Improvement of intelligibility and naturalness of speech in noise inspired by Lombard effect

Thuanvan Ngo

Japan Advanced Institute of Science and Technology

Doctoral Dissertation

Improvement of intelligibility and naturalness of speech in noise inspired by Lombard effect

Thuanvan Ngo

Supervisor: Professor Masato Akagi

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Information Science
September 2020

Abstract

In public announcements in train stations or airports, the presence of noise often smears presented speech, thus makes it hard for listeners to understand it. By reducing noise throughout the presentation, speech is still intelligible and natural to listeners. However, this seems impractical due to complex architectures of these places and costly installed devices. Besides, it is practical and less expensive to enhance speech itself before presenting to complement degradation in intelligibility and naturalness by the smears.

Lombard speech is intelligible speech produce by humans in noise due to the Lombard effect. Investigation of Lombard speech could reveal essential features to increase speech intelligibility and naturalness. Therefore, the purpose of this research was to improve the intelligibility and naturalness of speech in noise using conversion rules inspired by the Lombard effect. It came up with two sub-goals: (I) obtaining feature understanding and control which contributes to the intelligibility of Lombard speech under **noise-level-varying and various noise** and (II) identifying and applying the **effective feature control methods** for exceeding the intelligibility and naturalness of Lombard speech. From the previous research and the properties of Lombard speech that varies with noise levels and noise types, three problems arose to cover the search space (features, noise levels, feature variations, SNRs, and spectral-varied noise) for finding features and applying them.

(1) Contribution of acoustic features of Lombard speech has no consideration of their articulatory features at one noise level: For this problem, the modification of acoustic features of Lombard speech was often done without any considering the articulatory changes or challenging to be obtained in acoustical levels. The multiple contributions of the features to the intelligibility and naturalness were unclear.

(2) Control and contribution of acoustic features of Lombard speech in multiple noise levels of backgrounds: In this problem, acoustic features contributing to Lombard speech in various noise levels were difficult to be modeled and controlled by the conventional methods.

(3) Unclear effective features to the intelligibility and naturalness of speech varying noise levels and various types of noise: This problem followed the second problem. Acoustic features,

when varying that contribute to the intelligibility of speech in noise, remained unclear. Recent studies differently reported effective features for the intelligibility and naturalness of speech in noise. Thus, the precise set of effective features was unidentified.

Thus, the investigation has three steps: (1) Mimicking Lombard speech by controlling articulatory and acoustic features, (2) Effective features for the intelligibility and naturalness of speech in noise, (3) Application to improve the intelligibility of speech under noisy reverberant conditions.

Following these steps, this study obtained the articulatory and acoustical controls in mimicking Lombard speech. The contributive features, including spectral tilts, f_0, and formants, were also explored in the first step. In the second step, the effective features for intelligibility and naturalness in all kinds of noise were identified. In the final step, the effective features were successfully applied to increase intelligibility and naturalness of speech under noisy reverberant conditions with a pair of effective time-frequency features.

As a result, this research can enlighten the fields of speech enhancements, objective intelligibility measurements, voice conversion, and synthesis. It provides essential, necessary information for the areas of speech enhancement engineering.

Keywords: Lombard speech, mimicking, articulatory features, acoustic features, effective features, speech intelligibility and naturalness in noise.

Acknowledgments

First of all, I would like to express my deepest gratitude to my supervisor, Professor Masato Akagi, for his guidance, encouragement, and support during my master and Ph.D. studies. Without his guidance, this thesis could not have been finished. He also guides me on technical writing and English writing and makes a lot of publications in my master and Ph.D. courses. I was studying under his guidance, that is my best experience.

I would like to thank my associate supervisor, Professor Masashi Unoki, for his comments and advice in my master and Ph.D. studies. Especially in the laboratory meetings, he helps me to improve my knowledge, presentation skill, and Powerpoint skill.

I also would like to thank my sub-theme supervisor, Professor Peter Birkholz, from the Technical University of Dresden, Germany, for his supervision and supports during my study there and after finishing the study there. Notably, he helped me a lot in writing a scientific paper.

I thank all members of the acoustic information science laboratory at JAIST. In particular, thanks to Dr. Rieko Kubo, for her help and advice in my study. I would like to thank all Japanse and German friends, who helped me to do experiments.

Finally, I would like to express my gratitude for the financial support of my Ph.D. study scholarship by JAIST and Professor Masato Akagi. I also thank the SECOM Science and Technology foundation and JST-Mirai Program (Number: JPMJMI18D1) for their supports.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research motivation

The presence of noise in public announcements in train stations and airports often smears the speech spectra, thus making it hard for listeners to understand the announcement. Speech intelligibility could be maintained by reducing noise and reverberation throughout the presentation. However, this is impractical due to the complex architectures of such locations and the cost of installing the necessary devices. A more practical and efficient approach is to enhance the speech before presentation to compensate for degradation in intelligibility due to smearing. Therefore, in this study, the presented speech was modified to increase its intelligibility in noisy environments.

The intelligibility of the presented speech in noisy environments can be increased by (1) optimizing indexes of objective measures and (2) mimicking intelligible production and perception styles. The former method (1) tried to increase speech intelligibility in noise by optimizing indexes of objective measures [1–4] such as Speech Intelligibility Index (SII) [5], Speech Transmission Index (STI) [6] and high energy glimpse portion (HEGP [7]). It obtained considerable intelligibility improvement. However, the optimized speech has low naturalness because speech features for naturalness are often broken. The second method (2) in terms of intelligible production styles tried to mimic clear speech and Lombard speech. Clear speech [8, 9] is produced when humans speak as clearly as possible in quiet conditions. Lombard speech is produced in noise due to the Lombard effect [10]. Lombard speech is intelligible and natural in noise [11–15]. Summer et al. [15] reported that utterances originally produced in noise (produced at

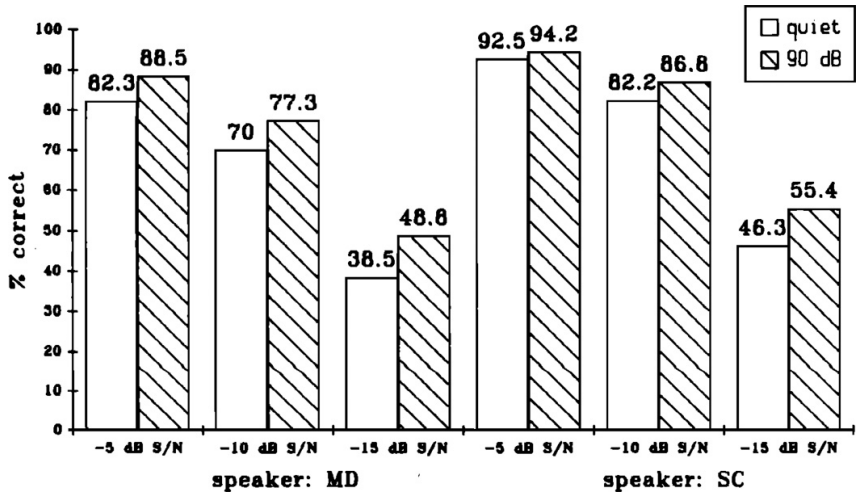


Figure 1-1: Intelligibility of words produced in quiet (plain speech) and 90 dB of masking noise (Lombard speech) taken from [15, Fig. 8]

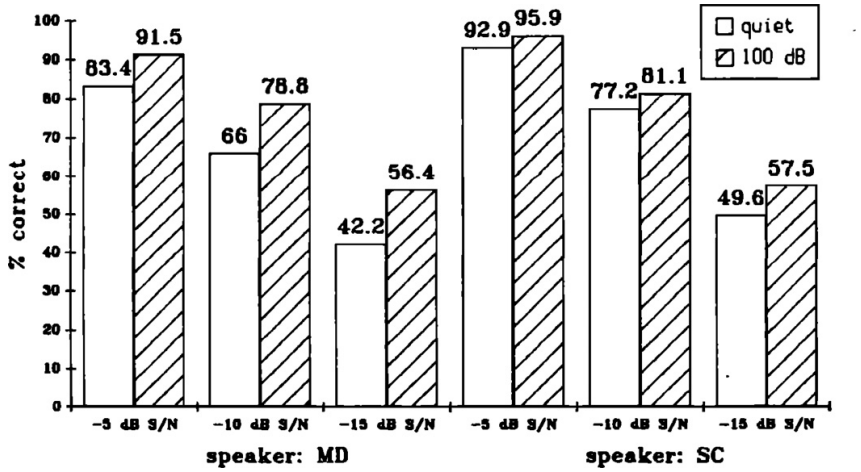


Figure 1-2: Intelligibility of words produced in quiet (plain speech) and 100 dB of masking noise (Lombard speech) taken from [15, Fig. 9]

90 dB and 100 dB level of noise), i.e., Lombard speech were found to be more intelligible than utterances produced in the quiet, i.e., plain speech. Figures 1-1 and 1-2 (taken from their study) group the intelligibility of plain speech and Lombard speech according to SNR (eliminating the intensity differences), and the intelligibility of Lombard speech was still higher than the intelligibility of plain speech at all SNRs. Also, in the study of Kubo et al. [14], besides intelligibility, annoyance was also investigated for the plain and Lombard speech at different levels of noise. They mentioned in the results that when Lombard speech is produced at the same noise level as the testing noise level, higher intelligibility and less annoyance are obtained, disregarding the intensity is normalized or not. These results emphasized that Lombard speech is still intelligible and natural in noise.

Therefore, by applying the properties of clear speech, and especially Lombard speech into modifying plain speech, intelligible and natural speech in noise can be synthesized. Thus, by making the transmitted speech *perceivable* as or better than clear speech and/or Lombard speech, i.e., by *perceptually* mimicking and exceeding clear speech and/or Lombard speech, listeners are expected to be able to capture the transmitted information. When the mimicked speech is obtained, it is also highly potential to proceed further manipulation to exceed the intelligibility of these speech. The second method (2) in terms of intelligible perception styles was about temporal perception. In a method by dynamic range compression (DRC), dynamic releases and attacks on the power envelope and a static range compression into the range of -40 dB - 0 dB (by deriving an input-output energy curve, i.e., IOEC) were applied in spectral shaping and dynamic range compression (SSDRC) method [16] to increase speech intelligibility. The DRC emphasizes the voice onsets and offsets, stops and nasals, and improves the co-articulation effect, thus increases speech intelligibility [16, 17]. However, it sometimes makes speech signals degraded, especially in naturalness [18]. In the present study, increases in the intelligibility and naturalness of speech in noise were mainly inspired by the Lombard speech due to the Lombard effect.

1.2 Research goal and problems

The goal of this study was to improve the intelligibility and naturalness of speech in noise using conversion rules inspired by the Lombard effect. It came up with two sub-goals: (I) obtaining feature understanding and control which contributes to the intelligibility of Lombard speech under **noise-level-varying and various noise** and (II) identifying and applying the **effective feature control methods** for exceeding the intelligibility and naturalness of Lombard speech. According to the properties of Lombard speech that varies with noise level and noise types, three problems arose with these sub-goals as follows to cover the search space (features, noise levels, feature variations, SNRs, and spectral-varied noise) for finding features and applying the features.

1. Contribution of acoustic features of Lombard speech has no consideration of their articulatory features (at a noise level): For this problem, the modification of acoustic features of Lombard speech was often done without any considering the articulatory changes or dif-

ficult to be obtained in acoustical levels. The multiple contributions of the features to the intelligibility and naturalness were unclear.

2. Control and contribution of acoustic features of Lombard speech in multiple noise levels of backgrounds: In this problem, acoustic features contributing to Lombard speech in multiple noise levels were difficult to be modeled and controlled by the conventional methods.

3. Unclear effective features to the intelligibility and naturalness of speech varying noise levels and various noise types: This problem followed the second problem. Acoustic features, when varying that contribute to the intelligibility of speech in noise, remained unclear. Recent studies differently reported effective features for the intelligibility and naturalness of speech in noise. Thus the precise set of the effective features was unidentified.

1.3 Research approach

Based on the goal and problems, the present study was built up by three steps:

(1) Mimicking Lombard speech to independently control the articulatory and acoustic features and investigate their contribution to the intelligibility and naturalness of speech in noise.

- On mimicking by articulatory features, it was to investigate the effects of individual *articulatory* features and their combinations on enhancing the intelligibility of speech in noise. To this end, an enhanced version of the articulatory speech synthesizer VocalTractLab [19, 20] was used to synthesize ten German words for digits in multiple variants that differ with respect to f_0 (mean and range), phonation type, formant frequency, and duration. In a perception experiment, the intelligibility benefit of these features for pink noise and babble noise was evaluated. From this Lombard speech mimicking by articulatory features corresponding with acoustical targets, several mechanisms to produce Lombard speech at the articulatory layer could be confirmed and applied to define the acoustic features to mimic Lombard speech at the acoustical layer.
- Expanding the search space from one noise level to multiple noise levels, on mimicking by acoustic features, to improve upon the analysis for feature tendencies in previous studies and based on the mechanisms extracted from the Lombard speech mimicking by articulatory features, *in-depth analyses were adequately performed* on

the acoustical features of plain speech and Lombard speech produced in backgrounds with various noise levels. Then, based on the analyzed feature tendencies, *a continuous rule generation model* of acoustic features was designed to precisely estimate the effects of noise. This model was expected to overcome the inaccuracy of the previous models and to increase the adaptability of mimicking speech. Lastly, to *flexibly and precisely control multiple features* with varying noise levels in a way that preserves the naturalness of synthesized speech, it was to apply a coarticulation model [21] and a modified-restricted temporal decomposition (MRTD) [22] with spectral-GMM [23] for the synthesis and modification. Due to the limitation of the dataset, the mimicked speech was compared with that of statistical BGMM-based methods (rather than that of DNN-based methods) and Lombard speech through subjective listening tests.

(2) Identify effective features for improving the intelligibility and naturalness of speech in variable noise

By the independent control of acoustic features and the rule generation model, the speech differed with varied values of acoustic features/feature variations was synthesized and evaluated in multi-levels of pink noise. The effective acoustic feature to vary with multiple noise levels was identified. Sequentially, the results of this effective acoustic feature were investigated along with the results of different studies on the intelligibility of speech in noise (not only from Lombard speech) [4, 16] in terms of time-frequency features under variable noise and SNRs. The concept based on modulation spectrum and modulation transfer function concepts in relationship with listening tests was proposed to model and extract intelligibly and naturally-correlated time-frequency features from the speech synthesized by these methods. Finally, the effective time-frequency was identified after performing perception experiments on the extracted modulation spectral features as time-frequency features.

(3) Considering applying the effective features to increase speech intelligibility in noisy reverberant conditions

The evaluation of the effective time-frequency features was performed in noisy reverberant conditions with the dataset provided by Hurricane challenge 2.0 (HC 2.0) [24]

1.4 Research originality and novelty

The originality of this study was the first investigation of the contribution of articulatory features to the intelligibility of speech in noise. Secondly, it was mimicking Lombard speech under various noise levels. Finally, this study presented a brutal-force method for extracting the effective acoustic features to vary to increase the intelligibility of speech in noise. Besides, the present study obtained novelty in two aspects. Firstly, it was the concept of applying rule-based methods and the Lombard effect model for the rule generation model to mimic Lombard speech concerning multiple noise levels. Secondly, it was the concept based on the modulation spectrum and modulation transfer function concepts in relationship with listening tests to identify the effective features to increase speech intelligibility and naturalness in noise.

1.5 Structure of Dissertation

The remaining parts of this thesis were as follows:

- **Chapter 2** summaries related work about the articulatory and acoustical investigation of Lombard speech. The intelligibility and naturalness improvements based on Lombard speech were then described. It also listed models to increase the intelligibility of speech in noise.

- **Chapter 3** describes Lombard speech mimicking by articulatory features at a noise level. The contribution of the multiple articulatory-acoustic features was figured out. This chapter corresponds to solving the first problem of articulatory features in step 1 in the research approach.

- **Chapter 4** presents the control methods and the rule generation model for mimicking Lombard speech at multiple noise levels. This chapter explains the solution for the second problem of mimicking at multiple noise levels by acoustic features in step 1 in the research approach.

- **Chapter 5** focuses on the investigations of effective features. In the first half, acoustic features were varied. We studied the contribution of these variations to speech intelligibility and naturalness in noise. In the second half, a model of effective feature extraction

was proposed to identify the final effective features. This chapter corresponds to step 2 in the research approach.

- **Chapter 6** shows an application and evaluation of some effective features under various noisy reverberant conditions and languages. This chapter reflects step 3 in the research approach.

- **Chapter 7** concludes this study with a summary of all the done work and the contribution of the findings to its research field and other research fields. Furthermore, future work was also discussed.

The connection among these major chapters based on the back bone of this dissertation for finding features to exceed Lombard speech and their application are illustrated in Fig. 1-3.

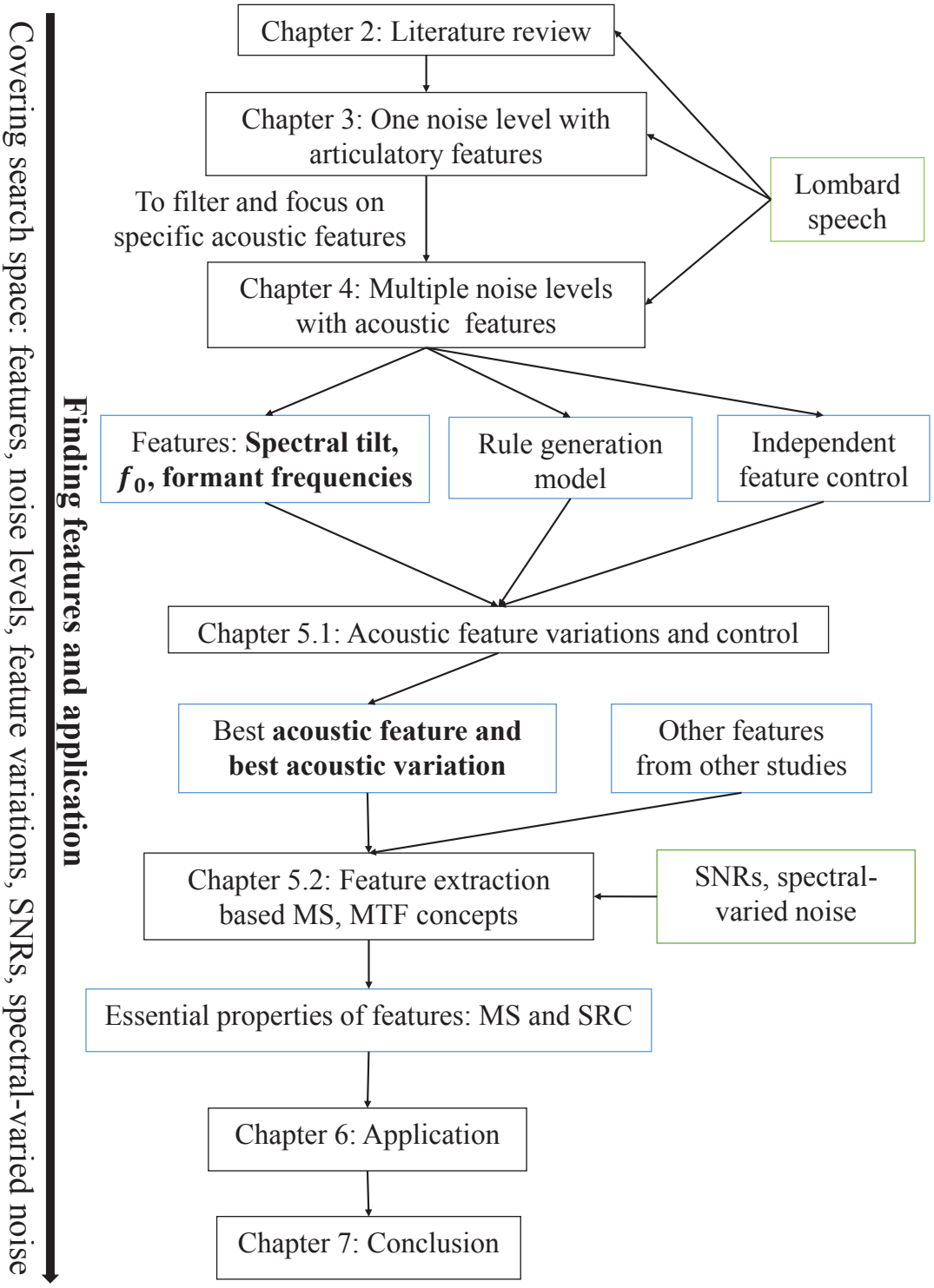


Figure 1-3: Connection among chapters and based on the back bone of this study for finding features and application

Chapter 2

Literature review

This chapter firstly reviews the investigation of Lombard speech about articulatory and acoustic features. Then, it describes the literature of the intelligibility and naturalness improvements based on Lombard speech. The models used to estimate speech intelligibility in noise were also explained. They were used to increase speech intelligibility. The problems stated in Chapter 1 are also allocated accordingly.

