# Acoustic analysis of adaptive tendencies in Lombard speech produced in a noise-level varying background

By NGO THUAN VAN

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Professor Masato AKAGI

March, 2017

# Acoustic analysis of adaptive tendencies in Lombard speech produced in a noise-level varying background

By NGO THUAN VAN (1510016)

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Professor Masato AKAGI

and approved by
Professor Masashi UNOKI
Professor Jianwu Dang
Associate Professor

February, 2017 (Submitted)

# Abstract

Intelligibility of speech is extremely important to understand information of a transmitted message in noisy reverberant environments. The noise and reverberation smear the speech content during its transmission. The degraded speech causes the failure of understanding for listeners. Many methods have been proposed to increase speech intelligibility in these circumstances. Some methods concentrate on eliminating all the negative effects of such environments. By applying estimation algorithms, they try to estimate the noise and reverberant segments to subtract them from the noisy reverberant speech or segregate speech portions in the combination with the noise and reverberation. Those techniques are impractical and infeasible in real application due to the high complication and constant dynamics the real environments and complex configuration of noise-reverberation removal devices. On another hand, many methods try to internally increasing intelligibility of transmitted speech. Ones base on the signal processing techniques consisting of dynamic range compression, formant enhancement, and score maximization of an objective measurement to obtain better speech intelligibility. Many others follow the knowledge studied from naturally-intelligible speech i.e. clear speech, shouted speech, and emphasized speech due to Lombard effect (or can be called Lombard speech) to increase speech intelligibility intrinsically. In those studies, it is very important to realize tendencies of increasing intelligibility from the intelligible speech then control them to apply in synthesizing mimicking intelligible speech. The synthesized speech is required to be both intelligible and natural. Moreover, their intelligibility need to be preserved or well adaptive in any circumstances of surrounding environments (noisy airports, noisy and crowded train stations with noise-level varying according to time, factory noise, reverberant train station). Especially, in the continuous variations of surrounding environments for an instance noise-level varying, it even becomes more difficult. The scope of our study belongs to applied knowledge of naturally intelligible speech. The final goal is to synthesize intelligible speech that maximally adapts with varying environments by knowledge learned from the emphasizing speech due to Lombard effect - Lombard speech.

The investigations on Lombard speech to explore mechanisms of improving speech intelligibility in noisy environments have been carrying out. In general, 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. More specifically, it was pointed out that the acoustic changes from the neutral speech: lengthening duration, increasing fundamental frequency (F0), and flattening spectral tilts are main contributing factors. Then, by mimicking Lombard speech, 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 realize adaptive intelligible tendencies studied from Lombard speech produced in a various noise-level background.

The analyses on $-\infty$ (neutral), 66, 72, 78, 84, 90 dB (Lombard) speech produced in the environment of pink noise have been doing by the following procedure.

- Acoustic feature extraction: The extracted features are selected to analyze basing on their representation of neutral-Lombard differences and speech intelligibility. Basic acoustic features showing the distinction of Lombard and neutral speech are extracted first (F0, duration, spectral tilts), then the others being recognized as intelligible features are analyzed (Formants, Modulation Spectrum, energy redistribution of phonemes). The set of techniques including HTK-toolkit for automatic segmentation before manual correction; STRAIGHT to extract F0, and spectrum; LPC-based and Spectral-GMM for extracting formants, a derivation from Modulation to extract Modulation spectrum, average normalized Spectrum on frequency domain to concern with energy redistribution.

- Adaptive tendencies realization, physiological-psychological-acoustical establishment: Values of acoustic parameters are organized in an order of increasing noise level. If the variation exists, it will show with noise level varying accordingly. A specific meaning of psychoacoustics or physiology might be reflected in each kind of the realized variations.

- Adaptive tendency modeling and deployments by mathematical functions and voice conversion techniques, and their evaluations in contribution to intelligibility: to propose an adaptation model for mimicking Lombard speech over noise level increasing and to construct or apply suitable voice conversion techniques for the synthesis. The adaptive tendencies can be suggested as the adaptation constraints to further optimization or intelligible rule derivation.

This thesis has completed the first two steps. 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, increasing amplitudes of valleys between formants, H1-H2 variation, and lifting modulation spectrum and energy redistribution

in specific frequency region on 0-5 kHz. Basing on the physiological and psychological knowledge we can reason their correlations with intelligibility. They are the intelligible patterns of louder talk, phonetic-contrast increase, better vowel recognition, and energetic-temporal masking release. 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.

On handling the third step, a discussion on adaptive modeling by log-linear evolution of mathematical dependences between values of acoustic features and noise levels is given. Simultaneously, a first employment of voice conversion techniques basing on the adaptive model for synthesizing the mimicking Lombard speech has been tackling. When successfully applying the voice conversion techniques to control the recognized features and doing evaluations on resynthesized speech, the third step can be finished. To overcome the remaining problems of intelligible maximization and reach the final goal, a detailed description of possible solutions for two cases of extrapolating Lombard speech is figured out. First, it is choosing the target acoustic features to control: energy redistribution, or spectral tilts, or formant frequency, amplitudes and valleys between formants, or modulation spectrum. Second, it is choosing an objective measure to have a criterion for the variations of the acoustic features. Finally, it is needed to mathematically model the calculation of the objective measure by the target acoustic features, or brute-force find the optimal value of the objective measure by varying the acoustic features by following recognized tendencies.

Since this study still has many problems to be remained, in future work, it is to finish them. Moreover, when the adaption and extrapolation for maximal intelligibility succeeds, because the current Lombard speech was produced in pink noise, it is also needed to verify the synthesized speech with other types of noise with different noise levels. When the verification finishes, it is to design an entire system for real application of public addressing in noisy environments. Besides noisy environments, reverberation often appears, therefore, it is to find the solution of maximally intelligible speech for both noisy and reverberant environments basing on the knowledge of naturally intelligible speech as Lombard speech.

**Keywords:** Speech Intelligibility, Lombard speech, noise level varying background, acoustic analysis, adaptive tendencies, Lombard mimicking, maximal intelligibility.

# Acknowledgments

It has been my great pleasure and opportunity to have Professor Masato Akagi as my supervisor at Japan Advanced Institute of Science and Technology (JAIST). During my master study, he taught me how to carry out a scientific research, instructed me how to organize the ideas for an official document, gave me a lot of valuable advice, explanation, and comments. Moreover, he also spent his precious time to help me whenever I got any troubles. Especially, he helped and encouraged me to determine and resolve research problems with a thorough instruction. In my opinion, I always feel his effort to connect our lab members all together and to make a comfortable and boundary-less environment for studying and researching in such national diversity as our lab. For me, I really appreciate, am grateful and thankful to his supervision, support, and his kindness. In my higher level of study, under his supervision, I wish to prove myself basing on what I have learned from him during the master course and his advice and comments at that time to become an effective researcher and a good student.

I would like to express my sincere thanks, appreciation, and gratitude to Professor Masashi Unoki for his enthusiastic advice, precise comments, and his encouragement. During my master study, he gave me a lot of instructive comments in our lab meetings. Moreover, he also helped me to carefully review any documents that needed his confirmation with contributive revisions. Besides, I really appreciate his support and guidance to resolve other problems of my study and survival at JAIST.

I would like to express my special thanks to Doctor Rieko Kubo, Assistant Professor Daisuke Morikawa, Doctor Maori Kobayashi for their helpful advice, comments, and contributions to my study. They gave me a lot of comments and instructions to deal with many arisen problems during doing research.

I would like to thank Assistant Professor Miyauchi, my tutor - Mr. Teruki Toya, Mr. Khanh Nguyen Bui, my ex-lab member - Mr. Tuan Anh Dinh and Professor Elbarougy Reda, Ms. Yawn Xue, Mr. Yongwei Li, Mr. Xing Feng, Mr. Zhi Zhu, Mr. Takuya Asai, Mr. Yuta Kashihara, Mr. Daisuke Ishikawa, Mr. Dung Kim Tran, Mr. Tuan Vu Ho, Ms. Hao Thi Nguyen and other lab members all most sincerely for their advice, help and inspiration to carry out my research. It is hard to mention all of them here, one again I greatly appreciate them for their contributions in making an excellent and supportive educational environment.

I would like to acknowledge Professor Jianwu Dang for his precious suggestions, advice, and comments in the midterm defense which helped me improving the final study.

I am infinitely thankful to Associate Professor Minh Le Nguyen for his advisor, comments, and instruction of my minor research. Additionally, I would like to express my great gratitude for his selection after interviewing me at Posts and Telecommunications Institute of Technology. If he had not chosen me, I would never have studied at JAIST and never had a chance to be supervised by Professor Masato Akagi.

Yet life becomes tough without financial aid. I would like to express my great gratitude, thanks, and appreciation to Professor Masato Akagi once more, and especially Professor Hiroyuki Iida for their support during my master study and advice to resolve my critical problems to survive in Japan. I deeply appreciate and especially thank Heiwa Nakajima Foundation and JAIST for granting me scholarships in the second year of my master. I also would like to express an enormous thank to Secom Technology Foundation for their support of my research. With all of their help, I have been able to continue pursuing my dream of studying. I could also concentrate much more on my study and research without worrying about many other problems.

I sincerely thank all my friends at JAIST who always supported me in times of need. To my many Vietnamese friends at JAIST, thanks for the good times over the past two years. Thank Khanh, Quyen, Cuong, Chien, brother Dai, Linh, Hao, Vu, brother Vu, Tam, Duc, Phuc, Nhien for sharing joys and sorrows with me. It is impossible to mention all of them here, yet they indirectly relate with my thesis.

JAIST offered me the greatest studying environment I have ever experiencing - the computing environment, the excellent lecturers, the profound faculties, the industrious students, and the opportunity to meet famous researchers all over the world. Among the friendly administrators, I gratefully thank the International Student Section, Secretary Section and other sections for the kind and constant assistance they provided. Without them, I would absolutely have run into many difficulties.

Ultimately, I have preserved the biggest for the last. I wish to express my eternal love and gratitude to my family, Mom, Dad, Nga, and Tung for always encouraging and supporting me throughout all my years of the academy. I am especially appreciative to my parents for everything they taught me and for all the sacrifices they made in my growth.

# Contents

**Abstract** **1**

**Acknowledgments** **4**

**1 Introduction** **11**
1.1 Literatures . . . 12
1.1.1 Lombard Effect - Discovery of Lombard speech . . . 12
1.1.2 Intelligibility of Lombard speech . . . 13
1.1.3 Some studies on acoustical and adaptive properties of Lombard speech 14
1.2 Problem Definition . . . 14
1.3 Purposes of the Thesis . . . 15
1.4 Structure of the Thesis . . . 16

**2 Lombard Speech Corpus** **18**
2.1 Recording Conditions, Settings, and Participants . . . 18
2.2 Data Preprocessing . . . 22
2.3 Segmentation Results . . . 22
2.4 Discussion . . . 23

**3 Acoustic analysis of Lombard Speech** **24**
3.1 Acoustic Feature Extraction . . . 24
3.2 Analysis Results . . . 29
3.2.1 Duration . . . 29
3.2.2 Fundamental Frequency (F0) . . . 30
3.2.3 Formants . . . 30
3.2.4 Spectral Tilts . . . 34
3.2.5 Modulation Spectrum . . . 36
3.2.6 Energy redistribution . . . 36
3.3 Discussion . . . 37
3.4 Conclusion . . . 41

**4 Lombard Mimicking** **42**
4.1 A discussion on Mathematical Modeling for Adaptive Tendencies and Voice Conversion . . . 42

4.1.1 Duration . . . . . 43
4.1.2 Fundamental Frequency . . . . . 43
4.1.3 Formants . . . . . 46
4.1.4 Spectral Tilts . . . . . 46
4.2 Conclusion . . . . . 49

**5 Summary and Conclusion** **51**
5.1 Summary of the Thesis . . . . . 51
5.1.1 Main Contribution . . . . . 52
5.1.2 Possible solutions of remaining problems . . . . . 52
5.2 Conclusion . . . . . 53
5.3 Future Work . . . . . 53

**Publications** **57**

# List of Figures


1.1 Transformation of Auditory Feedback (created by Teruki Toya) . . . . . 12
1.2 Intelligibility of Lombard speech (produced in 66, 78, 90 dB noise level background) and neutral speech (non) [11] . . . . . 13
1.3 Related studies on acoustic analyses of neutral-Lombard distinction and Lombard adaptation system . . . . . 15

2.1 Recording configuration [11] . . . . . 19
2.2 Structure of an utterance . . . . . 21
2.3 Model Training of Forced Alignment Using ATR database [29] . . . . . 22
2.4 Phonemic segmentation by using Praat . . . . . 23

3.1 F0 slope of a word . . . . . 26
3.2 Estimation of Formants based on LPC . . . . . 27
3.3 Estimation of Formants using the spectral GMM . . . . . 27
3.4 Estimation of Formant by using LPC . . . . . 28
3.5 Estimation of harmonic locations . . . . . 28
3.6 Harmonic estimation of vowel /a/ (male) Lombard speech 66 dB . . . . . 29
3.7 Consonant and Vowel Duration . . . . . 30
3.8 Fundamental frequency . . . . . 31
3.9 Vowel space . . . . . 31
3.10 Formant Amplitudes . . . . . 32
3.11 Amplitudes of valleys between formants . . . . . 33
3.12 Ratios of amplitudes of valleys between formants and amplitudes of formants (female) . . . . . 34
3.13 Ratios of amplitudes of valleys between formants and amplitudes of formants (male) . . . . . 35
3.14 Spectral tilt H1-H2 . . . . . 36
3.15 Spectral tilt A1-A3 . . . . . 37
3.16 Modulation spectral difference (Lombard 66-78 dB and Neutral speech) . . 38
3.17 Modulation spectral difference (Lombard 84-90 dB and Neutral speech) . . 39
3.18 Energy redistribution /a/ (both genders), /i/(female) on 0-5 kHz . . . . . 39
3.19 Energy redistribution /i/ (male), /u/(both genders) on 0-5 kHz . . . . . 40
3.20 Energy redistribution /e/ (both genders), /o/(female) on 0-5 kHz . . . . . 40
3.21 Energy redistribution /o/ (male) on 0-5 kHz . . . . . 40

4.1 Modeling Vowel Duration . . . . . 43
4.2 Duration Control /mukogane/ at 90 dB (Female) . . . . . 44
4.3 F0 contour type of the dataset . . . . . 44
4.4 Construction F0 contour by Fujisaki model (Fujisaki *et al.* [31]) . . . . . 45
4.5 Optimization of F0 contour based on acoustic parameters . . . . . 45
4.6 Modeling formant amplitude (Female) . . . . . 46
4.7 Formant Modification on /o/ at 84 dB . . . . . 47
4.8 Modification of spectral tilts first frame of vowel /a/ in the $1^{st}$ mora of /yamamayu/ to mimick 84 dB Lombard speech . . . . . 48

# List of Tables


2.1 Lists of uttered sentences . . . . . 20
2.1 Lists of uttered sentences . . . . . 21
2.2 Lombard corpus . . . . . 22
2.3 Statistics on vowels and consonant for each speech type over speakers . . . 23

3.1 Analyzed Acoustic Features . . . . . 25

4.1 Mathematical modeling of acoustic values depending on the noise level (Female) . . . . . 47
4.2 Mathematical modeling of acoustic values depending on the noise level (Male) 48

# Chapter 1

# Introduction

Improving speech intelligibility in noisy and reverberant environments is still a challenging topic attracting a lot of interests of researchers. On noise reduction, some methods try to eliminate all the negative effects of these environments. They are noise reduction or reverberation elimination techniques [1] [2] [3]. Yet those noise-reverberant removals are often impractical and infeasible because of the high complication and constant dynamics of the real environments and configuration of noise-removal devices. On the internal modification of speech for better intelligibility, some methods based on typical signal processing techniques: dynamic range compression [4][5], formant enhancements [6], maximizing intelligibility score of objective measures [7]. Many others follow the knowledge studied from naturally-intelligible speech i.e. clear speech, shouted speech, and emphasized speech due to Lombard effect (or can be called Lombard speech) to increase speech intelligibility intrinsically. In those studies, it is very important to realize tendencies of increasing intelligibility from the intelligible speech then control them to apply in synthesizing mimicking intelligible speech. The synthesized speech is required to be both intelligible and natural. Moreover, their intelligibility need to be preserved or well adaptive in any circumstances of surrounding environments (noisy airports, noisy and crowded train stations with noise -level varying according to time, factory noise, reverberant train station). Especially, in the continuous variations of surrounding environments for an instance noise-level varying, it even becomes more difficult. The scope of our study belongs to applied knowledge of naturally intelligible speech. The final goal is to synthesize intelligible speech that maximally adapts with varying environments by knowledge learned from the emphasizing speech due to Lombard effect - Lombard speech. To reach the final purpose, it is firstly required to achieve mechanisms of adaptive variation of the Lombard speech with the varying environments. Therefore, in the master study, we have conducted an acoustic analysis of adaptive tendencies in Lombard speech produced in a noise-level varying background. The next section - Literatures takes an overview of previous studies on properties of Lombard speech.

