Speech Prosody 2016
31 May - 3 June 2016, Boston, USA

Speech Produced in Noise: Relationship between Listening Difficulty and Acoustic and Durational Parameters

Simone Graetzer, Pasquale Bottalico, Eric J. Hunter

Voice Biomechanics and Acoustics Laboratory, Communicative Sciences and Disorders, Michigan State University

sgraetz@msu.edu, pb@msu.edu, ejhunter@msu.edu

Abstract

Speech produced in noise can be characterised by increases in intelligibility relative to conversational speech produced in quiet. The objectives of the study were to characterise the relationship between listening difficulty and speech produced in different noise and style conditions; and to evaluate the spectral and durational modifications associated with these noise and style conditions. 19 subjects were instructed to speak at normal and loud volumes in the presence of background noise at 40.5 dBA and babble noise at 61 dBA. The speech signal was amplitude-normalised, combined with pink noise to obtain a signal-to-noise ratio of -6dB, and presented to 20 raters who judged their listening difficulty. Vowel duration, fundamental frequency (*fo*, in semitones) and fo modulation, and the proportion of the spectral energy in high relative to low frequency bands, increased with the level of the noise, independently of the effect of style. Listening difficulty was lowest when the speech was produced in the presence of high level noise and at a loud volume, indicating improved intelligibility. The difference in spectral energy was observed to predict listening difficulty, and therefore, intelligibility scores (*IS*). These findings have implications for the improvement of communication in noisy environments.

Index Terms: listening difficulty, intelligibility, acoustics, noise

1. Introduction

Talkers modify their speech in the presence of noise to maintain a level that is sufficient for communication. The Lombard effect [1] is the involuntary tendency to increase the level of speech in the presence of noise. In noisy environments, speakers commonly increase not only their vocal intensity but also their fundamental frequency (*fo*), their first vowel formant (*F*1), and the energy in upper part of the spectrum (between 1 and 3 or 4 kHz) (e.g., [2], [3]).Speech produced in noise can also demonstrate changes in segment duration and/or a slowing of the speech rate (e.g., [4]). It appears that the spectrum flattens when the level of the speech increases (e.g., [5]). Hence it is necessary to separate any effect on the ratio of an increase in intelligibility from a mere increase in speech level.

The assessment of intelligibility in speech communication can be performed by means of several methods. The Listening Difficulty (LD) measure of intelligibility was developed by Sato, Morimoto, Bradley and colleagues as an alternative to the speech intelligibility measure (e.g., [6]). It is intended to be used when the transmission channel is relatively good, and it is desirable to avoid the potential confounding influence of the choice of speech material, and to minimise the effect of the cognitive process, to obtain an objective assessment of the speech transmission performance of the space. The LD measure has not previously been used in an attempt to evaluate which spectral or durational changes predict the intelligibility of continuous speech produced in noise.

In this paper, the modifications that occur in non-communicative laboratory speech in noisy environments are evaluated. In a previous publication [7], it was reported for the same talkers and experimental conditions that voice level or Sound Pressure Level (*SPL*) was higher when speech was produced in high level relative to low level noise; i.e., there was an observable Lombard effect [1]. SPL was also higher when the talkers were instructed to speak at a loud level relative to a normal level. The mean level for normal style in high level noise was lower than the mean level for loud style in low level noise.

The primary research questions were as follows: (1) Are vowel durations longer when speech is produced in high vs. low level noise and in the loud speech level or style relative to the normal style? (2) Is fo (in semitones) higher and more variable when speech is produced in high *vs*. low level noise and in the loud *vs.* the normal style? (3) Is there a greater proportion of high to low frequency energy, as indicated by the mean energy difference in dB between 0 – 1 kHz and 1 – 4 kHz bands (*spectrum balance*), in speech produced in high vs. low level noise and in the loud relative to the normal style? (4) Is speech produced in a noisy environment or in a loud style more intelligible as indicated by the listening difficulty or *LD* measure than speech produced in low ambient noise or in a normal style?