2.1 Articulatory-acoustic features of Lombard speech

Lombard [10] realized that humans involuntarily change their way of speaking in noisy conditions. This phenomenon is now called the "Lombard effect" or "Lombard speech" and has been shown to improve the intelligibility of speech in noise [11–15]. The articulatory and acoustic changes underlying this phenomenon have been thoroughly studied. Compared with "plain" speech (a term adopted from Bradlow and Alexander [25] to refer to "normal" speech produced in quiet conditions), Lombard speech mainly differs in terms of vocal intensity, spectral tilt, formant frequency, fundamental frequency (f_0), and the duration or speaking rate. Vocal intensity is usually increased in Lombard speech [15, 26]. The spectral tilt of Lombard speech is normally flatter than for normal speech, i.e., there is more energy at higher frequencies [27]. With regard to formant frequencies, multiple studies found a systematic increase of F_1 in Lombard speech [15, 26, 28, 29]. Some of these studies also reported an increase of F_2, e.g., Uemura et al. [29], but this increase was smaller and not as systematic as for F_1. With regard to f_0, both the f_0 mean and range increase with Lombard speech [26, 27, 29]. Finally, Junqua [26] found

that, in Lombard speech, the duration of vowels is significantly increased and the duration of consonants is slightly decreased. This leads to an overall increase of word durations and hence a lower speaking rate. Most of these features (spectral tilt, formant frequencies, f_0, duration) were shown to vary continuously with the background noise level [28].

The reasons for the acoustic changes with increasing background noise level are corresponding articulatory changes. Lombard speech is generally hyperarticulated, i.e., the spatial extent and the velocity of tongue, jaw and lip movements are increased [30–33]. The consequence is that the tongue position of vowels in Lombard speech is on average lower than during plain speech [34, 35]. Given the general inverse relationship between tongue height and F_1, this explains the increase of F_1 in Lombard speech. Garnier et al. [30] and Garnier et al. [36] demonstrated a correlation of the extent of tongue and lip movements not only with F_1 but also with F_2 and f_0. The flattening of the spectral tilt in Lombard speech is most likely explained by a change of the phonation type. According to Stevens [37], the spectral tilt increases (flattens) by about 6 dB/oct when phonation changes from modal to pressed. With regard to glottal articulation, a more pressed voice quality is achieved by a stronger adduction of the vocal folds.

All the studies mentioned above essentially analyzed the articulatory and acoustic features of naturally produced Lombard speech. However, it is also of great interest to learn which of these features contribute to what extent to the enhanced intelligibility of Lombard speech. This knowledge can help in developing suitable methods for synthesizing more intelligible synthetic speech [38–40] or modifying natural speech to make it more intelligible [41].

Currently, there are only few studies that clarified the potential intelligibility benefit of typical features of Lombard speech. Lu and Cooke [42] analyzed to what extent an increased f_0 and a flattened spectral tilt contribute to enhanced intelligibility in noise. With natural speech recordings as the basis, they used the vocoder STRAIGHT [43] to increase f_0 and a digital filter with a specific magnitude response to flatten the spectral tilt. They found that a flattened spectral tilt had a strong positive effect on the intelligibility, while an increase of mean f_0 had no significant effect. In a similar way, Cooke et al. [44] analyzed the effect of increased phone durations (besides a flattened spectral tilt) on the recognition of speech in noise. To modify the duration of natural basis material, the PSOLA algorithm implemented in Praat [45] was used. However, no beneficial effects of durational increases were found. In a later study, Cooke and Aubanel [46] found that increasing durations may still have a positive effect on the intelligibil-

ity but only when the background noise is fluctuating (as opposed to stationary). Common to the pioneering studies by Lu and Cooke; Cooke et al. [42, 44] and Cooke and Aubanel [46] is that the acoustic features of interest were modified at the *acoustic* level on the basis of natural speech recordings. While this is effective for the manipulation of the features f_0, spectral tilt, and duration, it would be more difficult to explicitly modify individual formants in natural recordings. Furthermore, acoustic manipulations are not explicitly related to articulatory and physical mechanisms. These limitations were related to the problem 1 of mimicking Lombard speech with modifying acoustic targets by articulatory features.

2.2 Intelligibility and naturalness improvements based on Lombard speech

Mimicking and exceeding are two typical ways to increase the intelligibility of speech in noise based on Lombard speech. In mimicking Lombard speech, state-of-the-art methods are based on statistical Bayesian GMM (BGMM) [47] or DNN techniques [48]. In these methods, speech features or waveforms are automatically mapped from plain speech to these of Lombard speech by trained models, and the resulting mimicked speech is quite similar to Lombard speech. However, since the characteristics of Lombard speech are varied according to noise levels and different types of noise [28, 49–51], these state-of-the-art methods, especially DNN-based methods, require an extremely huge dataset to train, thus rendering them impractical. The target of these methods is just mapping plain to Lombard speech. While proceeding a further target like exceeding Lombard speech, the mimicking has to reveal the mechanism of the production of Lombard speech.

The other way is thus to understand the Lombard speech more by analyzing it to extract mimicking rules. This method is often known as rules-based synthesis. Figure 2-1 shows a typical flow for a rules-based system to covert speech to speech. To convert an input speech to an output speech, a vocoder is needed to extract features from the input speech. Then these features are modified by rules to obtain modified features. Next, using the vocoder again to resynthesize the modified feature to obtain the output speech. The major problem for this system is to obtain rules for specific tasks. To create rules, it requires some analyses to understand how the features from the input speech different from these of the output speech. In other words, the modifica-

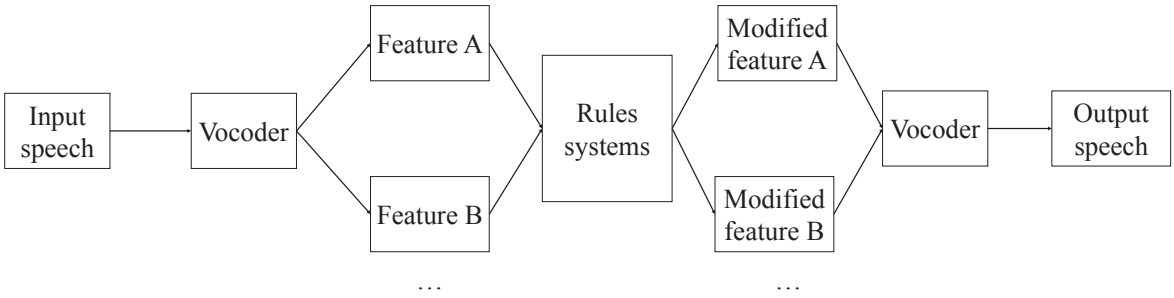


Figure 2-1: A typical flow in rules-based systems

tion should contribute to target properties in the output speech. Similarly, researchers with the target of Lombard speech extracted acoustic features correlated with the intelligibility benefit of Lombard speech in noise to modify these of the plain speech. By using these rule-based methods, some insight into how Lombard speech is really produced can be obtained. Thus, Lombard speech can be synthesized systematically. In these methods, feature control is more robust to change and can be manipulated for the further purpose of exceeding Lombard speech yet requires detailed analyses and advanced modification techniques. Acoustical analyses [42, 44] have mainly revealed that, compared with plain speech, the distinctive acoustic features of Lombard speech include increased duration, increased fundamental frequency (f_0), flattened spectral tilt, and increased vocal intensity [52]. In addition, Ngo et al. [28] found that these distinctive features of flattened spectral tilt, increased power envelope (or rises in modulation spectrum in specific frequencies), increased f_0, increased F_1, and increased vowel duration continuously vary with increasing noise levels. On the basis of the analysis results, methods such as those by Huang et al. [53, 54] and Rottschaefer et al. [55] make acoustical rules to directly modify plain speech to Lombard speech. Huang et al. mimicked Lombard speech with fixed adjustments for features and obtained acceptable similarity and naturalness. The problem is that these fixed adjustments are suitable for just one noise level rather than multiple noise levels. Feature modification methods that perform adjustments according to noise level are still hard to control. For example, formants have been modified by weighting on frequency bands, which led to increased sensitivity to errors and affected naturalness. Rottschaefer et al. constructed an online Lombard adaptation in incremental speech synthesis to present Lombard speech with noise levels of environments continuously. This online model achieved good results in adapting voice intensity and spectral emphasis (mainly, vocal intensity) but failed with other features (e.g., f_0 and duration). It was claimed that a more subtle and advanced Lombard-adaptation model

did not have any effect on intelligibility or perceived naturalness. In other words, this method is limited in terms of both the quantity (the number of controlled features) and the quality of mimicked speech. The reasons are probably that the incremental synthesis was not suitable for modifying many features precisely and flexibly, and that the adaptation model was still inaccurate. Solving these limitations was also related to the problem 2 of mimicking Lombard speech at multiple noise levels.

In exceeding Lombard speech, although the method by speech shaping and dynamic range compression (SSDRC) [16] can produce better intelligibility than Lombard speech [56], few studies have especially concerned problems in exceeding intelligibility and preserving naturalness of Lombard or clear speech. In SSDRC, the Lombard inspired features were combined with other features to obtain better intelligibility. It modified spectral features to decrease spectral tilts and increase formant amplitudes, which was shown in Lombard speech. It also modified speech amplitude by DRC. This SSDRC is only applicable to a fixed noise level, some further investigation of Lombard speech to find out effective features might help to obtain better intelligibility under more variable environments.

2.3 Models estimating speech intelligibility

The models estimating speech intelligibility seeks for intelligible factors. Then it is to give speech an index score of intelligibility in range 0 to 1. When 0 means totally unintelligible, 1 means completely intelligible. For environment phenomena such as SNR and noise levels, the models, including perceptual models and room acoustic models, can estimate the equivalent amount to compensate degradation of intelligibility. Using them to increasing speech intelligibility is also able to adapt to environments.

2.3.1 Perceptual models

The models helped to increase speech intelligibility by the way speech is modified, such that their indexes were optimized mathematically.

a. Speech Intelligibility Index

   SII is an objective index calculating the intelligibility of speech in noise. The speech spectrum, the noise spectrum, and the hearing threshold of the listener are used in the calculation

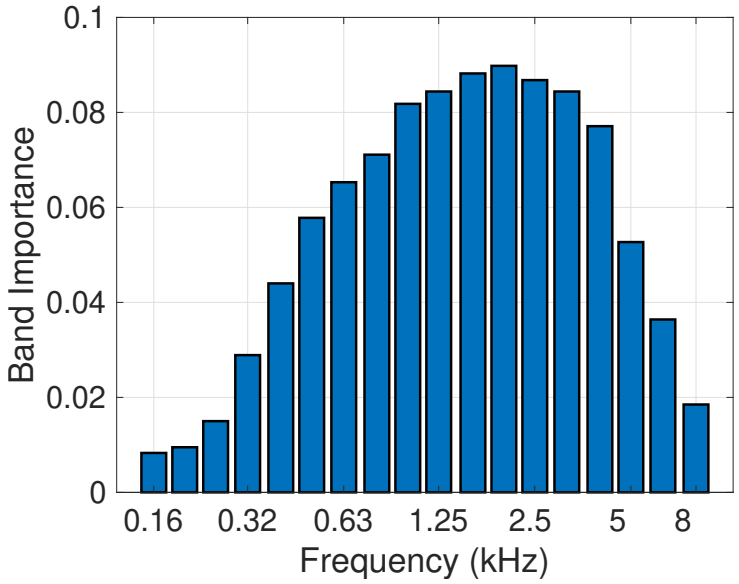


Figure 2-2: Band important function used in the calculation of SII (at word level)

of SII. Speech and noise are filtered into band-limited signals. For each frequency band, a band auditory is estimated. The auditory band function is weighted by band-importance function. The band-importance function indicates the degree frequency band affecting intelligibility (e.g., Fig. 2-2). The SII is obtained by summing up all the weighted values, which produces the result between 0 and 1. The methods using SII [1, 2] to increase intelligibility adjusted the auditory band function to obtain the maximal index of SII.

b. Speech Transmission Index

STI is an objective index measuring the intelligibility of speech on a transmission channel. The basis in the calculation of STI is the estimation of the modulation transfer function (MTF). The methods using STI [57] to increase intelligibility adjusted locations of source speech to maximize STI.

c. High energy glimpse portion

A glimpse model was proposed by Cooke et al. [58]. The glimpse is the proportion of time-frequency regions assuming to be audible due to exceeding equivalent regions of the masker by a specific amount of energy (e.g., 3 dB). HEGP [7] is defined as the time-frequency regions to be glimpses providing that the local excitation pattern for the noisy speech at a frequency region is higher than the average level in that frequency region. The methods using HEGP [4] to increase intelligibility adjusted the weights on the time-frequency spectra

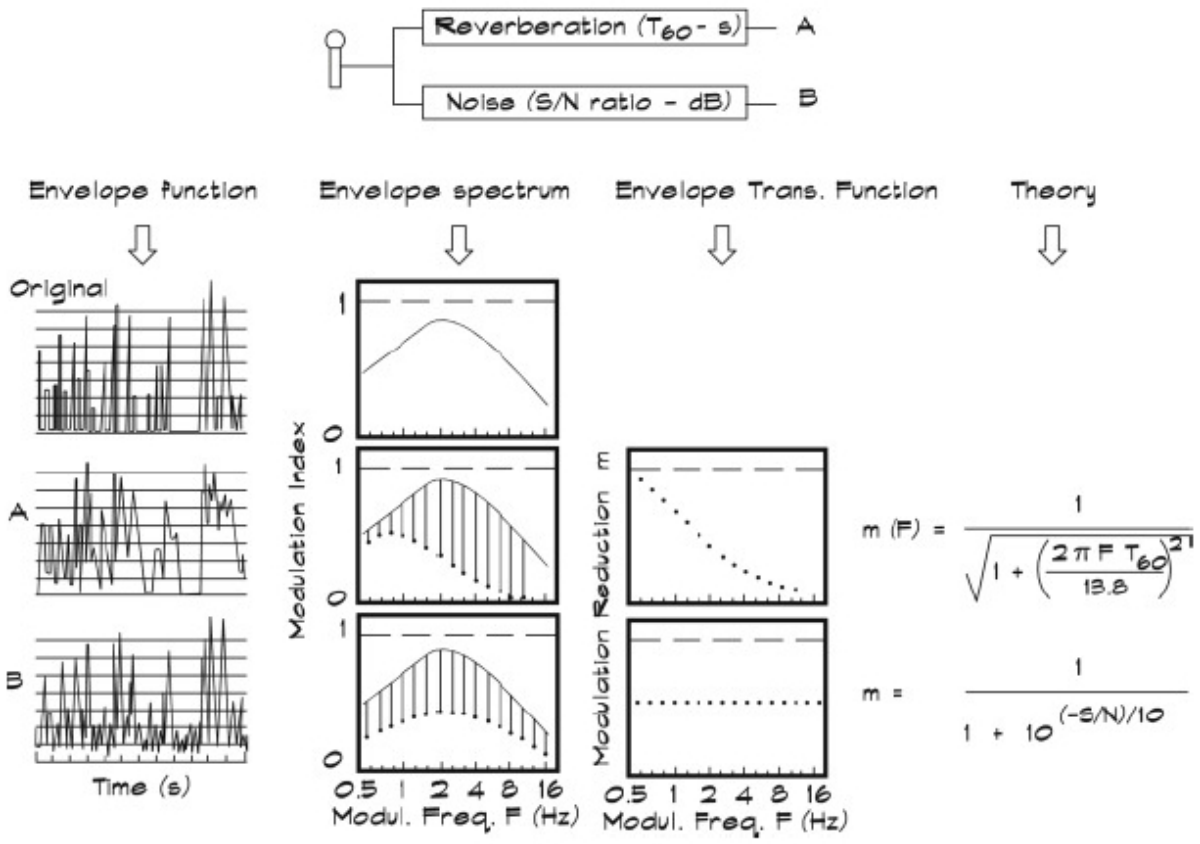


Figure 2-3: Basis of the Modulation Transfer Function, taken from [60, Fig. 4.27]. "The reduction of the fluctuations in the (octave band specific) envelope of an output signal (A or B) relative to the original signal can be expressed as Modulation Transfer function." In Houtgast and Steeneken [61]

of speech to maximize HEGP.

2.3.2 Room acoustic models

A model of MTF proposed by Houtgast and Steeneken [59] is the most famous room-acoustic model. The modulation reduction of an output signal is imitated from an input signal during transmission in rooms.

The basic concept of MTF is shown in Fig. 2-3. As speech consists of modulated bands of noise, i.e., a band of noise is originally produced by the vibration of vocal cords and finally modulated by the mouth at various frequencies to generate words. Houtgast and Steeneken [61] simulated this process by an octave band of noise modulated with a low-frequency pure tone, (i.e., multiplying the carrier with a sinusoidal function). The resulting modulated signal is a source signal shown on the left side of Fig. 2-4. When transmitting this signal to a listener,

the environment smears it, which results in decreased speech intelligibility. The distortion in background noise and reverberation is shown like that the bottom of the signal is raised above zero and it is reflected in the same way as a delayed and/or distorted replication of the signal is added back to it. When the signal arrived at the listener, it thus looks like the signal on the right side of Fig. 2-4. The received signal is less modulated than the source signal. The degree of modulation is defined by the depth/index of the modulation envelope, which is later referred to as modulation spectrum. The reduction in modulation is defined by a modulation reduction factor, $m(f_m)$, which is a function of the modulation frequency of f_m. The modulation reduction factor varies from 0 to 1, where 1 means no reduction and 0 mean total reduction. m depends on f_m as curves shown on the bottom of Fig. 2-4. These modulation reduction curves are generally understood as MTF.

The MTF model for noise is independent of modulation frequencies and dependent on signal-to-noise ratios. The noise raises all levels for the received signal, thereby equally reduces modulation on the signal. The MTF model for reverberation has the shape of a low-pass filter. It can be calculated from modulation frequency and reverberation time (T_{60}), where T_{60} is the time it takes for the MTF index to reduce by 60 dB. The stepper the reduction curve, the more complicated the effects of reverberant environments. The MTF model for noisy reverberation is the product of the MTF model for noise and the MTF model for reverberation. MTF is often used in the calculation of STI and speech restoration [62].