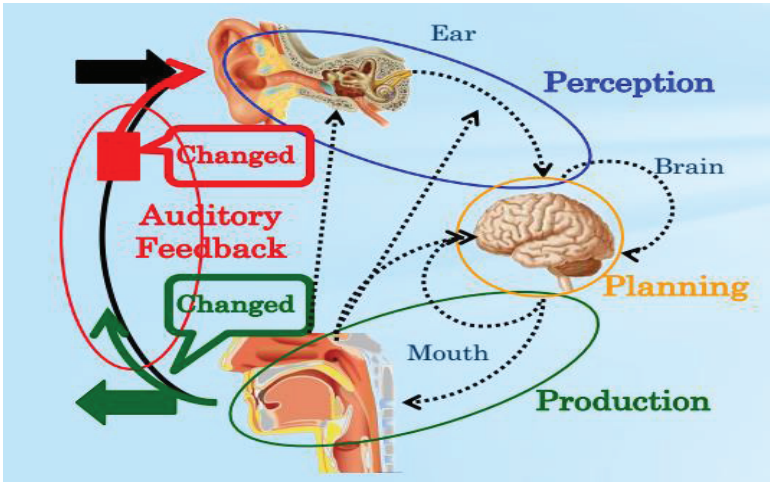



Figure 1.1: Transformation of Auditory Feedback (created by Teruki Toya)

## 1.1 Literatures

The intelligible human voice or speech still contains a lot of mysteries to discover. A special type among them known as Lombard speech [8] also possesses much of interesting knowledge. Lombard speech is a naturally intelligible speech produced when humans speak in noisy environments such as airports, train stations or noisy factories. In the field of studying speech intelligibility, by investigating the human production of Lombard speech, it is expected to explore mechanisms of increasing intelligibility in noise. Analysis on Lombard speech exists in many levels of physiology, psychology, and acoustics. A set of psycho-acoustical studies were referred in "The evolution of the Lombard effect: 100 years of psychoacoustic research" [9]. The following content provides recognized knowledge of Lombard speech with various levels of the acoustical analysis in a general comparison with neutral speech - the speech spoken in a quiet environment and the contribution into its intelligibility.

### 1.1.1 Lombard Effect - Discovery of Lombard speech

Lombard effect or Lombard reflex [8] is the phenomenon that humans increase vocal efforts when speaking in loud noise to enhance their voice audibility. Some acoustic features are found to be changed: loudness, pitch, rate, and duration of syllables. It was discovered in 1909 by Etienne Lombard. With regard to auditory feedback, Lombard effect can be considered negative auditory feedback (Figure 1.1). It shows that humans try to resolve any negative effect from the ambient environments. For instances, when they are in noise, their minimum audible threshold becomes larger. The loudness of speech relatively decrease. They try to make it louder by increasing of intensity (It is also the role of "negative feedback"). Lombard speech are produced by consequences of the Lombard effect.

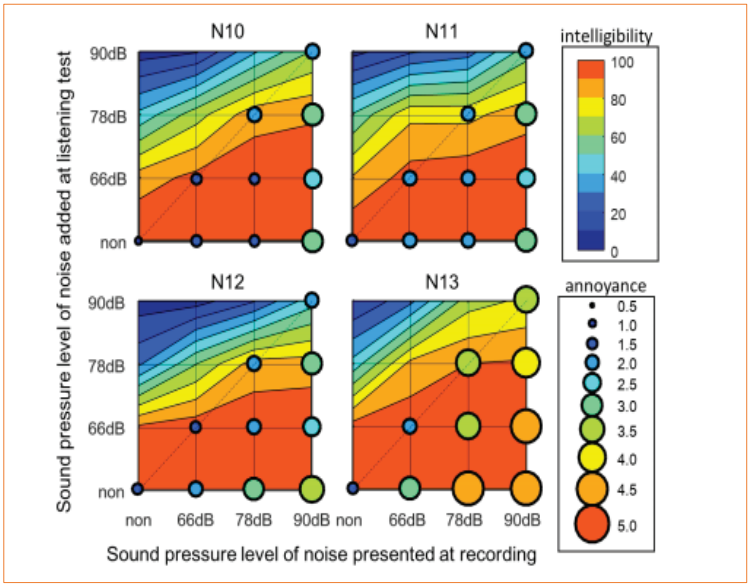



Figure 1.2: Intelligibility of Lombard speech (produced in 66, 78, 90 dB noise level background) and neutral speech (non) [11]

### 1.1.2 Intelligibility of Lombard speech

Lombard speech can be recognized the naturally intelligible speech. It is experienced by everyone, yet scientific investigations also proved for its intelligibility in noise.

- **A theory of Masking Release**

  Cooke *et al.* [10] suggested that the intelligibility of Lombard speech is because of the ability to release from masking. He did an experiment on the speech produced in quiet and in backgrounds of speech-shaped noise, speech-modulated noise, and competing speech. The analysis results show that relatively to quiet, speech output level, fundamental frequency is increased, spectral tilt is flattened in proportion to the energetic masking capacity of the background. The phenomena might be to reduce substantially the degree of temporal overlap with the modulated noise. Or in another word, perhaps, reduction in the foreground-background overlap is to release from both energetic and informational masking for listeners.

- **Subjective and Objective Evaluation**

  Besides, the theoretical explanation by Cooke *et al.* [10], some practical experiments were carried out. Kubo *et al.* [11] did a subjective test on intelligibility of Lombard speech. The testing results indicated that Lombard speech has better intelligibility in noise than the neutral speech. Moreover, an objective measure by Godoy *et al.* [12] also pointed out the better intelligibility for Lombard speech comparing with the neutral speech in noise. The details can be seen in Figure 1.2.

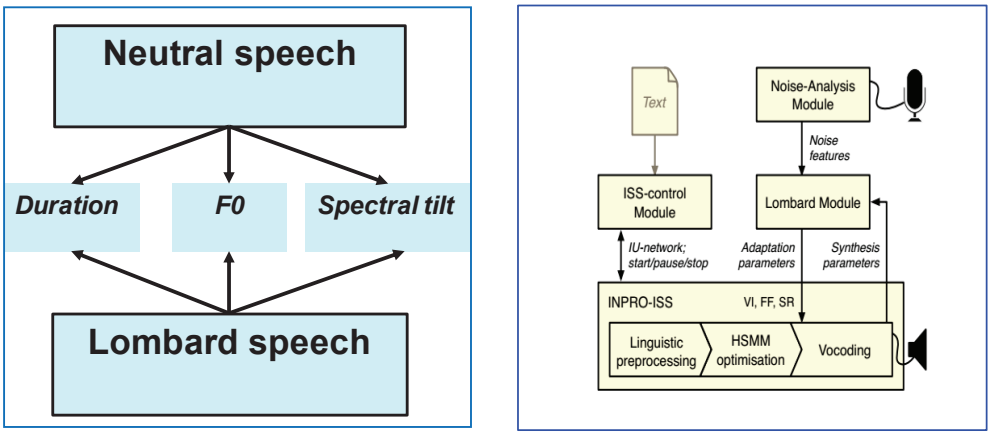
### 1.1.3 Some studies on acoustical and adaptive properties of Lombard speech

Researchers have investigated acoustic features that contribute to intelligibility improvement in Lombard speech and distinction with neutral speech and tried to deployed a Lombard adaptation system (Figure 1.3). On analyzing acoustical properties, typically, Cooke *et al.*[10] studied the contribution of durational and spectral changes to the Lombard speech intelligibility benefit. He found in Lombard speech that the duration is lengthened, The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise has been considered by Lu *et al.* [13]. Duration, amplitudes or intensity, formant frequencies, and voice onset time were also studied by Lau [14]. The effect of those typical acoustic features have been also applied in synthesizing the mimicking Lombard speech in the study of Huang [15] (imitating duration, formant frequencies, formant bandwidth, fundamental frequency, and energy in each frequency band), or Rottschafer [16] (by mimicking voice intensity, speaking rate and fundamental frequency). Many other studies have been shown those typical characteristics of Lombard speech among various languages [17] [18] [19]. In summary, a general distinction and intelligible characteristics of Lombard speech comparing with the neutral speech are due to Lengthening duration, F1 shifting, flattening spectral tilt, and increasing fundamental frequency. On an adaptive analysis, a preliminary investigation of the idea that Lombard Effect is different across noise types and noise levels were done by Hansen *et al.* [20]. Hansen's study was based on GMM classification (Gaussian Mixture Model classification), not on acoustic analyses. However, it can be considered the first concern with investigating adaptive tendencies of Lombard speech in different noise levels. A design of the online adaptation system to handle fundamental frequency (F0), intensity, speaking rate corresponding with voice intensity was firstly introduced by Rottschafer [16] but no any analysis had been performed.

## 1.2 Problem Definition

As the aforementioned knowledge in Literature section, researchers have been investigating Lombard speech [8] to explore mechanisms of improving speech intelligibility in noisy environments. In general, 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 [10]. More specifically, Lu *et al.* [13] pointed out that the acoustic changes from the neutral speech: lengthening duration, increasing fundamental frequency (F0), and flattening spectral tilts are main contributing factors. Then, by mimicking Lombard speech [15] [16] [21], 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




(a) Distinctive features

(b) Adaptive system [16]

Figure 1.3: Related studies on acoustic analyses of neutral-Lombard distinction and Lombard adaptation system

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 realize adaptive intelligible tendencies studied from Lombard speech produced in a various noise-level background.

With regard to the adaptive tendencies, a series of novel mathematical models of acoustical variations is needed to be considered. The models are required to appropriately represent for the adaptation with adequate credibility and simple complexity. The evaluations of these mathematical model are also needed. To do these tasks, a set of high-quality and reliable voice conversion techniques becomes important.

## 1.3 Purposes of the Thesis

Humans produce Lombard speech not only being intelligible but also perhaps being adaptable in noise. Some basic distinctions from the neutral speech basing on acoustic features were recognized for Lombard speech in general [10] [13]. Investigations among Lombard speech produced in the various noise level background are still unfulfilled.

Therefore, motivated by the hypothesis of the existence of adaptive tendencies, this study aims to perform acoustic analyses on Lombard speech produced in the noise level varying background to realize the adaptable acoustical variations. Simultaneously, a correlation of these variations with physiological and psychological aspects of intelligibility is discussed. By further modeling the tendencies, and evaluating their effects, we try to find out the acoustic features being robust and be able to intelligibly adaptable with noise

level increasing. They could be suggested as important features and rules in an adaptive synthesis system dealing with the various noise-level background. We hypothesize a preliminary realization of adaptation as follows: with noise level continuously increasing, the acoustic variations which preserve the same tendency among utterances and speakers: continuously decreasing or increasing, or abrupt changing can be considered adaptable. Then, they can be modeled and evaluated to become actual adaptive tendencies. On processed Lombard speech corpus, the step-by-step execution of the analysis is figured out as follows:

1. Acoustic feature extraction: The extracted features are selected to analyzed basing on their representation of neutral-Lombard differences and speech intelligibility. Basic acoustic features showing the distinction of Lombard and neutral speech are extracted first (F0, duration, spectral tilts), then the others being recognized as intelligible features are analyzed (Formants, Modulation Spectrum, energy redistribution (over frequency region) of phonemes). The set of techniques including HTK-toolkit (for automatic segmentation before manual correction); STRAIGHT [23] (to extract F0, and spectrum); LPC-based [24] and Spectral-GMM [25](for extracting formants), a derivation from Modulation Filter [26] (to extract Modulation spectrum), Average normalized Spectrum on frequency domain (to concern with energy redistribution).

2. Adaptive tendencies realization, physiological-psychological-acoustical establishment: Values of acoustic parameters are organized in an order of increasing noise level. If the variation exists, it will show with noise level varying accordingly. A specific meaning of psychoacoustics or physiology might be reflected in each kind of the realized variations.

3. Adaptive tendency modeling and deployments by mathematical functions and voice conversion techniques, and their evaluations in contribution to intelligibility: to propose an adaptation model for mimicking Lombard speech over noise level increasing and to construct or apply suitable voice conversion techniques for the synthesis. The adaptive tendencies can be suggested as the adaptation constraints to further optimization or intelligible rule derivation.

## 1.4 Structure of the Thesis

This thesis is organized as follows: Chapter 1 introduces the background and final goal of this study in improving speech intelligibility in noise, afterward, the literature overview among previous research in the field of the acoustical analysis of Lombard speech and the problems and objectives are defined. Chapter 2 describes the Lombard corpus used in the analyses and its preprocessing procedure. Chapter 3, the acoustical analysis are performed: the important tasks of acoustic feature extraction on Lombard speech produced in the noise level varying background, and their analysis results and discussions on adaptable acoustic variations are presented. Chapter 4 is expected to consider the

mathematical modeling of acoustic variations and their evaluations with the help of voice conversion techniques and intelligible measurement metrics. However, at this stage, it is only a discussion on mathematical modeling the adaptive tendencies with voice conversion techniques. Chapter 5 finally summarizes and concludes this thesis with respect to the research question and give a plan for future work.

# Chapter 2

# Lombard Speech Corpus

This chapter gives a description of Lombard speech corpus used in our analysis and their preprocessing tasks. First, we describe the details of how the corpus was created and credit to be a reliable and good dataset. Secondly, the preprocessing of segmentation the utterances into phonetic levels with its accuracy is discussed. The preprocessing step is to prepare segmented data for the analysis phase.

## 2.1 Recording Conditions, Settings, and Participants

Speakers and recording word lists were drawn from the previous study that examined intelligibility of Lombard speech [11]. A male and a female participated in the recording. Three familiarity-controlled word lists [22](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. According to the corpus description provided by the author [11], the detailed information is as follows:

- **Participants:** Participants in the recording are two speakers the same as the spoken ones used for listening tests [11]. They reported having been experienced speaking in noisy environments. At the time of doing the record, the female speaker (N12) was 33 years old, had been a voice training for theater actress. She worked in an amateur theater company. The male speaker (N10) was 21 years old, experienced a swimming instructor in noisy and reverberant environments. Both of them are native Japanese living in Ishikawa, Japan. They have no problems with hearing impairment. Moreover, they were certified to pass a hearing test at six frequencies: 250, 500, 1000, 2000, 4000, and 8000 Hz at a hearing level of 20 dB with two ears separately.

- **Noise:** The noise used as the background was pink noise, which was recorded in a CD "AUDIO TEST CD-1". Two 6000 ms portions were selected from the record.

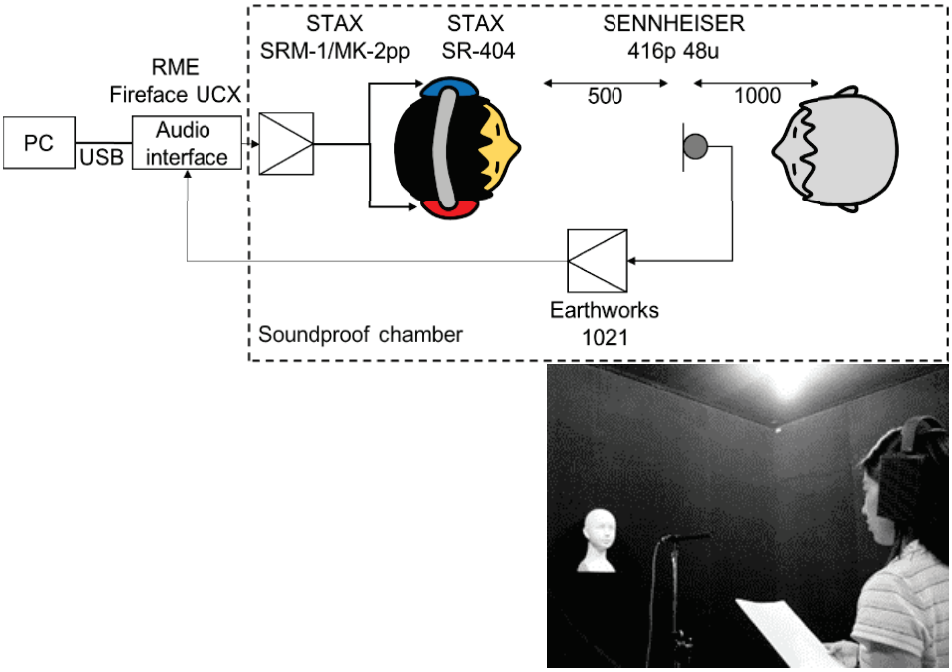



Figure 2.1: Recording configuration [11]

- **Recording conditions:** Six noise levels, without noise: *non* (in this study, it is re-denoted $-\infty$), with noise: 66, 72, 78, 84, and 90. The noise was presented at the ears of the speaker. A-weighted SPL of either of 66, 72, 78, 84 or 90 dB was performed.

