NATURAL AND SYNTHETIC SPEECH COMPREHENSION IN SIMULATED TONAL AND PULSATILE TINNITUS: A PILOT STUDY

Jacek Kudera, Marjolein van Os, Bernd Möbius

Department of Language Science and Technology, Saarland University
kudera@coli.uni-saarland.de

Abstract: This paper summarizes the results of a Modified Rhyme Test conducted with masked stimuli to simulate two common types of hearing impairment: bilateral pulsatile and pure tinnitus. Two types of stimuli, meaningful German words (natural read speech and TTS output) differing in initial or final positioned minimal pairs were modified to correspond to six listening conditions. Results showed higher recognition scores for natural speech compared to synthetic and better intelligibility for pulsatile tinnitus noise over pure tone tinnitus. These insights are of relevance given the alarming rates of tinnitus in epidemiological reports.

1 Introduction

More than a third of adults above 55 and 10-15% of worldwide population of all ages experience tinnitus: a common condition of sound perception in the absence of acoustic stimuli [1, 2, 3]. This hearing impairment symptom evokes the sensation of hissing noise or ringing sound which affects speech comprehension [4] and in consequence leads to a decrease of communication capabilities [5]. Studies on younger adults report an increased number of tinnitus patients over the years [6] and identify recreational noise exposure as a cause of hearing loss, hyperacusis and noise-induced tinnitus [7]. Therefore, the question of speech comprehension under conditions of hearing disfunctions remains valid.

The present study aims to answer the following questions: (i) how does speech comprehension differ under simulated tinnitus conditions in natural and synthetic speech?; (ii) which hearing impairment symptom (binaural tonal or pulsatile tinnitus) has a greater effect on speech intelligibility?; and (iii) in which position in the word (onset or coda) is recognition of consonants affected most by the masking sound? We hypothesize that speech comprehension across conditions differs and that the exposure to synthetic speech in noise causes higher comprehension scores in comparison to natural read speech. We assume that high concatenation cost in the unit selection system can contribute to speech intelligibility in noise in contrast to natural read speech signal characterized by smooth transitions. Furthermore, we expect to find better performance in tonal than in pulsatile tinnitus across conditions, due to the larger spectral overlap of speech and masker in the latter tinnitus noise. Finally, we expect that identification scores of consonants in onsets are higher than the ones in final positions, as a result of a higher initial functional load and onset prominence.

1.1 Epidemiology & Aetiology

Epidemiological reports concerning hearing disorders suggest that prolonged tinnitus affects 13% of the German population [8]. Most epidemiological reports show prevalence rates from 10 to 15% across societies [9]. In January 2021, the German DTL (Deutsche Tinnitus-Liga,

Gemeinnützige Selbsthilfeorganisation gegen Tinnitus, Hörsturz und Morbus Menière) organization reports over one million Germans severely suffering from tinnitus, suggesting a higher proportion. Prevalence in both sexes is similar [10]. Age distribution, however, appears difficult to estimate, because children seem to be less distressed by noise perception than adults [11], so it does not get reported as often. Also the pathophysiology of tinnitus remains unclear. It is hypothesized that spontaneous firing of neurons in the central auditory system contributes to the sensation. Genetic predisposition was estimated as a small risk factor [12]; whereas ototoxic drugs intake, hypertension, high frequency hearing loss, Meningitis, Otitis media, Ménière's disease, vestibular vertigo, epilepsy, along with psychological conditions such as depression and emotional trauma contribute to the aetiology of tinnitus [9].

1.2 Related work

Several previous studies introduced the simulation of hearing deficit conditions by manipulation of the stimuli [13]. The rationale for such procedures lies in the ethical considerations of inducing the hearing loss by the means of more invasive methods. A recent investigation by Marrufo-Pérez et al. [14] showed that adaptation to noisy environments during human speech recognition depends on the noise level distribution, and that in continuous noise, for example, neural dynamic range adaptation can lead to improvement of speech envelope decoding [15, 16]. This suggests that not all types of noise conditions lead to decreased speech comprehension.