Envelope spectrum as shown in Fig. 2-3, also referred to as modulation spectrum (MS) can be produced by doing spectral analysis on the temporal amplitude envelope of frequency spectra. The component lying between 1 and 16 Hz modulation frequencies (with a peak around 4 Hz) dominates MS of continuous speech [59, 63, 64].

The methods using MTF [64] to increase intelligibility performed directly inverting effects by MTF on the MS of speech. Inverting effects of the environments by MTF on the MS of the original speech is to keep indexes of the dominant components of MS still high enough to have the same intelligibility as the original speech in quiet. A problem with these methods is that it is challenging to estimating the inverse MTF under realistic variable environments because it requires estimating MTF. This estimation is complicated and tolerant under these variable environments. Another perspective in this study was to base on MS and MTF concepts to propose an effective feature extraction model from differently enhanced speech, which was

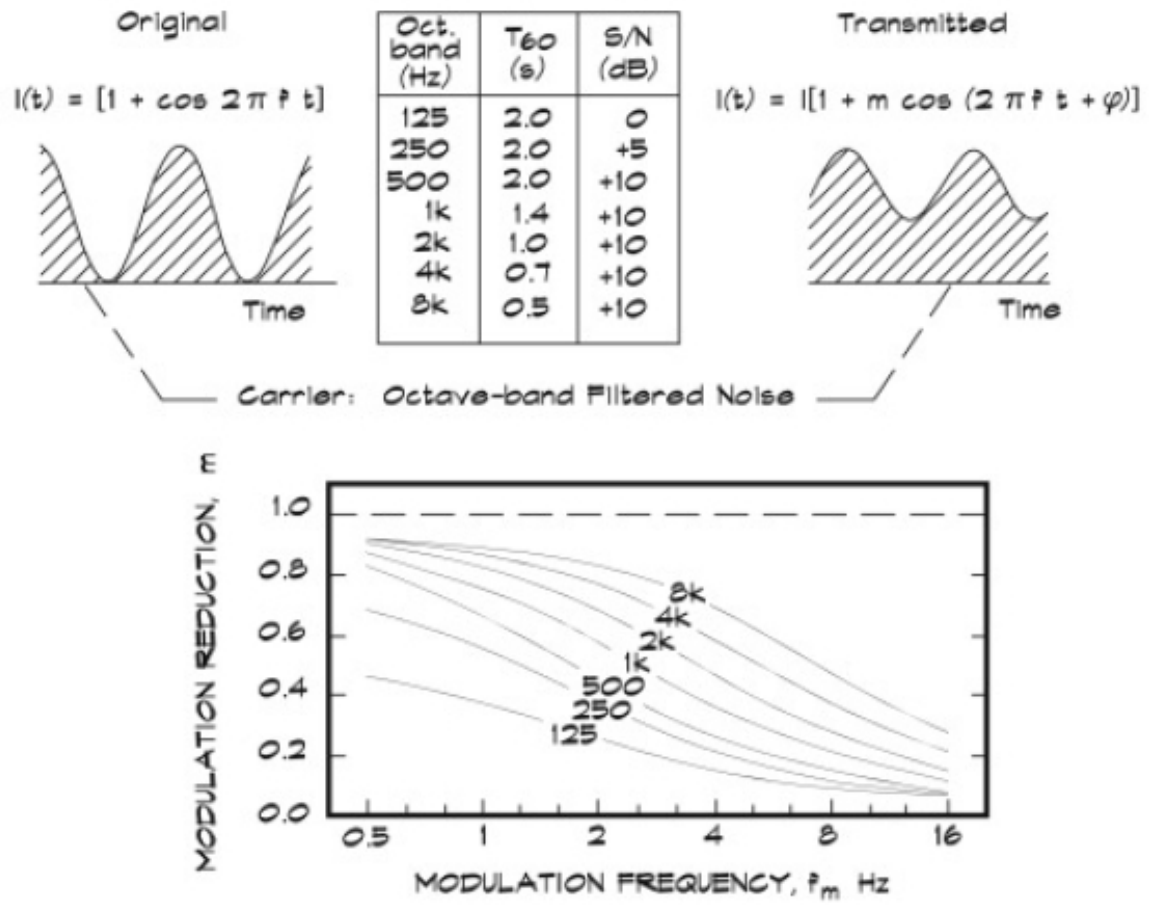


Figure 2-4: Modulation Transfer Function, taken from [60, Fig. 4.28].“Illustration of how an MTF analysis is performed by using an octave-band filtered noise carrier, 100% intensity modulated, for each modulation frequency successively. This leads to a family of MTF curves. Each curve is calculated for the data given in the table.” In Houtgast and Steeneken [61]

also related to the problem 3 of finding out effective features or feature control methods.

Chapter 3

Mimicking Lombard speech by articulatory-acoustic features

This chapter describes the mimicking Lombard speech by using articulatory features to control acoustic feature targets aiming to understand how the contribution of articulatory-acoustic features on the intelligibility and naturalness of speech in noise at a noise level is. In the end, several mechanisms to produce Lombard speech by articulatory features, which could be used to define and focus on acoustic features to investigate further in the next chapter, were summarized.

The increased intelligibility by Lombard speech is enabled by the change of multiple articulatory and acoustic features. While the major features of Lombard speech are well known from previous studies, little is known about their relative contributions to the intelligibility of speech in noise. This section describes the usage of an analysis-by-synthesis strategy to explore the contributions of multiple of these features. To this end, an articulatory speech synthesizer was used to synthesize the ten German digit words "Null" to "Neun", for all 16 combinations of four binary features, i.e., modal vs. pressed phonation, normal vs. increased F_1 and F_2 formant frequencies, normal vs. increased f_0 mean and range, and normal vs. increased duration of vowels. Subjects were asked to try to recognize the synthesized words in the presence of strong pink noise and babble noise. Compared to "plain" speech, the word recognition rate was most improved by pressed phonation, followed by an increased f_0 mean and f_0 range, and increased formant frequencies. Increased duration of vowels slightly reduced the recognition rate for pink noise but had no effect for babble noise.

3.1 Method

Using the articulatory speech synthesizer VocalTractLab, each of the ten German words for the digits “0” to “9” (0 /nʊl/, 1 /aɐ̯ns/, 2 /tsʏaɐ̯/, 3 /dʁaɐ̯/, 4, /fiːᵊ/, 5 /fʏnf/, 6 /zɛks/, 7 /zˈiːbn̩/, 8 /axt/, 9 /nɔœ̯n/) was synthesized in 16 variants. The 16 variants represented all combinations of four binary features, namely, phonation type, formant frequency, f_0, and duration. Each feature had two possible settings, one typical for plain speech and one typical for Lombard speech. Hence, for each digit word, there was one variant with all features of plain speech, one variant with all features of Lombard speech, and 14 variants with a mixture of features typical for plain speech and Lombard speech. To analyze the potential intelligibility benefit of the different feature combinations, a group of listeners was asked to identify the digit words in the presence of pink noise and in the presence of babble noise. In the following, the articulatory speech synthesizer, the creation of the stimuli, and the procedure of the perception experiment are presented in more detail .

3.1.1 Articulatory speech synthesizer VocalTractLab

VocalTractLab (VTL) is an articulatory speech synthesizer that is capable of generating a full range of speech sounds in high quality while providing full control of time-varying glottal and supraglottal articulation. Supraglottal articulation is modeled by means of a 3D geometric model of the vocal tract of an adult male speaker [20]. This model is controlled by 23 parameters that specify the shape and position of the articulators. The glottis is modeled by means of a geometric vocal fold model [65], which is a recent extension of VTL 2.2 and allows more precise control of the glottal geometry than the self-oscillating bar-mass model used in previous studies [66]. The vocal fold model is controlled by ten parameters, which specify subglottal pressure, fundamental frequency, the shape of the glottis at rest, and oscillatory features such as the phase lag between the lower and upper vocal fold edges and the skewness of the glottal area pulses. For the synthesis of speech, the models of the vocal tract and the vocal folds are transformed into a unified 1D tube model of the vocal system (also including the subglottal system and the nasal cavity). This tube model is the basis of an aero-acoustic simulation in the time domain [67, 68].

The parameters of the vocal tract and vocal fold models are controlled by means of a gestural

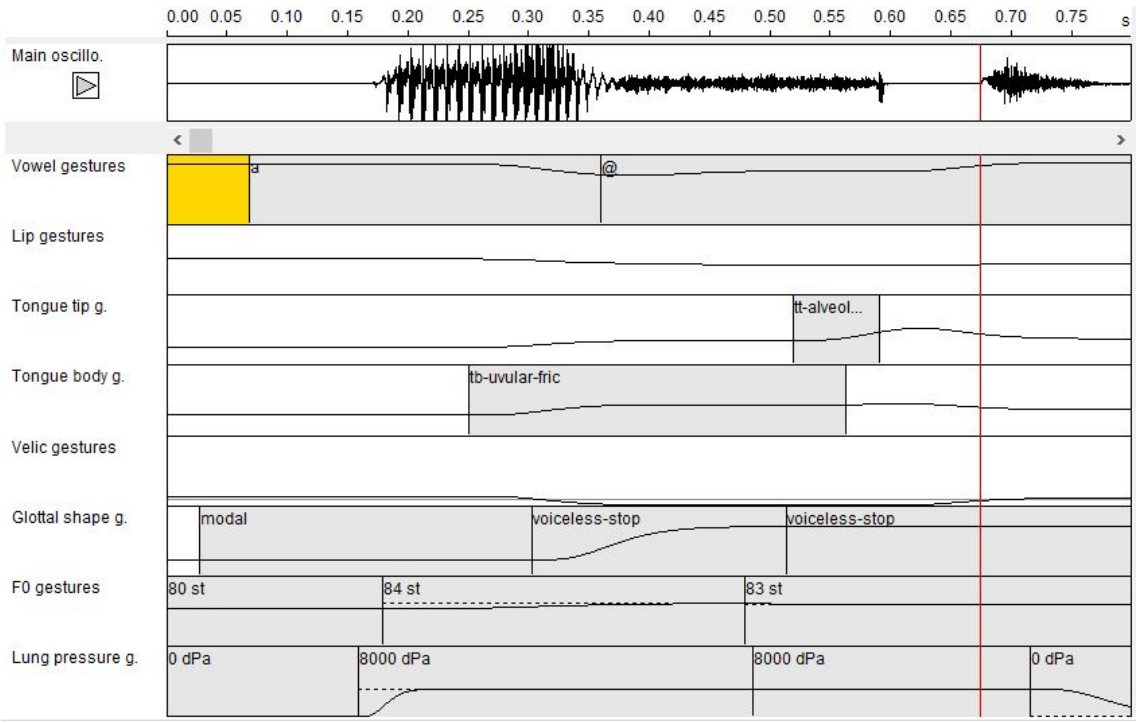


Figure 3-1: Gestural score and synthesized waveform for the plain speech variant of the German word for digit 8 (/axt/).

score (similar to a musical score) [69], which is a high-level concept for speech movement control based on the ideas of articulatory phonology [70]. In these scores, the articulatory gestures required to generate an utterance are specified and temporally coordinated. In VTL, gestural scores are created by means of a graphical editor as shown in Figure 3-1. The score has eight tiers, five of which define the supraglottal gestures (vowel gestures, lip gestures, tongue tip gestures, tongue body gestures, and velic gestures) and three of which define the laryngeal and pulmonary gestures (glottal shape gestures, f_0 gestures, and lung pressure gestures). As an example, Figure 3-1 shows the temporal coordination of the gestures required for the German word "acht" (/axt/, engl.: eight).

3.1.2 Creation of stimuli

The stimuli for the perception experiment were created in three steps:

(1) For each German digit word, a recorded natural utterance of that word spoken with a "plain" speaking style was resynthesized in terms of a gestural score similar to [71]. In the resynthesized utterances, the phone durations and the f_0 contours were closely fitted to those of the natural utterances. The exact acoustic realization of the individual phones was determined

Table 3.1: Plain speech settings and Lombard speech settings of four examined features used for articulatory speech synthesis.

Feature	Setting for plain speech	Setting for Lombard speech
Phonation type	Modal voice & 800 Pa lung pressure	Pressed voice & 1600 Pa lung pressure
Formants	Standard formant values in VTL	F_1 increased by 25%, F_2 increased by 5%
Fundamental frequency	Reproduced from natural plain speech	f_0 mean increased by 5 st, f_0 range increased by the factor 1.3
Phone durations	Reproduced from natural plain speech	Durations of vowels increased by 30%

by the corresponding predefined settings (shapes) of the vocal tract and vocal fold models. For all ten words, modal phonation and a subglottal pressure of 800 Pa was used. As a result, the resynthesized words had all typical features of plain speech.

(2) The gestural scores created in (1) were used to generate the remaining 15 variants of each digit word by changing the phonation type, f_0, formant frequencies, and phone duration (either individually or jointly) to a setting typical for Lombard speech. How exactly the features were adjusted is detailed in the adjustments below. Table 3.1 gives an overview of the plain speech settings and the Lombard speech settings for the four features. All speech items were synthesized as 16-bit mono signals with a sampling frequency of 22,050 Hz. The amplitude of the synthetic items was *not* normalized so that the inherent amplitude differences between the items due to the different feature settings (e.g., modal vs. pressed voice) were maintained.

(3) All 160 speech items (10 digit words $\times$ 16 variants) were combined with two types of noise as detailed below to create the stimuli for the perception experiment.

a. Adjustment of phonation type

In Lombard speech, the spectral tilt is flatter than in plain speech, i.e., the higher-frequency components are enhanced. To cause spectral flattening for the synthetic words, the parameters of the vocal fold model [65] were adjusted to generate a more pressed voice quality. The main vocal fold model parameters that affect the voice quality on the continuum from a modal to a pressed voice are the (pre-phonatory) rest displacement x_{rest} of the vocal folds at the level of the arytenoids, the area A_{chink} of a permanent glottal chink between the ary-

tenoids, and the subglottal pressure P_{sub}. The settings for modal phonation (for plain speech) were the "standard" values $x_{\text{rest}} = 0.3$ mm, $A_{\text{chink}} = 2$ mm^2, and $P_{\text{sub}} = 800$ Pa. In contrast, pressed phonation was generated with $x_{\text{rest}} = 0$ mm, $A_{\text{chink}} = 0$ mm^2, and $P_{\text{sub}} = 1600$ Pa, i.e., with no glottal rest area and twice the subglottal pressure used for modal phonation. In the absence of any published measurements of subglottal pressure during Lombard speech, the value of 1600 Pa was chosen to be clearly higher than that of plain speech to reflect the higher vocal effort, but also clearly below the maximum lung pressures of around 6 kPa that humans can produce in extreme situations [72].

With these settings, the average spectral tilt across all ten of the digit words (as approximated by a regression line to the long-term average spectral magnitude in dB between 0 and 4 kHz) was -9.23 dB/oct for the modal voice and -3.55 dB/oct for the pressed voice. Hence, from modal to pressed voice, the spectral tilt increased by 5.68 dB/oct, which is close to the typical difference of 6 dB between these phonation types [37]. Furthermore, the average sound pressure level (SPL) of the digit words synthesized with pressed voice increased by 10.05 dB compared with modal voice, which is in the range of an 8 to 15-dB difference in SPL between plain and Lombard speech as observed by [14].

b. Adjustment of fundamental frequency

Compared to plain speech, Lombard speech is characterized by both an increased f_0 mean and range. The data of [28, Fig. 2b] indicate that for very high background noise levels, f_0 increases by about 5 st compared with plain speech. Furthermore, the data by [27] show that the f_0 range of Lombard speech is 1.2 to 1.8 times the f_0 range of plain speech (on the Hz scale). Accordingly, to model the change in f_0 due to the Lombard effect in the synthesizer, the mean f_0 and the f_0 range of the reference gestural scores were increased by 5 st and a factor of 1.3, respectively. This change was implemented by modifying the f_0 target offsets of the target approximation model [73], which is the f_0 model used in VTL.

c. Adjustment of formant frequencies

Multiple studies found that Lombard speech has increased frequencies of F_1 and F_2 compared with plain speech. The values estimated from [28] and used here are a 25% increase of F_1 and a 5% increase of F_2. To implement a change of the formant frequencies in VTL, the vocal tract target shapes of the corresponding vowels had to be adjusted. The target shapes

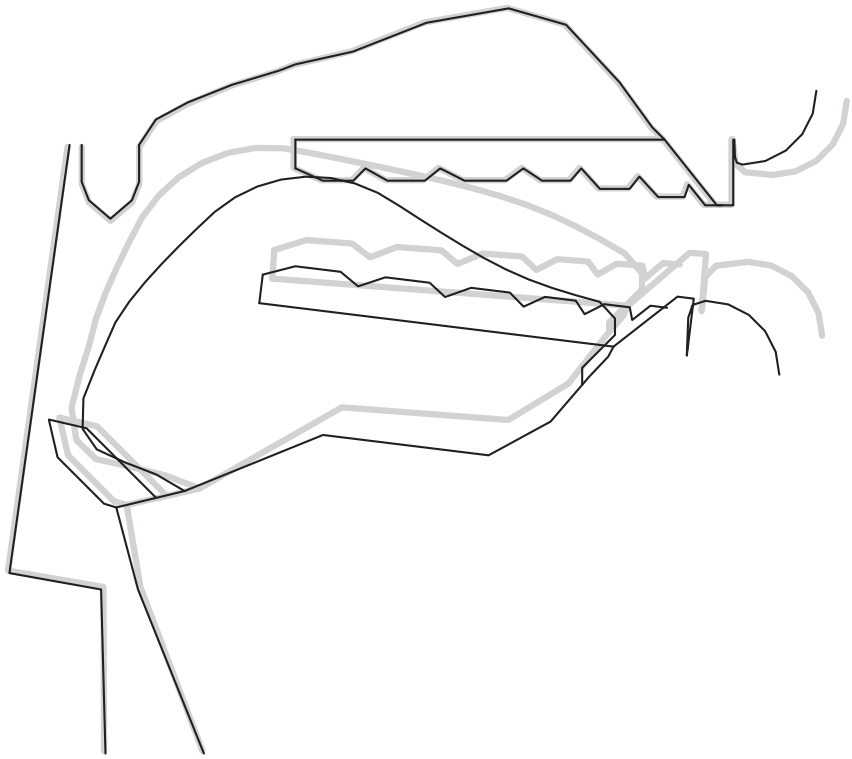


Figure 3-2: Midsagittal shapes of the vocal tract model for the vowel /a/ with normal (standard) articulation (gray lines) and Lombard articulation (black lines).

of the vowels occurring in the ten digit words were adjusted in two steps. First, starting with the standard shapes of the vowels, the mouth cavity was opened more (as for shouting) by manually adjusting the vocal tract parameters of the jaw, the lips, and the tongue. The jaw angle was decreased by 3° (parameter JA), the distance between the lower and upper lips was increased by 5 mm (LD), and the tongue position was lowered by 5 mm (TTY, TBY, TCY). While this manual adjustment increased F_1 for all vowels, the increase was never exactly 25%, and the change of F_2 was rather unpredictable. Therefore, as the 2^{nd} step, a greedy optimization algorithm [20] was used to fine-tune all vocal tract model parameters in such a way that the resulting formant frequencies assumed the intended values. After optimization, the deviation of the formant frequencies from the intended values was 1.5% ± 1.5% across all vowels. As an example, Figure 3-2 shows the standard vocal tract shape (corresponding to plain speech) of the vowel /a/ in gray and the adjusted shape with F_1 increased by 25% and F_2 increased by 5% in black.

d. Adjustment of duration

In Lombard speech, vowels are on average longer, and consonants tend to be slightly shorter

than in plain speech [26, 28]. According to [28, Fig. 2a], the increase of vowel duration converges to about 30% for very high background noise levels. As this change is much greater than the change of consonantal durations, it was chosen to model the durational changes due to the Lombard effect by increasing the durations of all vowels in the gestural scores by 30%. This was achieved by "stretching" the scores by the appropriate durations around the acoustic midpoints of the vowels contained in the utterances.

e. Addition of noise

As each German digit word was synthesized in 16 variants, there were 160 (clean) synthetic speech items in total. A second set of 160 items was generated by adding pink noise [74] to the clean speech items, and a third set of 160 items was generated by adding babble noise [75] to the clean speech items. Both types of noise are common in daily life and have a speech-like overall spectral shape (see Figure 3-3). Furthermore, they represent both a kind of stationary noise (pink noise) and a kind of non-stationary noise (babble noise). The babble noise was generated by 100 people speaking in a cafeteria with individual voices still slightly audible. For each kind of noise, the amplitude was chosen in such a way that the sound pressure level (SPL) of the noise was 20 dB higher than the average SPL of the ten synthesized digits in the plain speaking style. The SPL difference of 20 dB is equivalent to the 84 dB absolute noise level that was used to induce Lombard speech in the database of [28] and [14]. For the perception experiment, the audio files of the speech items superimposed with noise had a total length of 2 s with the target words embedded in the middle.

