ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 4aSCa: Auditory Feedback in Speech Production I

4aSCa3. Speech production in amplitude-modulated noise
Ewen N. MacDonald* and Stefan Raufer

*Corresponding author's address: Department of Electrical Engineering, Technical University of Denmark, Kgs. Lyngby, DK-2800, Denmark, Denmark, emcd@elektro.dtu.dk

The Lombard effect refers to the phenomenon where talkers automatically increase their level of speech in a noisy environment. While many studies have characterized how the Lombard effect influences different measures of speech production (e.g., F0, spectral tilt, etc.), few have investigated the consequences of temporally fluctuating noise. In the present study, 20 talkers produced speech in a variety of noise conditions, including both steady-state and amplitude-modulated white noise. While listening to noise over headphones, talkers produced randomly generated 5-word sentences. Similar to previous studies, talkers raised the level of their voice in steady-state noise. While talkers also increased the level of their voice in amplitude-modulated noise, the increase was not as large as that observed in steady-state noise. Importantly, for the 2 and 4 Hz amplitude-modulated noise conditions, talkers altered the timing of their utterances, reducing the energetic overlap with the masker by approximately 2%. However, for the 1 Hz amplitude-modulated condition, talkers increased the overlap by approximately 4%. Overall, the results demonstrate that talkers are sensitive to the temporal aspects of noisy environments and will alter their speech accordingly.
INTRODUCTION

The Lombard effect refers to the phenomenon where talkers automatically increase their level of speech in a noisy environment (Lombard, 1911). Several studies have investigated how the Lombard effect alters other aspects of speech production and, in general, have found that it leads to an increase in fundamental frequency (F0), a reduced spectral tilt (i.e., more energy at higher frequencies), and an increase in word duration (Dreher and O'Neill, 1957; Webster and Klumpp, 1962; Charlip and Burk, 1969; Summers et al., 1988; Junqua, 1993, 1996; Lu and Cooke, 2008). Further, it has been suggested that the Lombard effect is due to talkers adapting their speech to be more intelligible in noise (Summers et al., 1988). This is supported by the observation that noise-induced speech (i.e., speech recorded when the talker is exposed to noise) is more intelligible than speech recorded in quiet (Pittman and Wiley, 2001).

While many studies have investigated the Lombard effect in steady-state noise, few have explored how talkers adapt in temporally fluctuating noise. It is well known that when faced with a competing masker, talkers are better able to understand speech when the masker is temporally fluctuating than when it is constant (e.g., Festen and Plomp, 1990; Gustafsson and Arlinger, 1994). This phenomena has been attributed to “listening in the dips,” where a listener takes advantage of short periods where the SNR is improved. Thus, if talkers are adapting their speech to maintain intelligibility, the Lombard effect may be reduced in the presence of a temporally fluctuating masker, which is less effective than a constant masker.

Recently, Cooke and Lu (2010) reported that talkers reduce the temporal overlap of their own speech with that of a fluctuating background masker. By “talking in the dips,” shifting speech to temporal regions where the masker is reduced, talkers could, presumably, improve the intelligibility of their utterances. In that study, the fluctuating background maskers were speech-modulated noise or competing talkers. Thus, to reduce temporal overlap, talkers had to predict when dips would be present in the non-deterministic background masker.

The purpose of the present study was to investigate how talkers adapt their speech in the presence of amplitude modulated background noise. Talkers produced sentences in broadband noise that was amplitude modulated at different rates, ranging from 1 to 16 Hz. Measurements of speech level and F0 from these conditions were compared to those from speech recorded in quiet and unmodulated noise. As well, the distribution of speech energy over time was examined to determine if talkers shifted speech energy into periods where masking energy was reduced. In contrast to the study by Cooke and Lu (2010), the modulations in the present study were deterministic, so it should be easier for talkers to predict when dips in the masker would occur.

METHODS

A total of 20 talkers (mean age: 26.8 years) from the Technical University of Denmark community participated in this experiment. None of the talkers reported any history of hearing or speech disorders. Three were native Danish speakers and spoke Danish during the experiment. Five were native German speakers and spoke German during the experiment. The rest (three native English and nine non-native English speakers) spoke English during the experiment.

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 experiment, talkers produced speech in quiet and in each of 12 different noise conditions. Five modulation frequencies (1, 2, 4, 8, and 16 Hz) and unmodulated (UM) Guassian
Table 1: Word sets used to generate random five-word sentence prompts for each language.

