2017 RISP International Workshop on Nonlinear Circuits,
Communications and Signal Processing (NCSP'17)
Guam, USA, February 28th to March 3rd, 2017

Acoustical analyses of Lombard speech by different background noise levels for tendencies of intelligibility

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

Japan Advanced Institute of Science and Technology
1–1 Asahidai, Nomi, Ishikawa 923–1292 Japan
Phone: +81-761-51-1149
E-mail: {vanthuanngo, rkubo, morikawa, akagi}@jaist.ac.jp

Abstract

This paper investigates acoustic variations for producing Lombard speech under the effect of environmental dynamics to identify adaptive tendencies of intelligibility. Analyses of acoustic features: duration, F0, formants, spectral tilts, and modulation spectrum on the dataset of speech: $-\infty$(neutral), 66, 72, 78, 84, and 90 dB noise level were carried out. Analysis results show that the recognized tendencies (neutral-Lombard distinction) including lengthening vowel duration, increasing F0, shifting F1 and decreasing spectral tilt (A1-A3) still preserve among Lombard speech produced in a various noise-level background. Besides, new findings are abrupt changing in F0 at 84 dB, increasing formant amplitudes, H1-H2 variation, and lifting modulation spectrum. Basing on the physiological and psychological knowledge we can reason their correlations with intelligibility. Moreover, those variations are continuously varying with noise level increasing. As a result, it can be suggested that they are related to the adaptive tendencies of intelligibility.

1. Introduction

Researchers have been investigating Lombard speech [1] to explore mechanisms of improving speech intelligibility in noisy environments. The better intelligibility is recently explained by release from masking. The reduction in foreground-background overlap causes release from both energetic and informational masking for listeners [2]. More specifically, Lu *et al.* [3] pointed out that the acoustic changes from neutral speech - speech spoken in quiet: lengthening duration, increasing F0, and flattening spectral tilts are main contributing factors. Then, by mimicking Lombard speech [4], the intelligible speech can be synthesized from human or synthetic one with high stability and preservation of naturalness.

However, when changing environments are considered, some limitations arise. First, it has still lacked analyzing Lombard speech in a noise-level-varying background. Consequently, a convincing explanation for correlation of Lombard speech production with physiological and psychological meanings in intelligibility improvement has also remained. Second, in re-synthesis, the problem of maximally intelligible adaptation has been unresolved. They have still limited to the capability of Lombard speech itself or unadapted when the noise level is changing. Therefore, if we want to present intelligible speech maximally adapting with noise, findings of optimal solution on noise-level adaptation need to be done. Then, it is required to perform these analyses and manipulate intelligible tendencies studied from Lombard speech produced in a various noise-level background.

In this study, we conducted analyses on acoustical properties of neutral and Lombard speech produced in the various noise-level environments. The set of acoustic features predicted to have a strong relationship with intelligibility were chosen to analyze. By putting acoustic parameters of all investigated speech under the order of noise level increasing, it could better realize tendencies for producing Lombard speech under the effect of environmental dynamics. It is also easier to argue which acoustic variations could be reasonable for being intelligible.

2. Analysis Procedure

2.1 Speech Corpus

Speakers and recording word lists were drawn from the previous study that examined intelligibility of Lombard speech [5]. A male and a female participated in the recording. Three familiarity-controlled word lists [6](60 words - Type of pitch accent pattern) with lowest familiarity rank (1.0-2.5) were used. Each word contains 4 morae (e.g. sa sa wa ra). It was embedded in a carrier sentence as a target word: ”Tsugi ni yomu tango wa” word ”desu”. The speech was different from the ones used in the listening tests, yet their intelligibility can be implied.