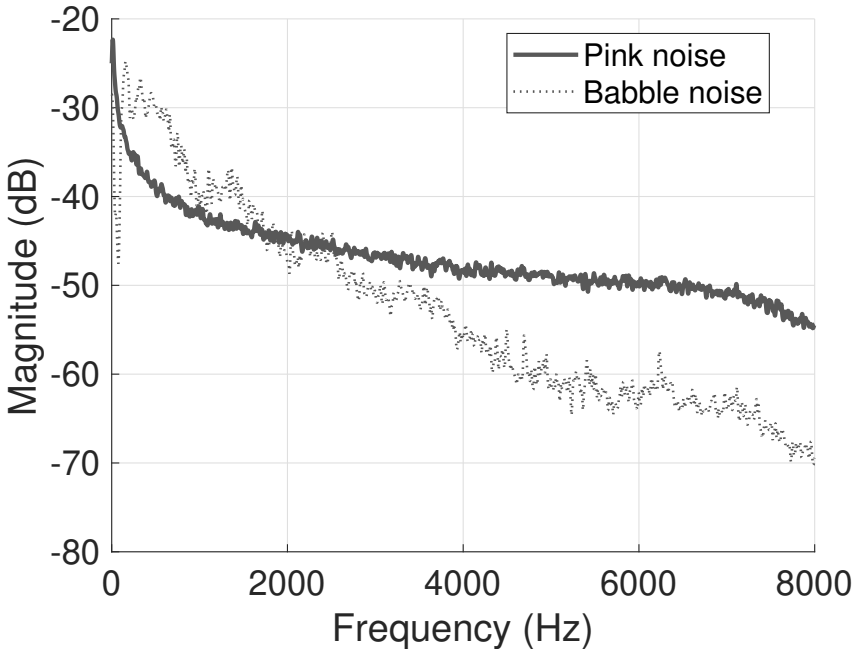


Figure 3-3: Long-term average spectra of additive noise from 0-8000 Hz with frequency resolution of 15 Hz.

3.1.3 Perception experiment

The perception experiment had two tasks. The first task was an evaluation of the naturalness of the synthetic utterances (without background noise), and the second task was a test for intelligibility. Seventeen native Germans, including 13 men with an average age of 32.5 years and a standard deviation of 9.8 years, and 4 women with an average age of 40.3 years and a standard deviation of 11.1 years, participated in the tests. All participants gave informed consent and reported no hearing problems. Each participant performed the two tasks in two consecutive sessions.

For the evaluation of the naturalness of the synthesis (first task), each participant listened to the 160 stimuli without added background noise in random order using high-quality headphones (AKG K240) connected to a desktop computer via a FireWire audio interface (MOTU 896HD) in a soundproof room. The volume for the headphones was adjusted to make all stimuli without added noise comfortably audible for the subjects. After each stimulus was played, the participants were asked to rate the naturalness among four options, 1 - unnatural, 2 - rather unnatural, 3 - rather natural, or 4 - natural, by clicking on one of four buttons with the respective labels. After choosing an answer to the current stimulus, the next stimulus was automatically played (repetitions were not possible). The session took about 12 minutes.

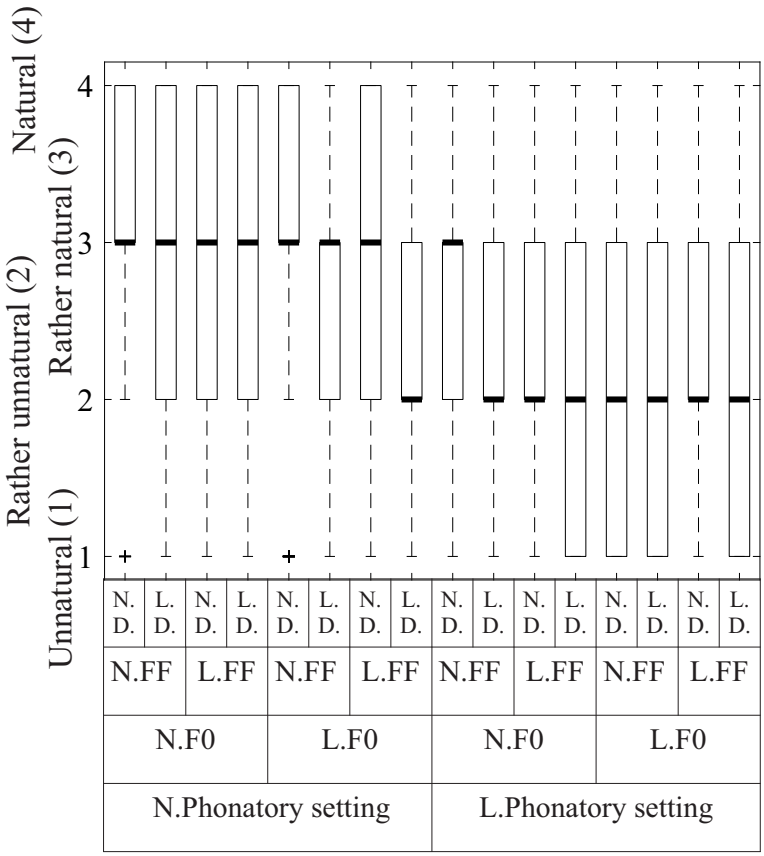


Figure 3-4: Box plots of naturalness ratings of all 16 word variants, i.e., feature combinations. The labels below boxes indicate feature combinations, where "N." stands for the neutral setting of a feature (as in plain speech), "L." stands for the Lombard setting of a feature, "D." stands for duration, "FF" for formant frequency, and "F0" for fundamental frequency.

After a short break, the participants started the second session for the second task. In this task, each participant listened to all 480 stimuli (160 speech items in three conditions: pink noise, babble noise, and no noise) in random order by using the same equipment as in the first session. After each stimulus was played, the participants had to click on one of ten buttons on the computer screen that represented the perceived digit word. If they could not clearly understand the spoken digit, they were allowed to randomly choose one of the digits. After choosing the answer to the current stimulus, the next stimulus was automatically played (repetitions were not possible). Halfway through this session, the participants took a short break of 2 minutes. The whole session lasted about 35 minutes.

Table 3.2: Feature combinations of the groups of stimuli that were compared with respect to the perceived naturalness of the stimuli. The stimuli in group A represent plain speech. The stimuli in the groups B, C, D, and E differ in one feature each from group A.

Group	Type of stimuli
A	Modal phonation, normal f_0, normal formants, and normal durations.
B	**Pressed phonation**, normal f_0, normal formants, and normal durations.
C	Modal phonation, **Lombard f_0**, normal formants, and normal durations.
D	Modal phonation, normal f_0, **Lombard formants** and normal durations.
E	Modal phonation, normal f_0, normal formants, and **Lombard durations**.

3.2 Results and discussion

3.2.1 Perceptual test of naturalness

The perceptual ratings of the naturalness of the synthetic stimuli are shown in Figure 3-4, with one boxplot for each of the 16 feature combinations. The leftmost boxplot represents the ratings of the stimuli with all features of plain speech (median = 3), and the rightmost boxplot represents the ratings of the stimuli with all features of Lombard speech (median = 2). As can be seen, the feature combinations affected the ratings, and the stimuli with the settings for plain speech were among the most natural sounding items.

To test the effect of individual features on the naturalness ratings, five groups of stimuli were formed (see Table 3.2). Four two-tailed Mann-Whitney U tests were performed to compare the response distributions of groups A vs. B, A vs. C, A vs. D, and A vs. E. The response distributions of groups A and B differed significantly (Mann-Whitney $U = 9739.0$, $N1 = N2 = 170$, $p < 0.001$), i.e., the stimuli with pressed phonation were perceived to be more unnatural than the stimuli with modal phonation. The response distributions of groups A and C did *not* differ significantly (Mann-Whitney $U = 13393.0$, $N1 = N2 = 170$, $p > 0.05$), i.e., the f_0 setting had no effect on the naturalness. The response distributions of groups A and D (Mann-Whitney $U = 12494.5$, $N1 = N2 = 170$, $p < 0.022$) and groups A and E (Mann-Whitney $U = 12537.5$, $N1 = N2 = 170$, $p < 0.025$) were both significantly different; hence, the Lombard-typical settings of the formants and durations also caused a slight decrease of the naturalness.

The decreased naturalness of the stimuli with the Lombard-typical phonation type, formants, and durations may be in part due to the fact that the raters listened to the stimuli in the absence

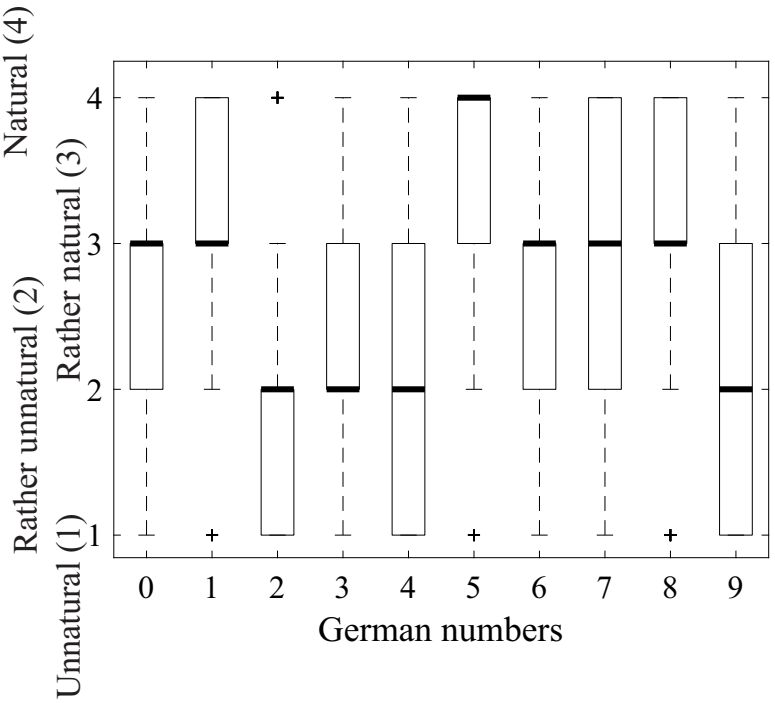


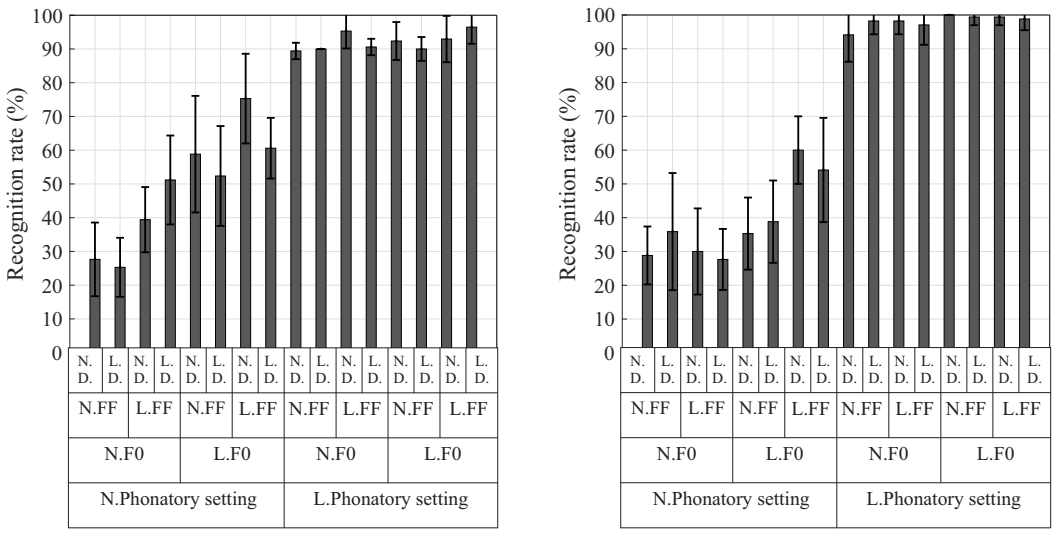
Figure 3-5: Naturalness ratings of individual digit words, pooled across all 16 variants.

of noise, while they are normally used to perceive Lombard-typical features under noisy conditions. Another reason may be a non-perfect simulation of the according features.

Figure 3-5 shows how the naturalness varied across the individual digit words. According to the median response values, six of the ten words were perceived as "natural" or "rather natural", while four words were perceived as "rather unnatural". The reasons for the ratings of the words for 2, 3, 4, and 9 as rather unnatural are hard to tell. A retrospective informal comparison of the synthetic words indicated that the modeled f_0 contour of the four words with the lower ratings might have been somewhat atypical.

3.2.2 Perceptual test of intelligibility

The results of the intelligibility test are shown in Figures 3-6 and 3-7 in terms of recognition rates. Figure 3-6 shows how the 16 different feature combinations affected the recognition rates in the presence of pink noise and babble noise, respectively. The recognition rates of the digit words without background noise are not explicitly shown as they were very close to 100%. In general, the more Lombard-typical features the stimuli contained under the noisy conditions, the higher the recognition rates.



(a) Pink noise

(b) Babble noise

Figure 3-6: Recognition rates for all 16 feature combinations in presence of pink noise and babble noise, pooled across all digit words and listeners. The labels indicate feature combinations, where "N." stands for the neutral setting of a feature (as in plain speech), "L." stands for the Lombard setting of a feature, "D." stands for duration, "FF" for formant frequency, and "F0" for fundamental frequency.

To study the effects of the four features on the recognition rate in more detail, a four-way repeated measures ANOVA was performed. The features phonation type, f_0, formant, and duration were the four factors, each having two levels (the normal setting and the Lombard-typical setting). The dependent factor was the recognition rate. Two individual ANOVAs were performed, one for the case of pink background noise and one for the case of babble background noise.

In the case of *pink noise*, there were significant main effects for all four factors, that is, phonation type [$F(1,16) = 724.5$, $p < 0.001$], f_0 [$F(1,16) = 101.5$, $p < 0.001$], formant [$F(1,16) = 82.3$, $p < 0.001$], and duration [$F(1,16) = 4.7$, $p = 0.045$]. However, while the Lombard settings of phonation type, f_0, and formant had a *positive* effect on the recognition rate, the Lombard-typical (i.e. longer) durations *reduced* the mean recognition rate (from a mean of 71.40% across all samples with normal vowel durations to a mean of 69.56% across all samples with increased vowel durations). The effect size was strongest for phonation type ($\eta^2 = 0.978$), followed by f_0 ($\eta^2 = 0.864$), formant ($\eta^2 = 0.837$), and duration ($\eta^2 = 0.229$). In addition, there were multiple significant interactions between factors, namely, between phonation type and any of the f_0, formants, and durations, between f_0 and duration, and between phonation

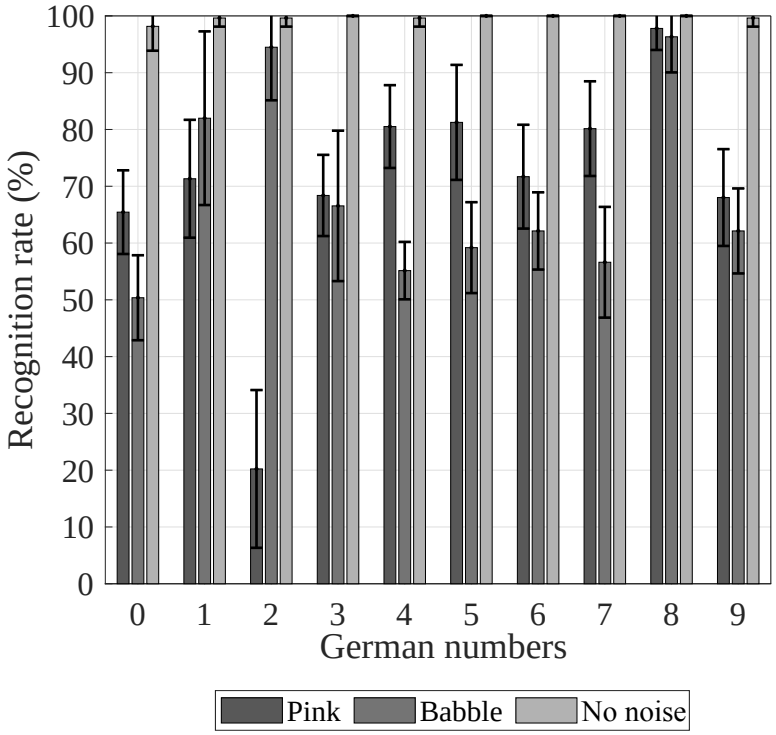


Figure 3-7: Recognition rates of individual German digit words across all speech variants, separated by noise conditions: pink noise, babble noise, and no noise. Bar heights indicate mean values, and error bars indicate standard deviations.

type, formant, f_0, and duration. Among these, the interaction between phonation type and f_0 was strongest ($F(1, 16) = 108.0$, $p < 0.001$).

In the case of *babble noise*, there were significant main effects for three of the factors, that is, phonation type [$F(1, 16) = 1284.8$, $p < 0.001$], f_0 [$F(1, 16) = 124.4$, $p < 0.005$], and formant [$F(1, 16) = 17.8$, $p = 0.001$], i.e., their Lombard-typical settings increased the recognition rate. However, in contrast to the pink noise case, there was no significant effect for the factor duration ($p > 0.5$). The effect size was strongest for phonation type ($\eta^2 = 0.988$), followed by f_0 ($\eta^2 = 0.886$) and formant ($\eta^2 = 0.527$). In addition, there were significant interactions between phonation type and formant, between phonation type and f_0, between formant and duration, and between f_0 and phonation type. Among these, the interaction between phonation type and f_0 was strongest ($F(1, 16) = 75.4$, $p < 0.001$).

Figure 3-7 shows the recognition rates separated by the digit words and the noise conditions. It illustrates that the digit words were almost perfectly recognized without background noise and that the recognition rates were generally higher in the presence of pink noise than in babble

noise. A notable exception is the word /tsʏ̥aɛ̯/ (digit 2), which was badly recognized in pink noise. The reason is probably that the initial consonant cluster /tsʏ̥/ mainly consists of wideband noise which is hard to perceive in pink noise due to similar noise characteristics.

