学位論文の骨子（情報科学）　　　　　　　　　　　　　　　　年　　月　　日

Dissertation Outline (Information Science)

氏名 Name	NGO Thuan Van	学生番号 Student Number	1720010

主指導教員 Supervisor	Masato AKAGI	印 Seal	副指導教員 Second Supervisor	Masashi UNOKI	印 Seal
■ 副テーマ指導教員 Advisor for Minor Research Project □ インターンシップ指導教員 Advisor for Internship				Jianwu DANG	印 Seal

＜博士論文題目（仮）＞ (Tentative) Title of Doctoral Dissertation

Improvement of intelligibility of speech with naturalness in varying noisy reverberant conditions based on speaking styles of Lombard and clear speech.

＜研究の目的と効果＞　Research Aim and Impact

In the public announcements provided in train stations, airports, and factories, the presence of noise and reverberation often smears transmitted speech and makes it difficult for listeners to understand. There are wo possible approaches to solving this problem. One is reducing effects on both noise and reverberation and another one is making the speech itself more intelligible. It is impractical to reduce the noise and reverberation due to the complex structures of the facilities. It should be, therefore, possible and more flexible to increase the intelligibility of transmitted speech, from the second approach.

Under noisy conditions, intelligibility of transmitted speech can be increased by (1) optimizing indices of objective measures and (2) mimicking intelligible speaking styles. The former (1) tried to improve speech intelligibility in noise by optimizing indices of objective measures like speech intelligibility index (SII) (Taal et al., 2013; 2014, Cooke et. al., 2018) Although this method obtained considerable intelligibility improvement, the modified speech has low naturalness. Because speech features for naturalness are often broken. The latter (2) tried to mimic clear speech and/or Lombard speech. Clear speech is produced when humans speak as clearly as possible under quiet conditions. Lombard speech is a kind of intelligible speech that is produced in noise due to the Lombard effect. By applying their properties into modifying neutral speech (uttered in quiet conditions), intelligible and natural speech under noise can be synthesized. State-of-the-art methods to mimic Lombard speech are based on statistical methods like Gaussian Mixture Model (GMM) or Deep Neural Network (DNN) techniques (Lopze et al., 2017, Bollepalli et al., 2017). In these methods, speech features are automatically mapped from neutral speech to these of Lombard speech by trained models, and the resulting mimicked speech resembles Lombard speech. However, when considering multiple noise levels, since the characteristics of Lombard speech are varied according to noise level, these state-of-the-art methods require an extremely huge dataset to train. Thus it becomes impractical to render them. As another way, rule-based methods such as spectral shaping and dynamic range compression (SSDRC) (Zorila et al., 2012) and Lombard effect mimicking (Huang et al., 2010) produced the transmuted speech with excellent intelligibility and acceptable naturalness. The problem is that these methods with fixed control rules are suitable for just one noise level rather than multiple noise levels. Feature modification methods that perform adjustments according to noise level are still hard to control. Although SSDRC can produce mimicked speech with better intelligibility than Lombard speech (Cooke et al., 2013), few studies has especially concerned problems in exceeding intelligibility and preserve naturalness of Lombard or clear speech. In reverberation, few studies have clarified the best way to increase intelligibility in reverberant conditions. Modulation filtering (MF), steady-state suppression (SSS) and "clear" speech mimicking (the speech is clear under reverberant conditions) are mainly applied. In MF and SSS (Kusumoto et al., 2005, Arai et al., 2007), intelligibility is increased by improving consonant identification. Meanwhile, little information was reported for the clear speech in reverberation. This kind of clear speech is still expected to provide some good solutions as Lombard speech.

Considering pros and cons of the previous methods, objectives of this study is to develop the concept that rules studied from Lombard speech and clear speech are applied to obtain both intelligibility and naturalness for synthesized speech. Previous methods are unable to adapt to changing environments, and intelligibility improvement based on only Lombard speech and clear speech is still limited. Instead, our proposed rule-based method (Ngo et al., 2019) aims to adapt to any changes of noises and to exceed intelligibility of Lombard and clear speech. To this end, the following procedures were performed with Lombard speech and were considered performing with the clear speech in reverberation.