Firstly, the effects of noise and style on spectral and durational speech parameters will be reported. Secondly, the extent to which any of these parameters predict listening difficulty, or, relatedly, speech intelligibility scores (*IS*), will be discussed.

2. Method

19 young adult native American English speaking subjects (9 male, 10 female) with self-reported normal speech and hearing were recorded reading the Rainbow passage in a semi-reverberant room (5.8m x 6m x 2.7m) in two different styles corresponding to normal and loud voice levels and in the presence of background (ventilation) noise at 40.5 dBA or children's babble at 61 dBA, in the talker position. The babble was emitted by a directional loud speaker. Speech was acquired by an omnidirectional head-mounted microphone (Glottal Enterprises M-80) at 5 cm from the mouth (less than the critical distance) and recorded by a Roland R-05 digital

10.21437/SpeechProsody.2016-102

recorded (WAV 16-bit, 44.1 kHz sampling rate). The mid-frequency reverberation time was 0.53 s (s.d. 0.04). MATLAB 2014b and Praat 5.4.01 [8] were used for signal processing. Post-processing and statistical analysis was conducted in R version 3.1.2 [9]. For the speech intelligibility assessment, in MATLAB, 2 sentence extracts of the Rainbow passage (the 2nd and 3rd sentences) produced by each talker in each condition were amplitude normalised and combined with pink noise to obtain a signal-to-noise ratio (SNR) of -6 dB. The onset of noise preceded the onset of the signal by 500ms. The modified signals were presented via Sennheiser HD205 headphones to 20 listeners in a sound-attenuated booth. The background noise level in the listener position in the booth was 25.1 dBA. The words in the read sentences were highly familiar ones. Randomisation of the order of presentation and the recording of LD ratings was obtained via a custom Praat script. The 20 young adult normal hearing native American English speaking listeners (10 female, 10 male) were audiometrically assessed between 250 Hz and 6 kHz before stimulus presentation.

Recorded words were manually segmented in Praat. Individual vowels were segmented in Python version 3.4 by means of the FAVE-align [10] and HTK toolkits and visually inspected for errors. Normalised vowel duration was calculated by dividing each vowel duration in ms associated with a given subject by that subject's mean in the low noise and normal style. Durations were analysed by means of Welch-corrected one way tests for equal means. The Welch correction for non-homogeneity of variance to denominator degrees of freedom (*df*) results in a lower denominator df.

Fo was calculated in Praat using the autocorrelation method with Hanning windows with a length of 0.043s, a 0.01s time step, a pitch floor of 70Hz, an octave cost of 0.0025 (favouring lower frequency candidates), a voiced/unvoiced cost of 0.14 and a pitch ceiling of 400Hz. Fo was then converted to semitones in R with bases for males and females equal to their mean fo (Hz), which was 128Hz for males and 203Hz for females. These values are representative of typical adult males and females, the difference relating primarily to differences in membranous vocal fold length [12, 12]. When an increase in intensity occurs with an increase in glottal flow and hence subglottal pressure, fo will typically rise [13]. Fo was converted to semitones using the *f2st* function in the hqmisc package in R.

Spectral analysis was conducted in order to determine whether high frequency spectral emphasis occurred in high relative to low level noise. The spectral parameter, which is a form of the spectrum balance or α ratio measure, is in this context a measure of the energy difference between the 1-4kHz and 0-1 kHz regions or *bands* (*i.e.*, the mean energy computed for the upper band minus the mean computed for the lower band, in dB). The claim is that in intelligible speech produced by normal talkers the energy difference between the lower and the upper bands becomes smaller (resulting in a flatter spectrum). However, this difference can also be affected by the level and fo of the speech.

The concatenated words (the words inside the read sentences, with silences between words removed) that were produced by each talker in each condition were subjected to Long Term Average Spectrum (LTAS) analysis performed in Praat. After Fast Fourier analysis, each LTAS (without loss of frequency resolution) was calculated and the modified α ratio was derived via the Get slope function with the lower band limits of 0 and 1kHz, and the upper band limits of 1 and 4kHz, where the energy is averaged over the concatenated signal in dB, based on the mean power in Pa2/s of the signal. An evaluation of the effects of noise, style and interactions of noise and style, noise and gender and style and gender on the response variable, α ratio, was conducted by means of a linear mixed effects or LME model fitted by Restricted Maximum Likelihood with the random effects term of talker (lme4 and lmerTest R packages).

