NOISE DOES NOT AFFECT WEIGHING OF SPEAKER INFORMATION IN SPOKEN-WORD RECOGNITION

Helen Reese & Eva Reinisch

Acoustics Research Institute, Austrian Academy of Sciences
hreese@kfs.oeaw.ac.at, ereinisch@kfs.oeaw.ac.at

ABSTRACT

Information about the speaker, such as gender, is known to affect phoneme categorization. However, studies examining this effect have mostly been conducted in quiet laboratory conditions. In addition to more closely resembling realistic listening, background noise has been shown to impact listeners reliance on phonetic detail relative to lexical information. Noise is therefore a useful manipulation for testing how different parts of the speech signal are processed. We tested the effect of noise on the weighing of speaker gender. Speaker gender is encoded in the acoustic signal but provides information about the speaker rather than what is being said. Its processing thus may pattern with 'lower-level' phonetic or 'higher-level' indexical information. The results of a minimal pair categorization experiment indicate that listeners' use of speaker gender during phoneme categorization is not affected by speech-shaped noise masking the speech signal. This suggests that speaker information does not pattern with phonetic detail.

Keywords: speech perception, speaker information, adverse conditions, noise

1. INTRODUCTION

Listeners are able to recognize words with remarkable accuracy, despite the immense variability in speech across speakers and listening conditions. Some factors underlying this variability have been shown to systematically affect speech perception. For instance, the identity of the speaker, such as their (inferred) age or gender, has been shown to affect how listeners perceive individual sounds [4, 6, 10]. This was first demonstrated in an experiment testing the effect of speaker gender on the perception of ambiguous stimuli along an /s/-/ʃ/ continuum [17]. Listeners were more likely to categorize stimuli along the continuum as an /s/ when it was paired with a male voice and as /ʃ/ when paired with a female voice [17]. Similar effects have since been replicated with other contrasts, such as /ʊ/-/ʌ/ [6] and /s/-/θ/ [10], and via other manipulations aimed at isolating the effect of gender from other general acoustic effects. This was achieved, for example, by presenting the same audio stimulus with images of male and female faces [10, 18].

Importantly, it has been suggested that these gender effects arise from unconscious expectations about how men and women differ in the production of certain phonemes. For example, with regards to the /s/-/ʃ/ contrast, listeners appear to expect fricatives produced by a man to have lower spectral centers of gravity than when the same phoneme is produced by a woman, corresponding to the typical characteristics of female versus male speakers. Therefore, listeners report hearing /s/, which has a higher spectral center of gravity than /ʃ/, at lower values along an /s/-/ʃ/ continuum if the speaker is perceived as male [17].

This speaker gender effect has traditionally been studied in 'ideal' laboratory conditions, that is, without any background noise or distracting tasks. More recently, however, interest has increased in how listeners fare in so-called 'adverse' conditions that more closely resemble everyday listening situations [8]. Those are often characterized by noise disrupting the speech signal or listeners whose attention is dived among multiple tasks. Not only are these conditions more realistic, but they have been shown to also allow for examining how listeners weigh different parts of the speech signal during perception, that is, to what extent they rely on acoustic properties versus, for instance, lexical information to recognize words.

In general, listening under increased cognitive load in the form of a concurrent visual search task has been shown to increase listeners' reliance on lexical information while decreasing attention to acoustic cues. That is, in experiments testing the lexical bias (i.e., "Ganong") effect [5], participants showed a larger effect, that is, they were more likely to perceive an ambiguous sound as forming part of a real word as opposed to a non-word when completing a simultaneous visual search task with the phoneme categorization task [9]. In contrast, when listening in noise, listeners up-weighed (i.e., relied more on) phonetic as opposed to lexical information [7]. In an experiment testing word segmentation, participants were more willing to rely on acoustic cues to boundaries even if it led to semantically unacceptable combinations when listening in noise than when presented with the same stimuli in clear listening conditions [7].

By examining the effect of speaker gender on phoneme categorization with the additional manipulation of adverse listening conditions, the present study aimed at better understanding the processing of speaker information during speech perception. The role of speaker gender in speech perception is difficult to categorize in this respect. It deals with higher-level information about the speaker, but it is conveyed in the acoustic signal. Specifically, we asked whether the effect of speaker gender would pattern with the processing of higher-level (e.g., lexical) information or lower-level acoustic information.

One previous study [15] already examined the role of additional cognitive load on the speaker gender effect. It was found that performing a visual search task while listening to the audio stimuli did not affect the magnitude of speaker effect in either direction. That is, listeners' reliance on speaker gender when categorizing an /s/-/ʃ/ continuum neither increased nor decreased as compared to the control condition without or a lower added cognitive load. A possible explanation for this is that speaker information is processed as acoustic information as opposed to higher-level indexical or lexical information.

