Intelligibility of speech produced in temporally modulated noise

Ewen N. MacDonald[1], Stefan Raufer[2]
[1] *CHeSS, Technical University of Denmark, Lyngby, Denmark, Email: emcd@elektro.dtu.dk*
[2] *Jade-Hocshule, Oldenburg, Germany, Email: stefan.raufer@web.de*

Introduction

Talkers automatically increase the level of their voice in the presence of a noisy environment, a phenomenon referred to as the Lombard effect [1]. Beyond an increase in level, studies have also demonstrated that talkers alter other aspects of speech production, increasing their fundamental frequency (F0), reducing spectral tilt (i.e., increasing the energy produced at higher frequencies), and increasing word duration [2, 3, 4, 5, 6, 7, 8]. It has been suggested that talkers perform these adaptations to make their speech more intelligible in noise [5] and studies have shown that noise-induced speech (i.e., speech recorded when the talker is exposed to noise) is more intelligible than speech recorded in quiet [9].

While many studies have investigated the Lombard effect in steady-state noise, few have explored how talkers adapt in temporally fluctuating noise. With regards to speech intelligibility, it is well known that listeners are better able to understand speech in the presence of a fluctuating vs. steady-state masker (e.g., [10, 11]). In the presence of a fluctuating masker, a listener can take advantage of short periods where the SNR is improved — "listening in the dips." Similarly, in a noisy environment, talkers can take advantage of a fluctuating background noise by aligning speech with temporal dips and some evidence of this has been observed recently [12]

In the present study, we present preliminary results from an investigation into the possible temporal adaptation of speech produced in the presence of amplitude modulated background noise and the consequence of this adaptation on intelligibility. Talkers produced sentences in broadband noise that was amplitude modulated at different rates, ranging from 1 to 16 Hz. Speech reception thresholds (SRTs) were then measured using the recorded speech mixed with amplitude modulated noise where the phase of the envelope was either the same as or the opposite to that which the talker experienced. As the two maskers differ only in the phase of the envelope, any difference in SRT can be attributed to the temporal alignment of speech and the masker envelope.

Methods

A total of 9 native Danish talkers (average 22.8 years; 5 female, 4 male) from the Technical University of Denmark community participated in this experiment. None of the talkers reported any history of hearing or speech disorders.

The talkers were seated in a sound attenuated booth. Sennheiser HD 201 headphones were used to provide the subjects with noise stimuli mixed with auditory feedback of their own voice. Speech was recorded using an Audix OM 5 dynamic microphone. Talkers were instructed to maintain a constant microphone-mouth distance of about 25 cm. An RME UCX fireface external soundcard was used to play back noise mixed with the subjects own voice. The input gain of the microphone, as well as the voice feedback level were held constant.

In the production phase of the experiment, talkers produced speech in quiet and in each of 6 different noise conditions: five were amplitude modulated (with modulation frequencies 1, 2, 4, 8, and 16 Hz) and one was an unmodulated Gaussian noise condition. The rms level in each noise condition was 80 dB SPL. The order in which each noise condition was tested was randomized across talkers.

In each condition, talkers were presented with 50 sentence prompts that appeared individually on a computer screen. The talkers were instructed to read the sentence aloud when it appeared on the screen. Further, they were instructed to speak clearly so that the operator, listening outside the booth via headphones, could understand their utterances. The sentence prompts consisted of five-word sentences with a fixed grammar. Each word in a sentence was randomly selected from a set of five words based on Dantale II [13].

Name	Verb	Number	Adjective	Object
Anders	ejer	ti	gamle	jakker
Birgit	købte	fem	røde	kasser
Ingrid	solgte	syv	pæne	ringe
Kirsten	valgte	tre	fine	skabe
Linda	finder	seks	flotte	masker

Table 1: Word sets used to generate random five-word sentence prompts for each language.

Along with the speech, two sets of noise were also recorded. The first set, N_0, consisted of the noise that the talker heard over the headphones while speaking. For the quiet condition, a noise equivalent to that in the unmodulated condition was recorded. The second set of noises, N_π, were the same but with an amplitude modulation where the phase of the envelope was opposite that of N_0. For the quiet and unmodulated conditions, a second unmodulated noise was recorded.