- **Instruction:** The speakers were asked to read the sentences as if they were talking with a listener at the distance of 1.5 m from them. These listeners were assumed in the same noisy conditions as the speakers. The list of sentences for each speaker uttered is described in Speech materials portion.

- **Recording Procedure:** The utterances were recorded in a sound-proof booth with a background noise of smaller than A-weighted 21 dB sound pressure level. Figure 2.1 shows the configuration of the speaker, the microphone, and the simulated listener's head. The microphone was setup at 0.5 m from the speaker's mouth. The head model was at the location of 1.5 m from the speaker. Speakers heard the noise by the headphones STAX SRM-1/MK-2pp and STAX SR-404. The speakers were guided in advance to the noise's spread. After the noise presented, speaker started to speak. Utterances were recorded with a SENNHEISER 416p 48u microphone, and amplifier of Earthworks 1021, and audio-interface of RME Fireface UCX. They were digitized in a 44.1 kHz sampling frequency with 16-bit quantization. Each speaker recorded 60 utterances for each noise level randomly. For a list, the recording was fixed in the order as in Table 2.1.

- **Speech materials:** Three familiarity-controlled word lists (Table 2.1) [22](60 four-

mora words - Type of pitch accent pattern) with lowest familiarity rank (1.0-2.5) were used. Each of them was carried in the order: “Tsugi ni yomu tango wa” word “desu” to make a sentence for the speakers to speak. (Figure 2.2) shows the structure of a recording sentence.

Table 2.1: Lists of uttered sentences

| **ID** | **Content** | **Translation ( space-separated mora)** |
|---|---|---|
| 1201 | ササワラ | sa sa wa ra |
| 1202 | ヤママユ | ya ma ma yu |
| 1203 | ロウダン | ro u da N |
| 1204 | ブンダイ | bu N da i |
| 1205 | コヤガケ | ko ya ga ke |
| 1206 | カワホネ | ka wa ho ne |
| 1207 | オカボレ | o ka bo re |
| 1208 | ソトワニ | so to wa ni |
| 1209 | ノマオイ | no ma o i |
| 1210 | アオダチ | a o da chi |
| 1211 | タマヨビ | ta ma yo bi |
| 1212 | ウラドシ | u ra do shi |
| 1213 | ムコガネ | mu ko ga ne |
| 1214 | ゴフショウ | go fu shyo u |
| 1215 | リンガク | ri N ga ku |
| 1216 | ツボガリ | tsu bo ga ri |
| 1217 | トクハツ | to ku ha tsu |
| 1218 | ドウニン | do u ni N |
| 1219 | ボクタク | bo ku ta ku |
| 1220 | ユウモン | yu u mo N |
| 1501 | フウブン | fu u bu N |
| 1502 | トップヤ | to p pu ya |
| 1503 | ブツジョウ | bu tsu jyo u |
| 1504 | コトブレ | ko to bu re |
| 1505 | キュウサン | kyu u sa N |
| 1506 | ゴウフク | go u fu ku |
| 1507 | クネンボ | ku ne N bo |
| 1508 | ガンニン | ga N ni N |
| 1509 | ハンヅラ | ha N zu ra |
| 1510 | ショウモノ | shyo u mo no |
| 1511 | オモガイ | o mo ga i |
| 1512 | ウマノス | u ma no su |
| 1513 | スソワタ | su so wa ta |
| 1514 | シャクブク | shya ku bu ku |
| 1515 | ヨツダマ | yo tsu da ma |
| 1516 | ワルドメ | wa ru do me |

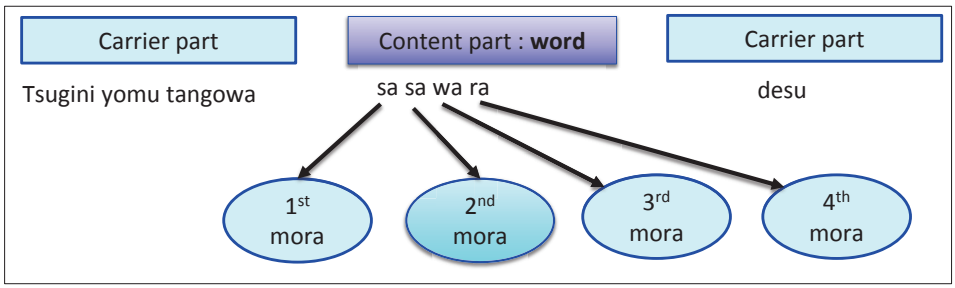



Figure 2.2: Structure of an utterance

Table 2.1: Lists of uttered sentences

| ID | Content | Translation ( space-separated mora) |
|---|---|---|
| 1517 | カワソウ | ka wa so u |
| 1518 | ソラドケ | so ra do ke |
| 1519 | ユウギン | yu u gi N |
| 1520 | モクネン | mo ku ne N |
| 1601 | ビリケン | bi ri ke N |
| 1602 | タキジマ | ta ki ji ma |
| 1603 | マイハダ | ma i ha da |
| 1604 | ミササギ | mi sa sa gi |
| 1605 | シバハラ | shi ba ha ra |
| 1606 | サビアユ | sa bi a yu |
| 1607 | モンサツ | mo N sa tsu |
| 1608 | ザンカン | za N ka N |
| 1609 | ナミバン | na mi ba N |
| 1610 | ダイワレ | da i wa re |
| 1611 | ジリダカ | ji ri da ka |
| 1612 | クチナワ | ku chi na wa |
| 1613 | スミガネ | su mi ga ne |
| 1614 | アミハン | a mi ha N |
| 1615 | ウラジロ | u ra ji ro |
| 1616 | メンキツ | me N ki tsu |
| 1617 | キュウカツ | kyu u ka tsu |
| 1618 | キクバン | ki ku ba N |
| 1619 | ヤキフデ | ya ki fu de |
| 1620 | バンシツ | ba N shi tsu |

Hence, with six different recording conditions, the number of utterances in our Lombard dataset is stated as in Table 2.2.

Table 2.2: Lombard corpus

| Speech | Neutral speech | Lombard speech | | | | |
|---|---|---|---|---|---|---|
| | #non ($-\infty$) dB | #66 dB | #72 dB | #78 dB | #84 dB | #90 dB |
| Female (N12) | 60 | 60 | 60 | 60 | 60 | 60 |
| Male (N10) | 60 | 60 | 60 | 60 | 60 | 60 |

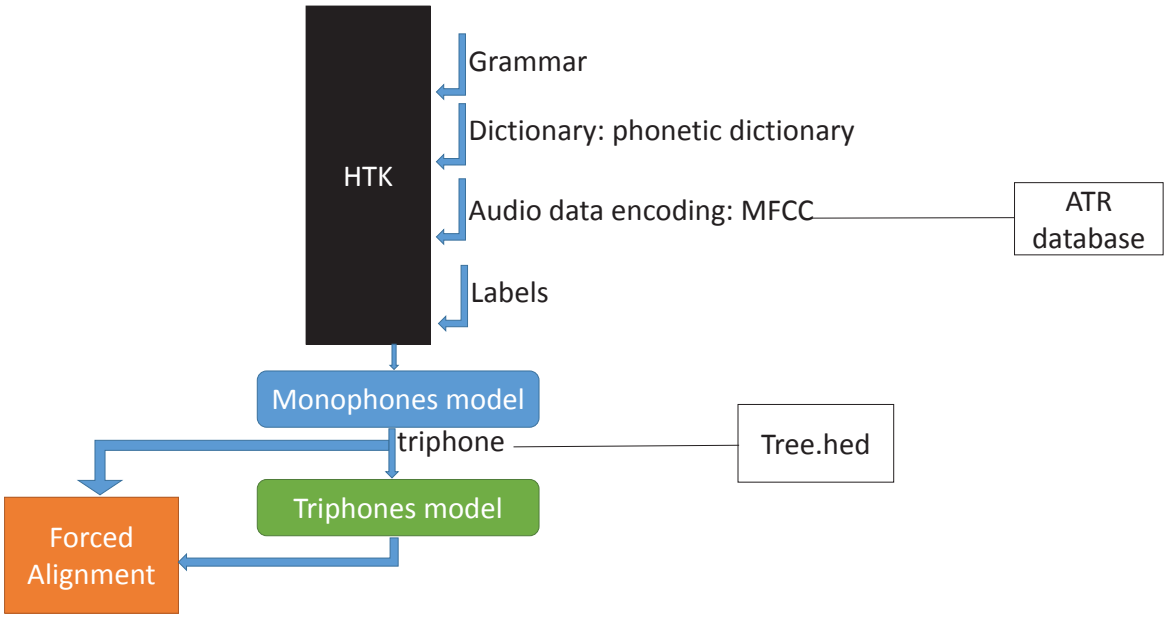



Figure 2.3: Model Training of Forced Alignment Using ATR database [29]

## 2.2 Data Preprocessing

Only target words are objectives of our analyses. Before start acoustically analyzing, each target word was separated from the whole utterance and segmented at phonetic level. It was firstly done with HTK toolkit (Forced Alignment - Figure 2.3). We used ATR database (A set) in training of HTK, and monophones model to algin the phonemes with their audios. Those alignments still contained much of errors (30-40 ms tolerance). Then it was needed to manually correct each segment with the help of Praat basing on the knowledge of its spectrogram, especially formant information, and spectral transition information [27] [28].

## 2.3 Segmentation Results

Figure 2.4 shows an example of segmentation. The detailed number of segmented phonemes are listed in Table 2.3 Each audio was resampled at 16 kHz. Some phonemes for instances

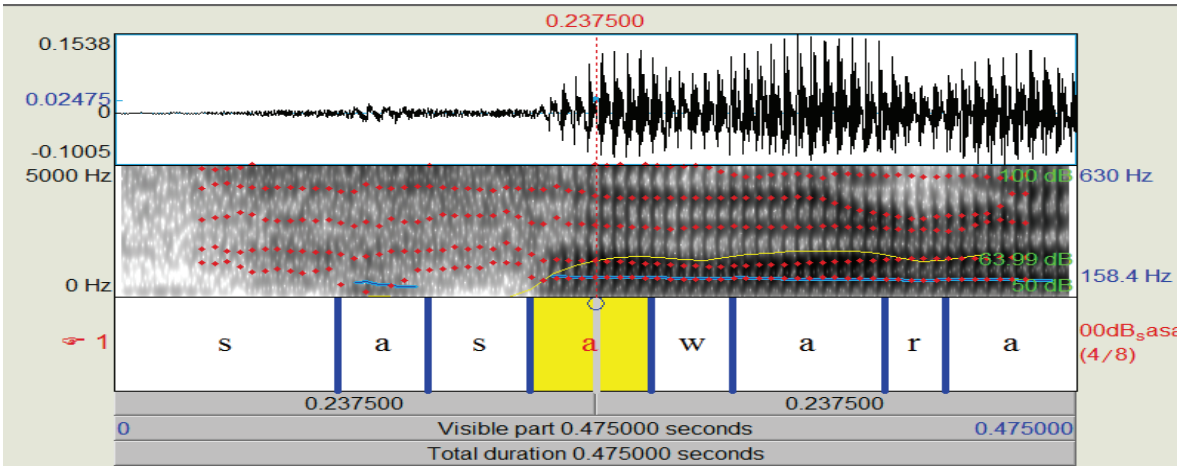



Figure 2.4: Phonemic segmentation by using Praat

Table 2.3: Statistics on vowels and consonant for each speech type over speakers

| Phonemes | /a/ | /i/ | /u/ | /e/ | /o/ | Consonants | Unidentified phonemes (fu. ou, uu) |
|---|---|---|---|---|---|---|---|
| Female (N12) | 73 | 33 | 56 | 14 | 41 | 217 | 0 |
| Male (N10) | 73 | 33 | 37 | 14 | 34 | 194 | 27 |

/u/, /o/ in the combination /fu/, /ou/, /uu/ of the female speaker, still were not segmented.

## 2.4 Discussion

With the procedure as describing above (automatic segmentation, then manual correction by reading spectrogram), the segmented data of Lombard speech corpus can be credited to be a good dataset and used in the analysis. With the number of phonemes, the analysis results somehow can be believed, though the number of vowel /e/ is small a little bit.

# Chapter 3

# Acoustic analysis of Lombard Speech

This section discusses 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, modulation spectrum, and energy redistribution 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, increasing amplitudes of valleys between formants, H1-H2 variation, and lifting modulation spectrum and energy redistribution in specific frequency region on 0-5 kHz. 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.

In details, 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.

## 3.1 Acoustic Feature Extraction

This analysis 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 and the mechanism of redistributing energy over frequency domain were investigated. More-

Table 3.1: 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 [23] |
| Formants | Frequencies, bandwidths, amplitudes, and amplitudes of valleys between formants | LPC, Spectral-GMM based spectra [25] |
| 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.* [26] |
| Energy Redistribution | Distribution over 0-5 kHz | Average normalized Spectrum |

over, 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 3.1 and the explanation below.

- **Duration** From segmented data (already credited with realizability), the duration of vowels and consonants were calculated (Eq. 3.1 and 3.2).

$$Vowel\,Duration = \frac{1}{N} \sum_{v\,is\,a\,vowel} length(v) \tag{3.1}$$

  It is average length in millisecond from all segmented vowels: /a/, /i/, /u/, /e/, /o/, measured for each type of speech, over all utterances within a speaker.

$$Consonant\,Duration = \frac{1}{N} \sum_{v\,is\,a\,consonant} length(v) \tag{3.2}$$

  It is average length in millisecond from segmented consonants and measured for each type of speech, over all utterances within a speaker. All vowels and consonants of the utterances by the male speakers were used to calculated duration. On the female speaker, because the lengths of some phonemes are approximately zeros, the number of used vowels is (210/215 i.e. 98%), and used consonants is (192/194 i.e. 98%).

- **F0**: The F0 contour of each target word was extracted by using STRAIGHT ($V40\_005b$)[23] with frame length 40 ms, frame shift 1 ms and the boundary of F0: 77 Hz - 482 Hz (male), 137 Hz - 634 Hz (female) [30] with high realizability. F0 mean - Mean of F0 contour of a word (Eq. 3.3) 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 (Eq. 3.4 and Figure 3.1). Similarly, F0 slope from F0 at the center of vowel of the $2^{nd}$ mora to F0 at the center of the $4^{th}$ mora is also defined.

$$F0_{mean} = \frac{1}{N} \sum_{u\,is\,a\,word} mean(F0\,contour\,of\,u) \tag{3.3}$$

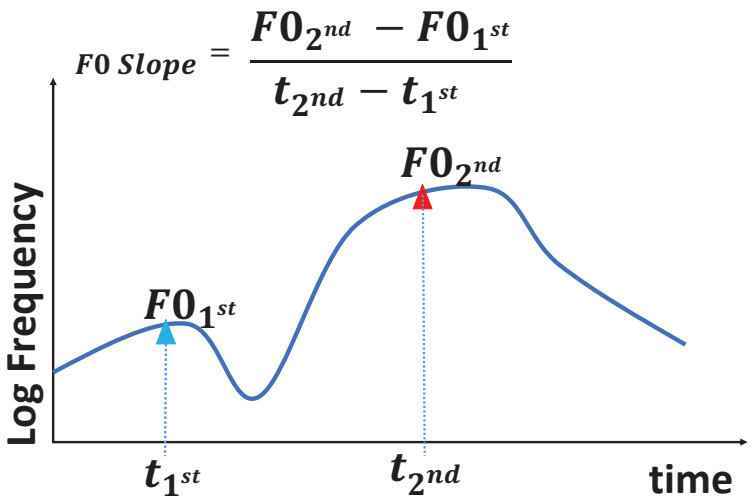



Figure 3.1: F0 slope of a word

where N is the number of words, for each type of speech, N is equal to 60.

$$F0_{slope} = \frac{1}{N}\sum u \frac{F0_{2^{nd}\,mora} - F0_{1_{st\,mora}}}{t_{mid\,point\,2^{nd}\,mora} - t_{mid\,point\,1^{st}\,mora}} \tag{3.4}$$

where $N = 37$ (male) and 42 (female). Only the words with nonzero F0 at the centers of the second and first morae were calculated their slopes.