The current study set out to further investigate this issue by examining the effect of background noise on the speaker gender effect. As the addition of noise to the speech signal has been shown to increase reliance on phonetic information, we hypothesized that when listening in noise participants will up-weigh speaker information together with acoustic detail, thereby increasing the size of the speaker effect. We thus conducted an experiment in which the same audio stimuli as used in [15] were overlayed with speech-shaped noise. Participants completed a phoneme categorization task under two conditions: with and without noise.

2. METHODOLOGY

2.1. Participants

Thirty native speakers of Austrian German were recruited via the online platform prolific.co [13] and compensated according to the site's payment scheme. They were aged between 18 and 40 and reported no hearing difficulties. The experiment was completed remotely. All participants reported using headphones.

2.2. Design and Materials

The experiment design followed that of Strand and Johnson's seminal study on the speaker gender effect in the perception of /s/-/ʃ/ continua [17]. Audio recordings were taken from [15] in which native speakers of Austrian German were recorded producing word initial /s/-/ʃ/ minimal word pairs (note that in Austria, as opposed to Germany, German word-initial /s/ is produced voiceless).

Recordings of four speakers (two female and two male) were selected for manipulation. Using a script in PRAAT [2], the word initial fricatives (/s/ and /ʃ/) were extracted from the recordings and interpolated to create a 15-step /s/-/ʃ/ continuum. This continuum was then spliced onto the words replacing the original word-initial fricatives. Multiple continua were created, and one was selected that the experimenters judged to sound the most natural when combined with all voices. Minimal pairs of all speakers with the same fricative continuum were then subjected to piloting. Two speakers (one female, one male) and two minimal pairs (*sein-schein* "to be" – "appearance", *senken-schenken* "to sink" – "to bestow") were selected for use in the final experiments. They showed the most typical s-shaped categorization curves suggesting that the continuum is perceived as intended, hence allowing to test for additional effects. Of the 15 continuum steps, the two endpoints and seven intermediary steps were selected for use in the experiment, for a total of nine stimuli per speaker and minimal pair. For more information on the creation and piloting of the continua as well as the spectral properties of the fricatives, see [15].

For the present study those stimuli were further manipulated. Speech shaped noise was created using a modified PRAAT script based on [19] that averaged the long-term average spectra (LTAS) of the speech stimuli and filtered a segment of white noise according to this LTAS. For the noise condition, the minimal pair continua were then overlayed with this speech shaped noise at a signal-to-noise ratio (SNR) of -3. This was the strongest distortion that still elicited typical categorization curves in a series of pilot experiments testing different SNRs. Fig. 1 illustrates the effect of the distortion on the speech signal by showing spectrograms of the same stimulus with and without the added noise.

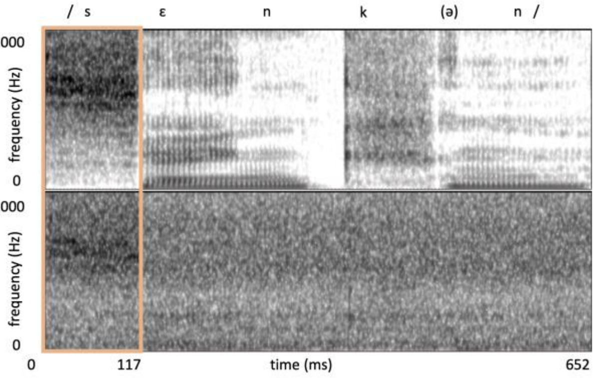

Figure 1: Spectrograms of the word *senken* (to sink) produced by the female speaker without (top) and with speech-shaped noise masking the signal (bottom).

Note that since the fricatives were also included in the noise mask, higher SNRs made it impossible for listeners to discern the acoustic properties of the fricatives.

2.3. Procedure

The experiment was created with PSYCHOPY3 [12] and hosted online via pavlovia.org [11]. It consisted of a two-alternative forced choice task in which participants were asked to identify the words along the /s/-/ʃ/ continuum. Per trial, participants were presented with a single audio stimulus, which they could hear only once per trial. They selected by button-press which of two words on the computer screen better matched the audio. In one condition, the audio files were presented overlayed with noise and in the other without noise. Context conditions were blocked, and condition order was counterbalanced across participants. Trial order was randomized for each participant. Each block consisted of 192 trials: for each minimal pair, the two end points were presented three times and the seven intermediary steps six times, thus 2 speakers × 2 minimal pairs × (2 end points × 3 repetitions + 7 intermediary steps × 6 repetitions) = 192. The experiment took approximately 15 minutes to complete.