Investigations on comprehension of synthetic speech in noise suggest that lower fluctuations of background amplitude may facilitate noise adaptation [17]. The adaptation strategies depend on the duration of sound perception. Longer exposure to tinnitus symptom may also result in sound habituation. Studies on short stimuli comprehension in speech masked with pulsatile noise showed that short temporal minima in fluctuating masking sounds can serve as speech cues [18]. This phenomenon called 'valley listening' or 'masking release effect' was previously tested on healthy and hearing-impaired subjects [17].

The relation between audiometric thresholds and tinnitus seems not to be straightforward. Some patients with diagnosed hearing loss do not experience tinnitus, whereas the other tinnitus patients exhibit intact hearing thresholds [9, 10]. Taken together, these findings suggest that simulated hearing deficits might not be directly comparable to the hearing abilities in particular of people with long-term histories with these deficits. Nevertheless, it is important to investigate how different hearing impairments affect patients also in early stages.

2 Method

In contrast to former invasive studies involving the noise induction technique [19], we modified the stimuli in order to reflect the hearing disorder and thus test speech comprehension using a non-invasive method. We simulated both binaural tonal and pulsatile tinnitus. A closed-set Modified Rhyme Test (MRT) [20, 21] was constructed in German. It consists of 15 sets of six words differing in the initial consonant clusters and 15 sets of six words differing in the final consonant clusters, leading to a total set of 180 items. In both sets syllable nuclei and final or initial clusters, respectively, were shared among all words of a set. See Table 1 for example stimuli.

Table 1 – Example of stimulus sets differing in initial and final consonants, respectively.

Initial	Hast	Last	Mast	Rast	Gast	passt
Final	Reis	Reif	Reim	Rhein	reit	Reiz

2.1 Stimuli

The total set of 180 German words was synthesized using the Mary TTS system [22] as well as read by a female German native speaker for natural recordings. We used the unit-selection speech synthesizer to test the effect of a lower degree of coarticulation on the intelligibility. The unit-selection mode, in contrast to current state-of-the-art synthesizers based on neural networks, often results in a more mechanic output due to the possible rapid transitions between units. This, however, can be seen as an advantage in acoustic surroundings which demand increased listening effort. One of the typical strategies of conveying a message in noise is to increase the volume to counterbalance the background intensity and to reduce the speech rate. Another possibility lies in hyperarticulation, comprising staccato speech. So, as a parallel to the behavioural phenomenon, the higher unit distortion might appear to have an advantage over smooth transitions at concatenation points in synthetic speech when perceived in noisy environments. The rapid spectral and waveform changes typical for unit-selection output may then result in a lower coarticulation effect and hence can lead to poorer MOS ratings but at the same time to higher speech intelligibility in noise like in tinnitus.

Pure tone tinnitus was simulated by overlapping the masking sound (SNR 15 dB, 5 kHz) with the samples. To reflect a binaural pulsatile tinnitus, a noise masker (AM depth 5 dB, centered at 6 kHz) was rendered with a 250 ms onset cosine transient stage and merged with spoken and TTS samples (both 65 dB SPL). Additionally, there was a control condition (Quiet) with no masking noise. The SNR levels that we used correspond to the thresholds most frequently reported in clinical studies and often applied in experiments on simulated hearing disfunctions (-10/-15 dB sensation level) [23, 24, 25]. Further evidence for the SNR levels in simulated hearing disfunctions is provided by measures of spontaneous otoacoustic emissions (SOAEs) and transient evoked otoacoustic emissions (TEOAEs) with an application of toneburst stimuli [26]. The results mentioned above and our experimental design required the application of maskers within the normal hearing thresholds.

2.2 Participants

We tested a total of 100 participants in our experiment (age range 18-61 years, mean age 30 years, 60 were male), half of which participated in the subset of items differing in the initial consonant, and the other half participated in the subset of items differing in the final consonant. All participants were native speakers of German recruited via a crowd sourcing platform and were paid for their participation. None of the subjects reported having hearing difficulties.

2.3 Procedure

A list of German meaningful words containing minimal pairs in final and initial positions was created. The items were then recorded, synthesized and masked to comply with different listening conditions, i.e., 2 speech modes (natural and synthetic) x 3 noise manipulations (pure tone tinnitus, pulsatile tinnitus, and quiet). We included fifteen items of each type and thus presented the same set of six words multiple times in different conditions, where the target word was always different. This resulted in a total of 90 items per list, which were given to the participants in random order. Completion of the entire experiment took approximately fifteen minutes.