<table>
<thead>
<tr>
<th>Language</th>
<th>Name</th>
<th>Verb</th>
<th>Number</th>
<th>Adjective</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Peter</td>
<td>owns</td>
<td>ten</td>
<td>old</td>
<td>jackets</td>
</tr>
<tr>
<td></td>
<td>Mary</td>
<td>had</td>
<td>five</td>
<td>red</td>
<td>boxes</td>
</tr>
<tr>
<td></td>
<td>Lucy</td>
<td>sees</td>
<td>two</td>
<td>nice</td>
<td>flowers</td>
</tr>
<tr>
<td></td>
<td>Maggie</td>
<td>bought</td>
<td>three</td>
<td>new</td>
<td>presents</td>
</tr>
<tr>
<td></td>
<td>Michael</td>
<td>won</td>
<td>six</td>
<td>fine</td>
<td>houses</td>
</tr>
<tr>
<td>Danish</td>
<td>Anders</td>
<td>ejer</td>
<td>ti</td>
<td>gamle</td>
<td>jakker</td>
</tr>
<tr>
<td></td>
<td>Birgit</td>
<td>købte</td>
<td>fem</td>
<td>røde</td>
<td>kasser</td>
</tr>
<tr>
<td></td>
<td>Ingrid</td>
<td>solgte</td>
<td>syv</td>
<td>pæne</td>
<td>ringe</td>
</tr>
<tr>
<td></td>
<td>Kirsten</td>
<td>valgte</td>
<td>tre</td>
<td>fine</td>
<td>skabe</td>
</tr>
<tr>
<td></td>
<td>Linda</td>
<td>finder</td>
<td>seks</td>
<td>flotte</td>
<td>masker</td>
</tr>
<tr>
<td>German</td>
<td>Peter</td>
<td>sieht</td>
<td>drei</td>
<td>große</td>
<td>Blumen</td>
</tr>
<tr>
<td></td>
<td>Kerstin</td>
<td>kauft</td>
<td>neun</td>
<td>kleine</td>
<td>Tassen</td>
</tr>
<tr>
<td></td>
<td>Tanja</td>
<td>gibt</td>
<td>acht</td>
<td>alte</td>
<td>Autos</td>
</tr>
<tr>
<td></td>
<td>Ulrich</td>
<td>schenkt</td>
<td>vier</td>
<td>nasse</td>
<td>Bilder</td>
</tr>
<tr>
<td></td>
<td>Britta</td>
<td>nahm</td>
<td>fünf</td>
<td>schwere</td>
<td>Dosen</td>
</tr>
</tbody>
</table>

noise were tested at 60 and 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 the language (see Table 1). The Danish word sets were based on Dantale II (Wagener et al., 2003). With the exception of the names, the English word sets were a translation of the Danish word sets. German sentences were based on the Oldenburger Satztest (OLSA; Wagener et al., 1999), a German equivalent of Dantale II.

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.

The distribution of speech energy varied across modulation frequency, and the pattern of results was similar in both 60 and 80 dB SPL conditions. In the 1 Hz modulation conditions, more speech energy was aligned with the peaks of the masker. Conversely, in the 2 and 4 Hz modulation conditions, more speech energy was aligned with the dips of the masker. Finally, in the 8 and 16 Hz modulation conditions, speech energy was equally aligned between the dips and peaks of the masker.

These observations were confirmed with a repeated measures ANOVA. A significant main effect of modulation frequency was found \([F(1.50, 28.52) = 16.05, p < 0.001]\). However, no significant main effect of level \([F(1, 19) = 0.262, p = 0.615]\) nor interaction were found \([F(1.83, 34.83) = \text{...}\]
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, little change in average speech levels were observed in the 60 dB SPL conditions. However, increases in average level were observed in the 80 dB SPL conditions. Further, a larger increase in average speech level was observed in the unmodulated 80 dB SPL condition compared to the other modulated 80 dB SPL conditions, which are very similar to each other.

These observations were confirmed by two repeated measures ANOVAs. When all noise conditions were included, significant main effects of level \([F(1, 19) = 97.21, p < 0.001]\) and modulation frequency were found \([F(2.88, 54.62) = 4.05, p = 0.012]\). However, when the unmodulated noise condition was excluded from the ANOVA, the main effect of level was again significant \([F(1, 19) = 410.82, p < 0.001]\) but no significant main effect of modulation frequency was found \([F(2.52, 47.82) = 0.73, p = 0.517]\).

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. Larger increases in F0 were observed in all of the 80 SPL conditions compared to the corresponding 60 dB SPL noise conditions. As well, a larger increase was observed in the unmodulated 80 dB SPL condition compared to the other modulated 80 dB SPL conditions, which are very similar to each other.