3. RESULTS

Fig. 2 illustrates the categorization data as proportion /ʃ/-responses over the continuum.

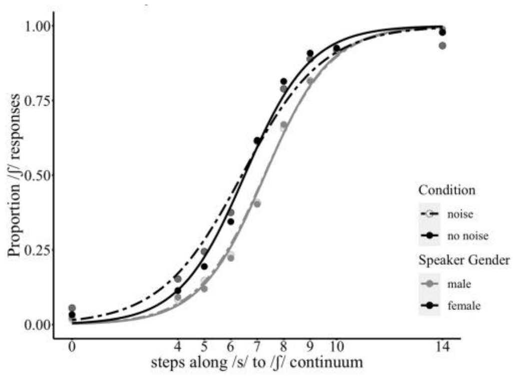

Figure 2: Proportion /ʃ/ responses over continuum steps for the two speakers (female in black, male in gray), by condition (solid lines -> noise; dashed lines -> no noise). Dots show the raw, lines the fitted data.

For statistical analyses the most extreme continuum steps were excluded due to near ceiling identification. Statistical analyses were conducted by means of a generalized linear mixed-effects model with a logistic linking function using the lme4 [1] package in R [14]. The dependent variable was response (/ʃ/ coded as 1, /s/ as 0). Fixed effects were Continuum Step (centered on 0), Condition (noise coded as 0.5, no noise as -0.5), Speaker Gender (female coded as 0.5, male as -0.5) and all interactions. The random-effects structure included an intercept for participants with random slopes for Continuum Step and Condition. This was the maximal random-effects structure that converged. Results are reported in Table 1.

Factor	b	z	p
Intercept	0.05	0.40	0.689
Step	0.95	18.52	<0.001
Condition	0.03	0.54	0.588
Gender	0.73	8.36	<0.001
Step:Condition	-0.07	-2.16	0.031
Step:Gender	-0.03	-0.91	0.362
Condition:Gender	-0.01	-0.05	0.959
Step:Condition:Gender	-0.10	-1.60	0.110

Table 1: Results of the statistical analysis. The model structure was: (phoneme classification ~ continuum step * condition * speaker gender + (1 + continuum step + speaker gender | participant)).

Significant effects were found for Continuum Step (more /ʃ/-responses the more /ʃ/-like the fricative) and for Speaker Gender, thus replicating the speaker effect that more /ʃ/-responses were given for the stimuli produced by the female speaker. The interaction Continuum Step x Condition was also significant, indicating that the effect of continuum step is smaller, that is, the categorization function is less steep in the noise than in the clear condition. No significant interaction was found for Step x Gender, indicating that the categorization function did not differ in steepness for the stimuli produced by the male and female speaker. Importantly, as for our main manipulation of listening condition, neither the interaction Condition x Gender nor the three-way interaction Step x Condition x Gender were significant, which signals that the noise condition did not affect the use of speaker gender in the categorization task.

4. DISCUSSION

The present study set out to further explore the effect of speaker gender in phoneme categorization, since the mechanism behind it is not clear. Information about the gender of the speaker is (mostly) contained in the acoustic signal (e.g., F0, voice quality etc.). At the same time, speaker gender provides higher-level indexical information about the speaker. In speech perception it could hence either pattern with the processing of lower-level acoustic or higher-level (e.g., indexical; also, lexical) information. While overall, listeners have been shown to use all types of information available (or accessible) to them for speech perception, differences in the weighing of different types of information have been demonstrated especially in adverse listening conditions [7, 8].

The present study made use of adverse listening conditions to assess the nature of the speaker gender effect. We asked whether the addition of speech-shaped noise masking the acoustic signal would increase listeners' reliance on speaker gender in spoken-word recognition. This is because previous studies have shown an increase in listeners' reliance on acoustic relative to higher-level lexical information when listening in background noise [7]. We hence tested the hypothesis that the effect of speaker gender in phoneme categorization patterns with low-level acoustic listening in noise.

To examine this, we conducted an experiment in which participants categorized /s/-/ʃ/ minimal pair words. The words contained an initial fricative taken from a single /s/-/ʃ/ continuum and then combined with word ends (i.e., the rest of the word following the fricative) produced by one male and one female speaker. In one condition, listeners heard the audio stimuli in clear listening conditions and in the other the stimuli were overlayed with speech-shaped noise. In both conditions we replicated the speaker effect as it was first established in [17], that is, that more /ʃ/ responses were given for the female than the male speaker. However, the magnitude of this effect did not differ significantly between the noise and the clear listening conditions. The hypothesis that the speaker gender effect would pattern with the up-weighing of acoustic relative to higher-level information in noise could hence not be supported.