- **Formants**: F1 and F2 were used to produce vowel space. Formant bandwidths and amplitudes at F1, F2, and F3 were also considered. Moreover, valleys between formants were investigated. In this study, the valleys between 0 Hz and F1, between F1 and F2, and between F2 and F3 were concerned. A hybrid method LPC-based [24] and spectral GMM [25] (Figure 3.2)were applied, and manually correction after applying LPC ( 15%). LPC can estimate the formant frequency quite accurately, but it fails to estimate the formant bandwidth and amplitude precisely. Spectral GMM is different. Its estimation of formant locations is not good. However, if the locations of formants are provided (estimated by LPC), it can estimate their bandwidth and formant amplitude with better accuracy than LPC can do. Therefore, a hybrid method by LPC and spectral GMM were employed. All the number of vowels are extracted formant information. Figure 3.4 presents an example of this feature estimation.

- **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. At different mora position, we have different values of spectral tilts. The identifications of the spectral tilts are figured as in Figure 3.5

  Figure 3.6 shows for results of estimate location and level of harmonics used in the calculation of the spectral tilts. It can be seen that the estimation is highly correct

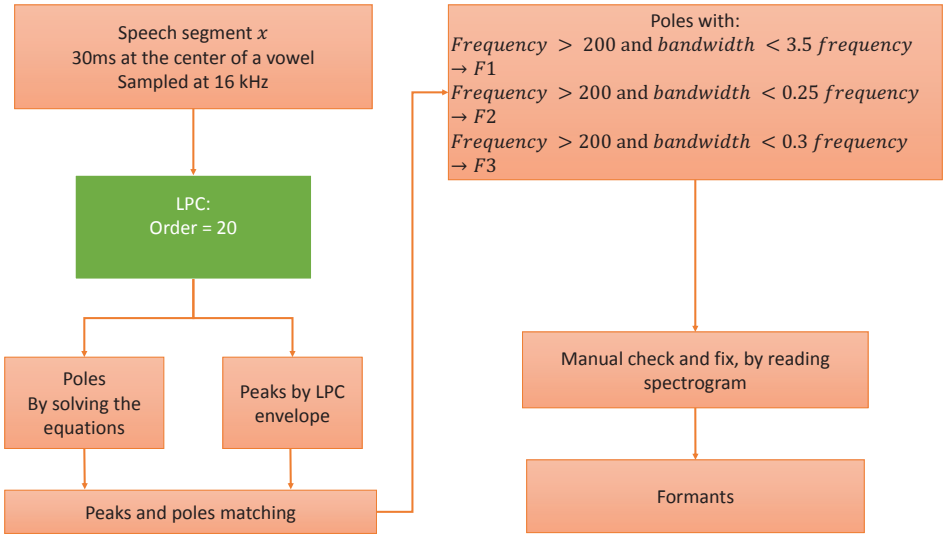



Figure 3.2: Estimation of Formants based on LPC

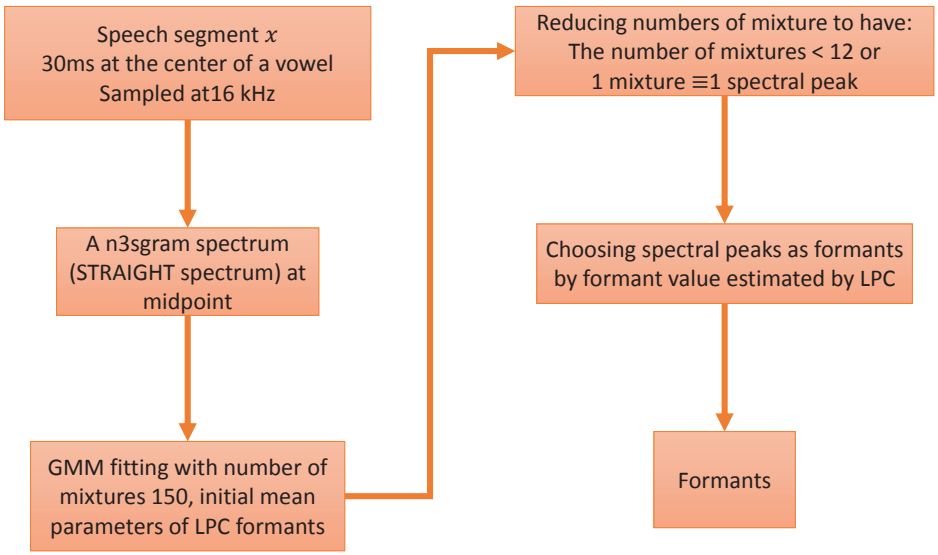



Figure 3.3: Estimation of Formants using the spectral GMM

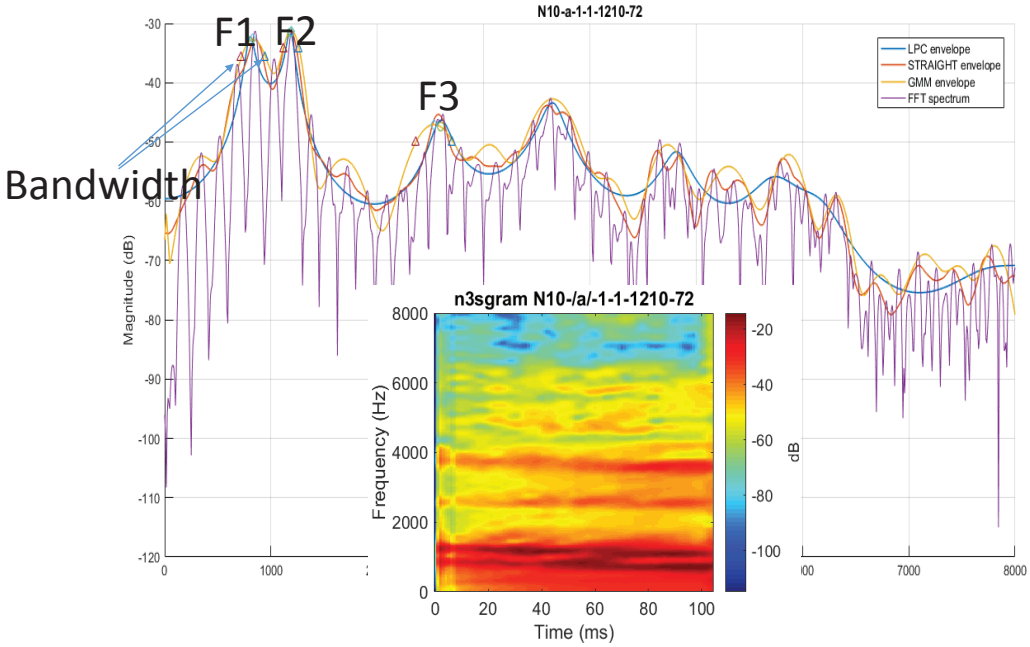



Figure 3.4: Estimation of Formant by using LPC

This representative estimated result (Figure 3.4) shows that we can basically estimate the formant bandwidths though some of them are bigger than the real one a bit. The formant amplitudes can also be used.

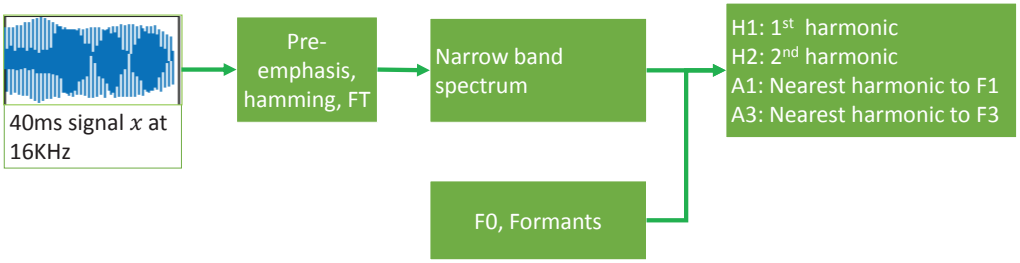



Figure 3.5: Estimation of harmonic locations

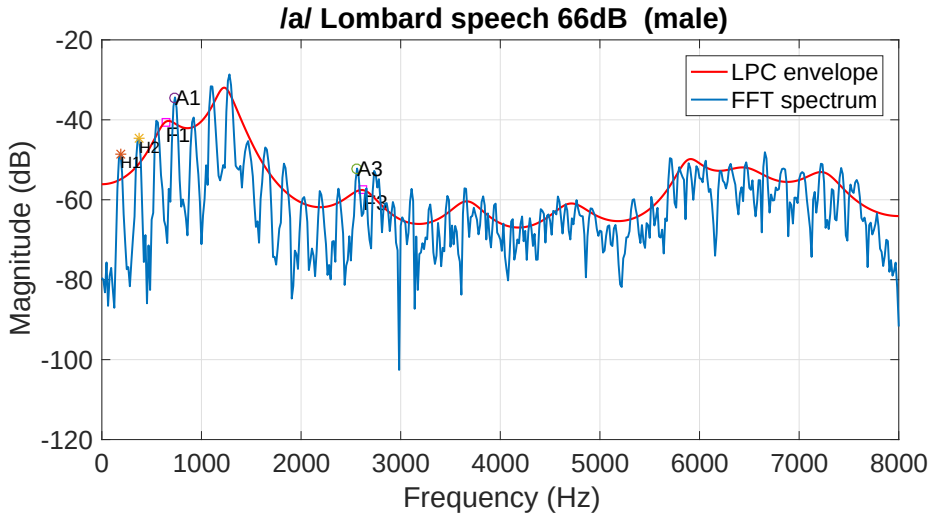



Figure 3.6: Harmonic estimation of vowel /a/ (male) Lombard speech 66 dB

if the location of formants and values of F0 are precise.

- **Modulation Spectrum**: Inspired by Modulation Filter [26], a method which can be used for both analyzing and modifying power envelope of the spectrum extracted by STRAIGHT [23] 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.

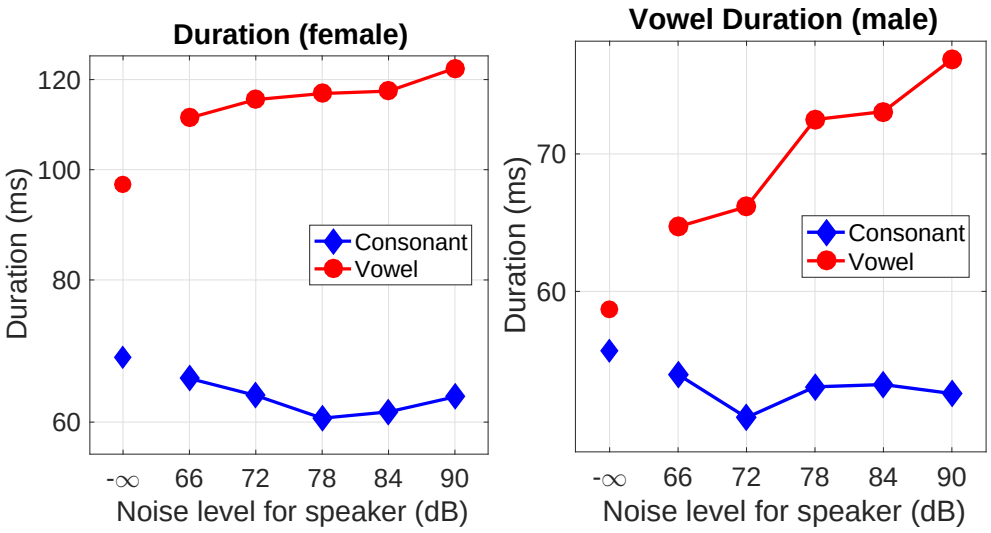
- **Energy redistribution**: For each phoneme, the frequency spectrum at the center was extracted, then normalized by its total energy. The average normalized spectrum was obtained by taking the average of all the normalized frequency spectrum of a phoneme.

## 3.2 Analysis Results

The values of acoustic parameters are observed under noise level increasing. The acoustic features: vowel duration, fundamental frequency, formant F1, formant amplitudes, spectral tilts, modulation spectrum, energy redistribution showing their variations with noise levels correspondingly were considered as follows.

### 3.2.1 Duration

Figure 3.7 shows that with noise level increasing, vowel duration is continuously lengthened for both male and female speakers (increasing about 5-10 ms on every transition of the studied noise levels). The consonant duration seems decreased a bit, can be considered unchanged.




(a) The female speaker

(b) The male speaker

Figure 3.7: Consonant and Vowel Duration

### 3.2.2 Fundamental Frequency (F0)

- **F0 mean**

  With noise level increasing, F0 mean is increased continuously (increasing about 20-30 Hz on every transition of the studied noise levels) or changed abruptly at 84 dB. The percentages of the continuity for the male and female speaker is 35% and 67% respectively. 65% in male, 33% in the female are the amount of abruptness occurs among their utterances. Figures 3.8(a) and 3.8(b) clearly demonstrate for this result.

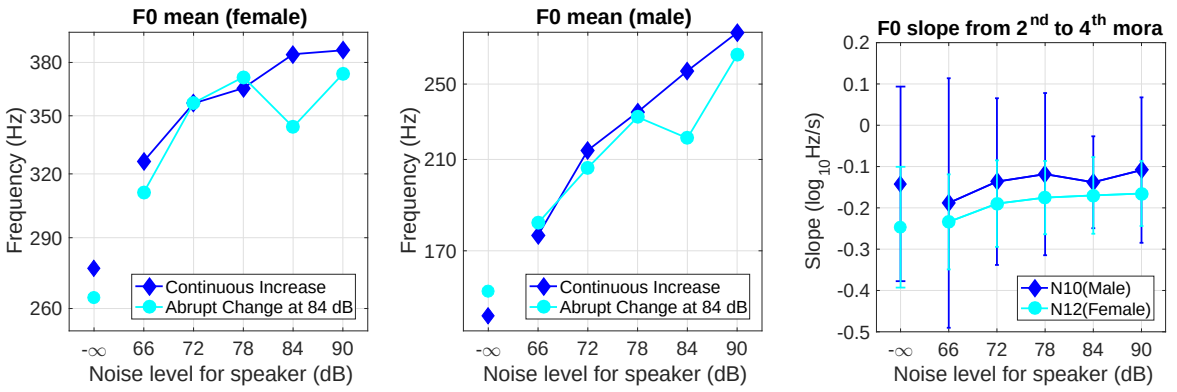
- **F0 slopes**

  F0 slope from the $1^{st}$ to $2^{nd}$ morae is seen mostly preserved among Lombard speech, so its result is not presented. Otherwise, the F0 slope from the $2^{nd}$ to $4^{th}$ morae (Figure 3.8(c)) is continuously increased with noise level increasing (clearly from 66 to 90 dB). The increasing slope is about $0.01 log_{10} Hz/s$ on every transition of the studied noise levels.

### 3.2.3 Formants

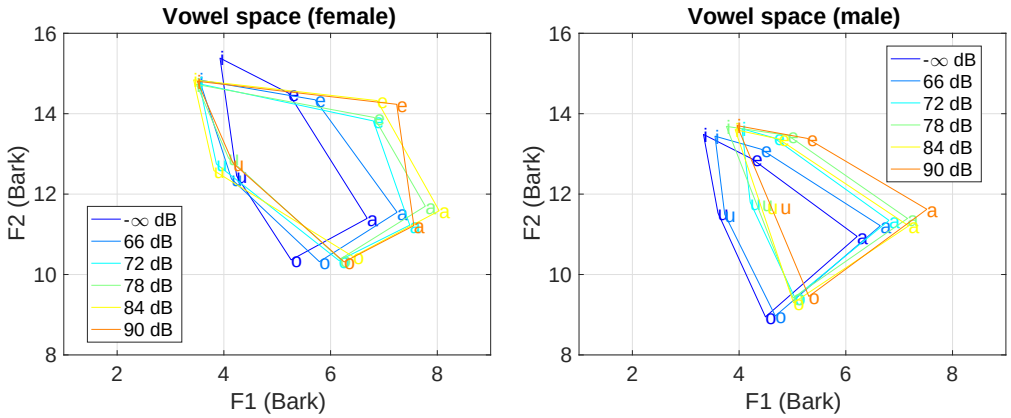
Formants were studied in terms of formant frequencies, bandwidths, and amplitudes. The adaptive variation only could be seen on formant frequency F1, formant amplitudes and amplitudes of valleys between formants. The details are presented in the following figures.