In summary, it was found that the change of phonation type from a modal to a pressed voice improved the intelligibility of the words most, independently from the type of background noise. Given that the effect of a more pressed voice quality is a flattening of the spectral tilt, this finding is in line with the previous findings of [42] and [44]. However, it was also found an increase of f_0 to be highly effective at increasing the intelligibility in both types of noise, which contradicts the findings by [42]. The reason may be that the greatest mean f_0 increase examined by [42] was only 2.5 st (from 148 Hz to 171 Hz), while 5 st were used basing on the data by [28]. The present study also proved that there was a beneficial effect of an increase of the first two formant frequencies of vowels, which had so far not been explicitly shown in analysis-by-synthesis experiments. Finally it was found that an increase of the duration of voiced sounds did not improve word intelligibility in noise. In fact, in pink noise, the durational increase even led to slightly worse intelligibility. However, what is the reason that increased durations are frequently observed in natural Lombard speech then? One reason is probably the wider extent of the articulatory movements in Lombard speech (e.g., generally lower tongue positions in vowels), which takes more time. Another reason is probably that longer phone durations *can* improve the intelligibility in noise but only for certain types of fluctuating noise [46].

The main limitations of this investigation are the following:

1. Only two values per feature were analyzed (one value for plain speech and one value typical for Lombard speech). This led to a ceiling effect of the recognition rates for certain feature combinations. For example, according to Figure 3-6b, the recognition rate was almost 100% for all stimuli with pressed phonation in the presence of babble noise, independently from the other feature values. Future studies could investigate the intelligibility benefit of multiple values in smaller steps along each feature dimension in more detail, or use multiple different levels of additive noise.

2. The speech material was limited to the ten German digit words, so the results may not directly translate to longer utterances or different languages. However, the used words do contain 8 different vowels and 11 different consonants, which cover roughly half of the German phonemes. Therefore it would be expected similar results for languages with

a phoneme system similar to German. For longer utterances, e.g., sentences, it would generally expect better recognition rates, because more context helps to disambiguate individual words that are strongly masked by noise.

3.3 Summary

In this chapter, articulatory speech synthesis as used to generate synthetic words with different combinations of articulatory-acoustic features and explored their individual and combined effects on the intelligibility of the words in pink noise and babble noise. It was found that using a pressed voice quality (i.e., flattening the spectral tilt), a higher lung pressure, increasing f_0, and increasing F_1 and F_2 all enhance the intelligibility to different degrees. Furthermore, the beneficial effect of these features is generally additive, e.g., increasing both f_0 and formant frequencies improves the intelligibility more than either feature alone. However, increasing vowel durations has no positive effect. As a result, these articulatory features confirmed the mechanisms that controlling the glottal source should be independent of the vocal tract as the phonation of pressed voiced was controlled independently of the formants. Furthermore, the adjustment of formants showed that the lower jaw and tongue for Lombard speech might make the vocal tract shorter and affect other formants not only F_1 and F_2. The articulatory features corresponded to the acoustic features of spectral tilt, f_0, formants, and some other temporal features (e.g., power envelope by lung pressure). Nonetheless, this result was still by one variation of Lombard speech under one noise level. It needed to expand the investigation in more noise levels and variable environments to verify whether this result was still valid and explore more, which was shown in the next chapters.

Chapter 4

Mimicking Lombard speech by acoustic features in varying noise levels

This chapter describes the mimicking Lombard speech by using acoustic features in varying noise levels aiming to independently control and understand how the contribution of acoustic features on the intelligibility and naturalness of speech in backgrounds with varying noise levels are.

To adaptively control the intelligibility of transmitted speech, *perceptually* mimicking Lombard speech under backgrounds with varying noise levels was performed. Other approaches map corresponding plain speech features to Lombard speech features, but as this can only be applied to one noise level at a time, it is unsuitable for varying noise levels because the characteristics of Lombard speech are varied according to noise level. Instead, a rule-based method that automatically generates rules and flexibly controls features with any change of noise level was utilized. Specifically, a feature tendency analysis was conducted with concerns of the results from the previous chapter for the mechanisms to produce Lombard speech. Then, a continuous rule generation model was proposed to estimate the effect of varying noise levels on features. The proposed techniques, which are based on a coarticulation model, MRTD, and spectral-GMM, can easily modify plain speech features by following the generated rules. Voices having these features are then synthesized by STRAIGHT to obtain Lombard speech fitting to noises with varying levels. To validate the proposed method, the quality of mimicking speech is evaluated in subjective listening experiments on similarity, intelligibility, and naturalness. In varying noise levels, the results show equal similarity with Lombard speech between the proposed method

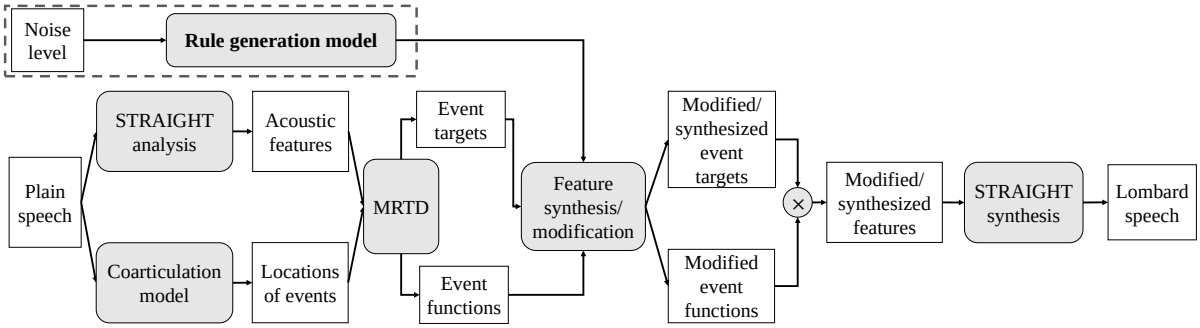


Figure 4-1: Outline of the analysis/synthesis methods used in the mimicking of Lombard speech at multiple noise levels.

and a state-of-the-art method. Intelligibility and naturalness are comparable with some feature modifications.

4.1 Methodology

In this investigation, an analysis and synthesis methodology outlined in Fig. 4-1 was used to convert plain speech to Lombard speech fit with varying noise levels. The approach is a combination of STRAIGHT [76], a coarticulation model, MRTD [22], and a newly designed rule generation model based on noise levels. STRAIGHT is a high-quality vocoder that extracts acoustic features from speech and then uses these features to synthesize speech. The coarticulation model (see Fig. 4-2), which is adopted from Nghia et al.'s study [21], represents the coarticulation effect between two adjacent phonemes, which is important for naturalness. In a phoneme, the model is the supposition of a nucleus interval and two coarticulated intervals at two sides. Five locations including two boundaries, two transitions, and a nuclei center are thus identified. The nuclei center is the minimal point of the spectral transition rate (STM) of the phoneme, i.e., the most stationary point of the phoneme. Transitions are respectively the minimal and maximal points of the derivatives of each half of the STM, i.e., the most transitional points of the phoneme. In this study, these locations are considered precise locations of acoustical events for further processing in MRTD. In addition, the transitions and nuclei center represent the transitional and stationary locations that are the best locations to extract spectral dynamics and static spectral targets of phonemes, and as such, they became the locations to analyze and modify/synthesize features.

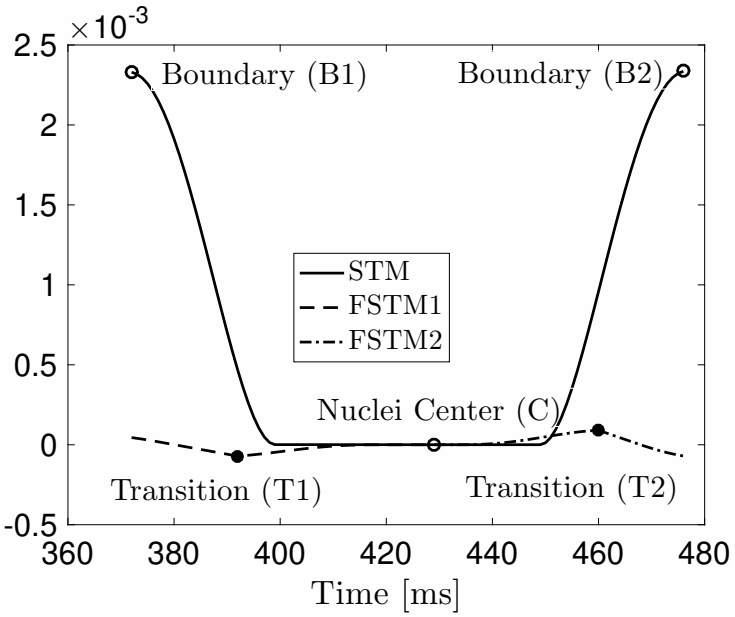


Figure 4-2: Locations to extract event targets in temporal decomposition, based on coarticulation model. STM indicates the spectral transition rate of a phoneme. FSTM1 and FSTM2 are respectively derivatives of STM on the first and second halves. These rates represent both spectral dynamics and static spectral targets. The spectral dynamics, known to be context-sensitive, contain a lot of phonetic information of speech, which is crucial for speech intelligibility [77]. The static spectral targets, which contain linguistic-phonetic and non-paralinguistic information of speech, are important for speech intelligibility and naturalness [77].

The MRTD [22] decomposes/interpolates acoustic features into/from temporal information (event functions) and acoustical parameters (event targets) at the event locations estimated by the coarticulation model. Regarding the acoustical events in the coarticulation model, MRTD event targets are related to the context-insensitive static features, and MRTD event functions are related to context-sensitive dynamic features. The context-sensitive transition movements are represented by two overlapping event functions, and thus they can be modified to fit with a new context such as lengthening or shortening. The static context-insensitive event targets, representing the context-independent characteristics of phonemes, are expected to be stable and reliable, and thus they have to be modified by following the characteristics of the decomposed features.

The MRTD is mainly used for spectral features to ensure coding efficiency. In this study, the model is extended to include other features. To represent co-articulation better, the event functions of a spectral feature are used to interpolate all features. To modify/synthesize acoustic features, the event targets of each feature are modified or synthesized according to the feature characteristics and the event functions of the spectral feature and then multiply them together .

In this way, an interpolation of temporal information can be also obtained for modified acoustic features to maintain naturalness.

The most important element here is the *rule generation model* (see at the end of Sect. 4.2.2) for adaptability, which takes a noise level as input to give corresponding values of acoustic features as output. This model is constructed after the acoustical analysis of plain and Lombard speech. With these components, the present method can deal with an adequate number of acoustic features, model acoustical events precisely, modify features more flexibly and easily, and continuously estimate and apply the effect of noise level to acoustic features. It shows great potential to obtain high-quality synthesized speech adapted to noisy backgrounds.

4.2 Feature analysis and rule generation for synthesis/modification

4.2.1 Speech dataset

Recorded speech by two speakers (one male, one female) in the pink noise levels of $-\infty$ (plain speech) and 66, 72, 78, 84, and 90 dB (Lombard speech) sampled at 16 kHz was taken from a previous study that examined the intelligibility of Lombard speech [28]. Three familiarity-controlled word lists, F0W7 [78] (60 words of 0-type of pitch accent pattern), with the lowest familiarity rank (1.0 – 2.5) were used. Each word consists of four morae (e.g., sa sa wa ra) and was embedded in a carrier sentence as a target word: "Tsugi ni yomu tango wa" word "desu". These four-mora words were manually segmented and then used in the analyses.

4.2.2 Feature analyses

The same dataset described above was used in the acoustical analysis in the previous study [28]. In that study, it was found that the well-known tendencies of lengthening vowel duration, increasing f_0, shifting F_1, and decreasing spectral tilts (A1-A3) were still present with increasing noise levels. Also, there was an abrupt change in F0 at 84 dB, increasing formant amplitudes, and H1-H2 variation, and a raised modulation spectrum. In the present study , three features (spectral sequence, f_0, and aperiodicity (Ap)) were extracted by STRAIGHT. Analyses of the acoustic features corresponding to these three features and other features in the previous studies

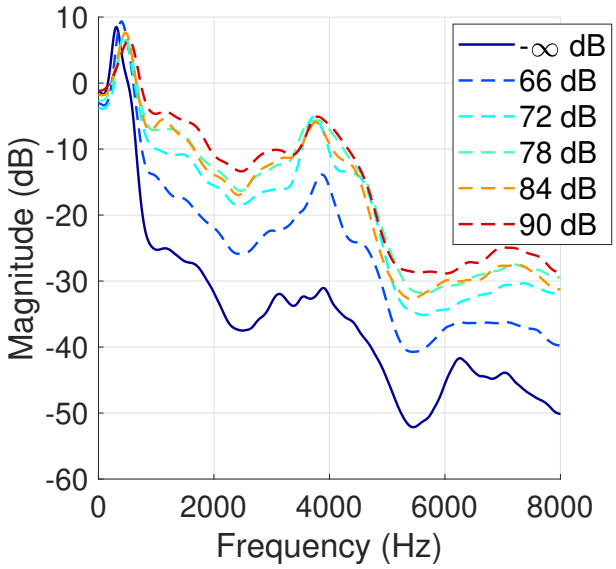


Figure 4-3: Spectral envelope of plain ($-\infty$ dB, solid line) and Lombard speech (66-90 dB, dashed lines) at the nuclei center (C) of vowels. A plateau between 2-6 kHz appears in Lombard speech.

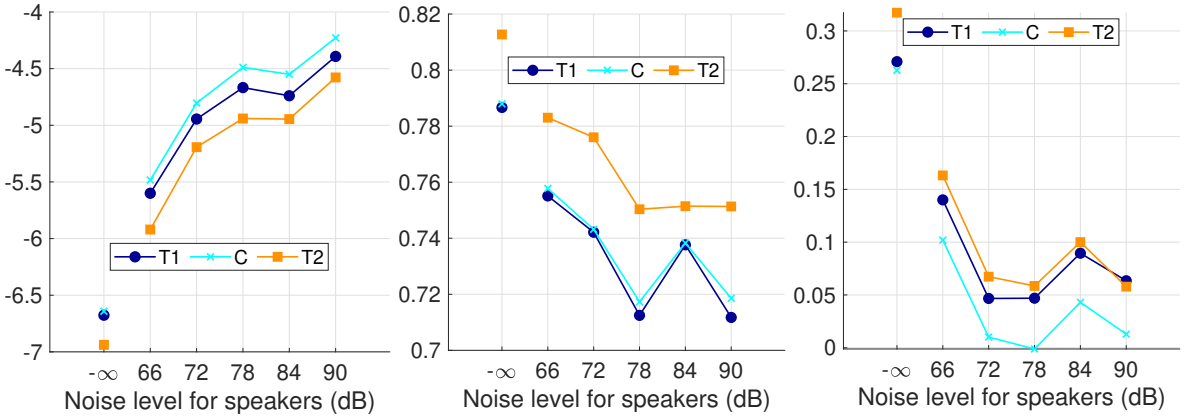
[28] with concerns of the results from the previous chapter in feature definitions were carried out.

- Spectral envelope

  The spectral envelope at three locations (transitions and nuclei center) was taken from the spectral sequence. As was the result that controlling the glottal source should be independent of the vocal tract, in order to independently analyze and control the features of the spectral envelope, spectral envelope $X(\omega)$ was decomposed into spectral tilt $T(\omega)$ and vocal tract spectra $\frac{B(\omega)}{A(\omega)}$ as a concept of the source-filter model, as

  $$log|X(\omega)| = T(\omega) + log\frac{B(\omega)}{A(\omega)}. \tag{4.1}$$

  – Spectral tilt: Further analysis indicated that a big plateau between 2-6 kHz existed in the Lombard speech (see Fig. 4-3). Under a 16-kHz sampling frequency, to include this information into spectral tilt, the first three cepstral coefficients were used, particularly the third cepstral coefficient. Spectral tilt $T(\omega)$ was thus a smoothed log magnitude spectrum estimated by cepstrum [see Eq. (4.2)], and each cepstral coefficient was estimated by discrete cosine transform type 2 (DCT-II) of log spectrum $log|X(\omega)|$ [see Eq. (4.3)]:



(a) c_0 (b) c_1 (c) c_2

Figure 4-4: Analysis results of cepstral coefficients of vowels. T1, T2 are the transitions. C is the nuclei center.

$$T(\omega) = c_0 + 2c_1 cos(\omega) + 2c_2 cos(2\omega), \tag{4.2}$$

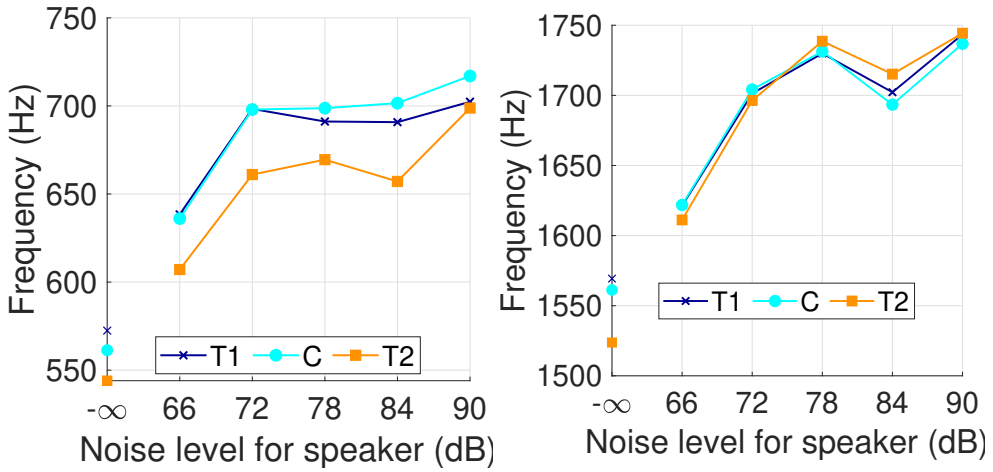
$$c_n = \frac{1}{N} \sum_m log|X(\omega)| cos(n\omega), \tag{4.3}$$

where $\omega = \frac{2\pi k}{2(N-1)}; k = 0, 1, ..., N-1$.

Figure 4-4 shows the analysis results of these cepstral coefficients at transitions and the nuclei center of vowels.

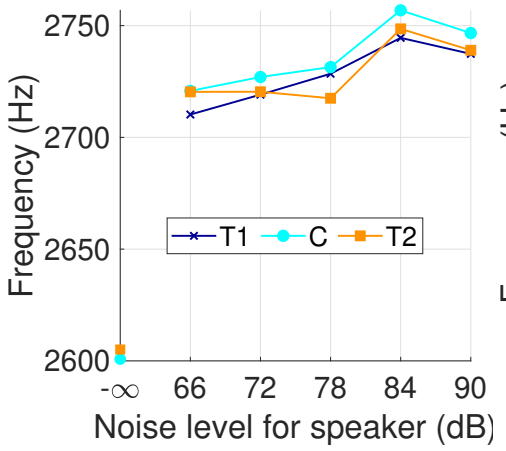
- Formants: The formant frequencies were estimated by KARMA [79]. In the previous study [28] and in the previous chapter, shifting in F_1 and F_2 to higher frequency regions was observed. With an assumption of the appearances of the plateau in the spectral tilt, the pattern of formant amplitudes was skipped. Further analyses on F_3 and F_4 were also conducted. Figure 4-5 shows the analysis results of these four formants.