1. Analysis of Lombard speech: Trends of distinctive features of Lombard speech with noise levels were explored.
2. Extract rules for mimicking: To adapt speech to varying noise, mimicking rules according to noise levels are generated.
3. Implement the rules into the system: Synthesis/modification methods were adopted. They can control features independently and flexibly. The expected feature values generated by a rule generation model can be precisely obtained.
4. Evaluation: Subjective listening tests of intelligibility and naturalness were performed to evaluate quality of mimicking.
5. Exceed Lombard speech: Mimicking speech had been achieved. Then, strategies to improve intelligibility and naturalness exceeding Lombard speech with any noise were explored. Effective features and their optimal values were identified.

Finally, this study achieved strategies to control speech quality to improve intelligibility and preserve naturalness of speech in noise and reverberation. It can be applied to public announcement and evacuation guidance systems. The novel points are rule generation model, strategies to improve speech intelligibility in various noise levels, types and reverberation. The found strategies and effective features can be used as criteria to measure speech intelligibility in noise and reverberation.

＜研究の概要＞ Research Outline

This study aims to construct a rule-based method in order to improve intelligibility and naturalness of speech according to noise and reverberation. Specifically, i**n noisy environment**, analyses were adequately performed on the acoustical features of neutral and Lombard speech produced in backgrounds with various noise levels. Then, on the basis of the analyzed feature tendencies according to noise levels, a continuous rule generation model of acoustic features was designed to precisely estimate with the effects of noise. This model is expected to overcome the inaccuracy of the previous models and to increase adaptability of mimicking speech. To flexibly and precisely control multiple features with varying noise levels in a way that preserves the naturalness of synthesized speech, some advanced synthesis and modification methods were implemented. The mimicked speech was compared with that of statistical Bayesian GMM-based methods (BGMM) and Lombard speech through subjective listening tests. In exceeding Lombard speech, effective acoustic features were identified by analysis-by-synthesis methods. On the basis of the rule generation model, independent control of multiple features and subjective evaluations, varying acoustic features affecting intelligibility were identified and optimized. In reverberation, aforementioned strategies are considered applying to the clear speech in reverberation. At last, evaluations of improved strategies in noisy reverberant conditions are performed. **In reverberant environment**, analysis results and effective acoustic features can be inherited from our colleagues' studies (Kubo et al., 2018, Kobayashi et al., 2017). Thus the current procedure is to exceed the clear speech in reverberation only. Especially, it is directly to find out optimal time-frequency features (acoustic frequency and modulation frequency) to improve intelligibility and preserve naturalness of mimicked speech under reverberant conditions. After all, the general discussion on strategies for both noisy and reverberant environments is going to be given. The application of these strategies is also being deployed. The details are given below.

Analysis of Lombard speech: By analyzing neutral and Lombard speech produced under various levels of pink noise, the distinctive features and their varying tendencies were discovered. They included decreasing spectral tilt, increasing mean and range of fundamental frequency (f_0), shifting formants F_1, F_2, F_3, and F_4, shortening vocal tract length, increasing average power and vowel-to-consonant ratio, and increasing vowel duration. These tendencies continuously varied with increasing noise levels.

Extract rules for mimicking Lombard speech: Using analyzed results and on the basis of the model reported by Hodgson et al. 2007 in which the Lombard effect represents the relationship between the constitutional factors of environments with noise levels, a rule generation model was proposed. This model represented the relationship between acoustical parameter values and noise levels. It is estimated to have a drastic change around 66 dB and a saturation starting from 90 dB. With very small fitting errors, the proposed model successfully captured the tendencies for all features.

Implement the rules into the system: To convert neutral speech to Lombard speech fit with varying noise levels, a combination of STRAIGHT, a coarticulation model, Modified Restricted Temporal Decomposition (MRTD), and the newly designed rule generation model based on noise levels was used. With these components, the proposed method can deal with an adequate number of acoustic features, model acoustical events precisely, modify features more flexibly and easily, and continuously estimate and apply the effect of noise level to acoustic features. It shows great potential to obtain high-quality synthesized speech adapted to noisy backgrounds.