- **Formant Frequencies - Vowel space**




(a) F0 mean (female) (b) F0 mean (male) (c) F0 slope from $2^{nd}$ to $4^{th}$ morae

Figure 3.8: Fundamental frequency



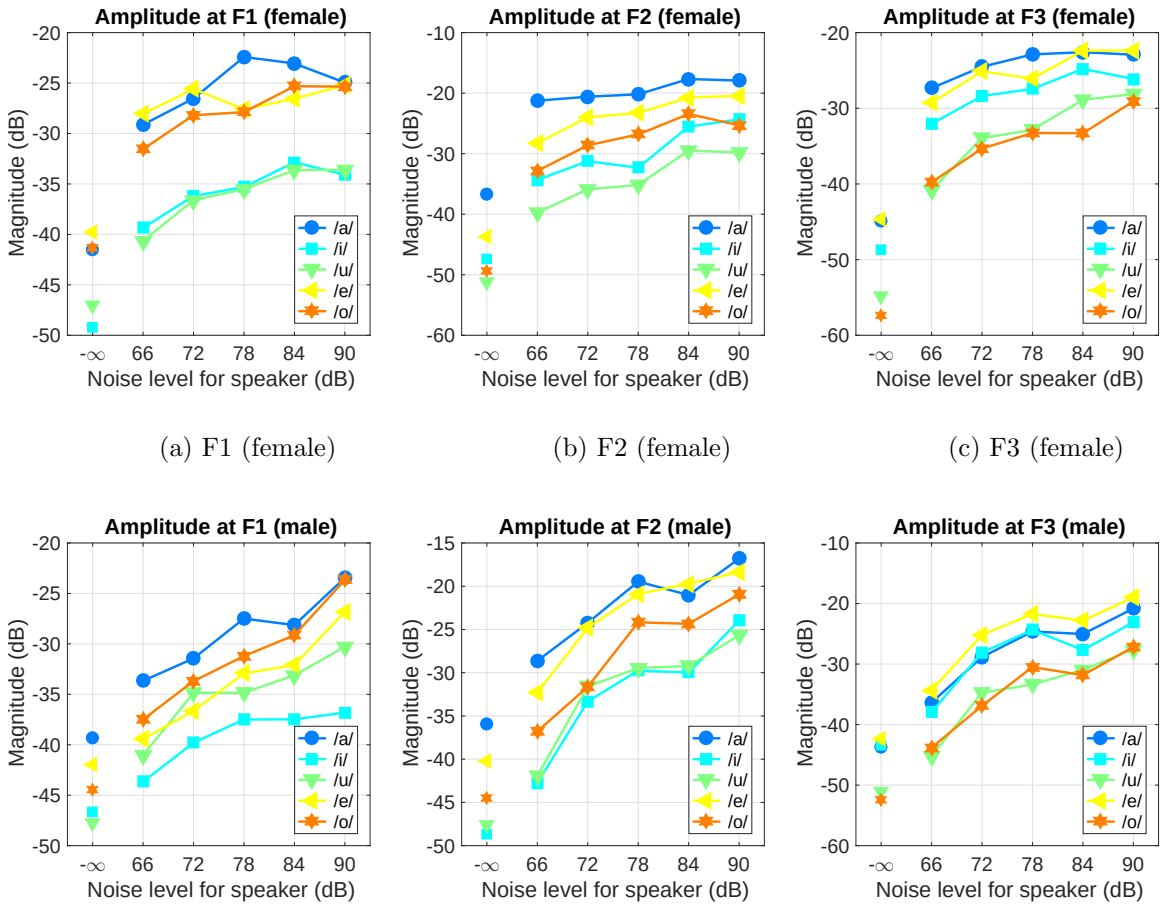

(a) The female speaker (b) The male speaker

Figure 3.9: Vowel space

With noise level increasing, shifting F1 can be seen: /a/, /e/, /o/ forward and /i/, /u/ backward (Figures 3.9(a)) and all vowels forward (Figure 3.9(b))

- **Formant Amplitudes: at F1, F2, and F3**

  From Figure 3.10, all formant amplitudes at F1, F2, and F3 are increased with noise level increasing. The levels of increasing are different among vowels and speakers. From neutral to 66, up to 78 dB, the increasing level is about 5-10 dB for each transition. From 78 to 90 dB, the increasing is constant, it is shown with less than 5 dB. Otherwise, it is also presented in another phenomenon. From 78 to 84 dB, it can be decreased (the same as the abruptness is seen on F0). Then, from 84 to 90 dB, the increasing level of around 5 dB is obtained again. F3 and F2 can be seen




(a) F1 (female) (b) F2 (female) (c) F3 (female)

(d) F1 (male) (e) F2 (male) (f) F3 (male)

Figure 3.10: Formant Amplitudes

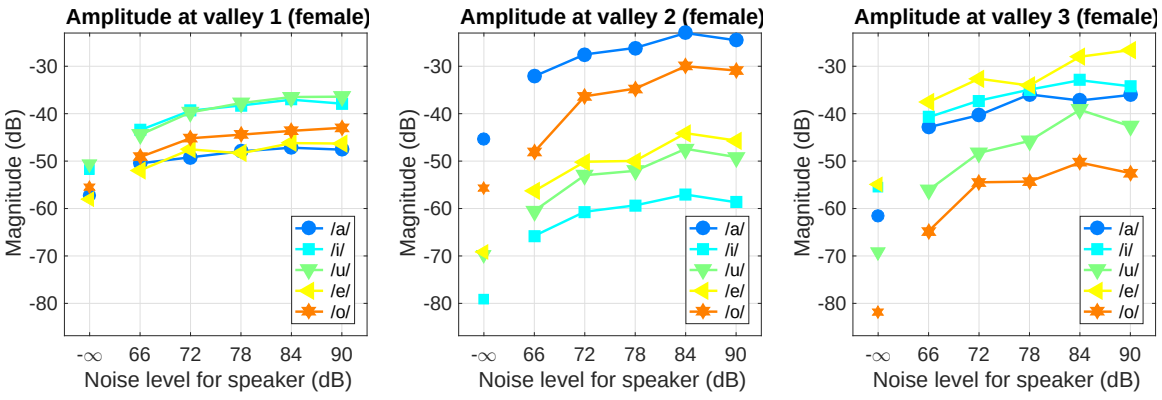
to be increased more than F1 can be.

- **Valleys between formants**

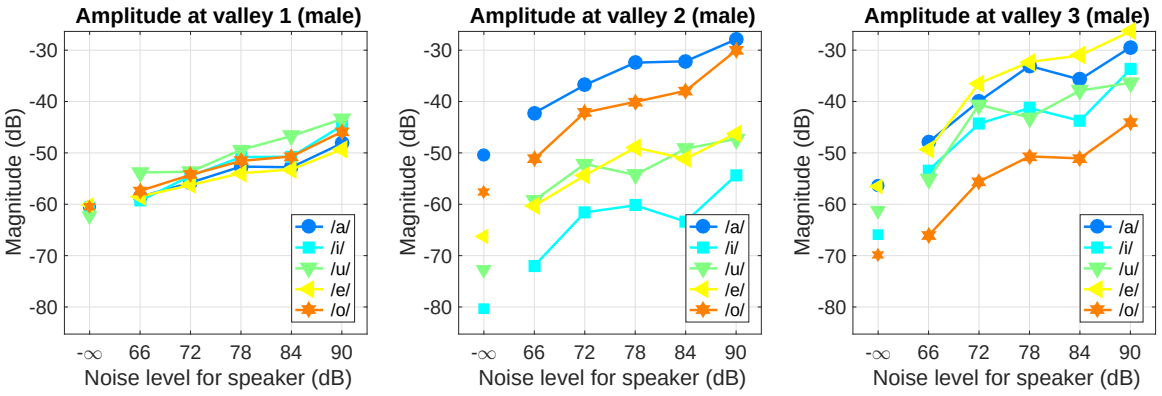
  In Figure 3.11, all amplitudes valleys between 0 Hz and F1 ( called valley 1), between F1 and F2 (called valley 2), and between F2 and F3 (called valley 3) are increased with noise level increasing. The level and ratio of increasing of the valleys 2 and 3 are much higher than the valley 1.

- **Amplitude ratios at valleys between formants and formants**

  Figures 3.12 and 3.13 show comparisons with formant amplitudes, ratios of amplitude increasing of formant amplitude at F1, F2, F3, and valley 1, 2, and 3 from the neutral speech (Ratio is equal to 1 for the neutral) among vowels and speakers. All the ratios are increasing with noise level increasing. Importantly, among all vowels and speakers, it is seen that ratios of increasing amplitude at the valley 2 and 3

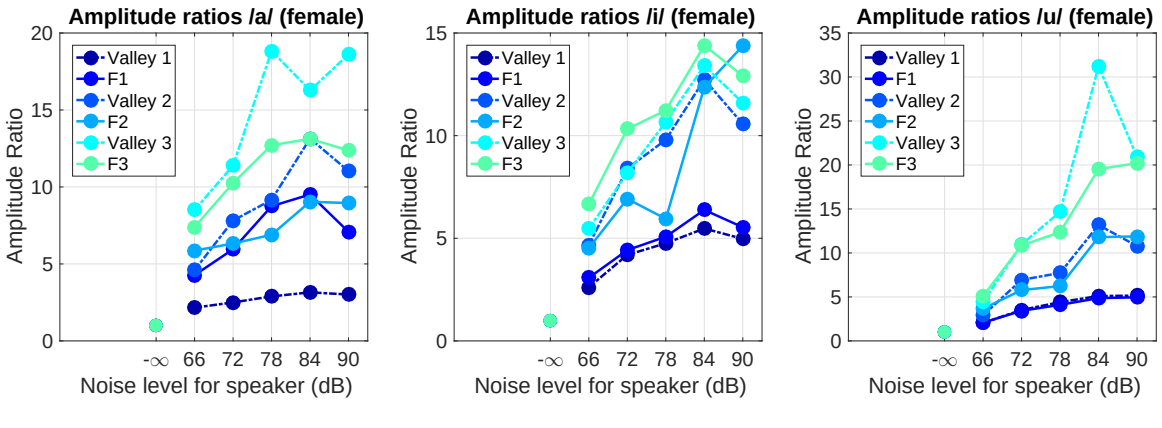



(a) Valley 1 (b) Valley 2 (c) Valley 3

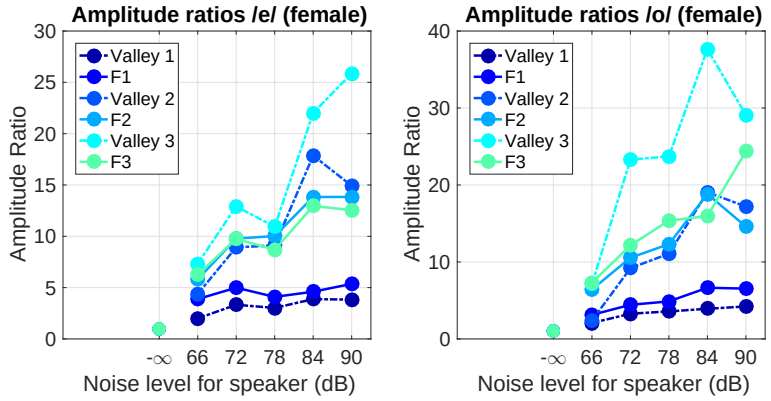



(d) Valley 1 (e) Valley 2 (f) Valley 3

Figure 3.11: Amplitudes of valleys between formants




(a) /a/ (b) /i/ (c) /u/




(d) /e/ (e) /o/

Figure 3.12: Ratios of amplitudes of valleys between formants and amplitudes of formants (female)

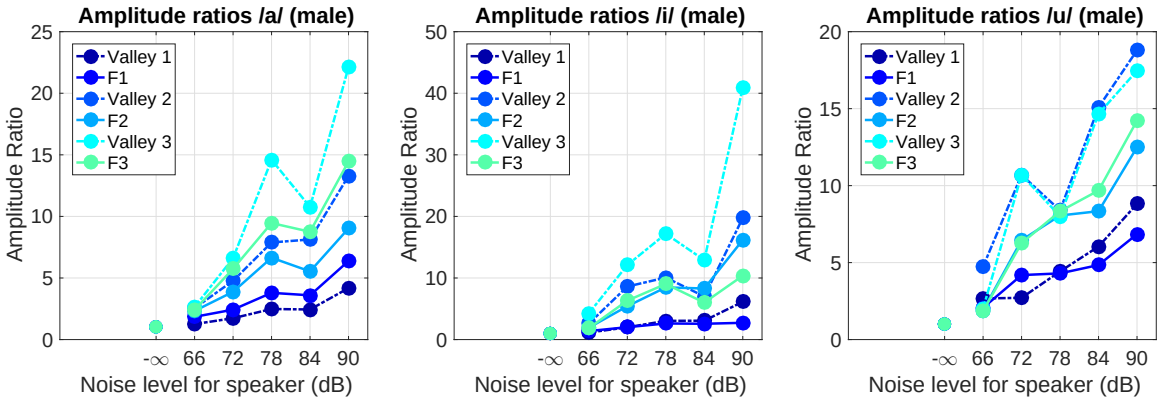
are much greater than or equal the ratios of increasing amplitude at F2 and F3. The ratio of increasing amplitude at the valley 1 is less than the ratio of increasing amplitude at F1.

### 3.2.4 Spectral Tilts

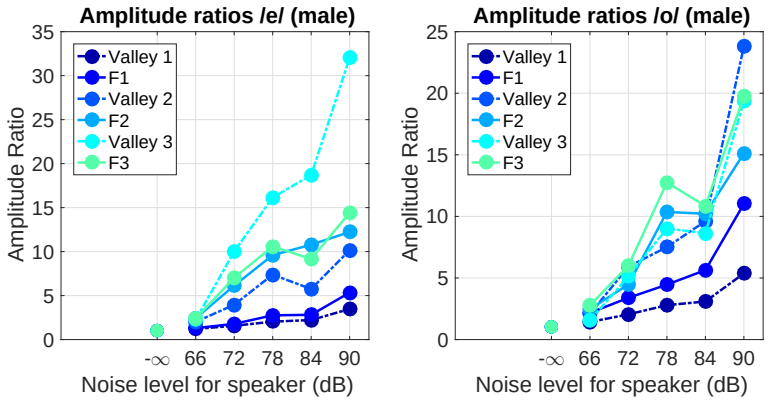
H1-H2 and A1-A3 were analyzed on 5 vowels with increasing noise level. The results of A1-A3 is similar for all the vowels crossing the speakers. H1-H2 showed in two groups of variation tendencies among vowels within a speaker. The demonstrations are figured as follows.