In the experiment, participants listened to the words and selected the correct option from the set of six words shown in written form on the screen. Before the start of the experiment, participants completed a short demographics questionnaire and were shown the task instructions. The instructions were followed by three practice items to help participants get familiar with the task. The sound played automatically once the trial was loaded and participants had the option

to replay the sound multiple times if they wished. There was no feedback on the practice items, nor during the experiment itself.

3 Results

In our analysis, we coded whether the word selected by the participant was the correct or incorrect word. Due to error in the experimental setup, the responses to two out of fifteen items in two synthetic conditions (pure tone initial and pulse tone final) were not recorded. For each condition, the mean accuracy scores were calculated and are presented in Table 1. Here, a score of 1 would mean 100% accuracy. As there were six response choices, chance was at 0.167. Figure 1 shows for each condition the amount of correct and incorrect responses. We used the quiet control condition to check whether participants had extreme difficulty with the task. Overall, we found incorrect responses in 2.35 out of 30 items (SD = 1.72, range 0-9) and chose not to exclude any participants based on their performance in the quiet condition.

For the analysis, we used general linear mixed models (GLMM), implemented in the lme4 package [27] in R [28]. We tested the participants' binomial response (0 = incorrect, 1 = correct) using a GLMM with a logistic linking function. The model included fixed effects of Speech Mode (categorical predictor with two levels using dummy coding, mapping Natural speech to the intercept), Tinnitus Manipulation (categorical predictor with three levels using dummy coding, mapping Quiet to the intercept), and Position (categorical predictor with two levels using dummy coding, mapping Final to the intercept). Additionally, the model included the interaction of Speech Mode and Position, the interaction of Noise Manipulation and Position, and the interaction of Speech Mode and Noise Manipulation. A by-Participant random intercept was included, as was a by-Item random intercept.

The model revealed a significant effect of Speech Mode, where synthetic speech is more difficult to identify correctly than natural speech ($\beta = -2.70, SE = 1.32, z = -2.05, p < .05$). There was a significant effect of Noise Manipulation, with the Quiet control condition being easier than either of the tinnitus simulations ($\beta = -1.64, SE = 0.18, z = -8.75, p < .001$) for Pulse tone tinnitus, and ($\beta = -2.02, SE = 0.18, z = -10.93, p < .001$) for Pure tone tinnitus. Pulse tone tinnitus and Pure tone tinnitus also differed significantly from each other, revealing that Pulse tone tinnitus is a slightly easier condition than Pure tone tinnitus ($\beta = 0.38, SE = 0.12, z = 3.29, p < .01$). We found no significant effect of Position ($p = .077$).

The model additionally revealed a significant interaction effect of Noise Manipulation and Position, showing fewer correct responses in Initial Pure tone items ($\beta = -0.44, SE = 0.17, z = -2.63, p < .01$). The interaction of Speech Mode and Noise Manipulation was also significant, indicating more correct responses in synthetic speech in Pure tone tinnitus conditions. ($\beta = 0.72, SE = 0.19, z = 3.70, p < .001$). Additionally, TTS output in Pulse tone tinnitus led to fewer correct responses than Pure tone tinnitus when Pure tone tinnitus is mapped to the intercept ($\beta = -0.66, SE = 0.12, z = -5.26, p < .001$).

Table 2 – Descriptive statistics for correctly identified words in each condition: M(SD).

	Initial			**Final**		
	Pure	Pulse	Quiet	Pure	Pulse	Quiet
Natural	.800 (.400)	.856 (.351)	.975 (.157)	.817 (.387)	.836 (.371)	.969 (.172)
Synthetic	.683 (.466)	.538 (.500)	.900 (.300)	.584 (.493)	.554 (.493)	.843 (.364)

