Journal of Signal Processing, Vol.21, No.4, pp.171-174, July 2017

SELECTED PAPER AT NCSP'17

Acoustical Analyses of Tendencies of Intelligibility in Lombard Speech with Different Background Noise Levels

Thuan Van Ngo, Rieko Kubo, Daisuke Morikawa, and Masato Akagi

Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
E-mail: {vanthuanngo, rkubo, morikawa, akagi}@jaist.ac.jp

Abstract

This study investigates acoustic variations when producing Lombard speech under the effect of a changing environment to identify adaptive tendencies of intelligibility. Analyses of the acoustic features of duration, F0, formants, spectral tilts and modulation spectrum in a dataset of speech at noise levels of $-\infty$, 66, 72, 78, 84, and 90 dB were carried out. The results show that the recognized tendencies (neutral-Lombard distinction), including lengthening vowel duration, increasing F0, shifting F1 and decreasing spectral tilts (A1-A3) are preserved among Lombard speech produced in backgrounds with a various noise levels. Our new findings are an abrupt change in F0 at 84 dB, increasing formant amplitudes, and H1-H2 variation, and a raised modulation spectrum. On the basis of physiological and psychological knowledge, we can give reasons for their correlations with intelligibility. Moreover, these variations continuously vary with increasing noise level. As a result, it is suggested that they are related to the adaptive tendencies of intelligibility.

1. Introduction

Researchers have been investigating Lombard speech [1] to explore mechanisms for improving speech intelligibility in noisy environments. Better intelligibility has recently been explained by a release from masking. A reduction in foreground-background overlap causes release from both energetic and informational masking for listeners [2]. More specifically, Lu and Cooke [3] pointed out that the acoustic changes from neutral speech (speech spoken quietly) which are lengthening duration, increasing F0, and flattening of spectral tilts are the main contributing factors. Thus, by mimicking Lombard speech [4], intelligible speech can be synthesized from human or synthetic speech with high stability and the preservation of naturalness.

However, when dynamical environments are considered, some limitations arise. First, Lombard speech has not been analyzed in a noise-level-varying background. Consequently, a convincing explanation for the correlation of Lombard speech production with physiological and psychological explanation of increased intelligibility also remains. Second, in resynthesis, the problem of maximally intelligible adaptation remains unresolved. These unresolved issues have limited the capability of resynthesized speech and prevented its adaptation when the noise level is changing. Therefore, to present maximally intelligible speech adapting to noise, the optimal solution for noise-level adaptation needs to be obtained. Then, it is necessary to perform such analyses and manipulate intelligible tendencies obtained from Lombard speech produced in backgrounds with various noise levels.

In this study, we conducted analyses on the acoustical properties of neutral and Lombard speech produced in environments with various noise levels. A set of acoustic features predicted to have a strong relationship with intelligibility were chosen for analysis. By placing the acoustic parameters of all investigated speech in the order of increasing noise level, we could more efficiently determine the factors producing Lombard speech under the effect of environmental dynamics. It is also easier to conclude which acoustic variations increase the intelligibility of Lombard speech.

2. Analysis Procedure

2.1 Speech corpus

Speakers and recorded word lists were taken from a previous study that examined the intelligibility of Lombard speech [5]. A female and a male participated in the recording. Three familiarity-controlled word lists [6] (60 words of 0-type of pitch accent pattern) with the lowest familiarity rank (1.0-2.5) were used. Each word consists of four morae (e.g., sa sa wa ra) and was embedded in a carrier sentence as a target word: "Tsugi ni yomu tango wa" word "desu". Although the speech was different from that used in the listening tests in Kubo's study [5], their intelligibility can be systematically inferred.

2.2 Feature extraction

This study aimed to find acoustic variations among Lombard speech that characterize intelligibility. Hence, a selection of distinctive features of Lombard speech and common intelligible features were considered. Specifically, we first considered analyzing the basic acoustic features of duration, F0, and spectral tilts, which represent the differences between Lombard and neutral speech. Also, formants that represent vowels were investigated. Moreover, we examined the modulation spectrum, which is well known to contribute to speech

Table 1: Analyzed acoustic features

Acoustic features	Acoustic parameters	Meanings
Duration	Consonant, Vowel duration	Length in time
F0	F0 mean, F0 slope	Glottal information
Formants	Frequencies, bandwidths, and amplitudes	Vocal tract information (articulation)
Spectral tilts	H1-H2, A1-A3	Glottal information
Modulation spectrum	Neutral-Lombard spectral difference	Temporal information of power envelope

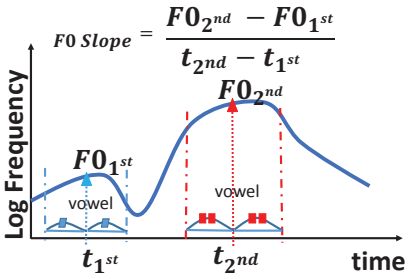


Figure 1: Calculation of F0 slope

perception. The features, parameters and their meanings are shown in Table 1. Their definitions and estimated methods are explained below.

