THE EFFECT OF TEMPORALLY FLUCTUATING MASKERS ON SPEECH PRODUCTION AND COMMUNICATION

Julie Saigusa, Valerie Hazan

Speech Hearing and Phonetic Sciences, UCL (University College London), UK
julie.saigusa.16@ucl.ac.uk, v.hazan@ucl.ac.uk

ABSTRACT

Recent work has found that neural oscillators entrain with envelope modulations of attended speech in noise, and that speakers produce more pronounced amplitude modulations in noisy conditions. This study aims to expand on this finding, and to investigate how speakers adapt to temporal fluctuations in masking noise. 16 adult female participants were recorded reading sentences to a partner who repeated back what they heard while both were in quiet, in speech-shaped noise (SSN), and in SSN modulated by 1Hz, 4Hz, or 8Hz square waves. The amplitude modulation spectra for speech produced each condition were calculated. More pronounced modulations were produced in noise, with most similar effects for the 1Hz and SSN conditions. This suggests that changes in amplitude modulations are adaptations to the specific environment rather than a generalised response to noise, and that adaptations are not simply a function of the energetic content of the masker.

Keywords: Speech production, speech in noise, temporal modulation.

1. INTRODUCTION

The presence of background noise in the environment is a large factor in how successfully we communicate with each other. In such environments, speech perception is often hindered, and speech production manifests the Lombard Effect, or an increase in acoustic features, especially loudness, in response to noise [7,11]. Since the discovery of the Lombard effect, several decades of research have been devoted to investigating how it manifests as a function of loudness, task type, spectrotemporal masker content, among other properties. Recently, the study of speech production in noise has expanded to consider the role of amplitude modulations.

The temporal envelope of a speech signal exhibits relatively slow amplitude modulations that are well-known to be important to speech perception [3,8]. Stress rate (1-2Hz), syllable rate (2-8Hz), and phoneme rate (8-40Hz) all contribute to the structure of the envelope [3]. Removing amplitude modulations from speech impairs speech perception, while 'repackaging' the envelope to be within an ideal 'theta' range (3-9Hz) improves it [4]. Recent neurophysiological work has found that neural oscillators 'entrain' to the envelope of an attended stream in noise, thought to be an important mechanism of successful speech perception [6,8].

Recent work has also shown corresponding changes in speech production. Peak amplitude and modulation rate have been found to differ across stress, syllable, and mora-timed languages [21]. In an examination of 1,445 sentences across 4 read-sentence corpora (all in intense steady energetic noise), [3] found more pronounced amplitude modulations for speech produced in noise, compared to speech produced in quiet, especially in the delta and theta ranges (1-8Hz) that reflect syllable rate and contribute to speech perception. Similar findings of increased modulation depth have been found for speakers instructed to speak clearly [10]. More pronounced modulations in the 3-4Hz range, for example, imply that speakers are more consistently conveying important stress and syllabic information when trying to be understood. This study aims to expand on this finding by adding a communicative element not present in [3] involving sentence reading between partners, and examines a larger number of speakers in a single corpus in a range of conditions.

There is other evidence from speech production research that demonstrates speakers can adapt the temporal structure of their speech, particularly in response to temporally fluctuating masking noise. In [15], energetic maskers fluctuating from 1 to 16Hz resulted in a small increase in speech energy in the 'dips' compared to regions where the masker was on. Temporally 'sparse' energetic maskers containing silent intervals of varying lengths (created by modulating speech-shaped noise with the envelope of one or more speakers) resulted in speakers timing their speech in a collaborative sudoku task with a 'wait and talk' strategy [2], timing their onsets to avoid coinciding with masker onsets. Since the pauses were not predictable, changes in speaker behaviour were seen as reactive and took place over several hundred milliseconds. The present study employs steadily fluctuating maskers, in an attempt to investigate the extent to which speakers can adjust when the disruption is predictable.