- **H1-H2**

  Figure 3.14 indicates that biasing H1-H2 can be seen by decreasing in /i/, /u/ and

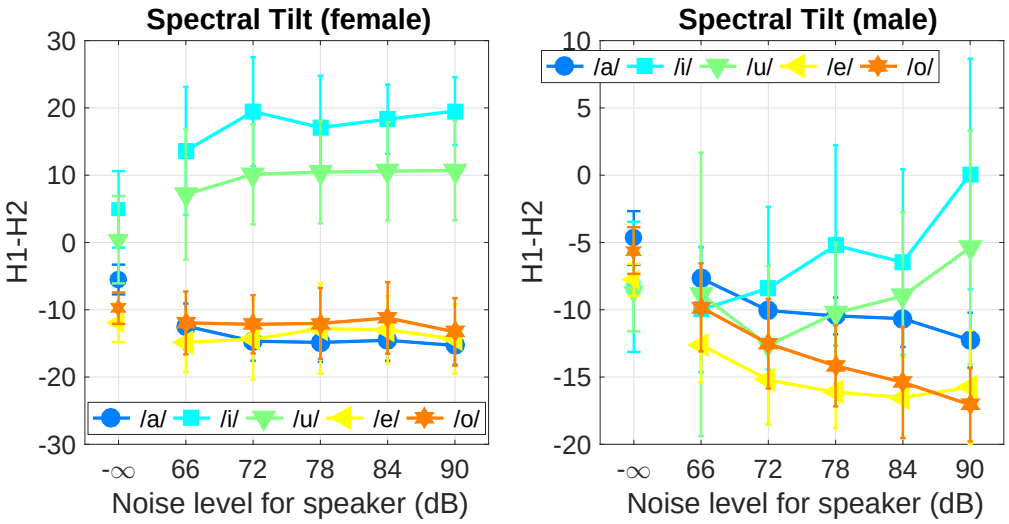



(a) /a/ (b) /i/ (c) /u/




(d) /e/ (e) /o/

Figure 3.13: Ratios of amplitudes of valleys between formants and amplitudes of formants (male)




(a) The female speaker

(b) The male speaker

Figure 3.14: Spectral tilt H1-H2

increasing in /a/, /e/, /o/ with increasing noise level. The division is clearer in the female speaker than in the male speaker. The level of tilting from neutral to 90 dB is about 20 dB. 5, or 10 dB or less is for each transition.

- **A1-A3**

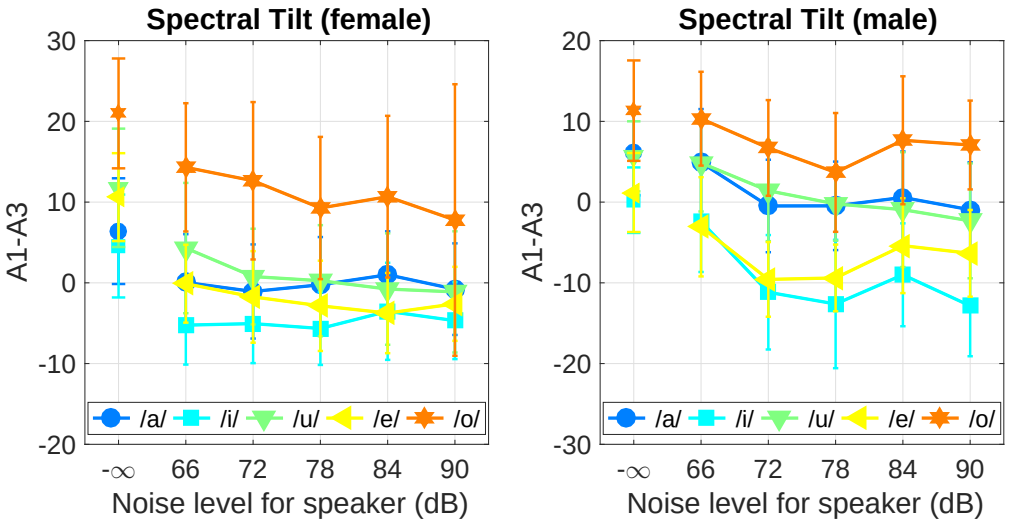
  With increasing noise levels, decreasing A1-A3 is observed in both speakers among all vowels (Figure 3.15). The decreasing tilts are continuous with the level of tilting of 10-15 dB from neutral to Lombard 90 dB.

### 3.2.5 Modulation Spectrum

With noise level increasing, lifting in modulation spectrum from 16 Hz - 128 Hz modulation frequency and below 1000 Hz acoustic frequency are presented (Figures 3.16 and 3.17). For the female speaker, it can also be seen in 2 kHz acoustic frequency. The higher noise level is, the stronger the lifting is.

### 3.2.6 Energy redistribution

Energy is seen to be redistributed over frequency region for vowels (Figures 3.18, 3.19 , 3.20, and 3.21) and vowel-like consonants (b, d, g, m, n, r, w, and y): decreasing in below 1 kHz, increasing in above 1 kHz, clearly in 2 - 5 kHz.




(a) The female speaker

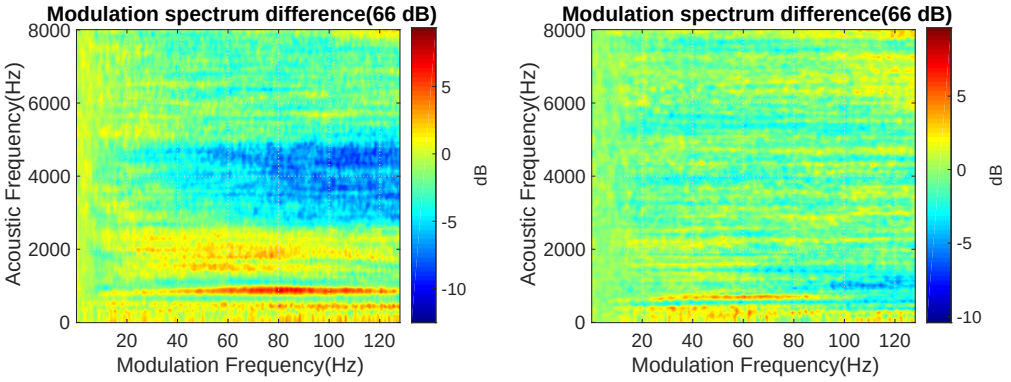
(b) The male speaker

Figure 3.15: Spectral tilt A1-A3

## 3.3 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 [10][13] of the neutral-Lombard distinction are still preserved among noise-level varying Lombard speech. On F0 mean increasing and abrupt changing at 84 dB (perhaps, 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. F0 slope from the $2^{nd}$ to $4th$ morae increasing might present the increasing effort of maintaining the speech during its uttering. Though Japanese are non-tonal language, it can be considered a similar pattern of the exaggeration at the end of F0 contour in Thai language [17]. 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. The amplitudes of valleys between formants increasing might show that the meaningful region is broadened, not only the peaks (formant amplitudes). Moreover, the ratios of increasing amplitude at the valley 2 and 3 are much greater than or equal the ratios of increasing amplitude at F2 and F3. The ratio of increasing amplitude at the valley 1 is less than the ratio of increasing amplitude at F1. They perhaps indicate that the meaningful region focuses on the region of F2, F3, valley 2 and 3. Those patterns might illustrate that more energy at vowels is expected.

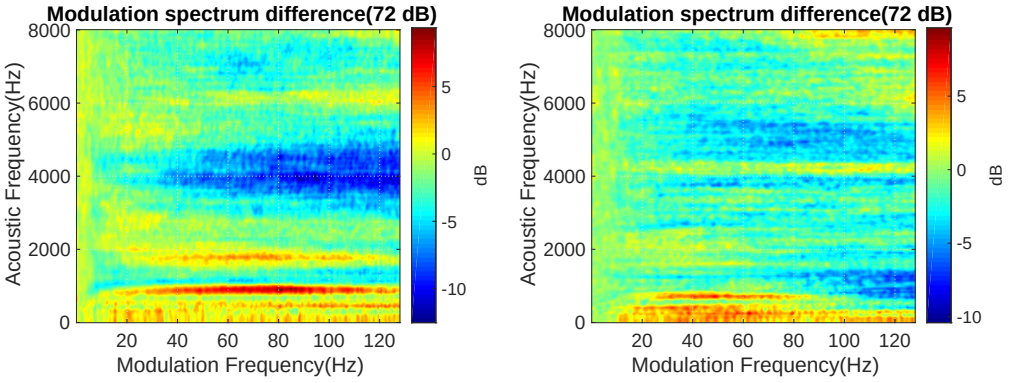
Besides, in previous study [13], spectral tilts are decreased for Lombard speech. However, in our results: H1 - H2 decreases in /i/, /u/ , increases in /a/, /e/, /o/ and A1




(a) The female speaker (66 dB)



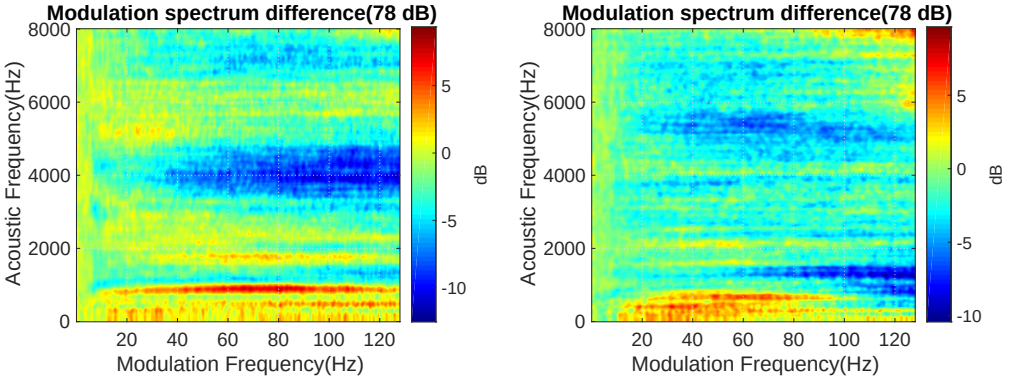

(b) The male speaker (66 dB)




(c) The female speaker (72 dB)




(d) The male speaker (72 dB)

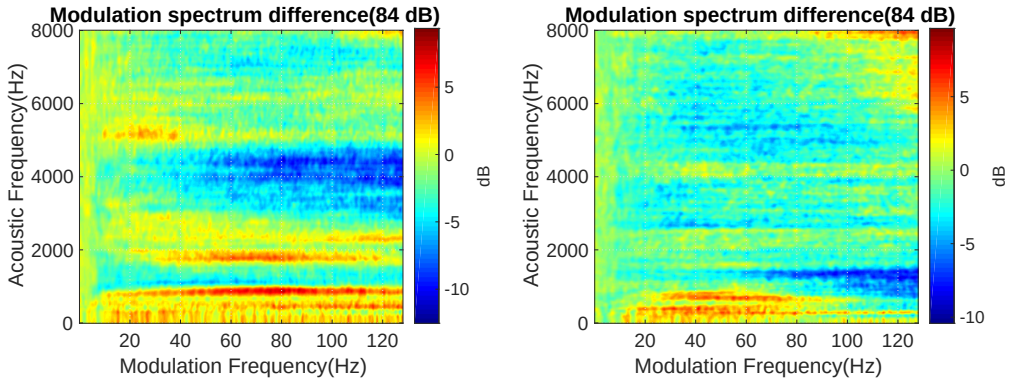



(e) The female speaker (78 dB)

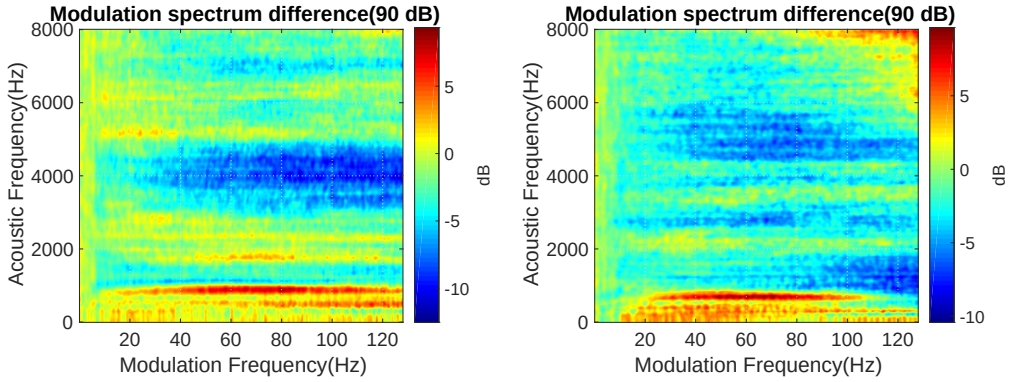



(f) The male speaker (78 dB)

Figure 3.16: Modulation spectral difference (Lombard 66-78 dB and Neutral speech)

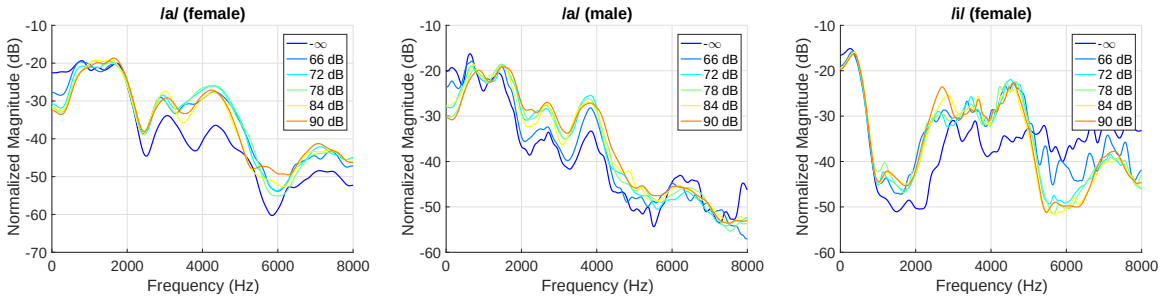



(a) The female speaker (84 dB) (b) The male speaker (84 dB)



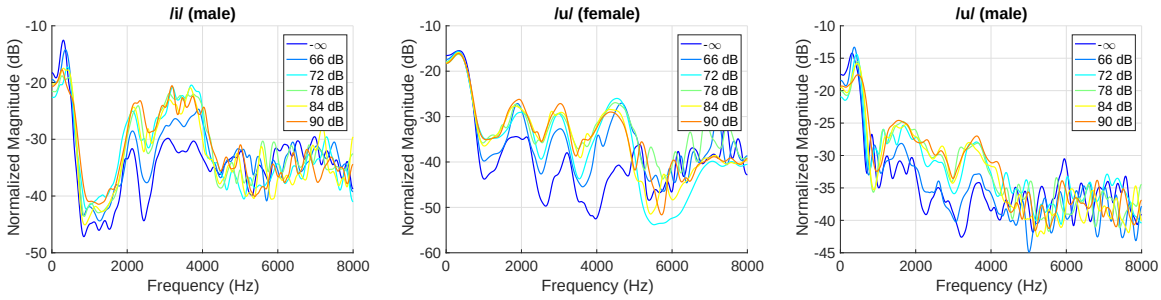

(c) The female speaker (90 dB) (d) The male speaker (90 dB)

Figure 3.17: Modulation spectral difference (Lombard 84-90 dB and Neutral speech)



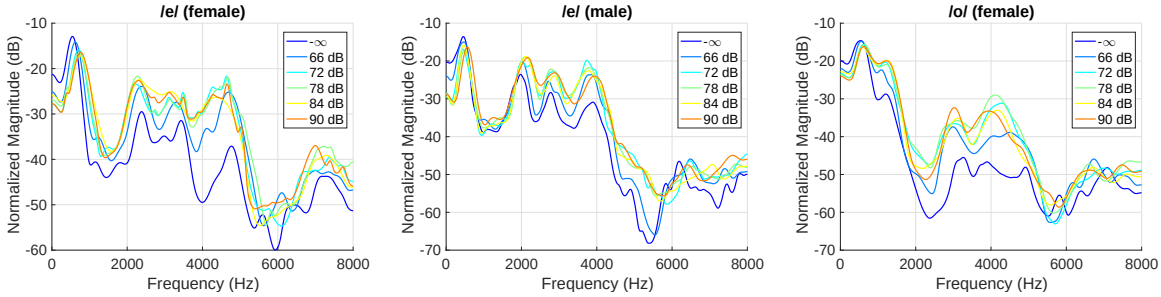

(a) /a/ (female) (b) /a/ (male) (c) /i/ (female)

Figure 3.18: Energy redistribution /a/ (both genders), /i/(female) on 0-5 kHz



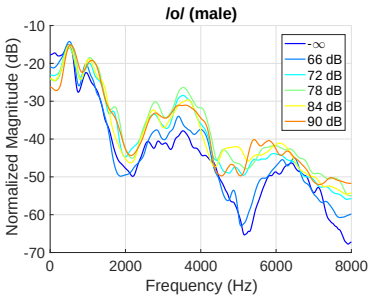

(a) /i/ (male) (b) /u/ (female) (c) /u/ (male)

Figure 3.19: Energy redistribution /i/ (male), /u/(both genders) on 0-5 kHz




(a) /e/ (female) (b) /e/ (male) (c) /o/ (female)

Figure 3.20: Energy redistribution /e/ (both genders), /o/(female) on 0-5 kHz




(a) /o/ (male)

Figure 3.21: Energy redistribution /o/ (male) on 0-5 kHz

- 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 the 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. The redistribution of energy on 0-5 kHz reflects the correspondence with decreasing spectral tilts. Moreover, those acoustic parameters can be seen continuously varying with noise level increasing.