Evaluation: Subjective listening tests of intelligibility and naturalness were carried out to compare the proposed mimicking method based on the rule generation model with a BGMM-based method optimally trained for each noise level. The results showed that the proposed model got comparable similarity and adaptively to the noise levels. Intelligibility and naturalness are comparable with spectral tilt modification. When noise levels are continuous, the BGMM-based method cannot adapt features to the noise levels, while in contrast, the proposed model can interpolate Lombard effect in any noise level. The most promising finding here was that the proposed method can control parameter values independently, thus enabling us to determine the most related parameters to intelligibility and improve intelligibility in noise more in the next step.

Exceed Lombard speech: Effective features and their optimal values for intelligibility and naturalness of speech with any noise types and levels were identified to establish strategies in exceeding Lombard speech.

- Effective features: By VocalTractLab (Birkholz, 2007), an articulatory analysis-by-synthesis method was applied to identify the effective features from all the mimicked features. The results showed that spectral tilt, f_0 and formants were effective. Here, power spectrum, spectral slope, plateau between 2-6 kHz of spectral tilt, f_0 mean, and F_1, F_2, F_3 and F_4 were involved. To identify more specific features when varied affecting intelligibility, experiments for intelligibility and naturalness were carried out on speech created from the variations of these features in multiple levels of pink noise. The results showed that varying amplitude of the plateau between 2-6 kHz and f_0 mean changed intelligibility.
- Optimal values of effective features: To estimate optimal values of the effective acoustic features to control, experiments for intelligibility and naturalness were conducted on various variations of these effective features and their mutated combinations. The results showed that increasing the plateau between 2-6 kHz about 13 dB for the noise level at 84 dB was the optimal spectral shaping value. Other advanced spectral shaping methods are in SSDRC and optimization of high energy glimpse portion (Cooke et al., 2018). A further process was thus performed to explore an optimal spectral control among all of these spectral shaping methods. Experiment of intelligibility and naturalness on the speech produced by these spectral shaping methods were conducted in various noise types and levels. On the basis of modulation spectrum (MS) and modulation transfer function (MTF) concepts (Houtgast et. al, 1985) and listening test results, it indicated that equally significant frequencies to improve intelligibility in noise were 1.25 kHz, 2.5-3 kHz, and 5-6 kHz acoustic frequency. Keeping the spectral dips in these frequency regions preserved naturalness. In addition, when analyzing dynamic range compression (DRC) in SSDRC, it was found that increasing amplitude of modulation spectrum around 4 Hz and from 8 Hz modulation frequency improved intelligibility and affected naturalness.

＜研究の概要（つづき）＞ (continued from previous page)

As reported in speech intelligibility tests and acoustic analyses of the clear speech in reverberation (Kubo et al., 2018), speech produced under long T60 (an abbreviation for reverberation time 60 dB, the time it takes for the sound pressure level to reduce by 60 dB) is more intelligible than neutral speech in reverberation. Vowel space of the clear speech in reverberation, compared with that of neutral speech, is expanded: high F_2 for front vowels and low F_2 for back vowels, and high F_1 for open vowels and low F_1 for close vowels. Brief formant transitions are found. Also, modulation spectra of the reverberant environment speech have higher modulation index than these of the neutral speech (Kobayashi et al., 2017). They can be effective acoustic features to intelligibility in reverberation. In exceeding intelligibility of the clear speech in reverberation, MS and MTF based analyses are applied on the same dataset used in the speech intelligibility tests of Kubo et al.'s study in order to find out an optimal control of time-frequency features. The analysis results are being collected.

In this study, Lombard speech produced in various noise levels was analyzed. It showed that the distinctive acoustic features continuously vary with increasing noise. The rule generation model to adapt acoustic features with noise levels was proposed. Effect of articulatory and acoustic features to intelligibility of speech in noise was clarified. On the basis of MS and MTF concepts and experiment results, the strategies to control features to improve intelligibility and preserve naturalness of speech in noise and reverberation are being suggested. The main contribution of this research was providing feature controlling strategies and methods to improve intelligibility and preserve naturalness of speech under noise and reverberation conditions. It can be applied for public announcement and evacuation guidance systems in noise and/or reverberation conditions.