F0: F0 was extracted by STRAIGHT [7]. F0 mean is defined as the mean of F0 contour of a word. F0 slope represents for the slope from F0 at the center of vowel of the 1^{st} mora to F0 at the center of the 2^{nd} mora (Fig. 1).

Formants: The frequencies, bandwidths and amplitudes at F1, F2 and F3 were estimated by linear predictive coding (LPC) and spectral Gaussian mixture model based spectra (spectral-GMM) [8]. F1 and F2 were used to produce the vowel space.

Spectral Tilts: H1-H2 is the magnitude difference between the first and second harmonics. A1-A3 is the magnitude difference between the nearest harmonics to F1 and F3.

Modulation Spectrum: A method that can be used to analyze the power envelope extracted by STRAIGHT [7] was employed. For each acoustic frequency (AF), A Fourier transform was applied to the power envelope with its mean value eliminated. The Fourier transform frequency can be considered as the modulation frequency (MF). The acoustic frequencies are coordinated with the modulation frequencies to produce a modulation spectrum. Then, the spectral difference between each Lombard speech and neutral speech were calculated in a decibel scale.

3. Results

The values of acoustic parameters were observed under increasing noise level. (Only the acoustic features showing variations with the noise level were considered. In this paper, if the figures for one gender are presented, the tendencies that they demonstrate were also observed for the other gender). Specifically, with increasing noise level, the vowel duration is increased (Fig. 2a). F0 increased continuously (observed in 35% of male speech, 67% of female speech) or changed abruptly at 84 dB (observed in 65% of male speech, 33% of speech) (Fig. 2b). A shift in F1 can be seen: /e/, /a/, and /o/ shift to higher frequencies in the female speech and all vowels shift to higher frequencies in the male speech (Fig. 2d). All the formant amplitudes at F1, F2 and F3 are increased (e.g., Fig. 2c). Decreases in H1-H2 for /i/, /u/ , increases for /e/, /a/ and /o/ can be seen (Fig. 2e, with the standard deviations clearly showing the variation). Decreases in A1-A3 were observed (Fig. 2f). Increases in amplitudes or rises in the modulation spectrum at 16 - 128 Hz (MF) and below 1000 Hz (AF) were also observed (Fig. 2g).

4. Discussion

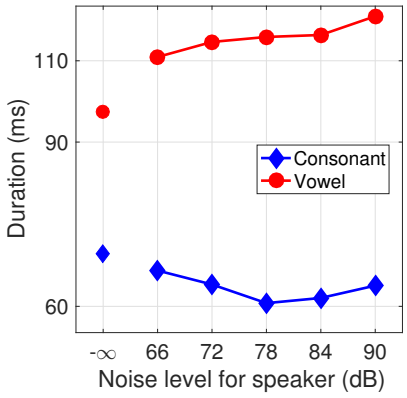
4.1 Acoustic variations

With increasing noise level, the acoustical variations continuously or abruptly changed at 84 dB. Specifically, in the recognized neutral-Lombard distinction, the vowel duration lengthened, F0 increased, spectral tilt decreased, F1 shifted higher. It was also newly discovered that F0 abruptly changed, the formant amplitude increased continuously, H1-H2 varied and the modulation spectrum rose. Additionally, formant bandwidths and F0 slope were found to be unchanged.

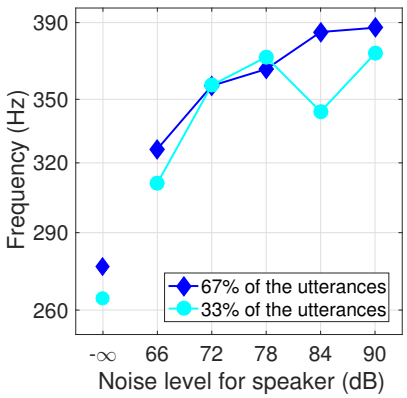
4.2 Physiological-acoustical-psychological correlation

Figure 3 shows that the above variations, which should be caused by physiological factors based on articulation, the glottal source, and energy, indicate the intelligible patterns of louder talk, increased phonetic contrast, better vowel recognition, and attracting attention from listeners. Namely, the increases or abrupt change in F0 (perhaps, changing speaking style from modal to scream) in both speakers and the shift in F1 in the male indicate mouth opens larger to speak louder. The shift in F1 in the female seems to expand the vowel space to increase phonetic contrast among vowels and to produce louder speech. The increasing formant amplitude means more energy at the vowels, which might make it easier recognize them.