2.2 Feature Extraction

Table 1: Analyzed Acoustic Features

Acoustic Properties	Acoustic Feature	Feature estimation method
Duration	Consonant, Vowel Duration	From segmented phonemes
F0	F0 mean, F0 slope	F0 extracted by STRAIGHT [7]
Formants	Frequencies, bandwidths, and amplitudes	LPC, Spectral-GMM based spectra [8]
Spectral tilts	H1-H2 (voice quality), A1-A3 (global tilt)	Harmonics in FFT spectrum
Modulation spectrum	Modulation Spectral Difference	A method based on Zhu *et al.* [9]

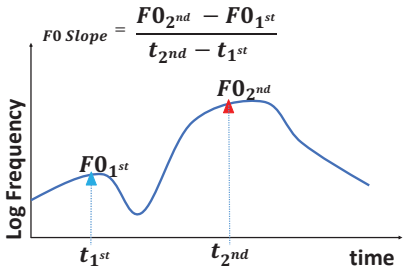

Figure 1: F0 Slope

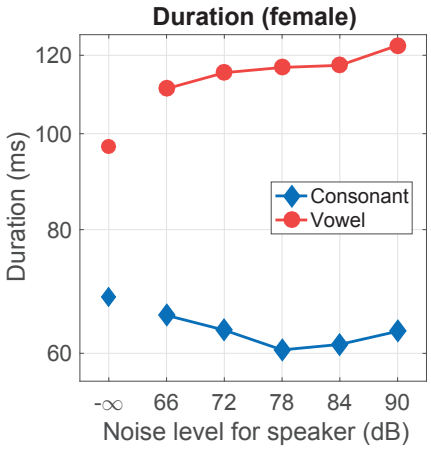

Figure 2: Consonant and Vowel Duration (female)

This study aimed to realize acoustic variations among Lombard speech which characterize for intelligibility. Hence, a selection of distinctive features of Lombard speech and recognizing intelligible features was concerned. Specifically, we first considered analyzing the basic acoustic features: duration, F0, spectral tilts, which represent for differences between Lombard and neutral speech. Besides, formants which stand for vowels were investigated. Moreover, the study extended to examine a new one - modulation spectrum which was known well contributing to speech perception. The features were extracted from all neutral and Lombard speech. The details are shown in Table 1 and the explanation below.

F0: F0 mean - Mean of F0 contour of a word. F0 slope - Slope from F0 at the center of vowel of the 1^{st} mora to F0 at the center of the 2^{nd} mora (Figure 1).

Formants: F1 and F2 were used to produce vowel space. Formant bandwidths and amplitudes at F1, F2, and F3 were also considered.

Spectral Tilts: H1-H2 - The spectrum-level difference between the first and second harmonic. A1-A3 - The spectrum-level difference between the nearest harmonics to F1 and F3.

Modulation Spectrum: Inspired by Modulation Filter [9], a method which can be used for both analyzing and modifying power envelope of the spectrum extracted by STRAIGHT [7] was employed. For each frequency (acoustic frequency), Fourier transform was applied on the power envelope eliminated its mean value. The Fourier transform frequency can be considered modulation frequency. The acoustic frequencies are coordinated with modulation frequencies to produce Modulation Spectrum.

3. Results and Discussion

Results: The values of acoustic parameters are observed under noise level increasing. (Only the acoustic features showing their variations with noise levels correspondingly were considered. If the figures standing for one gender are presented, the tendencies that they demonstrate also happen in another gender.) In particularly, with noise level increasing, vowel duration is lengthened (Figure 2). F0 is increased continuously (35% in male, 67% in female) or changed abruptly at 84 dB (65% in male, 33% in female) (Figure 3). Shifting F1 can be seen: /a/, /e/, /o/ forward and /i/, /u/ backward (Figures 4) and all vowels forward (Figure 5). All formant amplitudes at F1, F2, and F3 are increased (e.g. Figure 6). Biasing H1-H2 can be seen by decreasing in /i/, /u/ and increasing in /a/, /e/, /o/ (Figure 7). Decreasing A1-A3 is also observed (Figure 8). Lifting in modulation spectrum from 16 Hz - 128 Hz modulation frequency and below 1000 Hz acoustic frequency are presented (Figure 9). The higher noise level is, the stronger the lifting is.