In the intelligibility phase of the experiment, subjects listened to their own utterances mixed with either N_0 or N_π. SRTs for N_0 and N_π in each noise condition were adaptively measured using 20 sentences each. The sentences presented were picked randomly and the same sentence was never played back more than once. The

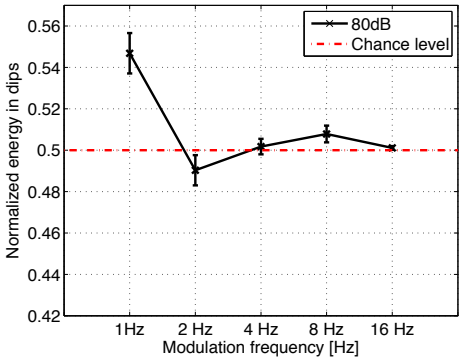

Figure 1: Normalized speech energy in the dips for speech produced in modulated masking noise conditions averaged across all talkers. The errorbars indicate one standard error. The horizontal line indicates the expected outcome if the energy is distributed uniformly.

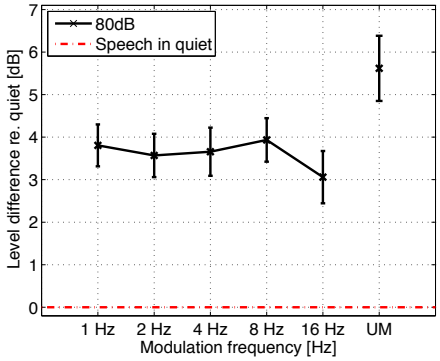

Figure 2: Speech level in each noise condition relative to that produced in quiet, averaged across talkers. Errorbars indicate one standard-error.

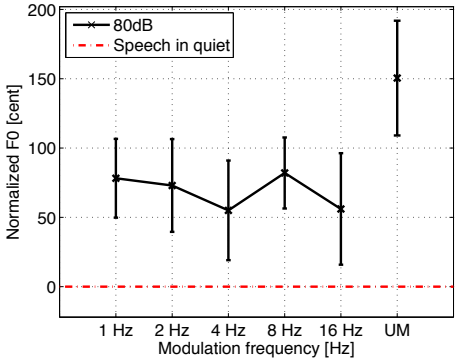

Figure 3: F0 in each noise condition relative to that produced in quiet, averaged across talkers. Errorbars indicate one standard-error.

order in which SRTs for the conditions in table 6 were measured was randomized and the order in which N_0 and N_π were tested was counterbalanced across conditions and subjects. Before the experiment, subjects received training with 40 sentences.

Results

To analyze the distribution of speech energy over time, each sentence was segmented based on the phase of the envelope of the modulated noise. Periods where the envelope of the masker was less than one half were considered noise dips. For each sentence, the speech energy that was present during these noise dips was summed and then normalized by the total speech energy of the sentence. The proportion of energy in the dips in each modulated noise condition, averaged across talkers, is plotted in Fig. 1.

In the 1 Hz modulation condition, talkers altered production, aligning more speech energy with the dips of the masker. However, no change in energy distribution was observed in the other modulated noise conditions. These observations were confirmed by an ANOVA [$F(4, 32) = 14.30$, $p < 0.001$]. Multiple comparisons with Bonferroni correction revealed that the 1 Hz condition differed significantly from all of the other conditions. No other differences were significant.

For each talker, the average speech level in each noise condition was normalized by the average level of speech produced in quiet. The results are plotted in Fig. 2. In general, talkers increased average speech levels in all of the noise conditions. The increase across modulation frequency was similar ($\approx$ 3.5 dB) but a larger increase was observed in the unmodulated condition ($\approx$ 5.5 dB).These observations were confirmed by two ANOVAs. When only modulated noise conditions were included, no significant effect was found [$F(4, 32) = 1.57$, $p = 0.21$]. However, when all noise conditions were included, a significant effect was found [$F(5, 40) = 7.82$, $p < 0.001$].

Using PRAAT (v5.3.35), the average F0 of each sentence was estimated. For each talker, the average F0 in each noise condition was normalized by the average F0 of speech produced in quiet. The results are plotted in Fig. 3. In general, the pattern of results for F0 is quite similar to the pattern of results for speech level. The increase in F0 was similar across all modulated noise conditions and largest in unmodulated noise. These observations were confirmed by two ANOVAs. When only modulated noise conditions were included, no significant effect was found [$F(4, 32) = 0.64$, $p = 0.64$]. However, when all noise conditions were included, a significant effect was found [$F(5, 40) = 4.367$, $p = 0.003$].

The SRT in each noise condition for both N_0 and N_π maskers is plotted in Fig. 4. While a large difference in SRT is observed across noise conditions, the SRT for both N_0 and N_π maskers is quite similar.

These observations were confirmed with a repeated measures ANOVA. When only modulated noise conditions were considered, a significant main effect of modulation frequency was found [$F(4, 32) = 11.48$, $p < 0.001$]. However, no significant main effect of envelope phase [$F(1, 8) = 2.28$, $p = 0.17$], or interaction was found [$F(4, 32) = 0.70$, $p = 0.60$].