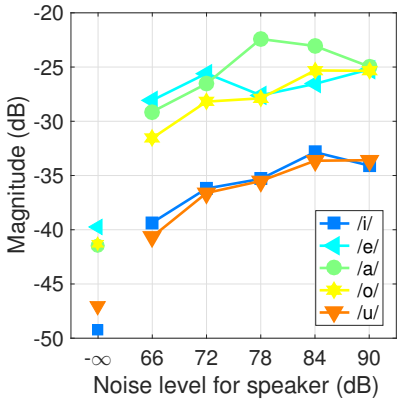
Previously, spectral tilts were found to decease for Lombard speech. However, in our results, H1-H2 decreased for /i/ and /u/, increased for /e/, /a/ and /o/, and A1-A3 decreased for all vowels. This implies two possibilities. First, /i/ and /u/ are in a different group and biased from the group of /e/, /a/ and /o/. This means that changes in the glottal source



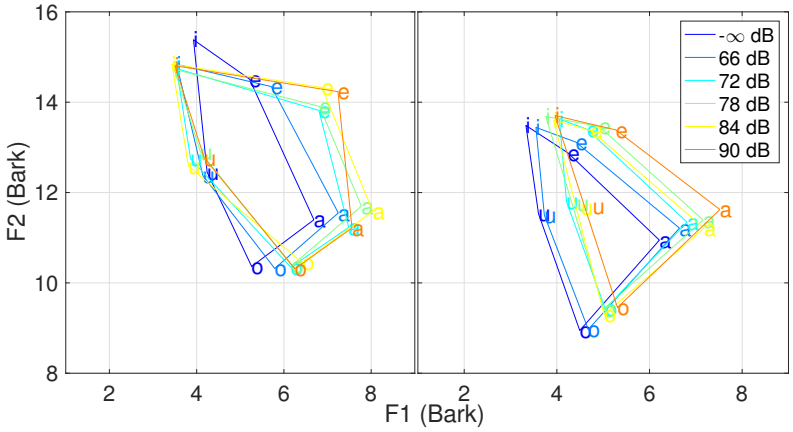
(a) Consonant and vowel duration (female)



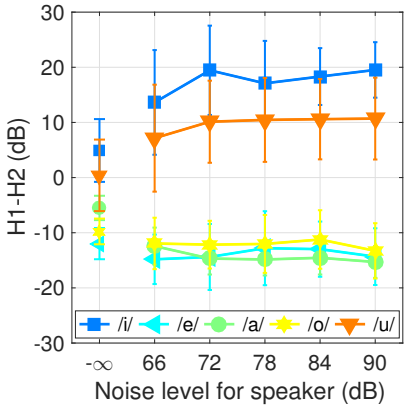
(b) F0 mean (female)



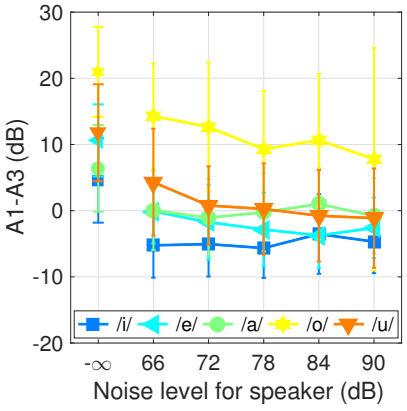
(c) Formant amplitude at F1 (female)



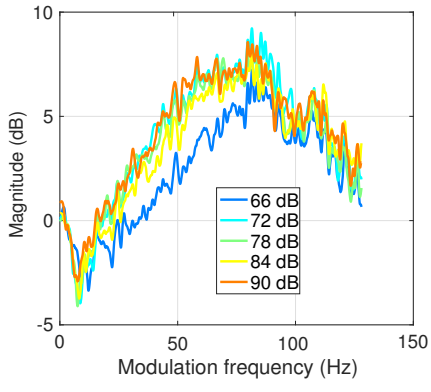
(d) Vowel spaces (left: female, right: male)



(e) H1-H2 (female)



(f) A1-A3 (female)



(g) Modulation spectral difference of Lombard speech from neutral speech at 890 Hz acoustic frequency (female)