The study sought to investigate to what degree speakers can adapt to a temporally fluctuating masker, what kind of adaptations they make, and the effects these adaptations have on the modulation spectrum. As speakers are aware of temporal fluctuations and can adjust to some extent [2,15], it was predicted that there would be more pronounced modulations across all noise types compared to quiet, and that speakers may try to adjust to get the most information across in the 'gaps', manifesting as peaks in the modulation spectrum at the rate of the masker, especially at the slower rates of maskers fluctuating at 1 and 4Hz, as speakers time speech in the opposite 'phase'. For example, a masker fluctuating at 4Hz may cause a speaker to adopt a 4Hz syllable rate to best take advantage of silent intervals.

2. METHOD

2.1. Participants

16 adult female speakers of Southern British English participated in the study (18-32 years, M=23.3, SD=4.7). All participants had normal hearing thresholds as assessed by pure-tone audiometry and reported no speech, hearing, or language impairments. Normal hearing was defined as a threshold <20dB between 0.25 - 8kHz. Participants were native SSBE speakers and were not bilingual from birth or a young age.

2.2. Procedure

2.2.1. Task

The task was designed to elicit read speech in a communicative situation. Each recording session included two participants who did not know each other, who were seated in separate sound-treated booths and communicated via Beyerdynamic DT297 headsets. One participant, 'Talker A', read sentences from the Harvard corpus to their partner, 'Talker B', who repeated back what they heard. The sentences were presented on a computer screen using the presentation software ProRec [14] and Talker A was able to click an arrow to move to the next sentence. When the task was complete, the participants switched places and Talker A became Talker B, and vice versa.

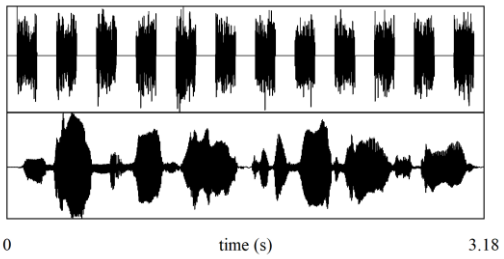
The first 10 lists of the Harvard corpus were used, with sentence and list order randomised and counterbalanced across participants. Therefore, the same sets, spoken by different participants, were analysed across conditions. Participants did not hear the same sentences more than once. Participants were told that the accuracy of Talker B's repetition would be scored, and that it was the job of Talker A to make sure they were understood. Talker A was instructed to read each sentence only once. Before the main task commenced, pairs were given 10 sentences as practice to familiarise them with the procedure.

2.2.2. Experimental conditions

The purpose of the study was to investigate the effects of temporally fluctuating maskers on speech production in noise. Four experimental noise conditions were used, and one quiet control (NORM) condition. Speech-shaped noise (SSN) generated using the long-term spectrum of female read speech was modulated by 1Hz, 4Hz, and 8Hz square waves with a modulation depth of one, which resulted in a stark on-off effect (Figure 1). These three noise conditions, along with unmodulated SSN, were used.

Both Talker A and B were presented with each noise at 80dB SPL, as measured with an artificial ear, inside a 'virtual room' Audio3D [1] to simulate the acoustics of a real room within headphones. In the quiet condition, there was no noise presented and they could communicate normally. Each condition consisted of a block of 30 Harvard sentences, and participants were able to rest between blocks. Each session commenced with the quiet block, followed by the four noise conditions in a random and counterbalanced order. Recordings were 2 channels and sampled at 44.1kHz.