These observations were confirmed by two repeated measures ANOVAs. When all the noise conditions were included, significant main effects of level \([F(1, 19) = 26.66, p < 0.001]\) and modulation frequency \([F(5, 95) = 3.71, p = 0.004]\) were found. Further the interaction of level and modulation frequency was found to be significant \([F(5, 95) = 4.31, p = 0.001]\). When the unmodulated noise condition was excluded from the ANOVA, the main effect of level was again significant but no significant main effect of modulation frequency \([F(3.27, 62.12) = 1.06, p = 0.377]\) nor significant interaction were found \([F(3.27, 62.08) = 1.55, p = 0.207]\).
The purpose of the present study was to investigate how the Lombard effect varies in amplitude modulated noise. As in previous studies, talkers increase speech level and F0 in the presence of background noise. However, this increase was smaller when the background noise was amplitude modulated. Interestingly, while a difference in level and F0 was observed between unmodulated and modulated conditions, no difference was found across modulation frequencies. Thus, while vocal effort was reduced in the modulated vs. unmodulated noise, it remained constant across noises that were amplitude modulated at different frequencies.

As mentioned previously, it has been proposed that talkers adapt their speech in noisy environments to maintain or improve intelligibility. 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 our observation that vocal effort was higher in the unmodulated vs. modulated noise conditions. However, over the range of modulation frequencies tested
in this study, the effectiveness of an amplitude modulated masker varies with modulation fre-
quency (Gustafsson and Arlinger, 1994). Thus, one would have expected that vocal effort would
vary across the modulation frequencies tested, but this was not observed.

In the presence of an amplitude modulated noise masker, a listener can “listen in the dips,”
i.e., take advantage of the short periods where the level of the masker is reduced. Cooke and Lu
(2010) demonstrated that talkers were able to reduce the temporal overlap of their speech with
a fluctuating background masker. In the present study, significant differences in the temporal
distribution of speech energy were observed across the modulated noise conditions. For the 2
and 4 Hz amplitude modulated noise conditions, talkers altered the timing of their utterances,
reducing the energetic overlap with the masker by approximately 2%. This suggests that talkers
may have attempted to improve the speech by temporally shifting speech energy. However, for
the 1 Hz amplitude modulated condition, talkers increased the overlap by approximately 4%.
Thus, for this condition, it appears that talkers became entrained by the modulated masker.

In order to reduce temporal overlap, a talker needs to predict when dips will occur in the
masker. In the study by Cooke and Lu (2010), the fluctuating background maskers were speech
modulated noise or a competing talker. Thus, the timing of the dips in these maskers was not de-
terministic. Conversely, in the present study, all of the modulation conditions were deterministic
in the sense that dips occurred at regular intervals. One of the reasons that this type of masker
was used in the present study was to test if talkers could take advantage of the deterministic
timing of the dips. However, the shifts in energy observed in the 2 and 4 Hz modulation condi-
tions are smaller than those observed in the Cooke and Lu study. Why then, did the talkers in
the present study not take more advantage of the fact that the dips were easier to predict?

The paradigm used in the present study differs significantly from that used by Cooke and
Lu. In the present study, the task of the talkers was to read aloud the sentences that appeared
on a computer screen. In contrast, talkers in the Cooke and Lu study were tasked with solving
Sudoku puzzles either alone or in pairs. In their study, Cooke and Lu found that 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). While talkers in the present study were
asked to produce speech that was intelligible to the operator listening over headphones outside
the booth, they received no intrinsic benefit in producing more intelligible speech.

In both studies, talkers repeated a set of words several times. In the present study, the speech
material consisted of five-word sentences with a fixed grammar, with five possible word options
in each position. As talkers were solving Sudoku puzzles in the study by Cooke and Lu, they
repeatedly produced utterances of the digits one through nine. However, in the present study,
there were long pauses between each sentence compared to normal speech. Both communicative
intent and an extended period of running speech may be important factors that lead to talkers to
temporally shifting speech towards dips in a fluctuating masker.

One should note that the notion that talkers would increase speech intelligibility by “talking
in the dips” tacitly assumes that speech intelligibility is correlated with speech energy. Indeed,
the success of some methods of predicting speech intelligibility such as the Speech Intelligibility
Index (SII; ANSI 1997) suggests that this may be true at a macroscopic level. However, phoneme
transitions, which may have less acoustic energy, may be of more use to a listener for under-
standing speech than the steady-state portions of vowels. Thus, it may be possible that talkers
are altering speech to improve intelligibility in a way that measuring energetic overlap does not
capture. Further studies along this line, such as measuring the speech intelligibility of speech
produced in the paradigm of the current study would be a stronger test of the hypothesis that
talkers adapt the timing of speech to maintain or improve intelligibility.

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 modulation frequencies of 2 and 4 Hz, talkers altered the timing of their utterances to reduce their energetic overlap with the masker.

ACKNOWLEDGMENTS

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