Interestingly, a previous study on the impact of another type of adverse listening condition on the speaker effect could not find support for its processing to pattern with higher-level lexical information either [13]. In this study, the addition of cognitive load in the form of a simultaneous visual search task was likewise shown not to affect the magnitude of the speaker effect [13]. In other words, the addition of a visual search task did not lead listeners to up-weigh speaker gender as a cue in the categorization task, while the same manipulation has been shown to cause listeners to rely more on lexical information by increasing the size of the lexical bias effect in experiments testing word segmentation and phonetic categorization under cognitive load [7, 9].

Evidently, the speaker effect does not pattern with effects found in previous studies implementing cognitive load and noise to test the differential weighing of acoustic versus higher-level information. Additional findings might speak to this apparent null result regarding the modulation of the speaker effect. First, it must be noted that noise "covers" the speech signal, hence acoustic cues are masked and not as easily available for listeners to use as in clear listening conditions. The supposed up-weighing of acoustic cues during listening in noise must therefore be seen relative to the use of other types of information [7]. In the present study, similarly to [15] in their study on cognitive load, the speaker effect was not directly compared to other (e.g., lexical) effects. Secondly, other studies that specifically targeted low-level acoustic processing effects such as speechrate-dependent phonetic categorization also failed to find modulating effects of cognitive load or noisy environments [3, 16].

Finally, while not affecting the magnitude of the speaker effect, the noise manipulation in the present study did show an effect on the perception of the continuum (as indicated by the Condition by Continuum Step interaction), namely resulting in a shallower categorization curve in the noise condition than in the clear listening condition. This suggests that listeners were less able to access acoustic information in the fricatives (also a kind of noise) in the categorization task when listening in noise. Similar effects on the categorization curve were found in the experiment on the effect of cognitive load on the speaker gender effect [13].

Where does this leave us with regard to the processing of speaker information? The role of speaker gender information in spoken-word recognition cannot be clearly categorized as patterning with acoustic or lexical information. One explanation for this could be that speaker gender information is processed independently of or in a different fashion than low-level acoustic information or higher-level (lexical) information. However, the exact mechanism remains to be described, possibly by using different experimental paradigms, such as by studying the time-course of when speaker (gender) information is employed during word recognition.

5. REFERENCES

[1] Bates, D. M., & Sarkar, D. 2007. lme4: Linear mixed-effects models using S4 classes (Version 0.999375–27) [Computer program]. http://www.r-project.org/: R Foundation for Statistical Computing.

[2] Boersma, P. & Weenink, D. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.50.

[3] Bosker, H.-R., Reinisch, E. & Sjerps, M. J. (2017). Cognitive load makes speech sound fast, but does not modulate acoustic context effects. Journal of Memory and Language, 94, 166-176.

[4] Drager, K. 2011. Speaker age and vowel perception. *Language and Speech* 54, 99-121.

[5] Ganong, W. F. 1980. Phonetic categorization in auditory word perception. *Journal of experimental*

psychology: Human perception and performance, 6, 110-125.

[6] Johnson, K., Strand, E. A., & D'Imperio, M. 1999. Auditory–visual integration of talker gender in vowel perception. *Journal of phonetics* 27, 359-384.

[7] Mattys, S. L., Brooks, J., & Cooke, M. 2009. Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive psychology* 59, 203-243.

[8] Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. 2012. Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, 27, 953-978.

[9] Matty, S. L., & Wiget, L. 2011. “Effects of cognitive load on speech recognition,” J. Mem. Lang. 65, 145–160.

[10] Munson, B. 2011. The influence of actual and imputed talker gender on fricative perception, revisited (L). *J. Acoust. Soc. Am*. 130, 2631-2634.

[11] Pavlovia, https://www.pavlovia.org/

[12] Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. 2019. PsychoPy2: experiments in behavior made easy. *Behavior Research Methods.*

[13] Prolific, https://www.prolific.co/.

[14] R Core Team. 2021. R: A language and environment for statistical computing. (version 4.1.2) R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.R

[15] Reese, H., & Reinisch, E. 2022. Cognitive load does not increase reliance on speaker information in phonetic categorization. *JASA Express Letters* 2.

[16] Reinisch, E., & Bosker, H. R. (2022). Encoding speech rate in challenging listening conditions: white noise and reverberation. Attention, Perception, & Psychophysics, 84, 2303-2318.

[17] Strand, E. & Johnson, K. 1996. Gradient and Visual Speaker Normalization in the Perception of Fricatives. *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference*, 14-26.

[18] Strand, E. 1999. Uncovering the role of gender stereotypes in speech perception. *Journal of language and social psychology* 18, 86-100.

[19] Quené, H., & Van Delft, L. E. (2010). Non-native durational patterns decrease speech intelligibility. *Speech Communication* 52, 911-918.