- Vocal tract length (VT length): As was indicated in the result of the previous chapter vocal tract length can be shortened due to increasing formant frequencies, to the best of the recent knowledge, changes of VT length can be calculated on the basis of the ratio between f_0 and the average of the first four formants [80, 81]. Figure 4-6 shows the analysis results of this relationship.

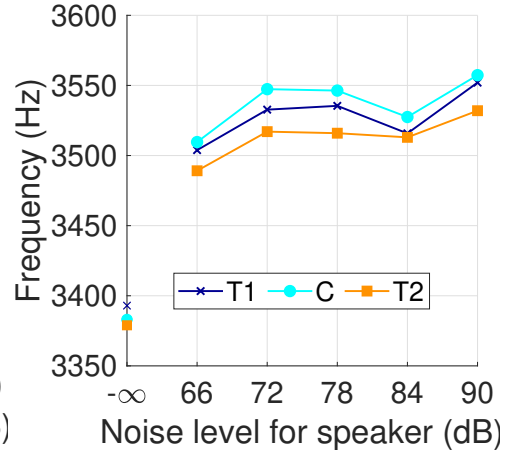


(a) F_1

(b) F_2



(c) F_3



(d) F_4

Figure 4-5: Analysis results of formants. T1, T2 are the transitions. C is the nuclei center.

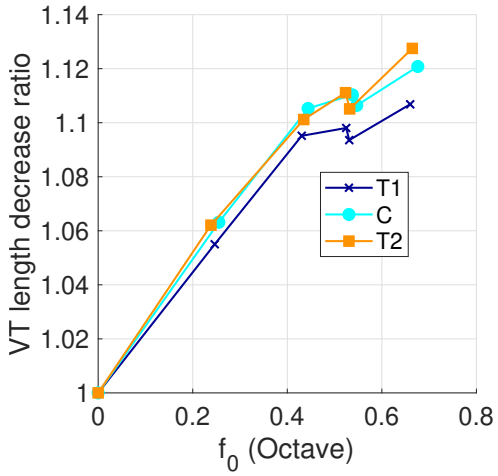
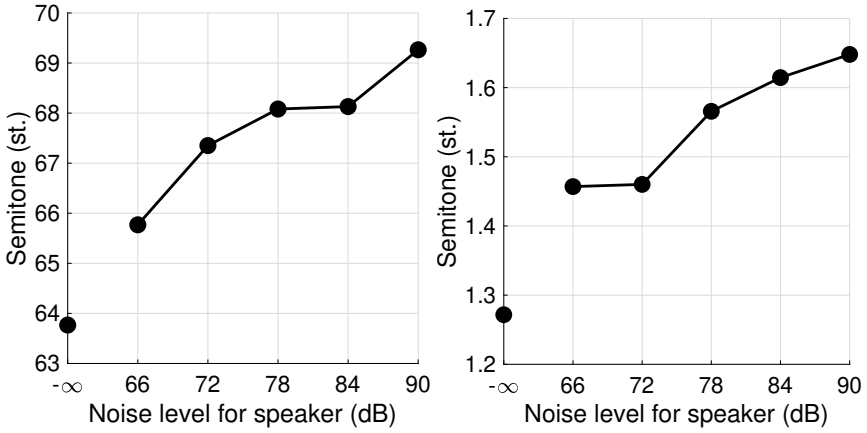
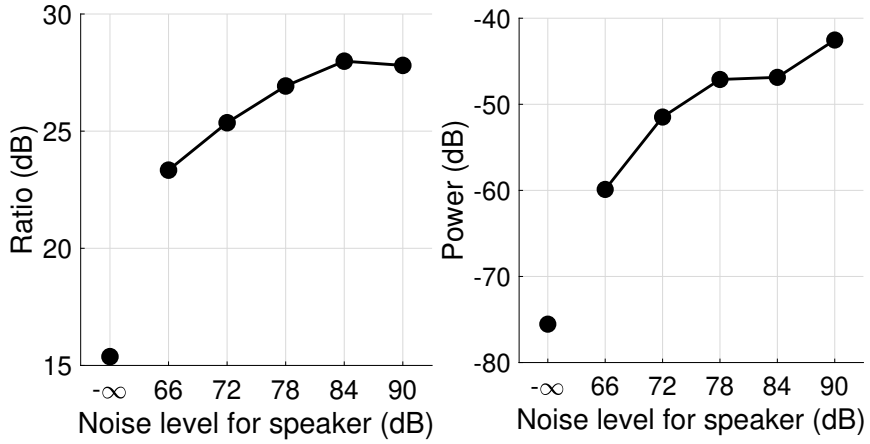


Figure 4-6: Analysis results of vocal tract length with f_0. T1, T2 are the transitions. C is the nuclei center.



(a) f_0 mean (b) f_0 range

Figure 4-7: Analysis results of f_0.



(a) Consonant-to-vowel ratio (b) Average power

Figure 4-8: Analysis results of power envelope.

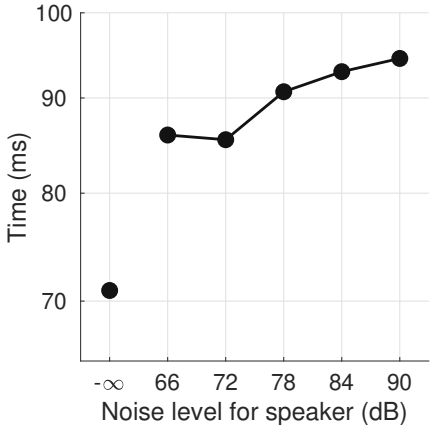


Figure 4-9: Analysis results of vowel duration.

- Fundamental frequency (f_0)

  In the previous study [28] and in the previous chapter, f_0 mean showed a clear correlation with increasing noise levels. In this study, the dynamic range of f_0 was also investigated. Figure 4-7 shows the analysis results of these acoustic features of f_0.

- Power envelope

  In the previous study [28], it was found that a rise in modulation spectrum appeared in high-frequency regions for Lombard speech. Modulation spectrum is an indirect way to understand the power envelope. Also, as was investigated in the previous chapter, higher lung pressure was obtained in Lombard speech, which also indicated some modification in power envelope. The direct analysis of the power envelope focused on the consonant-to-vowel ratio and the average power. The analysis results of these features are shown in Fig. 4-8. Also, a correlation between the power envelope and the f_0 contour was estimated.

- Duration

  The same as in the previous study [28] and as in the previous chapter, vowel duration was extracted, which is shown in Fig. 4-9.

- Rule generation model

  The analysis results are summarized in Table 4.1. For all the analyzed acoustic features (Figs. 4-3 to 4-9), all the features continuously varied with increasing noise levels. These variations nonlinearly increased or decreased with increasing noise levels, and in some acoustic features, they had an abrupt change at 84 dB. The abrupt change might be unexpected in the production of Lombard speech, so the feature variations could still be considered continuous through 84 dB. As perceived by humans, perceptual differences between plain and Lombard speech only start from a specific noise level. Furthermore, as produced by humans, saturation (i.e., the limitations to changing acoustic features) can exist, which can be seen as a logistic curve of the variations in approaching the limitations. These were the foundation to establish the rule generation model.

  Therefore, on the basis of the model reported by Hodgson et al. [82], in which the Lombard effect represents the relationship between the constitutional factors of environments

Table 4.1: Analysis results of acoustic features and their tendencies with increasing levels of pink noise.

Feature group	**Acoustic feature and tendency**
Spectral tilt	Increased c_0, decreased c_1 and c_2 (see Fig. 4-4)
Formants	Increased F_1, F_2, F_3, F_4 (see Fig. 4-5)
Vocal tract length	Decreased vocal tract length (see Fig. 4-6)
f_0	Increased f_0 mean and f_0 range (see Fig. 4-7)
Power envelope	Increased consonant-to-vowel ratio and average power (see Fig. 4-8), and a positive correlation of 0.47 with f_0
Duration	Increased vowel duration (see Fig. 4-9)

with noise levels, a rule generation model is proposed. This model represents the relationship between acoustical parameter values and noise levels. It is estimated to have a drastic change around 66 dB [82] and a saturation starting from 90 dB, as shown in Fig. 4-10 and Eq. (4.4).

$$\psi(x) = \frac{K}{(C + e^{-B(x-x_0)})^{1/v}} \tag{4.4}$$

By applying this model, for each acoustic feature, relative to plain speech, a model function is estimated by non-linear least square fit (by the function lsqcurvefit in Matlab) with initial values of $(K, C, B, v) = (K_0, 1, B_0, 1)$. Here, K_0 is the maximum of the estimated values if the changing tendency of the values with noise levels is increased, and vice versa (i.e., the minimum of these values). B_0 is set equal to the linear slope estimating these values. The lower and upper bounds are $(-\infty, 0, 0, 0)$, $(\infty, \infty, \infty, \infty)$, respectively, with a step size of 10^{-6}. The fitting errors were very small compared with the relative values of acoustic features of Lombard speech to plain speech: 3.8%, 5.8%, and 0.63% for c_0, c_1, and c_2, respectively; $3.8 - 6.0\%$ for F_1, F_2, F_3 , F_4; 3.7% for VT length; 6.8% for f_0 mean and range; 1% and 3% for consonant-to-vowel ratio and average power; and 2% for duration. As an example, Fig. 4-11 represents the modeling for c_0 at the nuclei center (C) of vowels, shown in Fig. 4-4a, *relative to plain speech*, with the estimated coefficients $K = 9.97$, $C = 2.22$, $B = 0.19$, $x_0 = 66$, and $v = 0.55$.

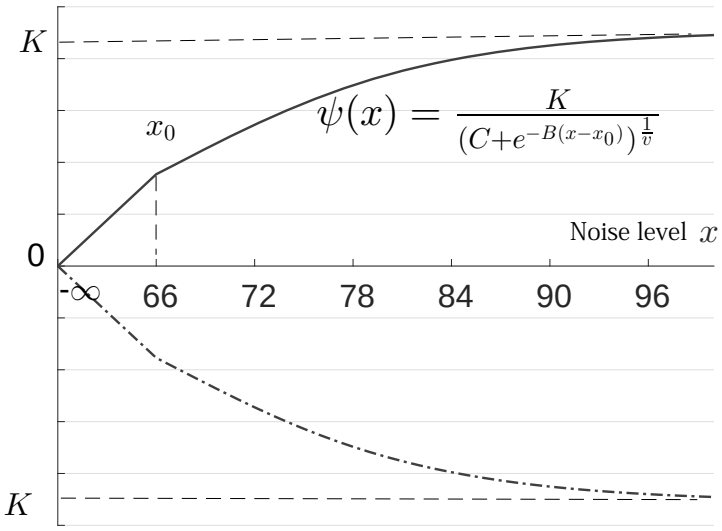


Figure 4-10: Rule generation model of acoustical parameter values ψ in log scale depending on the noise level x. K indicates the upper or lower limit to which the saturation approximates. x_0 indicates the noise level at which drastic change to Lombard speech occurs.

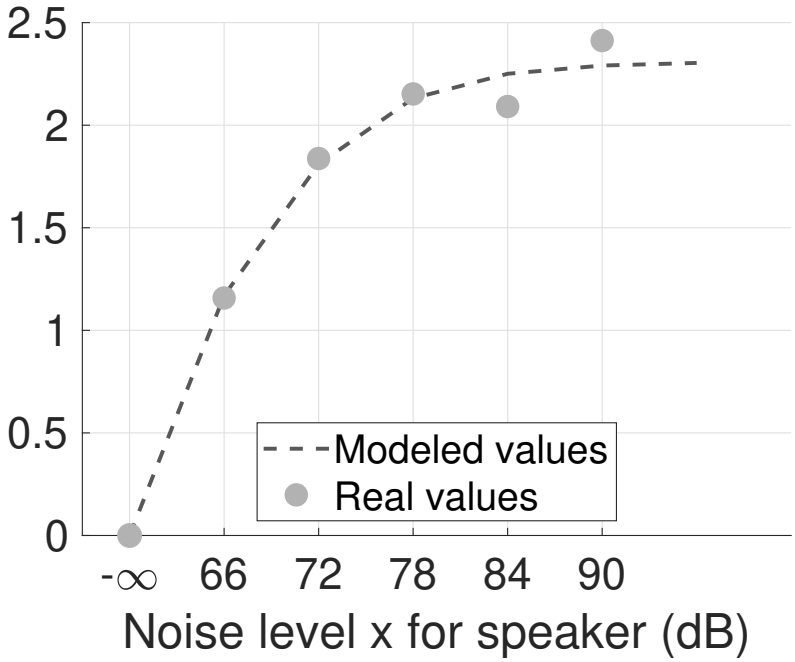


Figure 4-11: Rule generation model for c_0 at the nuclei center of vowels, shown in Fig. 4-4a, *relative to plain speech*.

4.2.3 Feature modification/synthesis

In this study, to modify/synthesize the acoustic features in mimicking Lombard speech, the following procedures and methods were utilized for each feature.

- Spectral envelope

  To synthesize the spectral envelope, the following steps were performed. Synthesis of the spectral tilt was easily done by modifying the three cepstral coefficients directly. To synthesize formant spectra, the original c_0, c_1, and c_2 were subtracted from the spectral envelope to obtain vocal tract spectra $log\frac{B(\omega)}{A(\omega)}$. These spectra were then divided into two parts: positive (peaks, which imply $A(\omega)$) and negative (dips, which imply $B(\omega)$) components. They were further modeled by spectral-GMM [23]. Because the peaks vary with noise levels, they are closely related to formants. The modification of F_1, F_2, F_3, F_4 and the vocal tract length were thus performed on the positive component, while the negative one was unchanged to preserve speaker individuality. The synthesized spectral tilt and modified vocal tract spectra were finally added together to produce the synthesized spectral envelope.

- Fundamental frequency (f_0)

  To synthesize f_0 contour, f_0 contour was parameterized and controlled by the Fujisaki model [83]. In this model, f_0 baseline (Fb) and amplitude of accent commands (Aa) were increased and the amplitude of phase commands (Ap) were varied to obtain the target f_0 mean and range by non-linear optimization. After that, modified f_0s at event-target locations were taken as modified event targets of f_0.

- Power envelope

  To synthesize the power envelope, the power envelope was parameterized by second-order damping modeling. In this model, the parameter *target* was used to control the power envelope portions to expected powers and to maximize the expected correlation with the modified f_0 contour. The *target* was extracted using a target prediction model [84, 85]. After that, the modified power envelope at event-target locations was taken as the modified event target of the power envelope.

- Duration

In order to modify the duration of vowels, event functions (EF) were scaled and then multiplied with the event targets of other synthesized features. On the basis of Bush and Kain's study [86], two halves of an EF (left and right halves, separated by the location of event function) were modeled by Eqs. (4.5) and (4.6). Expected scales in vowel duration were obtained by controlling N and s.

$$EF_L(t) = |1 - \frac{2}{1 + e^{s\frac{t}{t-t_N}}}| \tag{4.5}$$

where $t = 0, 1, ...N-1$ and N and s are respectively the duration and slope of the left half of an event function.

$$EF_R(t) = |1 - \frac{2}{1 + e^{s\frac{t}{t-t_1}}}| \tag{4.6}$$

where $t = N-1, N-2, ...0$ and N and s are respectively the duration and slope of the right half of an event function.

Finally, on each feature/noise level, all modified event targets were multiplied with the modified event functions. These multiplications created completely modified acoustic features. They were then synthesized by the STRAIGHT synthesizer to produce Lombard speech.

4.3 Perception experiments

In order to validate the proposed rule generation model, two main experiments: similarity, and intelligibility and naturalness were carried out to compare the proposed method with BGMM-based methods and Lombard speech produced by humans in the typical noise levels of 66, 72, 78, and 84 dB.

4.3.1 Experiment for similarity

The purpose of this experiment was to compare the proposed model of the rule-generation model with BGMM trained optimally for each noise level in terms of resembling Lombard speech in noise-free conditions.

a. Setup

- Speech material: Speech material was drawn from recorded speech (both Lombard speech produced at 66, 72, 78, and 84 dB noise levels and plain speech) [28]. 105 Japanese words (4-mora) spoken by one male and one female were used.
- Speech types: Two BGMM-based methods [47]: GlottalDNN-based (called **Glottal_BGMM**) and STRAIGHT-based (called **STRAIGHT_BGMM**) synthesis were used. In both, the modified features were spectral tilt, f_0, duration, and power envelope. The BGMM models were trained for each noise level. In addition, two more types were synthesized by the proposed method: **Rule_F0_Tilt** and **Rule_F0_Tilt_Formant**. The former's modified features were *spectral tilt*, f_0, duration, and power envelope, and the latter's were *spectral tilt*, f_0, *formants* , duration, and power envelope. In total, it had four mimicking types: Gottal_BGMM, STRAIGHT_BGMM, Rule_F0_Tilt, and Rule_F0_Tilt_Formant with equal root mean square (RMS) at a noise level.
- Listeners: 12 native Japanese: nine males and three females aged 23 to 25 years (mean: 24) with no reports of hearing problems.
- Procedure: The complete set was 105 words in both Lombard speech and the four mimicking types stated above produced at four noise levels: 66, 72, 78, and 84 dB. A stimuli is a pair of concatenated mimicked speech and Lombard speech with the same content. There were 1680 stimuli in total (105 words $\times$ 4 pair types $\times$ 4 noise levels). Each listener was assigned 64 pairs at a specific noise level using balanced design. Each pair type/noise level was heard by the same number of listeners. The listeners were asked to evaluate how well the mimicked speech resembled Lombard speech on a five-point scale (1: not at all, 2: a little, 3: moderately, 4: a lot, 5: quite a lot) by clicking the corresponding buttons.

  The experiment was carried out in a sound-proof room with high-quality headphones (STAX SL51-2216) connected to a desktop computer via an amplifier (STAX SRM-1/MK-2). The amplifier was used to set an exact noise level for the test, which was measured by a calibrated sound level meter (hand-held analyzer type 2250, Bruel & Kjar). Before initiating the experiment, listeners were familiarized with Lombard speech by listening to examples of Lombard speech and plain speech.

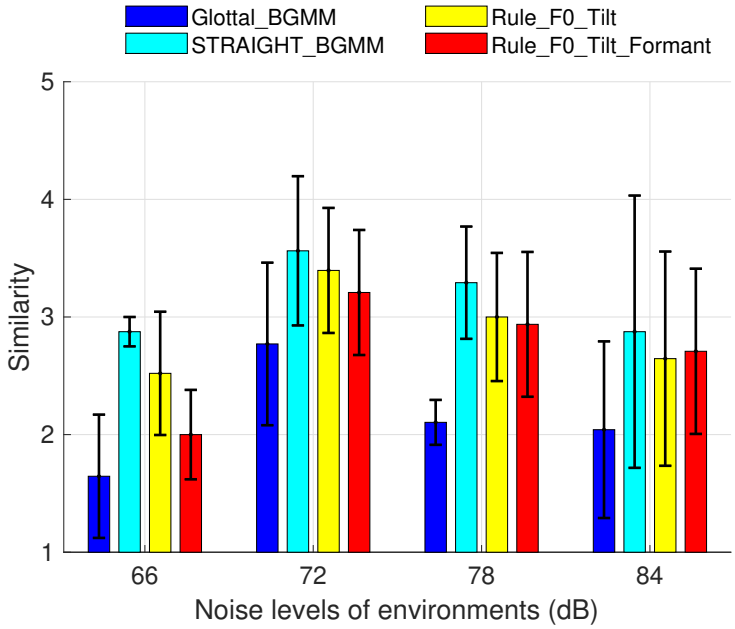


Figure 4-12: Similarity of the mimicked speech. The bar and error values indicate the mean and standard deviation among listeners. The values of similarity mean 1: not at all, 2: a little, 3: moderately, 4: a lot, and 5: quite a lot similar to Lombard speech.

b. Results and discussion

Figure 4-12 shows the similarity results of the mimicked speech to Lombard speech. For all noise levels, the similarity scores decreased in the order of STRAIGHT_BGMM, Rule_F0_Tilt, Rule_F0_Tilt_Formant, and Glottal_BGMM. Rule_F0_Tilt seemed comparable with STRAIGHT_BGMM. This finding indicates that the proposed method could obtain similar results to the statistical methods.