＜論文の構成案＞ Dissertation Structure

1. Introduction
    - 1.1 Research background
    - 1.2 Research approach
    - 1.3 Dissertation organization
2. Literature review
    - 2.1. Speech intelligibility improvement
    - 2.2. Clear speech
    - 2.3. Lombard speech
    - 2.4. Research purpose and problem
3. Mimicking Lombard speech
    - 3.1. Acoustical analysis of Lombard speech
        - 3.1.1. Feature extraction methods
        - 3.1.2. Analysis results
        - 3.1.3. Rule generation model
    - 3.2. Synthesis of Lombard speech
        - 3.2.1. Feature modification and synthesis methods
        - 3.2.2. Experiment results and discussion
4. Exceeding Lombard speech
    - 4.1. Effective features and strategies to improve intelligibility under various noises
        - 4.1.1. Articulatory-acoustic features
            - 4.1.1.1. Articulatory studies of speech intelligibility
            - 4.1.1.2. Effect of articulatory-acoustic features to intelligibility
        - 4.1.2. Acoustic features
            - 4.1.2.1. Identification of effective acoustic features
            - 4.1.2.2. Modulation spectrum and modulation transfer function
            - 4.1.2.3. Modulation spectrum analysis
        - 4.1.3. Discussion on strategy to control effective features in noise
5. Reverberant environment speech
    - 5.1. Inherited and shared acoustic features with Lombard speech
    - 5.2. Modulation spectrum analysis
    - 5.3. Discussion on strategy to control effective features in reverberant and noisy reverberant conditions
6. Application of strategies into noisy reverberant conditions
    - 6.1. Evaluation speech and noisy reverberant conditions
    - 6.2. Experiments
    - 6.3. Results and discussion
7. General discussion
8. Conclusion

＜研究業績＞ Publication List

Itemize your publications according to the type of publication.
After the name(s) of the author(s), list the name of the publication, the issue/volume number(s), page number(s) and date of publication.

❖ Journal

[1] ○ Thuan Van Ngo, Rieko Kubo, Daisuke Morikawa, Masato Akagi, Acoustical Analyses of Tendencies of Intelligibility in Lombard Speech with Different Background Noise Levels, Journal of Signal Processing, 2017, Volume 21, Issue 4, Pages 171-174.

❖ International conference paper

[1] ○ Thuan Van Ngo, Rieko Kubo, Daisuke Morikawa, and Masato Akagi. "Acoustical analyses of Lombard speech by different background noise levels for tendencies of intelligibility." In 2017 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP'17). Research Institute of Signal Processing, Japan, 2017.

[2] ○ Thuan Van Ngo, Rieko Kubo and Masato Akagi, Evaluation of the Lombard effect model on synthesizing Lombard speech in varying noise level environments with limited data, APSIPA 2019 (Appeared in Nov, 2019)

❖ Domestic conference paper

[Not refereed paper, poster presentation]

[1] ○ Thuan Van Ngo, Rieko Kubo, Masato Akagi, Acoustical control method for increasing intelligibility based on Lombard speech uttered in background noises with various levels, , Acoustic society of Japan, Fall 2017, 2-Q-29, pp. 313-316.

[2] ○ Thuan Van Ngo, Rieko Kubo, Masato Akagi, Acoustical rules for mimicking Lombard speech produced in a various noise level background, IEICE Technical Report, Engineering Acoustics, vol. 117, no. 170, 2017.

[3] ○ Thuan Van Ngo, Rieko Kubo, Masato Akagi, Speaker-independent control model for mimicking Lombard speech uttered in background noises with various levels, Acoustic society of Japan, Spring 2018, 3-P-33, pp. 1371-1374.

[4] ○ Thuan Van Ngo, Rieko Kubo and Masato Akagi, Improved quality and intelligibility of mimicking Lombard speech by source-filter and coarticulation model-based synthesis, Acoustic society of Japan, Spring 2019.

❖ Others

[1] NGO, Thuan Van, 日本音響学会第 15 回（2017 年春季研究発表会）学生優秀発表賞

＜現在の単位修得状況＞ Courses I have obtained credits

	発展科目 Intermediate	先端科目 Advanced	その他 Others	合計 Total	必修 B 科目(S50x) Required Courses
科目数 Number of courses	4	2		6	■ S503 □ S501 / S502
単位数 Number of credits	8	4		12	