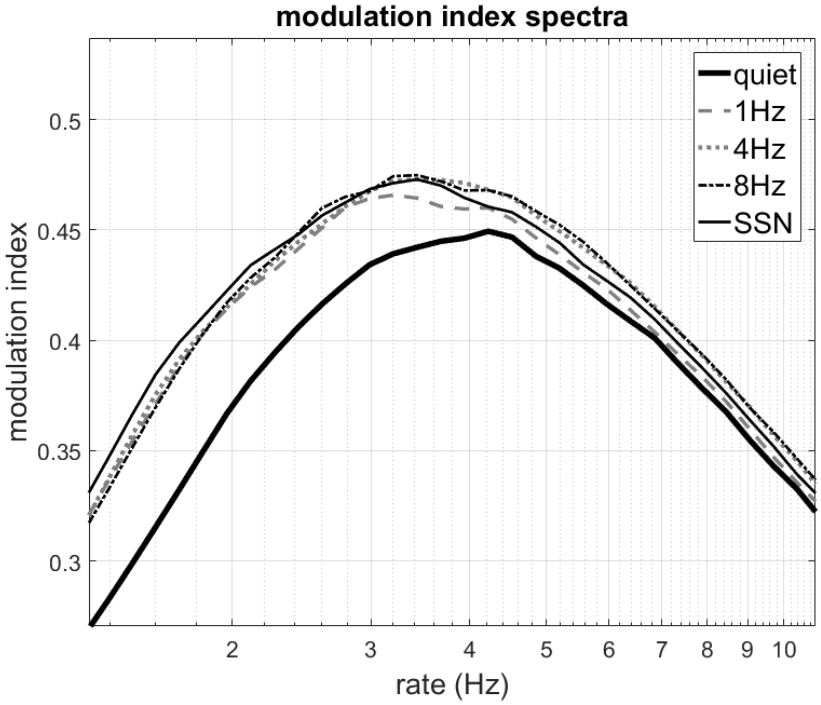
Figure 1: Waveforms of 4Hz square wave modulated speech-shaped noise (top) and a sentence spoken in the 4Hz noise condition (bottom).



2.3. Data processing

Only Talker A's speech was analysed as the focus was read speech intended to achieve communicate successfully rather than repeated speech. Orthographic time-aligned transcriptions, including pauses, were obtained using a speech recognition service [19], which were converted to Praat TextGrids [13]. These were manually checked against the recordings for transcription and alignment accuracy. Recordings were then segmented into sentence-level .wav files for analysis.

Figure 3: Amplitude modulation spectrum for each condition, showing the modulation index by modulation rate in Hz on a logarithmic scale.



2.3.1. Acoustic-phonetic measures

Several acoustic-phonetic measures for features known to vary in Lombard speech were taken for each recording. Median f0 and mean energy between 1-3kHz were extracted in Praat. Articulation rate, measured as syllables/second, was measured using Praat and the qdap package in R [18] The same analysis methods were used in [9], where they are described in more detail.

2.3.2. Amplitude modulations

Amplitude modulation spectra, similar to the familiar speech spectrum showing the frequency content of a signal, were calculated in MATLAB [16] following the method used in [21] using a gammatone filterbank, the output of which is analysed for the frequency components of the temporal envelope. The maximum modulation index and location of that peak for each sentence were calculated, as in [21], and the results averaged for each participant in each condition. Modulation index, which corresponds to modulation depth, is a ratio usually between 0 and 1 representing the amount of modulation at a certain rate (e.g., 4Hz) compared to the that of the whole envelope.

3. RESULTS

Statistical analysis on the acoustic Lombard measures and amplitude modulation spectrum measures was carried out with linear mixed effects models using the *nlme* package in R [17] with 'condition' as a fixed effect (5 levels: NORM, 1Hz, 4Hz, 8Hz, and SSN) and participant as a random effect. Tukey post-hoc tests were conducted using the *glht* function [20].