The equivalent results of similarity with Lombard speech between the proposed method and a BGMM-based method indicate that the current mimicked speech of this study could successfully adapt to noise levels. This demonstrates that the proposed model can correctly represent Lombard speech with varying noise levels.

4.3.2 Experiments for intelligibility and naturalness

The purpose of this experiment was to evaluate the intelligibility and naturalness of the mimicked speech by the proposed model compared with BGMM-based methods when different sets of features were modified. This might reveal clues as to how to improve intelligibility and naturalness for speech in noise and clarify the mimicking ability of different features.

Table 4.2: Speech types used in experiments for intelligibility and naturalness.

Speech type	**Modified feature**
Rule_Tilt	**Spectral tilt**, duration, and power envelope
Rule_F0_Tilt	**Spectral tilt**, f_0, duration, and power envelope
Rule_Tilt_Formant	**Spectral tilt**, **formants**, duration, and power envelope
Rule_F0_Tilt_Formant	**Spectral tilt**, f_0, **formants**, duration, and power envelope
STRAIGHT_BGMM	**Spectral tilt**, f_0, duration, and power envelope

a. Setup

- Speech material: Speech material was drawn from the ATR dataset [87]. A total of 384 words (3-mora) of plain speech from six different speakers (3 males, 3 females) was used.
- Speech types: Table 4.2 lists the five examined speech types. Four types were synthesized using the proposed rule generation model with prefix "Rule" considering the contribution and the mimicking ability from three main features: spectral tilt, f_0, and formants . One type was STRAIGHT_BGMM as a reference, as it uses the same vocoder . All had equal RMS at each noise level. No Lombard speech was used in these experiments because the speech was from the ATR dataset. If Lombard speech had been used, its results of intelligibility and naturalness would probably have been the best among these speech types.
- Listeners: Ten native Japanese: 8 males and 2 females aged 22 to 25 years (mean: 23.4, standard deviation: 1.5) with no history of hearing problems joined the tests.
- Maskers: Pink noise [74] at four noise levels (66, 72, 78, and 84 dB) was used; thus, there were four maskers.
- Procedure: The complete set included 7680 stimuli (384 words $\times$ 5 speech types $\times$ 4 noise level maskers). In each intelligibility or naturalness test, 60 unique words were assigned to one listener at each noise level. Each listener listened to all four noise levels in increasing order. Intelligibility and naturalness tests were performed in sequence. During the intelligibility test, the stimulus was played only one time. The listeners

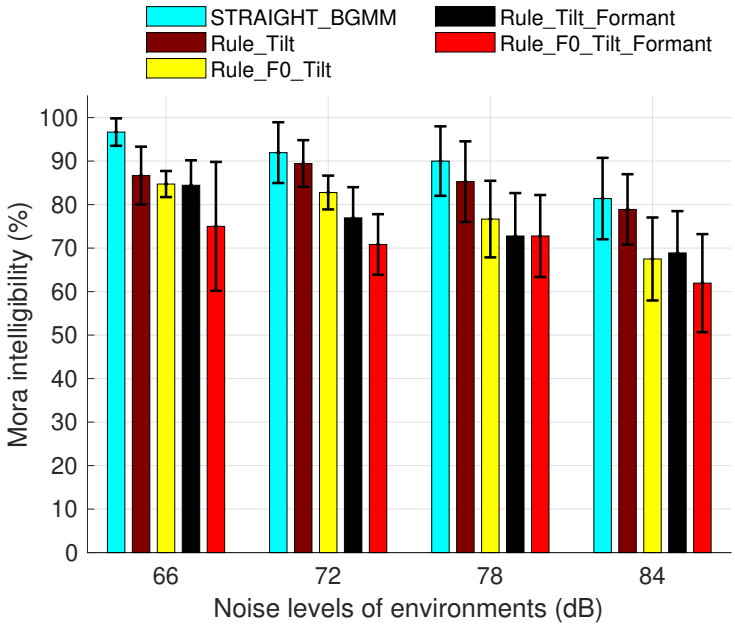


Figure 4-13: Intelligibility of speech when various features are mimicked, i.e., percentage of correctly perceived mora in a word. The bar and error values indicate the mean and standard deviation among participants.

were asked to write down the word they heard using a keyboard. They clicked on the "next" button to continue. During the naturalness test, the stimulus could be played again. The listeners were asked to evaluate their feeling of naturalness (human voice) on a four-point scale (1: unnatural, 2: rather unnatural, 3: rather natural, 4: natural) by clicking the corresponding buttons. The next stimulus would be played immediately after that.

b. Results and discussion

Figure 4-13 shows the results of speech intelligibility when various features were mimicked. It was found that the method of this study obtained a comparable result with STRAIGHT_BGMM only with the modification of spectral tilt. With the modification of the other feature sets, it obtained lower intelligibility. For all noise levels, the scores were varied in a similar way. Figure 4-14 shows the results of naturalness. Similar to the results of intelligibility, the current method with the modification of spectral tilt showed comparable naturalness with STRAIGHT_BGMM. With the modification of the other feature sets, it obtained lower naturalness. High Pearson correlation of 0.85 between intelligibility and naturalness scores also indicated these similar results.

The equivalent intelligibility and naturalness when spectral tilt was mimicked indicates that

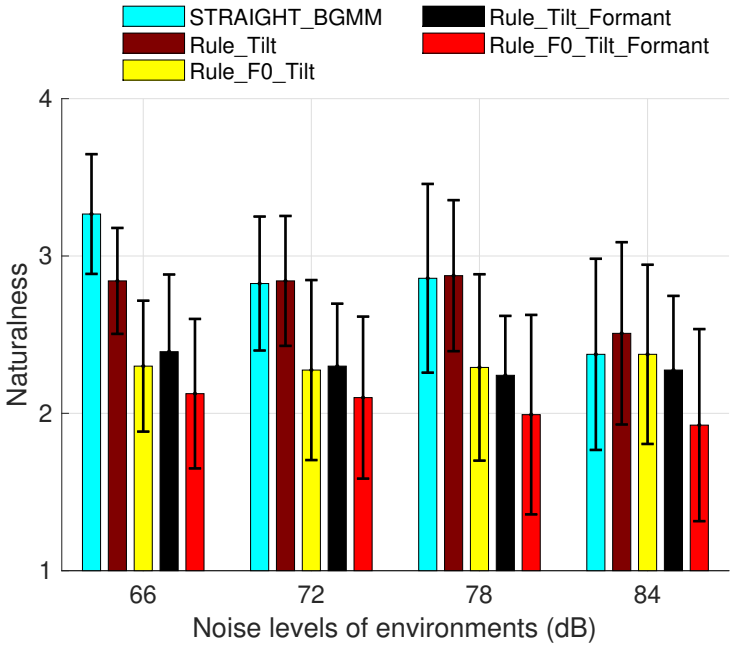


Figure 4-14: Naturalness of speech when various features are mimicked. The bar and error values indicate the mean and standard deviation among participants. The values of naturalness are 1: unnatural, 2: rather unnatural, 3: rather natural, 4: natural.

the proposed model contributes well to the intelligibility and naturalness adaption and with varying noise levels in which the most effective and successful feature to mimic was spectral tilt. It also demonstrates ability in terms of independent control of features.

The decreases of intelligibility and naturalness with f_0 and formants might stem from the effects of various feature modifications rather than the proposed model itself. Specifically, the modified f_0 range might cause incorrect pitch accents during the optimization process, thus reducing the naturalness. Although formants were flexible to control due to being modeled by GMMs, a GMM formant often has quite a large bandwidth. When it is shifted, some dips that are important for the intelligibility of phonemes might be erased, which in turn would reduce the naturalness and intelligibility. However, it should be possible to overcome these limitations without too much difficulty. It can be found constraints to preserve original pitch accents during f_0 modification, and a threshold can be chosen to avoid erasing dips when formants are shifted. On balance, it can be concluded that the proposed model performed adequately in all of the evaluations.

4.4 General discussion

Under the rule generation model, the rules in Lombard effect mimicking regarding noise levels were generalized. The values of model parameters were generated from a Lombard speech dataset to evaluate the reconstructions in both that Lombard speech dataset and another dataset (ATR dataset) without Lombard speech. The evaluated results were excellent as expected in both the datasets. Thus, it could be seen that the model was data-independent and could be applied for any other datasets. Regarding different noisy environments and not only concerning to SNR/noise levels, to the best of the current knowledge, it could affect the directional changes of acoustic features. For instance, formant frequencies in one kind of high-pass noise which masks from the F_2 region, F_2 decreases, not increases as in pink noise [49]. To apply for these different directional changes of these features, it is only needed to fit the rule generation model to several increasing noise levels, in which the Lombard speech is produced. The rule generation model can automatically adapt and adjust with the continuously decreasing or increasing directions of acoustic features with noise levels. Sometimes, other factors of noise drastically affect the features not only the directional changes. Then, some adjustable model parameters can be further investigated and derived to improve the adaptation of the rule generation model.

In the compared BGMM-based methods, the BGMM models have to be trained for each noise level, while in contrast, the proposed method can be controlled by the generated rules according to variation of noise levels. Thus, the proposed method represents an advancement because it can be applied to any noise level without additional training. This is done by explicitly modeling the tendencies of each acoustic feature with varying noise levels and independently controlling these features. On the basis of this achievement, if any other factors of adverse backgrounds are introduced, the proposed model can be improved to cover these factors. In addition, robustness in the independent control of acoustic features creates an opportunity to preliminarily investigate the effects of each parameter feature on the intelligibility of speech in noise.

4.5 Summary

In summary, in this chapter, analyses of Lombard speech were conducted and the rule generation model under backgrounds with varying noise levels for adaptively controlling the intel-

ligibility of transmitted speech in public announcement systems was presented. The proposed modification-synthesis method was described, which is based on co-articulation, MRTD, and spectral-GMM to easily control acoustic features with varying noise levels. Listening experiments were carried out to compare a state-of-the-art method and the proposed mimicking model. The proposed model showed comparable similarity and adaptivity to the noise levels. Intelligibility and naturalness are comparable with spectral tilt modification. When noise levels are continuous, the state-of-the-art method cannot adapt features to the noise levels, while in contrast, the proposed model can interpolate Lombard effect in any noise level. In order to obtain better intelligibility and naturalness, it is aimed to improve the modification in terms of f0 contour and formants. The most promising finding here is that the proposed method can control parameter values independently, thus enabling us to determine the most related parameters to intelligibility and improve intelligibility in noise more in the next step.

Combining the results from the previous chapter, mimicking Lombard speech has been achieved for articulatory features and acoustic features from one to multiple noise levels. The feature understanding and control for the intelligibility and naturalness of Lombard speech were obtained for the first sub-goal. In particular, the articulatory and acoustic features for the intelligibility and naturalness of speech in noise were thoroughly understood and identified, which were spectral tilts, f_0, and formants. The method with the rule generation model for the Lombard effect that was able to control acoustic feature variations independently was also obtained. Based on these features and the control method, in the next chapter, expanding the search space to more feature variations (not only the mimicked values with multiple noise level as this chapter), SNRs and spectral-varied noise, the variations of contributive features affecting intelligibility and naturalness most would be identified in the second sub-goal of identifying effective features to exceed Lombard speech.

Chapter 5

Effective features and strategies to improve the intelligibility and naturalness of speech in various noises

This chapter aimed to find out effective features under various noise to exceed the intelligibility of Lombard speech by expanding the search space of finding features to more feature variations, SNRs and spectral-varied noise from the resulting group of the acoustic features by the previous chapter (spectral tilt, f_0, and formants). The approach followed two consecutive subgoals:

In the first subgoal, it was aimed to identify the most effective acoustic features to vary with speech intelligibility and naturalness and their best-varied values. Listening tests were conducted for the speech synthesized by the feature variations at multiple noise levels of pink noise. The synthesized speech at a noise level was normalized at the same root mean square, which was estimated by the rule generation model for mimicking Lombard speech. All the task was divided into two phases:

(1) Finding out individual effective acoustic features to vary, affecting changing intelligibility significantly.

(2) Finding out joint variations of all individual effective acoustic features contributing to increase intelligibility and naturalness most.

In the second subgoal, the most effective acoustic features in the previous phase had been obtained. Meanwhiles, other methods also reported other effective features. A method to extract the effective feature from these was proposed .

In the final section, the discussion on the achievement of the exceeding of the intelligibility and the naturalness of Lombard speech was presented.

5.1 Effect of varying acoustic features

5.1.1 Features

The acoustic features used to generate their variations included:

- Spectral tilt: variations of decreasing c_1 and c_2 were investigated.

  This feature was decomposed into three components: increases in c_0 representing increases in average power spectra, a decrease in c_1 showing decreases in spectral slopes, and decreases in c_2 showing the plateau in mid-high frequency regions of spectra. It could be seen that increasing c_0 is similar to linearly increasing energy of speech signal. In addition, the synthesized speech was normalized at the same root-mean-square at a noise level. Therefore c_0 was excluded in this investigation because no variations could be made while variations of c_1 and c_2 were investigated.

- f_0: variations of increasing f_0 mean was investigated.

  As investigated in the previous section, increases in f_0 mean and f_0 range were the features that contributed to speech intelligibility. Because changing f_0 range could reduce naturalness, only variations of increasing f_0 mean were investigated.

- Formants: variations of increasing F_1, F_2, F_3, and F_4 were investigated.

  Due to the quite similar increasing ratios for F_2, F_3, F_4 in mimicking Lombard speech, variations of F_2, F_3, F_4 was applied in the same rate, while variations of F_1 was at a different rate. Vocal tract length was modified with the increases in formants accordingly.

5.1.2 Varying method

Based on the concept that variations under the Lombard effect constraints, it was assumed that the initial value for the drastic change from plain speech to Lombard speech at 66 dB could be varied, providing that the Lombard effect was still preserved. In other words, these variations were still by the Lombard effect; the mission was similar to perform searching in the region

of the Lombard effect. In the implementation, it meant that given a different initial value of an acoustic feature at 66 dB, the rule-generation model for the Lombard effect would be used to generate variations of acoustic features at other higher noise levels (72, 78, 84, 87 dB). In this study, the parameter A in the formula of the rule generation model was varied to change the feature value at 66 dB then interpolate other feature values at remaining noise levels accordingly.

5.1.3 Experiments

Perceptual tests for variations of single features

This experiment corresponded to the first phase of finding out individual effective acoustic features to vary. Subjective listening tests of intelligibility and naturalness in noise were conducted with the synthesized speech by variations of any single feature.

a. Feature variations

   Figures 5-1, 5-2, 5-3, and 5-4 show the variations for each feature: c_1, c_2, f_0 mean, and F_1, F_2, F_3, F_4 used in this experiment. It could be seen that increases in c_1 were varied, which yielded the compensated spectrum varied from 3 dB up to 15 dB. Increases in c_2 were varied, which yielded the compensated spectrum varied from 2 dB up to 11 dB. Also, increases in f_0 were varied from 2 st. up to 10 st., and formants were increased from 5 % up to 30 % for F_1 and from 4 % up to 15 % for F_2, F_3, and F_3. These increases were empirically targeted to cover three regions: above the region of the mimicking Lombard speech (up to 400%, depending on features), at the region of the mimicking Lombard speech, between the region of the mimicking Lombard and the regions of plain speech (excluding the region of plain speech).

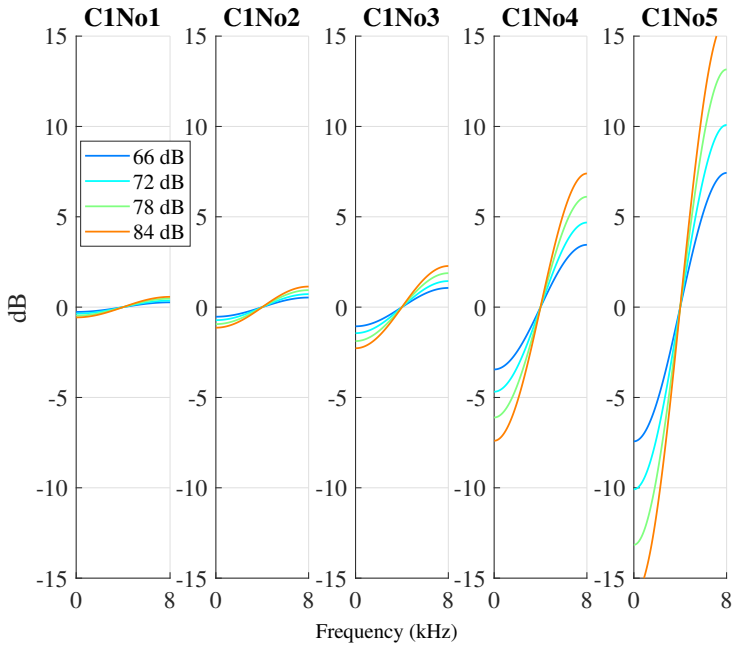


Figure 5-1: Compensated spectra corresponding to variations for c_1 for the experiment of variations of single features. The y-axis indicates the increased amount of the spectrum by c_1 from the normal spectrum (i.e., *C1Normal* meant no increasing or decreasing) of plain speech.

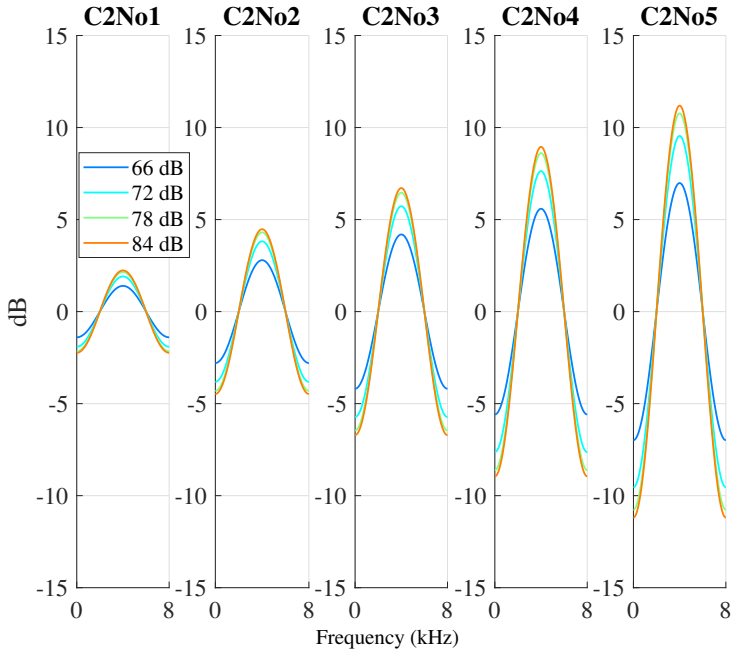


Figure 5-2: Compensated spectra corresponding to variations for c_2 for the experiment of variations of single features. The y-axis indicates the increased amount of the spectrum by c_2 from the normal spectrum (i.e., *C2Normal* meant no increasing or decreasing) of plain speech.