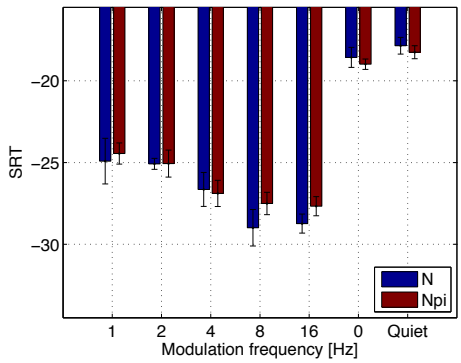

Figure 4: SRT in each noise condition averaged across subjects. The blue bars (left) indicate SRTs for noise recorded in phase with production, N_0. The red bars (right) indicate SRTs for noise where the phase of the envelope was opposite that which the subjects heard when producing speech, N_π. Errorbars indicate one standard-error.

Discussion

The purpose of the present study was to investigate how talkers alter speech production in the presence of amplitude modulated noise and to test the intelligibility of any temporal adaptation. As in previous studies, talkers increased both speech level and F0 in all noise conditions. However, while the largest increase was observed in unmodulated noise, the results across the modulated noise conditions were similar. Thus, vocal effort did not appear to change as a function of modulation frequency.

It has been proposed that talkers adapt their speech in noisy environments to maintain or improve intelligibility [5]. Since fluctuating maskers are less effective than constant maskers, one would expect that less vocal effort would be necessary to maintain intelligibility. This is consistent with the observation in the present study that while talkers increased vocal effort in the modulated noise conditions, it was not as large as in the unmodulated noise condition. However, previous work has demonstrated that the effectiveness of an amplitude modulated masker varies with modulation frequency [11]. Thus, it is surprising that no difference in vocal effort was observed across modulation frequencies. However, as discussed below, it is possible that the paradigm used in the present study did not encourage talkers to maintain intelligibility as they might in a real conversation in a similar adverse environment.

One way that talkers might improve intelligibility without increasing vocal effort would be to temporally align portions of speech to coincide with dips in a modulated background masker. In the present study, this was assessed by measuring the proportion of speech energy that was produced during dips in the masker. In the 1 Hz modulated noise condition, talkers produced speech with approximately 5% more energy in the dips than the peaks. However, in the other noise conditions, the speech energy was uniformly distributed between masker dips and peaks. The results here suggest that talkers may not be able to temporally align speech with a masker modulated at a rate of 2 Hz or higher.

Previous work has demonstrated that talkers can reduce the temporal overlap of their speech with a fluctuating background of speech modulated noise or a competing talker [12]. In contrast with [12], all of the fluctuating noise conditions in the present study were deterministic in the sense that dips occured at regular intervals. Thus, it should have been easier for talkers to temporally align speech with masker dips in the present study. However, there are other significant differences between the two studies that may have led to different results.

The task of the talkers in the present study was to read aloud the sentences that appeared on a computer screen. In contrast, talkers in [12] were tasked with solving Sudoku puzzles either alone or in pairs. Importantly, talkers adapted more when they were producing speech with communicative intent (i.e., when they were solving puzzles co-operatively with a partner rather than alone). In the present study, talkers were asked to produce speech that was intelligible to the operator listening over headphones outside the booth. However, they received no intrinsic benefit in producing more intelligible speech. While talkers repeated a set of words several times in both studies, in the present study, the speech material was limited. All sentences consisted of five-words with a fixed grammar, with five possible word options in each position. Further, the paradigm imposed long pauses between the production of each sentence compared to running speech. In [12], talkers were solving Sudoku puzzles, so they also produced a limited set of utterances repeatedly (i.e., the digits one through nine). However, these utterances were produced in a normal way with no artificially extended pauses between utterances. Thus, both communicative intent and an extended period of running speech may be important factors that lead to talkers to temporally aligning speech towards dips in a fluctuating masker.

One of the goals of the present study was to examine if talkers' adapted the temporal alignment of their speech to improve intelligibility. This was tested by comparing SRTs between two amplitude-modulated maskers: one where the phase of the envelope was the same, and a second where the phase of the envelope was opposite to that present when the speech was produced. As described earlier, talkers were able to align more energy with the dips of the masker in the 1 Hz modulated noise condition. However, no difference in SRT was found between the in-phase vs. opposite-phase maskers, implying the adaptation had no effect on intelligibility. Conversely, a trend towards improved intelligibility was observed for the 8 and 16 Hz noise conditions (i.e., the SRT for the in-phase masker was lower than that of the opposite-phase masker) even though no change in energetic overlap was observed.