Discussion: These varying tendencies show the patterns of louder talk, phonetic-contrast increase, better vowel recognition, and energetic-temporal masking release. They help to increase intelligibility for Lombard speech. In details, the recognized tendencies [2][3] of the neutral-Lombard distinction are still preserved among noise-level varying Lombard speech. On F0 increasing and abrupt changing at 84 dB (per-

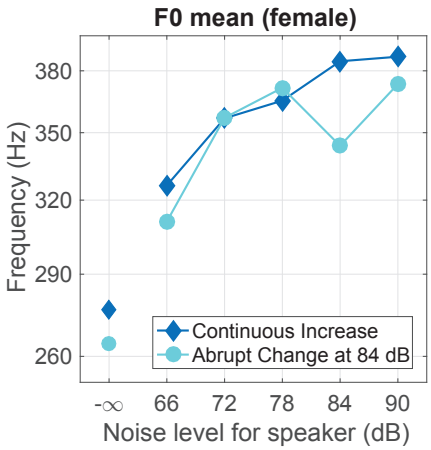

Figure 3: F0 mean of the female speaker

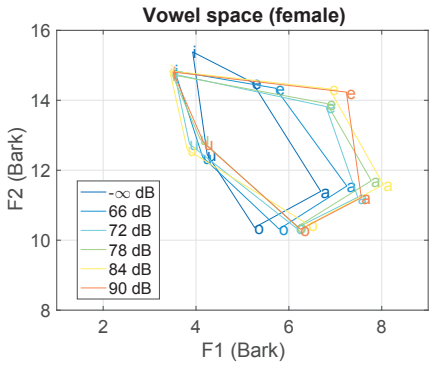

Figure 4: Vowel space of the female speaker

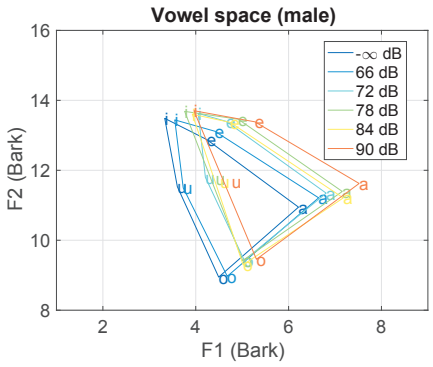

Figure 5: Vowel space of the male speaker

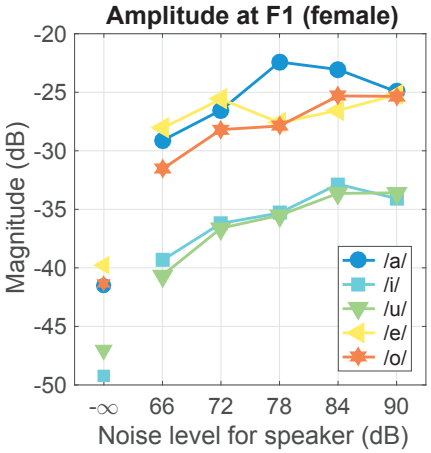

Figure 6: Formant amplitude at F1 (female)

haps, changing speaking style from modal to scream) in both speakers, and F1 shifting (all vowels to higher region) in the male speaker mean mouth opens more to speak louder. The F1 shifting in the female to lower region (/i/, /u/) and higher range (/a/, /e/, /u/) seems to expand vowel space, then increase phonetic contrasts among vowels. Formant amplitude increasing means more energy at vowels, which might be easier recognize vowels.

Besides, in previous study [3], spectral tilts are decreased for Lombard speech. However, in our results: H1 - H2 decreases in /i/, /u/ , increases in /a/, /e/, /o/ and A1 - A3 decreases in all vowels. Therefore, it could have two possibilities. First, /i/, /u/ are actually grouped and biases from the group of /a/, /u/, /e/. It means changes in glottal source might go in different directions for each group. Secondly, they still can be considered in the same group - one direction of change for the glottal source. It might be evidenced by the strong effect of vocal tract during producing /i/ and /u/. Specifically, the H1 and F1 of /i/, /u/ are close together. Formant amplitude at F1 is seen to be increased, which much affects to the increasing in H1. H2 is increased, yet it is negligible comparing with increasing in H1. Then, it leads to the increasing in H1-H2 for /i/, and /u/. Otherwise, the effect of vocal tract on /a/, /e/, 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 of decreasing spectral tilt. In our hypotheses, the second one is more feasible. The tendency of A1-A3 decreasing also shows the redistribution of energy from low to high frequency region and promote significant formants, perhaps to release from energetic masking and increase vowel realization. The H1-H2 variation still reflects emphases in formants.Vowel lengthened increases contrast from the background. Modulation spectrum (16Hz - 128 Hz modulation frequency) lifted shows power envelope fluctuating more rapidly, more contrast with noise. Both seem to release from temporal masking. Moreover, those acoustic parameters can be seen continuously varying with noise level increasing.