## 3.4 Conclusion

In this chapter, we have completed the analysis of the planned acoustic features including duration, F0, formants, spectral tilts, modulation spectrum, and energy redistribution on the neutral and various Lombard speech. The set of the analyzed acoustic features covers most of the aspect of the neutral-Lombard differences and intelligible characteristics. The analysis results were produced on the credit dataset with two genders various utterances and at the phonetic level. They confirm 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. Additionally, new tendencies have been found: abrupt changing in F0 at 84 dB, increasing formant amplitudes, increasing amplitudes of valleys between formants, H1-H2 variation, lifting modulation spectrum and energy redistribution among specific frequency regions (0-1 kHz and 2-5 kHz) over 0-5 kHz. 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.

# Chapter 4

# Lombard Mimicking

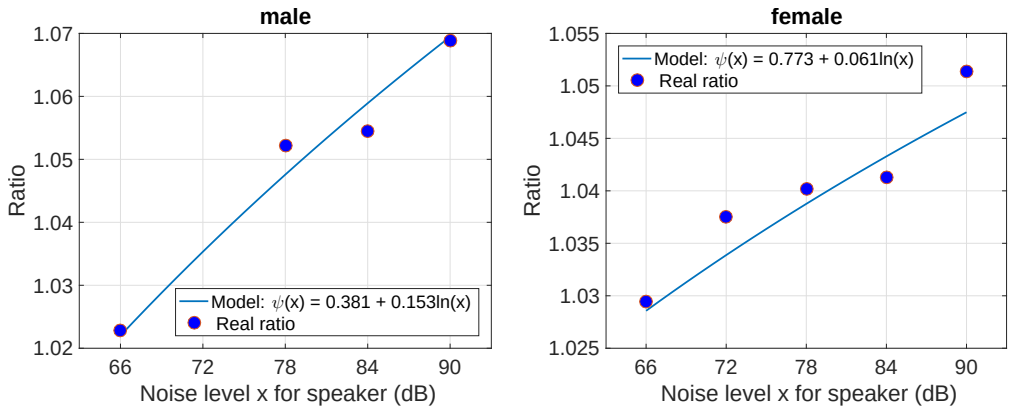
So far, by analyzing Lombard speech produced in the various noise-level background, significant feature variations corresponding with noise level increasing have been extracted. In this chapter, the final purpose is to verify the validity of those tendencies and their ability to adapt with noise-level varying on resynthesized speech. The following content presents a preliminary investigation on adaptive modeling and feasibility of applying voice conversion for obtaining the mimicking Lombard speech from the neutral speech.

## 4.1 A discussion on Mathematical Modeling for Adaptive Tendencies and Voice Conversion

The average values of acoustics features are modeled by a set of mathematical functions. These functions take noise levels as input variables to produce acoustical parameter values. The functions are selected to characterize to these variations studied in the analysis section above. Under the condition of noisy environments, the tendencies can be reasoned to nonlinearly changing with noise level increasing. We firstly consider a logarithmic based model (Eq. 4.1) of dependence between log values of acoustic parameters and log values of the noise level. The values produced by the model are further used in voice conversion to synthesize the mimicking Lombard speech. The following content of this section describes how the mathematical model represents for each feature and contribute to conversion procedures.

$$\psi(x) = k + a \times ln(x) \tag{4.1}$$

In the equation, $\psi(x)$ represents the acoustic values in log scale. The variable $x$ is noise level which is calculated in dB. And, $k$, $a$ are two characteristic coefficients. For converting, we intended to apply some common technique and tools: STRAIGHT [23] for extracting and synthesizing, Temporal Decomposition (TD), Spectral-GMM, a slope filter to spectral tilt.




(a) The female speaker

(b) The male speaker

Figure 4.1: Modeling Vowel Duration

### 4.1.1 Duration

The purpose is to convert neutral speech - the speech is spoken in quiet into a noise-level specific Lombard speech. Therefore, the relative values are taken. In equation 4.1, $\psi$ is the ratio are also kept between vowel durations of Lombard speech and neutral speech. From the analysis results, vowel duration of the male and female speakers are modeled as in the equations in Figures 4.1(b) and 4.1(a). The conversion of duration is lengthening vowel duration according to the ratio by the model. The 40 ms transition from a consonant to a vowel are kept unmodified. The consonants are also kept unchanged. A time trajectory is mapping between the previous length and a new length of the modified speech. One example of duration-converted speech is shown in Figure 4.2.

### 4.1.2 Fundamental Frequency

Fundamental frequency is one of the important properties of Lombard speech. Modeling of fundamental frequency with sufficient parameters is to characterize for F0 contours. The mimicking F0 contour of Lombard speech needs to be re-generated by the model. To create a suitable F0 contour for the synthesized speech, Fujisaki model [31] is considered. The Fujisaki model is known that it can obtain natural sound when modifying F0 by prosodic rules. In Figures 4.3 and 4.4, three main parameters: F0 baseline (Fb), phrase command (T0, Ap), and accent command (T1, T2, Aa) of Fujisaki model can be closely related to some acoustic properties of F0 contour. Fb can be handled by F0 mean. T0 and Ap are dependent on F0 slopes, and F0 highest value. Otherwise, Aa, T1, T2 are strongly connected with F0 mean, F0 highest values and F0 slopes. Therefore by an optimizing these properties: F0 mean, F0 highest value, F0 slopes (from $1^{st}$ to $2^{nd}$ mora, from $2^{nd}$ to $4^{th}$ mora) on the estimation of F0 contour of the neutral speech by Fujisaki model (an automatic extraction were used [32]), the mimicking F0 contour can be produced. The process is to minimize the mean square error in terms of these properties between

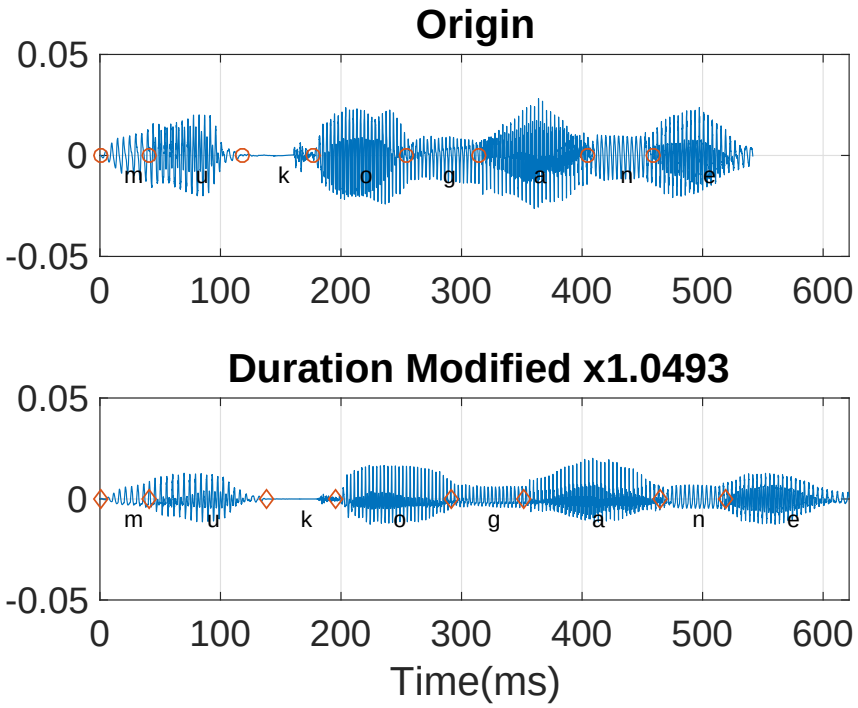



Figure 4.2: Duration Control /mukogane/ at 90 dB (Female)

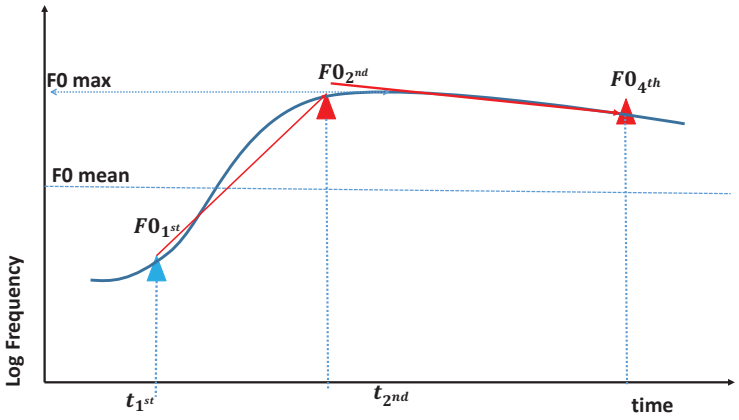



Figure 4.3: F0 contour type of the dataset

an equivalent extracted properties of F0 contour generated Fujisaki model after every iteration and the designed F0 contour. Similarly, the mathematical models of F0 mean, F0 max, F0 slopes are obtained by following the function 4.1. Figure 4.5 shows that if the adaptive model works well, the optimized F0 contour will fit quite well with F0 contour of Lombard speech.

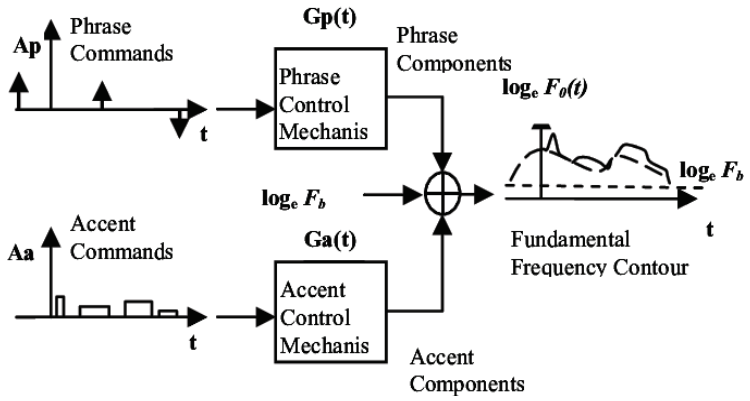



Figure 4.4: Construction F0 contour by Fujisaki model (Fujisaki *et al.* [31])

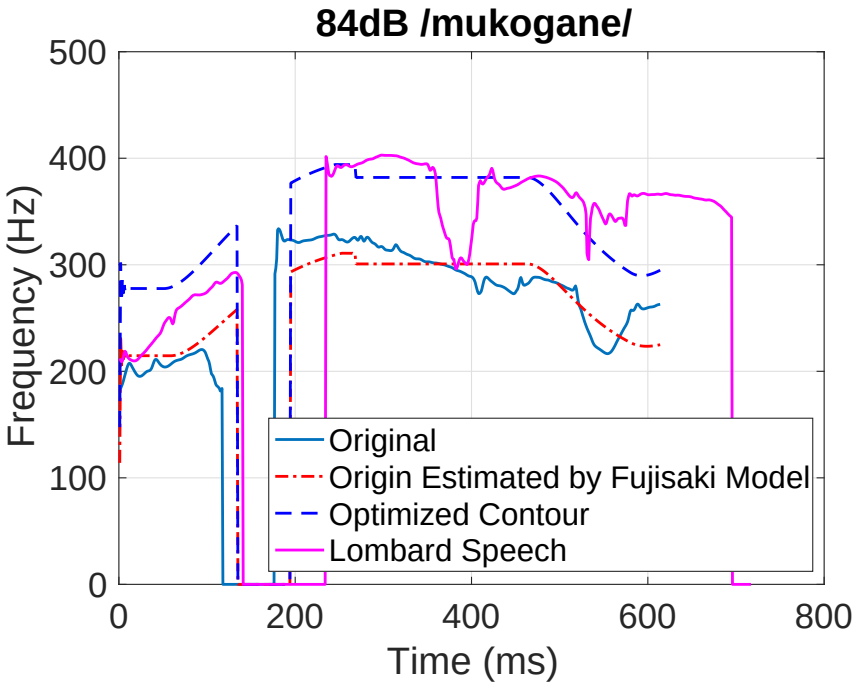
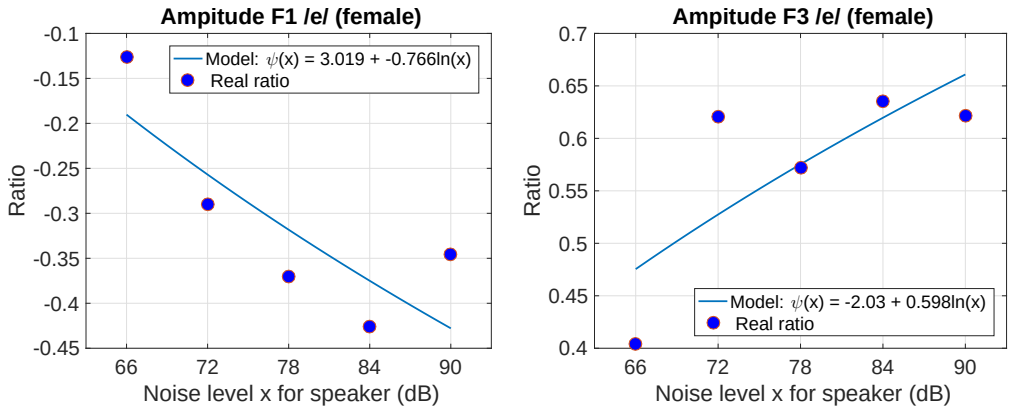



Figure 4.5: Optimization of F0 contour based on acoustic parameters




(a) F1 (b) F3

Figure 4.6: Modeling formant amplitude (Female)

### 4.1.3 Formants

F1 shifting, Formant amplitude increasing can be important for producing the louder sound. In the analysis section, the results show the scale up of formant amplitudes and formant shifting. The scale values of formant amplitude and the ratio of formant shifting between Lombard speech and neutral are modeled. Firstly, the formants of 5 basic vowels are noticed. The model is still strictly followed the equation 4.1 The spectrum of STRAIGHT is modified by applying spectral GMM and TD [25]. With spectral GMM, formant amplitudes and frequencies can be modified independently. Restricted temporal decomposition is applied to expect to gain a better voice quality after voice conversion. Although STRAIGHT allows 600 % of spectral modification, the analysis results of formant amplitudes can not be used directly for the voice conversion because of their larger ratios. Another analysis on normalized utterances is considered before concerning with spectral modification (Figure 4.6). On the normalized data, the modification occurs more slightly, help to avoid over converted problems. Figure 4.7 is an example to control formant frequencies and amplitudes by using Spectral GMM and TD.