The LD measure used to measure the intelligibility of the speech produced by the talkers in noise. In the current study, the discrete subjective scale was changed from 0 to 4 to 1 to 10 to improve the resolution from 1/4th to 1/10th. Listening difficulty can be converted to intelligibility scores (IS), *i.e.*, the percentage of a message understood correctly (e.g., [14]), which is a more widely used and understood measure. The equation is as follows:

$$IS = 124.2 - 11 \cdot LD \qquad (2)$$

The instruction was "Rate the level of listening difficulty for these sentences on a scale of 1 (not difficult, no effort required) to 10 (very difficult, considerable effort required)." A cumulative link mixed model (Laplace approximation; *ordinal* R package) was run with LD as the response variable and noise, style, their interaction and interactions of both noise and style with talker gender, with both the listener and the talker as random effects terms. To determine which, if any, of the acoustic and durational parameters predicted LD, a LME model fitted by REML was run with LD as the response variable and modified α ratio, fo (semitones), fo (semitones) standard deviation, and normalised vowel duration as independent variables, with an interaction of fo (semitones) and gender, and with talker as the random effects term. The response variable, LD, was averaged across listeners per signal.

3. Results

3.1. Normalised vowel duration

Welch-corrected one way tests for equal means indicated that there was an effect of noise ($F(1,18649) = 124.44$, $p<0.0001$), and gender ($F(1,18985) = 15.75$, $p<0.0001$) but not style ($p > 0.1$) on normalised vowel duration. Durations were longer when the speech was produced in high level than low level noise, and when produced by males than females (Figure 1).

3.2. Fundamental frequency

Fo (in semitones) was higher when speech was produced in the presence of high level noise than low level noise ($\hat{\beta}$ = 0.76, SE = 0.03, df = 254566, t = 29.70, $p<0.0001$), as shown in Figure 2. Fo was higher in the loud style than in the normal style ($\hat{\beta}$ = 2.09, SE = 0.03, df = 254566, t = 81.83, $p<0.0001$). There was an interaction between noise and style ($\hat{\beta}$ = 0.33, SE = 0.03, df = 25466, t = 11.52, $p < 0.0001$) such that the effect of noise was stronger in the loud style. There was also an interaction between style and gender ($\hat{\beta}$ = -0.11, SE = 0.03, df = 25466, t = -3.85, $p < 0.001$), such that males increased their fo more than females in the loud relative to the normal style. There was no interaction between noise and gender ($p > 0.1$). Variation in fo (semitones) was greater when speech was produced in the presence of high level noise than low level noise ($\hat{\beta}$ = 0.33, SE = 0.14, df = 128, t = 2.35, $p<0.05$). Variation was lower in the loud style than in the normal style

($\hat{\beta}$ = -0.25, SE = 0.11, df = 128, t = -2.16, p<0.05). There were no interactions (p > 0.1). Very similar results were found when the fo values were subjected to outlier detection and removal using the Bonferroni method prior to analysis, indicating that these results were not due to fo artefacts.

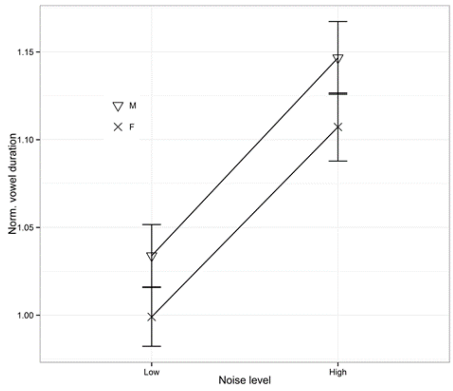

Figure 1: *Normalised vowel durations by noise (x-axis) and gender (symbol) condition. Means are shown with 95% confidence intervals.*