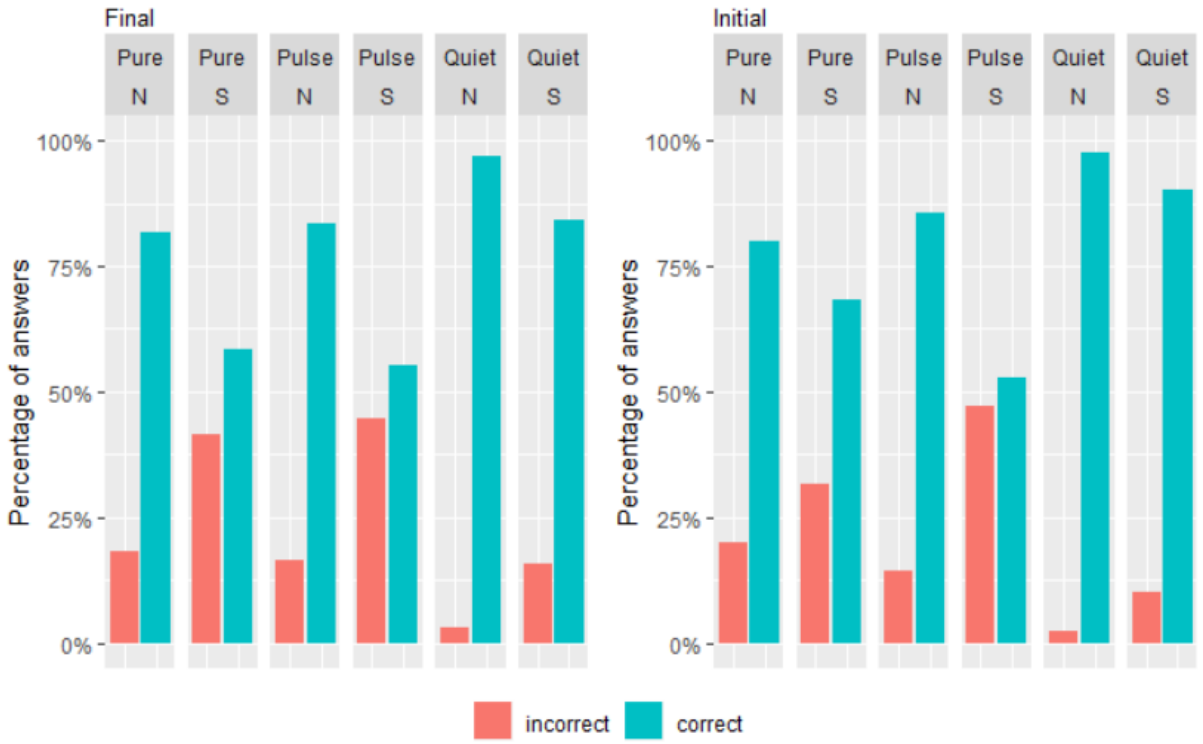

Figure 1 – Correct and incorrect responses in percentages for each condition (Pure, Pulse, Quiet in Natural (N) and Synthetic speech (S). Left - final position contrasts; right - initial position contrasts.

4 Discussion

In this study we investigated speech comprehension based on natural read and synthetic stimuli in simulated tinnitus conditions, aiming to shed light on the question of how tinnitus affects spoken language intelligibility.

4.1 Speech mode

The first question we addressed was to compare the effects of simulated tinnitus on natural speech on the one hand and TTS output on the other. We expected that synthetic speech might be easier to identify correctly in noisy environments than natural speech. However, contrary to these expectations, the results showed that synthetic stimuli were harder to understand than natural ones. A possible explanation for this finding is a degree of exposure to different speech modes. On average we are much more familiar with human voices than synthetic ones. This could make it easier to adapt to the natural speech in noise and identify the words correctly. Another factor which should be discussed is the quality of 'natural' speech in this study. The natural stimuli used here were carefully read lab tokens. The results might differ if stimuli were spontaneous speech recordings instead.

4.2 Tinnitus type

We compared simulated pure tone tinnitus with pulsatile tinnitus, expecting to find higher recognition scores in pure tone tinnitus. Again, the results disconfirmed our expectations. The short term noise adaptation and the degree of spectral masker–speech overlap resulted in a better performance after exposure to pulsatile masker. Generally, we are not as often exposed to pure tones in everyday life, so participants might have needed more time to adapt to this background. Also, exposure to pure tone maskers might cause irritation and lead to less careful listening.