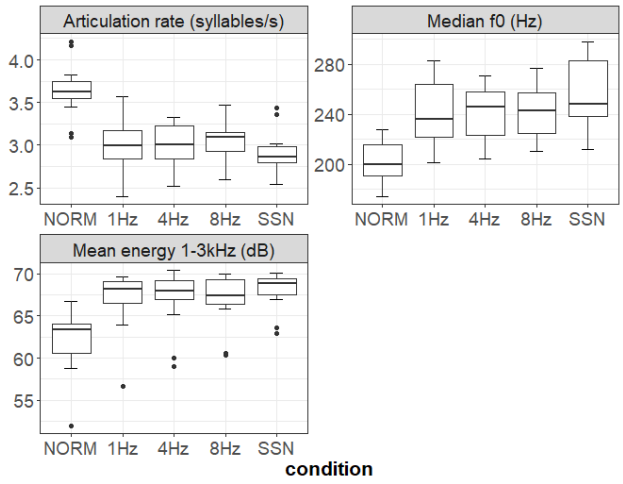
3.1 Acoustic phonetic measures

As predicted, talkers had a significantly higher median f0 ($\chi2(4)=111.79$, $p<.0001$) in noise than in quiet. All noise conditions were significantly higher than quiet ($p<.0001$), and SSN was significantly higher than the three fluctuating maskers ($p<.0001$), which did not differ from each other ($p>.9$). Talkers also had higher mean energy between 1-3kHz in noise than in quiet ($\chi2(4)=64.61$, $p<.0001$), indicating reduced spectral tilt. Post hoc revealed no significant differences between noise types, though all were significantly higher than quiet ($p<.001$ for all) (Figure 2). These changes in f0 and mid-frequency energy indicate a characteristic increase in physiological 'vocal effort' in noise [7]

Similarly, and consistent with previous research [5] talkers had a significantly slower articulation rate in all noise conditions than in quiet ($\chi2(4)=90.03$,

p<.0001), with post hoc tests revealing no differences between noise conditions.

Figure 2: Median f0, mean energy between 1-3kHz, and articulation rate in syllables per second for each noise condition.



3.2 Amplitude modulations

The maximum modulation index peaks in the modulation spectra were significantly higher in noise than in quiet ($\chi^2(4)=18.37$, p<.001). Post hoc tests revealed that SSN (M=0.56, p=.0014) and 1Hz (M=0.57, p<.001) were higher than quiet (M=0.51), while 4Hz (M=0.54, p=.16) and 8Hz (M=0.55, p=.67) were not, and none differed from each other. These findings of more pronounced modulations in at least steady-state noise are consistent with [3] (Figure 3).

There was a significant effect of condition on peak location in Hz ($\chi^2(4)=25.17$, p<.0001, quiet M=3.62). It was predicted that speakers would produce peaks near the rate of the masker, however post hoc tests revealed that 1Hz (M=3.24Hz, p=.039) and 4Hz (M=3.2, p=.0184) shifted to a significantly lower rate than quiet (M=3.62), and unexpectedly, SSN (M=3.84) did not differ from quiet (p=.47).

4. DISCUSSION

As predicted, the presence of noise significantly affected fundamental frequency, mean energy between 1-3kHz, and articulation rate. These changes are evidence that adaptations to noise were indeed being made and are in line with many clear and Lombard speech studies [e.g. 5, 7, 11].

One of the main aims of the study was to build on recent findings in [3] that speakers show more pronounced modulations when speaking in noise, across a larger number of talkers and noise types in a controlled, communicative task. Consistent with predictions and [3], peak modulation was significantly higher in several noise conditions than in quiet. This suggests that speakers employ a more consistent, pronounced speech rate in noise in order to communicate successfully. This fits with theories of neural entrainment where listeners follow temporal envelope modulations to perceive speech, especially in difficult conditions [6]. More consistent modulation at a certain rate might allow listeners to more easily entrain to the signal. In addition, there was no difference in peak modulation index between the condition with the least amount of energetic masking (EM) in time, the 1Hz condition, and the most EM, speech-shaped noise. This suggests that the amount of adaptation in envelope modulations is not dependent purely on the amount of energetic content in the masker. Rather, although the 1Hz condition provided longer silent intervals (250ms), speakers produced higher modulation peaks (than in 4Hz and 8Hz), suggesting that communication was as difficult as in steady noise, and that slow fluctuations did not disrupt their articulation rate.