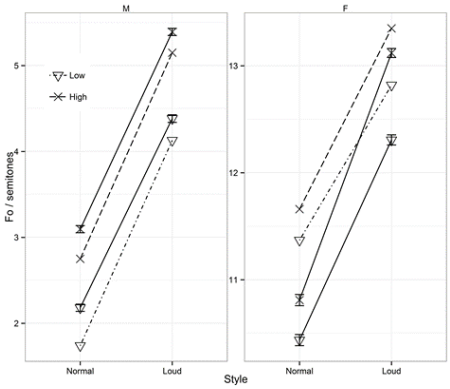

Figure 2: *Fo in semitones per style (x-axis), noise (symbol), and gender (Male, L, and Female, R panels) condition. Means are connected by solid lines and are shown with 95% confidence intervals. Medians are connected by dashed lines.*

3.3. Spectral analysis

The effects of noise, style and gender on the modified α ratio are reported in Table 1 and shown in Figure 3. When the speech was produced in high level noise *vs*. low level noise in the normal style, there was an increase in the α ratio, which indicates flattening of the spectrum. In addition, when the speech was produced in the loud style *vs*. the normal style in the presence of low level noise, there was an increase in the α ratio. In the normal style there was a greater difference between noise conditions than in the loud style (suggesting an achievement of the upper limit in the high level noise, loud style condition). There was also an interaction between style and gender: for males there was a much smaller difference between the style conditions than for the females (Figure 3).

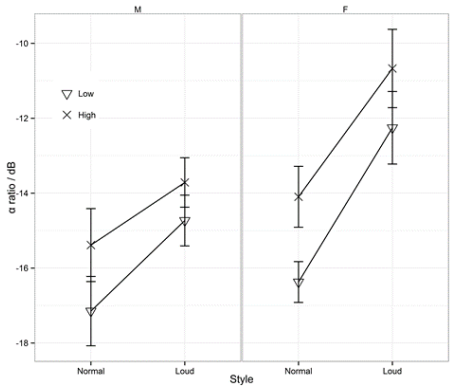

Figure 3: *Modified α ratio in dB per style (x-axis), noise (symbol), and gender (Male, L, and Female, R panels) condition, with means and 95% confidence intervals.*

Table 1. *LME model with the response variable modified α ratio and independent variables noise, and style and interaction terms for noise:style, noise:gender and style:gender (reference levels: Noise: Low, Style: Normal, Talker Gender Talker: Male or GTM), where *** p<0.0001; ** p<0.01; * p<0.05.*

Term	Estimate	SE	df	t
(Intercept)	-17.60	0.55	21	-31.26***
NoiseHigh	1.75	0.27	129	6.46***
StyleLoud	3.31	0.27	129	8.88***
NoiseHigh: StyleLoud	-0.72	0.31	129	-2.34*
NoiseLow: GenderF	0.76	0.75	20	1.02
NoiseHigh: GTF	1.30	0.75	20	1.74
StyleLoud: GTF	1.73	0.31	128	5.62***

Table 2. *Cumulative link mixed model (Laplace) output for listening difficulty by noise and style (reference levels are Noise: Low, Style: Normal, Gender Talker Male or GTM) where *** p<0.0001; ** p<0.01; * p<0.05; . p<0.1.*

Term	Estimate	SE	z
NoiseHigh	-0.86	0.12	-7.40***
StyleLoud	-1.17	0.12	-10.02***
NoiseHigh:StyleLoud	0.57	0.13	4.37***
NoiseLow:GTF	-0.61	0.35	-1.74.
NoiseHigh:GTF	-0.61	0.35	-1.73.
StyleLoud:GTF	-0.16	0.13	-1.2