4.3 Consonant position

The set of stimuli used in the present study consisted of words differing in the position of the sounds creating the minimal pairs. We expected to find more correctly identified items in the words differing in the initial position, due to a higher functional load and acoustic prominence of the initial segments than of codas. The results showed no significant difference between initial and final conditions, although there was a trend in higher recognition scores (overall means) for words differing in the initial consonants.

4.4 Outlook

The increasing diagnostic rates of tinnitus across all population strata are alarming. Findings of the present study may shed a light on a problem of speech recognition in noise, by simulating a common hearing disfunction. Several previous studies involving hearing-impaired listeners showed a lower degree of noise-adaptation in word recognition tasks [29] and pointed to differences in scores when exposed to natural and synthetic speech [15]. We found higher recognition scores for natural speech compared to synthetic speech, and better intelligibility for pulsatile tinnitus noise over pure tone tinnitus.

A possible continuation of this line of research could engage subjects reporting tinnitus symptoms in exposure to unmasked speech. To control such a design one could make use of available gradual scales, e.g., the Tinnitus Severity Index or the Tinnitus Handicap Inventory [2], and measure speech comprehension thresholds across available categories (none, mild, moderate and severe). Other possibilities for an extension of the current study are to further modulate the signal, to apply different speech synthesis techniques, or to investigate the degree of intelligibility in sentences rather than isolated words.

Funding

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

[1] CHAN, Y.: *Tinnitus: etiology, classification, characteristics, and treatment*. *Discovery medicine*, 8(42), pp. 133–136, 2009.

[2] HENRY, J. A., K. C. DENNIS, and M. A. SCHECHTER: *General review of tinnitus*. *Journal of speech, language, and hearing research*, 2005.

[3] MCCOMBE, A., D. BAGULEY, R. COLES, L. MCKENNA, C. MCKINNEY, and P. WINDLE-TAYLOR: *Guidelines for the grading of tinnitus severity: the results of a working group commissioned by the british association of otolaryngologists, head and neck surgeons, 1999*. *Clinical Otolaryngology & Allied Sciences*, 26(5), pp. 388–393, 2001.

[4] GILLES, A., W. SCHLEE, S. RABAU, K. WOUTERS, E. FRANSEN, and P. VAN DE HEYNING: *Decreased speech-in-noise understanding in young adults with tinnitus*. *Frontiers in neuroscience*, 10, p. 288, 2016.

[5] MCCORMACK, A., M. EDMONDSON-JONES, S. SOMERSET, and D. HALL: *A systematic review of the reporting of tinnitus prevalence and severity*. *Hearing research*, 337, pp. 70–79, 2016.

[6] HENDERSON, E., M. A. TESTA, and C. HARTNICK: *Prevalence of noise-induced hearing-threshold shifts and hearing loss among us youths*. *Pediatrics*, 127(1), pp. e39–e46, 2011.

[7] BEACH, E., W. WILLIAMS, and M. GILLIVER: *Estimating young australian adults' risk of hearing damage from selected leisure activities*. *Ear and hearing*, 34(1), pp. 75–82, 2013.

[8] PILGRAMM, M., R. RYCHLICK, H. LEBISCH, H. SIEDENTOP, G. GOEBEL, and D. KIRCHHOFF: *Tinnitus in the federal republic of germany: a representative epidemiological study*. In *Proceedings of the sixth international tinnitus seminar*, pp. 64–67. The Tinnitus and Hyperacusis Centre Cambridge, 1999.

[9] BAGULEY, D., D. MCFERRAN, and D. HALL: *Tinnitus*. *The Lancet*, 382(9904), pp. 1600–1607, 2013.

[10] DAVIS, A. and E. A. RAFAIE: *Epidemiology of tinnitus*. *Tinnitus handbook*, 1, p. 23, 2000.

[11] BAGULEY, D. M. and D. MCFERRAN: *Tinnitus in childhood*. *International journal of pediatric otorhinolaryngology*, 49(2), pp. 99–105, 1999.

[12] KVESTAD, E., N. CZAJKOWSKI, B. ENGDAHL, H. J. HOFFMAN, and K. TAMBS: *Low heritability of tinnitus: results from the second nord-trøndelag health study*. *Archives of Otolaryngology–Head & Neck Surgery*, 136(2), pp. 178–182, 2010.