### 4.1.4 Spectral Tilts

Spectral tilt A1-A3 and H1-H2 represents the properties of energy redistribution in the frequency domain. In the equation 4.1, $\psi(x)$ is standing for the level difference between the spectral tilt of Lombard speech and neutral speech. The modification of A1-A3 and H1-H2 is conducted after formant modification. Much of energy redistribution is expected by changing the spectral tilt. Its modification is made to affect the frequency region from H1 to 5 kHz by the calculated value of tilting (dB/oct) either from H1-H2 or A1-A3. The modifications based on H1-H2 or A1-A3 are both estimated with the vowels /a/, /e/, and /o/. For vowel /i/, and /u/, the modification based on H1-H2 seems unreliable, therefore

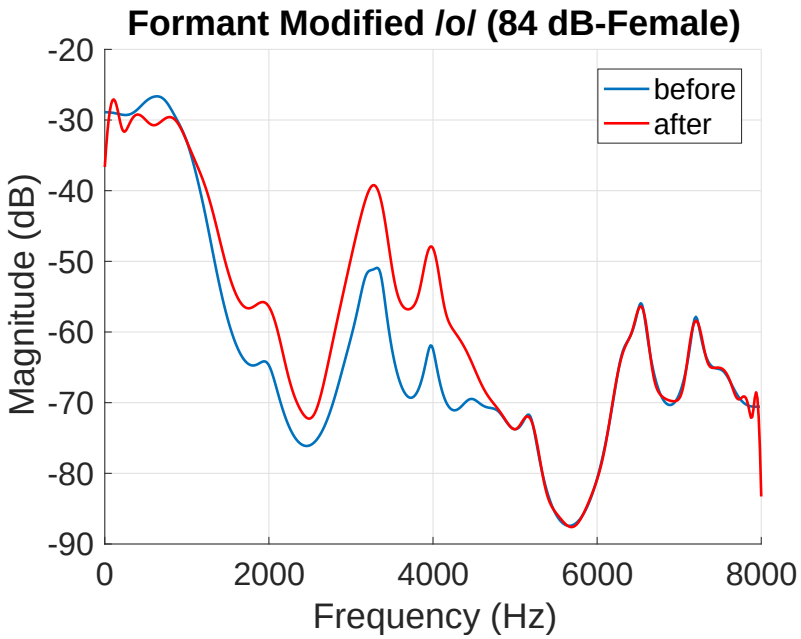
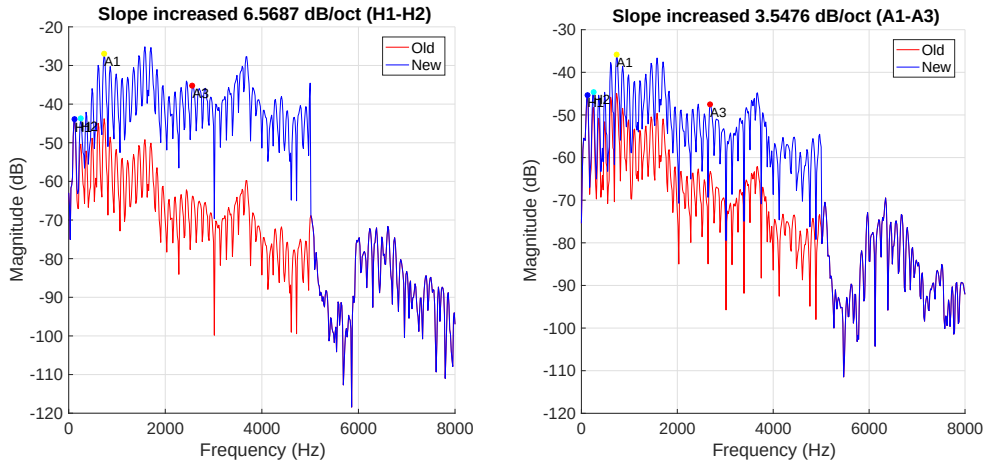



Figure 4.7: Formant Modification on /o/ at 84 dB

Table 4.1: Mathematical modeling of acoustic values depending on the noise level (Female)

| Acoustic Properties | Mathematical models of Acoustic feature | |
|---|---|---|
| | Acoustic Feature | Female |
| Duration | Vowel Duration ratio: | $\psi(x) = 0.773 + 0.0614 \times ln(x)$ |
| F0 | F0 mean ratio | $\psi(x) = 0.626 + 0.097 \times ln(x)$ |
| | F0 max ratio | $\psi(x) = 0.686 + 0.082 \times ln(x)$ |
| | F0 slop 1-2 ratio | $\psi(x) = 2.601 - 0.248 \times ln(x)$ |
| | F0 slop 2-4 ratio | $\psi(x) = 4.405 - 0.839 \times ln(x)$ |

the tilting value (dB/oct) of A1-A3 is used instead. A short time Fourier transform is used. Each vowel is framed with 40 ms frame length 0.5 ms frame shift. The modification is taken on each frame. The modified spectra are resynthesized by following overlap-added method. Figure 4.8 shows a modification on H1-H2 and A1-A3 to obtain a tilting difference from Lombard 84 dB on vowel /u/ of the female speaker.
The investigation how to apply voice conversion techniques to modify those acoustic features are still in the progress. The current application of voice conversion techniques still exists a lot of problems. It considerably affects to the evaluation of the adaptive model and limits the ability to mimic Lombard speech. To enclose this section the list of some of mathematical functions used in the adaptive model is presented as in Tables 4.1 and 4.2.




(a) Modification based on H1-H2

(b) Modification based on A1-A3

Figure 4.8: Modification of spectral tilts first frame of vowel /a/ in the $1^{st}$ mora of /yamamayu/ to mimick 84 dB Lombard speech

Table 4.2: Mathematical modeling of acoustic values depending on the noise level (Male)

| Acoustic Properties | Mathematical models of Acoustic feature | |
|---|---|---|
| | Acoustic Feature | Male |
| Duration | Vowel Duration ratio: | $\psi(x) = 0.381 + 0.153 \times ln(x)$ |
| F0 | F0 mean ratio | $\psi(x) = -0.179 + 0.292 \times ln(x)$ |
| | F0 max ratio | $\psi(x) = -0.091 + 0.27 \times ln(x)$ |
| | F0 slop 1-2 ratio | $\psi(x) = 1.782 - 0.247 \times ln(x)$ |
| | F0 slop 2-4 ratio | $\psi(x) = 7.341 - 1.464 \times ln(x)$ |

## 4.2 Conclusion

After carrying out the preliminary revision on currently used voice conversion techniques to modify acoustic features based on the constructed adaptive model, it can be realized that it is difficult to directly apply those techniques. More improvements and adjustments are needed to those techniques for a successful application. The details are explained as follows.

- **Duration**

  The modification of duration is basically satisfied with our expectation

- **Fundamental Frequency**

  By applying Fujisaki model and the optimization on the acoustic properties of F0 contour, the expected F0 contour can be generated. However, by experiencing the modification, it often takes much time to run the optimization. Hence, it is required to find out another solution for generating the F0 contour.

- **Formants**

  Although RMTD and spectral-GMM are flexible on modifying formant locations and amplitudes, they can not handle valleys between formants. Moreover, our first impression on the resynthesized speech of amplitude modification by those techniques is still not good. Therefore, it is needed to make adjustments on the original techniques to obtain the expected modification. Our first idea is applying a big Gaussian on concurrently handling formant amplitudes and valleys between formants, or separately estimating valleys between formants by polynomial fittings. Or another way is to control spectral tilts much.

- **Spectral tilts**

  The modifications of spectral tilts are taken on frequency spectrum of a frame. After modifying all designated frames, overlap-added rules are applied to resynthesize the speech. So far, our first impression on the quality of the resynthesized speech is not good. This technique is also needed to be revised and improved.

- **Modulation spectrum, energy redistribution**

  The ideas to modify those acoustic features are still in consideration. Our first outcomes are applying a filter on the modulation spectrum and doing spectral weighting for the energy redistribution.

After resolving the problems existing on the application of voice conversion techniques, we can obtain the good-quality synthesized speech. At that stage, they will be ready to be evaluated by objective tests (Intelligibility test by SII - Speech Intelligibility Index, or STI - Speech Transmission Index) and subjective tests (Reference tests for intelligibility and naturalness). Finally, the validation of each adaptive tendency can be answered to be justified or not. In our perspectives, the modification is to make the neutral speech the

same as Lombard speech as much as possible. Additionally, the set of modified acoustic features mostly covers every aspects of Lombard speech - temporal features, frequency features, and modulation frequency features. It is also verified that Lombard speech has better intelligibility than the neutral speech in terms of both subjective tests [11], and objective tests [12]. Therefore, it is expected that the resynthesized speech will also achieve this better intelligibility.

# Chapter 5

# Summary and Conclusion

## 5.1 Summary of the Thesis

The final goal of this study is to synthesize intelligible speech that maximally adapts with varying environments by knowledge learned from the emphasizing speech due to Lombard effect - Lombard speech. In this thesis, we have been tackling with acoustic analyzing the adaptive tendencies in Lombard speech produced in a noise-level varying background:

1. Acoustic feature extraction: The extracted features are selected to analyzed basing on their representation of neutral-Lombard differences and speech intelligibility. Basic acoustic features showing the distinction of Lombard and neutral speech are extracted first (F0, duration, spectral tilts), then the others being recognized as intelligible features are analyzed (Formants, Modulation Spectrum, energy redistribution (over frequency region) of phonemes).

2. Adaptive tendencies realization, physiological-psychological-acoustical establishment: Values of acoustic parameters are organized in an order of increasing noise level.

3. Adaptive tendency modeling and deployments by mathematical functions and voice conversion techniques, and their evaluations in contribution to intelligibility.

The acoustic feature extraction and realization of adaptive tendencies and physiological-psychological-acoustical correlation have been completed. Besides, we have still remained the problems:
The second phase of adaptive modeling and validation of adaptive model in contribution into intelligibility is still in progress.

- The processing tasks of adaptive tendency modeling and deployments for the mimicking Lombard speech.

- An evaluation of the mimicking Lombard speech the by objective and subjective tests.

In the way of approaching the final goal of synthesize intelligible speech maximally adapting with varying environments, it has still been remaining a lot of tasks. In addition to validation of adaptive tendencies on the mimicking Lombard speech, we also need to do further extrapolations.

- Extrapolating out of the region of 66-90 dB noise levels
- Extrapolating those features to maximize intelligibility on each noise level.

### 5.1.1 Main Contribution

By doing the feature extraction and analysis we have achieved the following contributions:

- Segmentation data of Lombard speech produced in the various noise level environments were created. The Lombard dataset has been used for acoustical analysis.
- Significant feature variations corresponding with noise level increasing have been extracted. They are lengthening vowel duration, increasing and abrupt changing at 84 dB in F0, shifting F1, increasing formant amplitudes, increasing amplitudes of valleys between formants, H1-H2 variation, decreasing A1-A3, and lifting modulation spectrum, energy redistribution over 0-5 kHz. 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.
- The first consideration of the log-linear evolution of acoustic variations has been considered with a set of mathematical modeling functions. A preliminary study on voice conversion techniques with handling of Lombard features. When the application of those techniques and knowledge of the model succeeds, the mimicking Lombard speech and extrapolated Lombard speech can be synthesized.

### 5.1.2 Possible solutions of remaining problems

To solve the problem of adaptive modeling and validation for the mimicking Lombard speech, and extrapolation out of the region of 66-90 dB noise levels, it is needed to do the following tasks.

- A consideration of suitable mathematical functions with evaluation by applying voice conversion techniques on objective and subjective tests
- Successful application of voice conversion techniques
- Doing the test by following a basic procedure on the synthesized speech

To solve the problem of extrapolating those features to maximize intelligibility on each noise level.

- Choosing the target acoustic features to control: energy redistribution, or spectral tilts, or formant frequency, amplitudes and valleys between formants, or modulation spectrum
- Choosing an objective measure to have a criterion for the variations of the acoustic features.
- Mathematical modeling the calculation of the objective measure by the target acoustic features, or brute-force finding the optimal value of the objective measure by varying the acoustic features by following recognized tendencies.

## 5.2 Conclusion

As the previous mention, this study has been carrying out to synthesize intelligible speech that maximally adapt with varying environments. So far, the acoustic analysis of the various Lombard speech produced in the noise-level varying background has been done with important feature variations being extracted. They would be the foundation for the next step of adaptation and extrapolation for achieving maximally intelligible speech in various noisy environments.

## 5.3 Future Work

- Many problems are still remaining, hence in future work, it is firstly to resolve the remaining problems described above by following the figures of possible solutions.
- When the adaption and extrapolation for maximal intelligibility succeeds, because the current Lombard speech was produced in pink noise, it is also needed to verify the synthesized speech with other types of noise with different noise levels.
- When the verification finishes, it is to design a entire system for real application of public addressing in noisy environments.
- Besides noisy environments, reverberation also often appears, therefore, it is to find the solution of maximally intelligible speech for both noisy and reverberant environments basing on the knowledge of naturally intelligible speech as Lombard speech.

# Bibliography


[1] Chen, Y., and Fomel, S., (2015)"Random noise attenuation using local signal-and-noise orthogonalization". *Geophysics*, 80(6), WD1-WD9.

[2] Shrawankar U., Thakare V., (2010) "Noise Estimation and Noise Removal Techniques for Speech Recognition in Adverse Environment". *Shi Z., Vadera S., Aamodt A., Leake D. (eds) Intelligent Information Processing V. IIP 2010. IFIP Advances in Information and Communication Technology, Springer, Berlin, Heidelberg*, 340.

[3] Yoshioka, T., Sehr, A., Delcroix, M., Kinoshita, K., Maas, R., Nakatani, T. and Kellermann, W., (2012)."Robustness against reverberation for automatic speech recognition". *IEEE SIGNAL PROCESSING MAGAZINE, November 2012*, 114-226.

[4] Schepker, H., Rennies, J., Doclo, S., (2012). "Improving speech intelligibility in background noise by SII-dependent amplification and compression". *AIA-DAGA 2013 Merano.*

[5] Zorila, C., Kandia, V., and Stylianou, Y., (2012). "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression". *In Proceedings of Interspeech, Portland, OR*, 635–638.

[6] Raitio, T., Suni, A., Pulakka, H., Vainio, M. and Alku, P., (2010). "Comparison of formant enhancement methods for HMM-based speech synthesis". *SSW.*, 334-339.

[7] Taal, C.H. and Jensen, J., (2013). "SII-based speech preprocessing for intelligibility improvement in noise". *INTERSPEECH*, 3582-3586.

[8] Lombard, E. (1911)."Le signe de l'elevation de la voix". *Annales des Maladies de l'Oreille, du Larynx, du Nez et du Pharynx*, 37, 101-119.

[9] Brumm, H., Zollinger, S. A., (2011)."The evolution of the Lombard effect: 100 years of psychoacoustic research". *Behaviour*, 148(11-13), 1173-1198.

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

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

[12] Godoy, E. , and Stylianou, Y., (2012). "Unsupervised acoustic analyses of normal and Lombard speech, with spectral envelope transformation to improve intelligibility". *Interspeech*, 1472-1475.

[13] 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.

[14] Lau, P., (2008). "The lombard effect as a communicative phenomenon".

[15] Huang, D.Y., Rahardja, S. and Ong, E.P., (2010). "Lombard effect mimicking". *In SSW.*, 258-263.

[16] Rottschafer, S., Buschmeier,H. , Welbergen, H. V., and Kopp, S., (2015). "Online Lombard adaptation in incremental speech synthesis". *Interspeech*, 80-84.

[17] Kasisopa, B., Attina, V., Burnham, D. K., (2014). "The Lombard effect with Thai lexical tones : an acoustic analysis of articulatory modifications in noise". *Proceedings of Interspeech 2014, 15th Annual Conference of the International Speech Communication Association, Singapore.*

[18] Zhao, Y., Jurafsky, D. (2009). "The effect of lexical frequency and Lombard reflex on tone hyperarticulation". *Journal of Phonetics*, 37, 231-247.

[19] Castellanos, A., Benedi, J.M. and Casacuberta, F., (1996). "An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect". *Speech Communication*, 20(1-2), 23-35.

[20] Hansen, J.H. and Varadarajan, V., (2009). "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition". *IEEE Transactions on Audio, Speech, and Language Processing*, 17(2), 366-378.

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

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

[23] Kawahara, H., Masuda-Katsuse, I. and De Cheveigne, A. (1999)."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 comm.*, 27(3), 187-207.

[24] Snell, R.C. and Milinazzo, F., (1993). "Formant location from LPC analysis data". *IEEE Transactions on Speech and Audio Processing*, 1(2), 129-134.

[25] 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.

[26] 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.

[27] Akagi, M., (2016). Lecture: "Speech Segmentation/Labeling Tour".

[28] Furui, S., (1986). "On the role of spectral transition for speech perception". *The Journal of the Acoustical Society of America*, 80(4), 1016-1025.

[29] Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H. and Shikano, K., (1990). "ATR Japanese speech database as a tool of speech recognition and synthesis". *Speech Communication*, 9(4), 357-363.

[30] Stemple, J. C., Glaze, L. E., Gerdeman-Klaben, B., (2000). "Clinical Voice Pathology, Theory and Management, 3rd Ed". *Canada: Singular Publishing Group.*

[31] Fujisaki, H. and Hirose, K., (1984). "Analysis of voice fundamental frequency contours for declarative sentences of Japanese". *Journal of the Acoustical Society of Japan (E)*, 5(4), 233-242.

[32] Rossi, P.S., Palmieri, F. and Cutugno, F., (2002). "A method for automatic extraction of Fujisaki-model parameters". *In Speech Prosody 2002, International Conference.*

# Publications

## International Conferences

[1] Ngo, V.T., Kubo, R., Morikawa, D., and Akagi, M., (2017). "Acoustical analyses of Lombard speech by different background noise levels for tendencies of intelligibility". *Proceedings of International Conference (NCSP'17)*, Guam, US.

## Domestic Conferences

[2] Ngo, V.T., Kubo, R., Morikawa, D., and Akagi, M., (2017). "Acoustic variation of Lombard speech produced in various noise-level environments". *ASJ Spring meeting*, Japan.