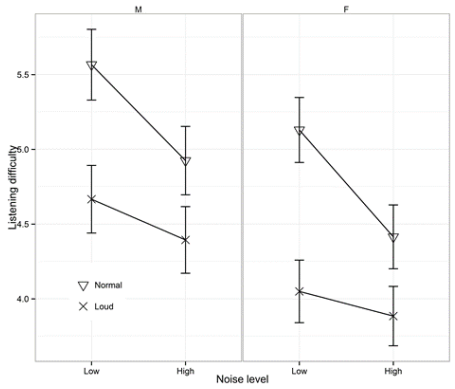

Figure 4: *Listening difficulty (1, lowest, 10, highest) by noise (x-axis) and style (symbol) condition, with means and 95% confidence intervals.*

3.4. Listening difficulty

3.4.1. Effects of noise, style and talker gender

20 listeners evaluated their difficulty in listening to the speech produced by the talkers in the two noise and two style conditions. The range of the LD responses was 2 – 8, where 8 indicated greater and 2 indicated lesser LD. As shown in Table 2, there was a decrease in LD when the speech was produced in high level (babble) noise in the normal style, and when the speech was produced in the loud style and in low level noise. There was an interaction of noise and style such that the difference between the styles was greatest in low-level noise, and the lowest LD scores occurred when speech was produced in high level noise in the loud condition (Figure 4).

3.4.2. Relationships between speech parameters and listening difficulty

A LME model was run with LD as the response variable and the acoustical and durational parameters as independent variables: α ratio, fo and fo modulation (semitones), vowel duration and an interaction of fo and talker gender. Of the parameters, only the modified α ratio reliably predicted LD ($\hat{\beta}$ = -0.29, SE = 0.03, df = 82, t = -10.87, p < 0.0001). However, there was a 0.14 difference in the slope fo-LD between females and males (Fo (semitones) by gender with reference level, male talkers: $\hat{\beta}$ = 0.14, SE = 0.06, df = 86, t = 2.20, p < 0.0001), such that for females there was a decrease in LD as fo increased (the slope can be derived from a simple linear regression: $y = 4.27 - 0.15 \cdot fo + \epsilon$).

After conversion to IS, a LME model was fit to IS with the independent variable of the α ratio, and compared with the same model including fo (semitones) and the interaction of fo (semitones) and gender. There was no observable difference between the models (X2 = 2.95, df = 1, p > 0.1). For the single independent variable model, each increase of 1 dB in the α ratio was associated with a 2.8% increase in IS ($\hat{\beta}$ = 2.76, SE = 0.21, df = 148, t = 13.28, p < 0.0001). Hence, an IS of 100% can be associated with a difference in mean energy between the low and high frequency bands of approximately 4.5 dB. While there was no difference between the nested models, there was a tendency for the increase in fo in females to be associated with an increase in IS. There was an average 7.5% IS increase in the normal style between low and high level noise conditions.

4. Discussion

In the current study, it was possible to identify effects of noise on speech that were independent of a mere effect of voice level: for vowel duration, there was an effect of noise but no effect of style; for fo, modified α ratio and LD, there was an additive effect of noise and style; and for fo modulation, the effect of style was inconsistent with the effect of noise. The effects of noise were an increase in vowel duration, an increase in fo and fo modulation, an increase in high relative to low frequency energy, and a decrease in LD, and hence an increase in IS. Arguably, these speech modifications, which occurred in a non-communicative context, are primarily associated with the increase in vocal intensity in the presence of multi-talker babble noise at 61 dBA, but could also reflect other modifications made to improve audibility for the talker at his/her own ear, at least once the SNR could no longer be improved (see, e.g., [14]).

The results concerning the relationship between durational changes and LD suggest that while vowel elongation can increase the amount of acoustic information available, the extent to which these changes can improve the intelligibility of speech masked by broadband noise appears to depend on other factors (see [2], [15]).

Shifts in the spectral energy distribution towards frequencies between 1 and 4 kHz were observed to predict LD when the signals were presented to listeners at the same SNR. The reported effects of these shifts on LD and IS are consistent with the results of Lu and Cooke [2] and Krause and Braida [16], who found that high frequency spectral emphasis contributes to the increased intelligibility of speech produced in noise. Such spectral emphasis appears to provide some release from energetic masking.