[13] SCHAETTE, R., C. TURTLE, and K. J. MUNRO: *Reversible induction of phantom auditory sensations through simulated unilateral hearing loss*. *PLoS One*, 7(6), p. e35238, 2012.

[14] MARRUFO-PÉREZ, M. I., D. DEL PILAR STURLA-CARRETO, A. EUSTAQUIO-MARTÍN, and E. A. LOPEZ-POVEDA: *Adaptation to noise in human speech recognition depends on noise-level statistics and fast dynamic-range compression*. *Journal of Neuroscience*, 40(34), pp. 6613–6623, 2020.

[15] MARRUFO-PÉREZ, M. I., A. EUSTAQUIO-MARTÍN, and E. A. LOPEZ-POVEDA: *Adaptation to noise in human speech recognition unrelated to the medial olivocochlear reflex*. *Journal of Neuroscience*, 38(17), pp. 4138–4145, 2018.

[16] AINSWORTH, W. and G. MEYER: *Recognition of plosive syllables in noise: Comparison of an auditory model with human performance*. *The Journal of the Acoustical Society of America*, 96(2), pp. 687–694, 1994.

[17] DUBNO, J. R., A. R. HORWITZ, and J. B. AHLSTROM: *Benefit of modulated maskers for speech recognition by younger and older adults with normal hearing*. *The Journal of the Acoustical Society of America*, 111(6), pp. 2897–2907, 2002.

[18] FÜLLGRABE, C., F. BERTHOMMIER, and C. LORENZI: *Masking release for consonant features in temporally fluctuating background noise*. *Hearing research*, 211(1-2), pp. 74–84, 2006.

[19] CHERMAK, G. D. and J. E. DENGERINK: *Characteristics of temporary noise-induced tinnitus in male and female subjects*. *Scandinavian Audiology*, 16(2), pp. 67–73, 1987.

[20] HOUSE, A. S., C. WILLIAMS, M. H. HECKER, and K. D. KRYTER: *Psychoacoustic speech tests: A modified rhyme test*. *The Journal of the Acoustical Society of America*, 35(11), pp. 1899–1899, 1963.

[21] HOUSE, A. S., C. E. WILLIAMS, M. H. HECKER, and K. D. KRYTER: *Articulation-testing methods: consonantal differentiation with a closed-response set*. *The Journal of the Acoustical Society of America*, 37(1), pp. 158–166, 1965.

[22] SCHRÖDER, M. and J. TROUVAIN: *The german text-to-speech synthesis system mary: A tool for research, development and teaching*. *International Journal of Speech Technology*, 6(4), pp. 365–377, 2003.

[23] SEREDA, M., D. A. HALL, D. J. BOSNYAK, M. EDMONDSON-JONES, L. E. ROBERTS, P. ADJAMIAN, and A. R. PALMER: *Re-examining the relationship between audiometric profile and tinnitus pitch*. *International journal of audiology*, 50(5), pp. 303–312, 2011.

[24] SHAILER, M., R. TYLER, and R. COLES: *Critical masking bands for sensorineural tinnitus*. *Scandinavian Audiology*, 10(3), pp. 157–162, 1981.

[25] PAUL, B. T., I. C. BRUCE, and L. E. ROBERTS: *Evidence that hidden hearing loss underlies amplitude modulation encoding deficits in individuals with and without tinnitus*. *Hearing Research*, 344, pp. 170–182, 2017.

[26] NORTON, S. J., A. R. SCHMIDT, and L. J. STOVER: *Tinnitus and otoacoustic emissions: Is there a link?* *Ear and hearing*, 11(2), pp. 159–166, 1990.

[27] BATES, D., M. MAECHLER, B. BOLKER, S. WALKER, R. H. B. CHRISTENSEN, H. SINGMANN, B. DAI, and F. SCHEIPL: *Package 'lme4'*. *CRAN. R Foundation for Statistical Computing, Vienna, Austria*, 2012.

[28] R CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

[29] BEN-DAVID, B. M., Y. VANIA, and B. A. SCHNEIDER: *Does it take older adults longer than younger adults to perceptually segregate a speech target from a background masker?* *Hearing research*, 290(1-2), pp. 55–63, 2012.