It was predicted that modulation spectra would show peaks at the rate of the masker as speakers tried to 'fill in the gaps'. However, although peak location was significantly affected by noise, as seen in figure 3, most peaks shifted downwards rather than towards the rate of the masker, and the peak for 1Hz did not lower significantly more than 4 or 8Hz. As the syllable rate/peak location in the quiet (NORM) condition was about 4Hz, the results suggest that rather than 'fill in the gaps' (especially in the 4Hz condition as predicted), which would have required speakers to increase their speech rate, talkers slowed down regardless of masker, as reflected by acoustic measures of syllable rate. This result is consistent with [15], who suggested that speakers may not be able to 'temporally align' to maskers faster than 2Hz. Different types of modulation spectra may be more useful for capturing the nuance of modulations not dominated by syllable rate by measuring local or multiple maxima, and sentence level analysis of temporal alignment and pausing behaviour may show changes in more detail. This is especially true of the 1Hz condition, which may have elicited more complex temporal reorganisation strategies. Further investigation of a greater sample size into amplitude modulations and communication accuracy could provide more insight into the relative difficulty of varying levels of fluctuation in masking noise.

ACKNOWLEDGEMENTS

This work was supported by the Economic and Social Research Council.

5. REFERENCES

[1] Audio3d: Spatial Audio Simulation System. https://www.phon.ucl.ac.uk/resource/audio3d/

[2] Aubanel, V., Cooke, M. 2013. Strategies adopted by talkers faced with fluctuating and competing speech maskers. *The Journal of the Acoustical Society of America* 134, 2884-2894.

[3] Bosker, H., Cooke, M. 2018. Talkers produce more pronounced amplitude modulations when speaking in noise. *The Journal of the Acoustical Society of America* 143, EL121-EL126.

[4] Bosker, H., Ghitza, O. 2018. Entrained theta oscillations guide perception of subsequent speech: behavioural evidence from rate normalisation. *Language, Cognition, and Neuroscience* 33, 955-967.

[5] Cooke, M., King, S., Garnier, M., Aubanel, V. 2014. The listening talker: A review of human an algorithmic context-induced modifications of speech. *Computer Speech and Language* 28, 543-571.

[6] Ding, N., Simon, J.Z. 2012. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences* 109, 11854-11859.

[7] Garnier, M., Heinrich, N. 2014. Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise? *Computer Speech and Language* 28, 580-597.

[8] Ghitza, O. 2011. Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology* 2.

[9] Hazan, V., Tuomainen, O., Kim, J., Davis, C., Sheffield, B., Brungart, D. Clear speech adaptations in spontaneous speech produced by young and older adults. *The Acoustical Society of America* 144, 1331-1346.

[10] Krause, J.C., Braida, L.D., 2004. Acoustic properties of naturally produced clear speech at normal speaking rates. *The Journal of the Acoustical Society of America* 115, 362-378.

[11] Lu, Y., Cooke, M. 2009. Speech modifications produced in the presence of low-pass and high-pass filtered noise. *The Journal of the Acoustical Society of America* 126, 1495-1499.

[12] MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.

[13] Praat, a system for doing phonetics by computer. www.praat.org

[14] ProRec: Speech Prompt and Record System. https://www.phon.ucl.ac.uk/resource/prorec/

[15] MacDonald, E.N., Raufer, S. 2013. Intelligibility of speech produced in temporally modulated noise. *The Journal of the Acoustical Society of America* 133, 3519.

[16] MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.

[17] Pinheiro J, Bates D, DebRoy S, Sarkar D and R Core Team (2018). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-137.

[18] Rinker, T. W. (2017). qdap: Quantitative Discourse Analysis Package. 2.3.0. Buffalo, New York. http://github.com/trinker/qdap

[19] Speechmatics. Cantab Research Ltd. www.speechmatics.com

[20] Torsten Hothorn, Frank Bretz and Peter Westfall (2008). Simultaneous Inference in General Parametric Models. Biometrical Journal 50(3), 346--363.

[21] Varnet, L., Ortiz-Barajas, M.C., Erra, R.G., Gervain, J., Lorenzi, C. 2017. A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America* 142, 1976-1989.