4. Conclusion

In this study, by analyzing Lombard speech produced in the various noise-level background, significant feature variations corresponding with noise level increasing were extracted.

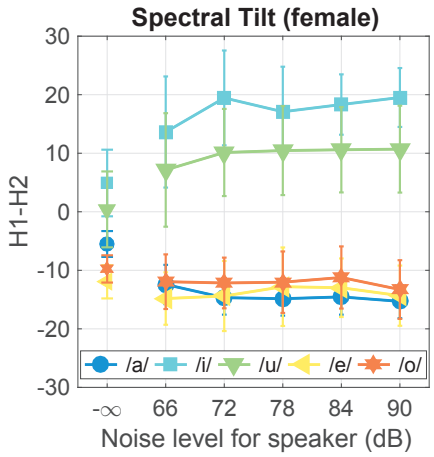

Figure 7: Spectral tilt H1-H2 (female)

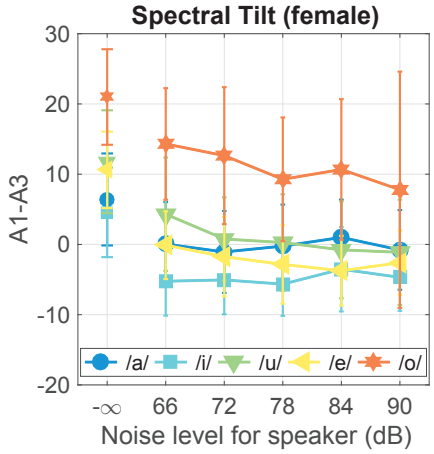

Figure 8: Spectral tilt A1-A3 (female)

They are lengthening vowel duration, increasing and abrupt changing at 84 dB in F0, shifting F1, increasing formant amplitudes, H1-H2 variation, decreasing A1-A3, and lifting modulation spectrum. Those variations can be physically and psychologically reasoned for the intelligible patterns of louder talk, phonetic-contrast increase, better vowel recognition, and energetic-temporal masking release. Moreover, they are continuously varying with noise level increasing. All of them are served as a foundation of intelligible adaptation in noise. In future work, it is to verify the validity of those tendencies and maximally-intelligible adapt with noise-level varying on resynthesized speech.

Acknowledgment

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

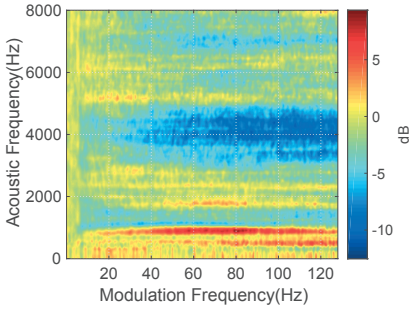

Figure 9: Modulation spectral difference (Lombard 90 dB and Neutral speech)

References

[1] Lombard, E. (1911). Le signe de l'elevation de la voix. Annales des Maladies de l'Oreille, 37, 101-119.

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

[3] Lu, Y. and Cooke, M. (2009). The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. Speech Commun., 51(12), 1253-1262.

[4] Junqua, J. C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. JASA., 93(1), 510-524.

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

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

[7] Kawahara, H., Masuda-Katsuse, I. and De Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech comm., 27(3), 187-207.

[8] Nguyen, B.P., Akagi, M. (2009). A flexible spectral modification method based on temporal decomposition and Gaussian mixture model. Acoust. Sci. Technol., 30(3), 170-179.

[9] Zhu, Z., Nishino, Y., Miyauchi, R., Unoki, M. (2016). Study on linguistic information and speaker individuality contained in temporal envelope of speech. Acoust. Sci. Technol., 37(5), 258-261.