Figure 2: Analysis results

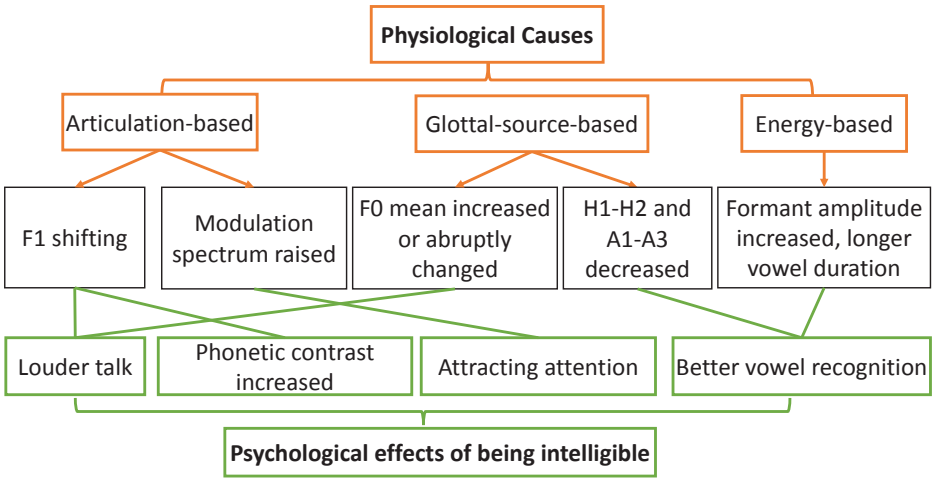


Figure 3: Physiological-acoustical-psychological correlation

might occur in different directions for each group. Secondly, they can still be considered to be in the same group i.e., one direction of change exists in the glottal source. This might be evidenced by the strong effect of the vocal tract during the production of /i/ and /u/. Specifically, in /i/ and /u/, H1 nears F1. The formant amplitude at F1 is seen to increase, which causes a large increase in H1. H2 is increased, but the increase is negligible compared with that in H1. This leads to an increase in H1-H2 for /i/ and /u/. The effect of the vocal tract on /e/, /a/ and /o/ is much smaller because F1 is far from H1 and H2. By this argument, all the vowels can be counted as one group with decreasing spectral tilts. Among our hypotheses, the second one is more feasible. In short, the H1-H2 variation and the decrease in A1-A3 still show the redistribution of energy from a low to high-frequency region and significantly promote formants, possibly increasing vowel realization. Increasing the vowel length requires more power and provides more time to recognize vowels. The raising of the modulation spectrum (16 - 128 Hz modulation frequency) by articulation shows that some parts of the power envelope rise more vertically. This might indicate the appearance of emphasizing points to attract attention from listeners.

5. Conclusion

In this study, by analyzing Lombard speech produced in backgrounds with various noise levels, significant feature variations corresponding to an increase in the noise level were extracted. They are increased vowel duration, an increase and abrupt change in F0 at 84 dB, an increase in F1, an increase in formant amplitudes and H1-H2 variation, a decrease in A1-A3 and raising of the modulation spectrum. These variations can be physiologically and psychologically explained in terms of the increased intelligibility of louder talk, increased phonetic contrast, better vowel recognition and attracting attention. Moreover, the continuous variation with increasing noise level are foundations for intelligible adaptation in noise. Future work is to verify these variations with the noise level and adapt them in a maximally intelligible manner to resynthesized speech.

Acknowledgment

Part of this research was supported by SECOM Science and Technology Foundation.

References

[1] E. Lombard: Le signe de l'elevation de la voix, Annales des Maladies de l'Oreille, Vol. 37, pp. 101–119, 1911.

[2] M. Cooke and Y. Lu: Spectral and temporal changes to speech produced in the presence of energetic and informational maskers, JASA, Vol. 128, pp. 2059–2069, 2010.

[3] Y. Lu and M. Cooke: The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise, Speech Commun., Vol. 51, No. 12, pp. 1253–1262, 2009.

[4] J. C. Junqua: The Lombard reflex and its role on human listeners and automatic speech recognizers, JASA, Vol. 93, No. 1, pp. 510–524, 1993.

[5] R. Kubo, D. Morikawa and M. Akagi: Effects of speaker's and listener's acoustic environments on speech intelligibility and annoyance, Proc. INTER-NOISE, 2016.

[6] K. Kondo, S. Amano, Y. Suzuki and S. Sakamoto: Japanese speech dataset for familiarity-controlled spoken-word intelligibility test (FW07), NII-SRC, 2007.

[7] H. Kawahara, I. Masuda-Katsuse and A. De Cheveigne: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds, Speech Commun., Vol. 27, No. 3, pp. 187–207, 1999.

[8] B. P. Nguyen and M. Akagi: A flexible spectral modification method based on temporal decomposition and Gaussian mixture model, Acoust. Sci. Technol., Vol. 30, No. 3, pp. 170–179, 2009.