2012. Linguistic contributions to speech-on-speech masking for native and non-native listeners: language familiarity and semantic content. *J. Acoust. Soc. Am*. 131, 1449-1464.

[5] Brungart, D. 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Am*. 109, 1101-1109.

[6] Brungart, D., Simpson B. 2004. Within-ear and across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty. *J. Acoust. Soc. Am*. 115, 301-310.

[7] Darwin, C., Brungart, D., Simpson, B. 2003. Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *J. Acoust. Soc. Am*. 114, 2913-2922.

[8] Darwin, C., Hukin, R. 2000. Effectiveness of spatial cues, prosody and talker characteristics in selective attention. *J. Acoust. Soc. Am*. 107, 970-977.

[9] Freyman, R., Helfer, K., Balakrishnan, U. 2007. Variability and uncertainty in masking by competing speech. *J. Acoust. Soc. Am*. 121, 1040-1046.

[10] Goldinger, S., Pisoni, D., Logan, J. 1991. On the nature of talker variability effects on recall of spoken word lists. *J. Exp. Psych.: Learn. Mem. Cogn*. 17, 152-162.

[11] Johnsrude, I., Mackey, A., Hakyemez, H., Alexander, E. Trang, H., Carlyon, R. 2013. Swinging at a cocktail party: voice familiarity aids speech perception in the presence of a competing voice. *Psych. Sci*. 24, 1995-2004.

[12] Mullennix, J., Pisoni, D., Martin, C. 1989. Some effects of talker variability on spoken word recognition. *J. Acoust. Soc. Am*. 85, 365-378.

[13] Nygaard, L., Pisoni, D. 1998. Talker-specific learning in speech perception. *Percep. Psychophys*. 60, 355-376.

[14] Soli, S., Wong, L. 2008. Assessment of speech intelligibility in noise with the hearing in noise test. *Intl. J. Audiology*. 47, 356-361.

[15] Van Engen, K., Bradlow, A. 2007. Sentence recognition in native- and foreign-language multi-talker background noise. *J. Acoust. Soc. Am*. 121, 519-526.