The results concerning the effects of fo and spectral energy distribution on LD and IS are consistent with the results reported by Lu and Cooke [2], who found that while changes in both fo and spectral energy distribution occur when speech is produced in noise, only the flattening of the spectrum contributes in a significant way to intelligibility (see also [15]). The fo increase may under most conditions merely accompany the increase in vocal intensity. One possible explanation of the finding for the slope fo-LD for female talkers may be that within the fo range of the females in this study (approximately 160 to 270 Hz), when fo increases, whether due to the presence of high level noise or the raised intensity of the loud style, there may be some release from energetic masking. Such a relationship between fo and speech intelligibility for female talkers may only occur at low SN-ratios [17].

In future studies, not only the level but also the type of noise in the environment of the talker will be manipulated during communicative tasks, to allow a clear separation of the effects on speech intelligibility of noise level from noise type.

5. Acknowledgements

Research reported in this publication was supported by the NIDCD of the NIH under Award Number 1R01DC012315. The content is solely the responsibility of the authors and does not necessarily represent official views of the National Institutes of Health.

6. References

[1] É. Lombard, "Le signe de l'élévation de la voix," *Annales des Maladies de L'Oreille et du Larynx*, vol. 37, no. 2, pp. 101-119, 1911.

[2] Y. Lu and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Sp. Comm.*, vol. 51, pp. 1253-1262, 2009.

[3] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech," *J. Speech. Lang. Hear. Res.*, vol. 28, pp. 96-103, 1985.

[4] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," *J. Speech. Lang. Hear. Res.*, vol. 29, pp. 434-446, 1986.

[5] M. Nordenberg and J. Sundberg, "Effect on LTAS of vocal loudness variation," in *Speech, Music and Hearing I*, Quarterly Progress Status Report 45 (Royal Institute of Technology, Stockholm), pp. 93-100, 2004.

[6] M. Morimoto, H. Sato, and M. Kobayashi, "Listening difficulty as a subjective measure for evaluation of speech transmission performance in public spaces," *J. Acoust. Soc. Am.*, vol. 116, no. 3, pp. 1607-1613, 2004.

[7] P. Bottalico, S. Graetzer, and E. J. Hunter, "Effects of voice style, noise level, and acoustic feedback on objective and subjective evaluations," *J. Acoust. Soc. Am.*, vol. 138, no. 6, 2015.

[8] P. Boersma and D. Weenink, "PRAAT: doing phonetics by computer Version 5.4.01" Computer program, retrieved 1 October 2015, from http://www.praat.org, 2015.

[9] R Development Core Team; R: a language and environment for statistical computing [Online]. Available: http://www.R-project.org

[10] I. Rosenfelder, J. Fruehwald, K. Evanini,, S. Seyfarth., K. Gorman, H. Prichard, and J. Yuan, J. FAVE 1.1.3. ZENODO. doi:10.5281/zenodo.9846, 2014.

[11] I. Titze, *Principles of voice production*, National Center for Voice and Speech, Iowa, 2000.

[12] I. Titze, "Vocal fold mass is not a useful quantity for describing F0 in vocalization," *J. Speech. Lang. Hear. Res.*, vol. 54, no. 2, pp. 520-522, 2011.

[13] G. Fant, "The voice source in connected speech," *Sp. Comm.*, vol. 22, pp. 125-139, 1997.

[14] G. Genta, A. Astolfi, P. Bottalico, G. Barbato and R. Levi, "Management of truncated data in speech transmission evaluation for pupils in classrooms," *Measurement Sci. Rev.*, vol. 13, no. 2, pp. 1335-8871, 2013.

[15] M. Cooke, S. King, M. Garnier, and V. Aubanel "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Comp. Sp. Lang.*, vol. 28, pp. 543-571, 2014.

[16] J. C. Krause and L. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," *J. Acoust. Soc. Am.*, vol. 115, no. 1, pp. 362-378, 2004.

[17] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Sp. Comm.*, vol. 49, pp. 402-417, 2007.