Previous work has shown that the effectiveness of an amplitude modulated masker varies with modulation frequency [11]. The SRT results in the present study

are consistent with this, with the lowest SRTs occurring for the 8 and 16 Hz modulated noise condition. Unfortunately, since it is possible that talkers may have produced speech that varied in clarity, it is difficult to directly compare SRTs across the noise conditions in the current study. However, it is interesting to compare the SRTs for speech produced in quiet with that produced in unmodulated noise since the same type of masker was used to measure SRTs. While the SRT for speech produced in unmodulated noise is slightly lower, this difference was not statistically significant. This is not consistent with previous work that has demonstrated noise-induced speech (i.e., speech recorded when the talker is exposed to noise) to be more intelligible than speech recorded in quiet [9]. However, as previously discussed, the talkers in the present study may not have been producing speech with communicative intent.

Due to time constraints, only a limited set of intelligibility tests could be performed in the present study. However, further intelligibility studies could be conducted using the speech collected in the production phase. One approach would be to measure intelligibility using maskers with different modulation frequency. Using this approach it would be possible to determine whether the lower SRTs observed in the 8 and 16 Hz modulated noise conditions are a result of talker adaptation (e.g., production of speech that is more clear) or due to differences in masking effectiveness related to the modulation frequency of a fluctuating masker. A second approach would be to edit the recordings to alter temporal alignment to minimize the temporal overlap with the masker. Using this method, it would be possible to test the extent to which this energetic alignment approach can improve intelligibility. In contrast, the recordings could be edited to align phoneme transitions with masker dips. For speech reception, information from phoneme transitions may provide more information than that available from higher energy, quasi steady-state portions of speech. This could explain why a trend for improved intelligibility for the in-phase masker was observed for the 8 and 16 Hz conditions, even though no change in energetic overlap was observed. Assuming this is true, a difference in intelligibility between the speech modified using the two different editing methods would be expected. This would be interesting from a modelling perspective as current models for predicting speech intelligibility would likely not be sensitive to this difference.

In summary, while talkers in this study increased vocal effort (i.e., increased speech level and F0) in amplitude modulated noise, the increase was smaller than in steady-state noise and did not vary across modulation frequency. Further, for a modulation frequency of 1 Hz, talkers altered the timing of their utterances to reduce their energetic overlap with the masker. Surprisingly, this change temporal alignment of speech did not lead to increased intelligibility. However, due to several aspects of the experimental paradigm used, talkers may not have adapted speech as much as they might have if they were having a conversation in the same adverse environment.

Acknowledgements

This experiment was conducted by Stefan Raufer as part of his bachelor thesis and supported by an Erasmus Mundus scholarship. This work was also supported by the Oticon Foundation.

References

[1] Lombard, E.: Le signe de le elevation de la voix. Ann. Malad. l'Orielle Larynx Nez Pharynx **84** (1911), 101-119.

[2] Dreher, J. J. and O'Neill, J.: Effects of ambient noise on speaker intelligibility for words and phrases. J. Acoust. Soc. Am **29** (1957), 1320-1323.

[3] Webster, J. C. and Klumpp, R. G.: Effects of ambient noise and nearby talkers on a face-to-face communication task. J. Acoust. Soc. Am **34** (1962), 936-941.

[4] Charlip, W. S. and Burk, K. W.: Effects of noise on selected speech parameters. J. Commun. Dis. **2** (1969), 212-219.

[5] Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A.: Effects of noise on speech production: Acoustic and perceptual analysis. J. Acoust. Soc. Am **84** (1988), 917-928.

[6] Junqua, J. C.: The lombard reflex and its role on human listeners and automatic speech recognizers. J. Acoust. Soc. Am **93**, 510-524.

[7] Junqua, J. C.: The influence of acoustics on speech production: A noise-induced stress phenomenon known as the lombard reflex. Speech Commun. **20** (1996), 13-22.

[8] Lu, Y. and Cooke, M.: Speech production modifications produced by competing talkers, babble, and stationary noise. J. Acoust. Soc. Am **124** (2008), 3261-3275.

[9] Pittman, A. L. and Wiley, T. L.: Recognition of speech produced in noise. J. Speech Lang. Hear. Res. **44** (2001), 487-496.

[10] Festen, J. M. and Plomp, R.: Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. J. Acoust. Soc. Am **88** (1990), 1725- 1736.

[11] Gustafsson, H. Å. and Arlinger, S. D.: Masking of speech by amplitude-modulated noise. J. Acoust. Soc. Am **95** (1994), 516-529.

[12] Cooke, M. and Lu, Y.: Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. J. Acoust. Soc. Am **128** (2010), 2059-2069.

[13] Wagener, K., Josvassen, J. L., and Ardenkjær, R.: Design, optimization and evaluation of a danish sentence test in noise. Intl. J. Aud. **42** (2003), 10-17.