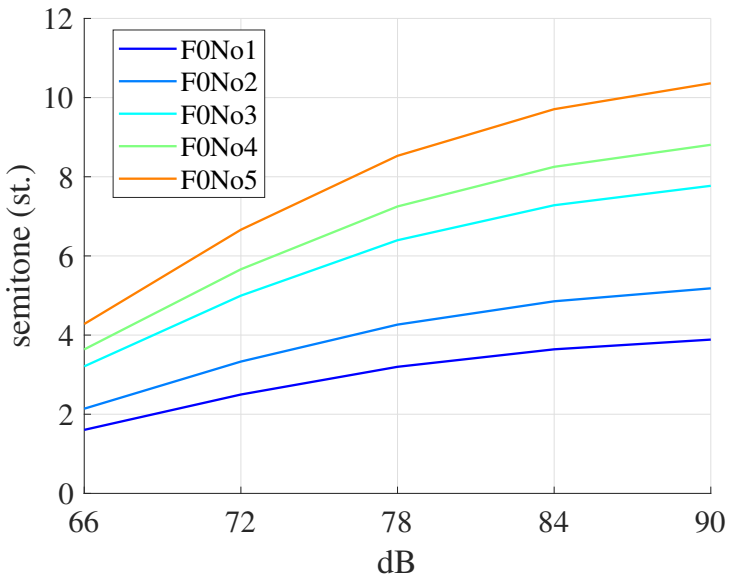


Figure 5-3: Variations for f_0 for the experiment of variations of single features. The y-axis indicates the increased amount of f_0 from normal values (i.e., *F0Normal* meant no increasing or decreasing) of plain speech.

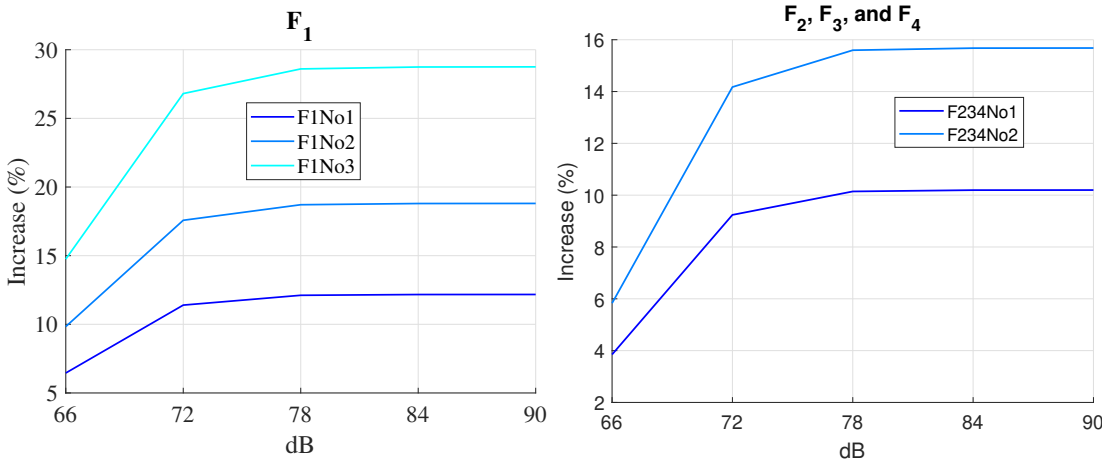


Figure 5-4: Variations for formants (F_1, F_2,F_3, and F_4). The y-axis indicates the increased amount of formants from normal values (i.e., *FFNormal* meant no increasing or decreasing) of plain speech.

b. Setup

- Speech materials were drawn the male and female speech in the ATR dataset, including 930 three-mora words.
- Speech types: 22 speech types: normal speech and 21 synthesized speech from c_1 variations (5 values), c_2 variations (5 values), f_0 variations (5 values), formant variations (6 values: 3 values of F_1 x 2 values $F_{2,3,4}$).
- Maskers: Pink noise at four levels: 66, 72, 78, 84 dB.
- Stimuli: 81840 stimuli (930 words $\times$ 22 speech types $\times$ 4 noise maskers) involved in the tests. The clean speech for testing at a noise level was amplified to an RMS related to that noise level. The RMS value was calculated by the original rule generation model for mimicking Lombard speech at a noise level.
- Participants: Eleven native Japanese including seven men and four women with an average age of 24.8 and the standard deviation of 3.8 with no report of hearing problems.
- Procedure: The participants evaluated all the speech types at all three noise levels of the pink noise. Each participant evaluated the stimuli by each feature in separated sets, i.e.,
  - c_1: 72 stimuli of unique words (6 speech types with 1 normal and 5 c_1 variations $\times$ 12 words)/noise level
  - c_2: 72 stimuli of unique words (6 speech types with 1 normal and 5 c_1 variations $\times$ 12 words)/noise level.
  - f_0: 72 stimuli of unique words (6 speech types with 1 normal and 5 f_0 variations $\times$ 12 words)/noise level.
  - Formants: 84 stimuli of unique words (7 speech types with 1 normal and 6 formant variations $\times$ 12 words)/noise level.

  They went through the evaluation for all four features in four designated times. For each feature, they did the intelligibility test and the naturalness test of the stimuli in sequence.

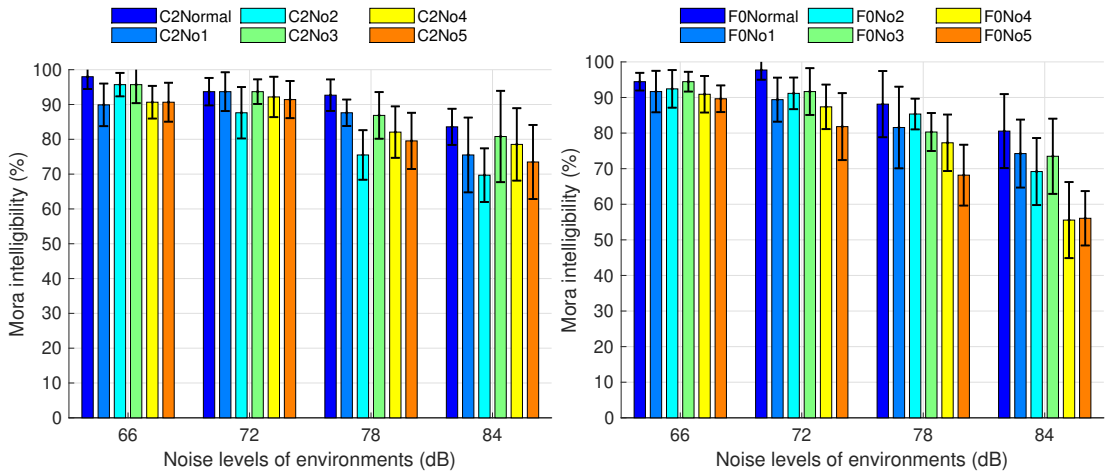
  In the intelligibility test, they went through 3 parts, where each part corresponded to a noise level. During this test, the stimulus was played only one time and in random

order. The listeners were asked to write down the word they heard by using a keyboard. They clicked on the next button to continue. After listening to twenty stimuli, they took a short break in 30 s; the test was continued after they clicked on the continue button. Between two parts, they also took a short break in 1 minute. This test took about 30 minutes.

In the naturalness test, they went through the same 3 parts as in the previous test. During this test, the stimulus was still played in random order, but it could be played again. The listeners were asked to evaluate their feeling of naturalness (human voices) in four scales (1: unnatural, 2: rather unnatural, 3: rather natural, 4: natural) by clicking the correspondent buttons. The next stimulus would be played immediately after that. After listening to twenty stimuli, they took a short break in 30 s; the test was continued after they clicked on the continue button. Between two parts, they also took a short break in 1 minute. This test took about 15 minutes.

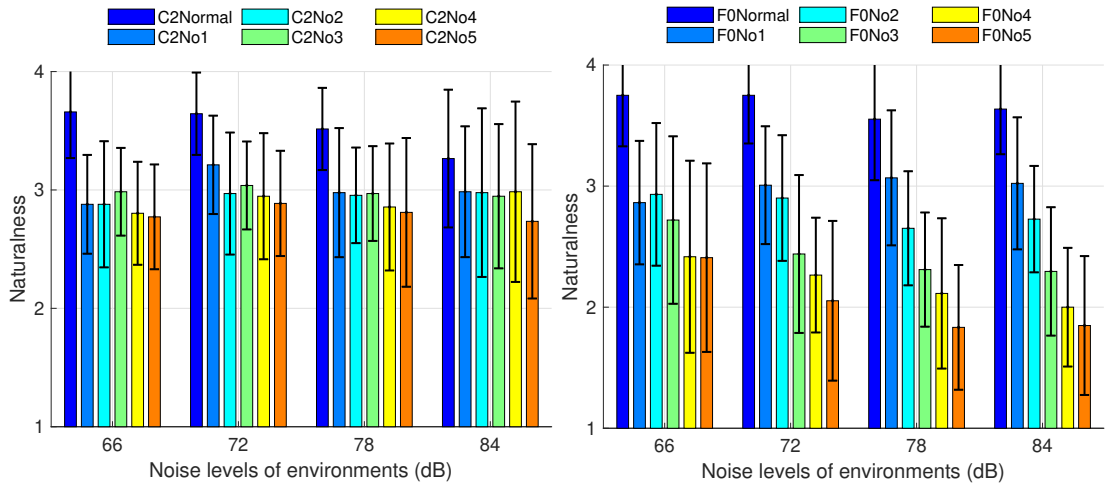
All the tests were carried out in a sound-proof room with a high-quality headphone (STAX SL51-2216) connected with a desktop computer via an amplifier (STAX SRM-1/MK-2). The amplifier was used to set an exact noise level of 80 dB SPL for the test, which was measured by a calibrated sound level meter (a hand-held analyzer type 2250 Bruel. & Kjar).

c. Results and discussion



(a) c_2 (b) f_0

Figure 5-5: Intelligibility scores (percentage of correctly answered morae in a word) of the synthesized speech differed by variations for c_2 and f_0 in the presence of pink noise at 66, 72, 78, and 84 dB noise levels.



(a) c_2 (b) f_0

Figure 5-6: Naturalness scores of the synthesized speech differed by variations for c_2 and f_0 in the presence of pink noise at 66, 72, 78, and 84 dB noise levels.

Figures 5-5 and 5-6 typically show the results for the intelligibility and naturalness of the speech by varying c_2 and f_0. Among these variations, the small different scores could be seen.

To identify the significant differences for the feature variations, two-way ANOVA analyses with two factors of noise levels and feature variations were conducted on all the listening test scores of intelligibility and naturalness for each feature. The posthoc tests were taken on the factor of feature variations.

The multiple comparisons among variations shown that varying c_2 and f_0 significantly affected to change the intelligibility ($p < 0.001; p < 0.05$) and naturalness ($p < 0.00; p < 0.05$) of speech. While there were no significant effects when varying the other features.

Perceptual tests for joint variations

As a result of the previous experiment, it was varying c_2 and f_0 significantly affected intelligibility and naturalness. This joint experiment was conducted to examine the single/joint combination of variations of c_2 and f_0 contributing to the intelligibility and naturalness most at multiple noise levels of the pink noise. Finally, it was to identify the most contributive feature variants to intelligibility and naturalness. When setting the low noise levels of 66, 72, and 78 dB (Fig. 5-5), the differences in the recognition rate among speech variants were small. The intelligibility scores were relatively high (80 %, similar to the ceiling effect coped with in Chapt. 3). In this experiment, higher levels of 84 and 87 dB were used. It was expected that this could make the recognition rate among speech types more distinguishable with clear improvement.

a. Joint variations of features

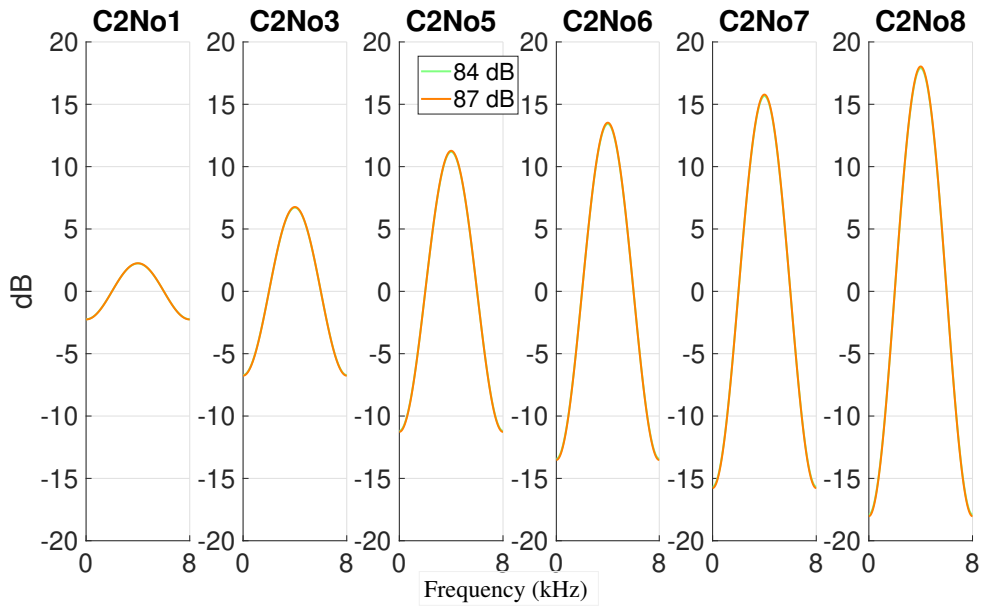


Figure 5-7: Compensated spectra corresponding to variations for c_2 for the joint experiment. The y-axis indicates the increased amount of the spectrum by c_2 from the normal spectrum (i.e., *C2Normal* meant no increasing or decreasing) of plain speech.

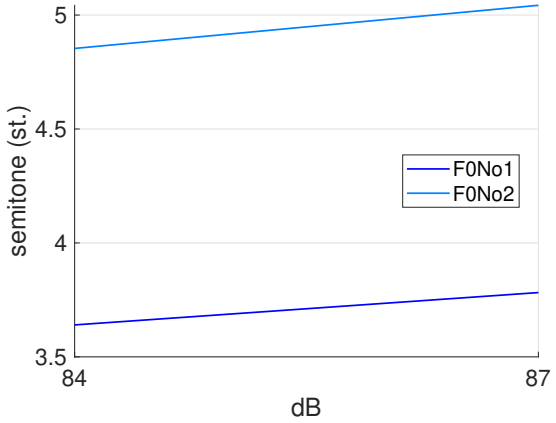
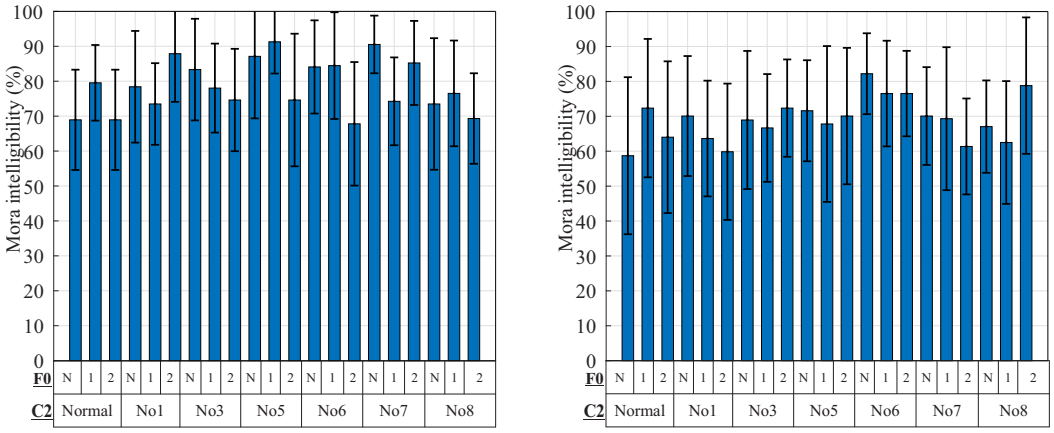


Figure 5-8: Variations for f_0 for the joint experiment. The y-axis indicates the increased amount of f_0 from normal values (i.e., *F0Normal* meant no increasing or decreasing) of plain speech.

Due to the uncertainty of more intelligibility benefits by a more reduction by c_2, and the drastic decreases in the intelligibility for high f_0 (Fig. 5-5). The varying regions of c_2 was expanded while the varying regions of f_0 was narrowed. As in Figs 5-7 and 5-8, c_2 was decreased more to obtain up to 20 dB of the increases in the compensated spectrum. f_0 was only at most 1.5 times (5 st.) more than the normal f_0.

b. Setup

- Speech materials were drawn from the male and female speakers in A set of the ATR dataset, including 276 three-mora words.



(a) 84 dB

(b) 87 dB

Figure 5-9: Intelligibility scores (percentage of correctly answered morae in a word) of the synthesized speech differed by **joint variations** between c_2 and f_0 in the presence of pink noise at 84 and 87 dB noise levels. The N, 1, and 2 in the row f_0 label corresponded to Normal, No1 and No2 of f_0 as defined in the f_0 variations.

- Speech types: 21 synthesized speech from 7 variations of c_2 (6 variations as Fig 5-7 and the normal value) $\times$ 3 variations of f_0) (2 variations as Fig 5-8 and the normal value).
- Maskers: Pink noise at three levels: 84, and 87 dB.
- Stimuli: 11,592 stimuli (276 words $\times$ 21 speech types $\times$ 2 noise level maskers) involved in the tests. As same as the previous experiment of variations of single features, the clean speech for testing at a noise level was amplified to the same RMS related to that noise level. The RMS value was calculated by the original rule generation model for mimicking Lombard speech at a noise level.
- Participants: twenty-two native Japanese including 16 men and 6 women with an average age of 24.3 and the standard deviation of 2.9 with no report of hearing problems.
- Procedure: The experiment was carried out with the same procedure and equipment as the experiment of variations of single features. Each participant evaluated the intelligibility and naturalness of all speech types at two noise levels in separated sections. A list of 84 stimuli of unique words (21 speech types x 4 words) was assigned to participants at each noise level. The participants did the intelligibility test first, then the naturalness test. It took about 50 minutes for the participant to complete the experiment.

c. Results and discussion

- Perceptual test of intelligibility

Figure 5-9 shows the results of the intelligibility test for 21 speech types at the multiple levels of the pink noise. In general, for all noise levels, the most improved intelligibility seemed to be located around the regions of $C2No3$ to $C2No6$ and $F0Normal$ or $F0No1$.

To study the effect of the two examining features over multiple noise levels in more detail, a three-way repeated-measures ANOVA was performed. The three factors were c_2 with 7 levels (from $C2Normal$ to $C2No8$), f_0 with 3 levels (from $F0Normal$ to $F0No2$) and noise levels with 2 levels (84 and 87 dB). The dependent factor was intelligibility scores.

In detail, there were significant main effects for all three factors, that is, c_2 [$F(6, 16) = 7.21$, $p = 0.001$], f_0 [$F(2, 20) = 6.73$, $p = 0.006$], noise levels [$F(1, 21) = 45.032$, $p < 0.001$]. The effect size was strongest for c_2 ($\eta^2 = 0.730$), followed by noise levels ($\eta^2 = 0.682$), and f_0 ($\eta^2 = 0.402$). In addition, there was multiple significant interactions between factors, namely, between c_2 and any other factors (two-way or three-way interactions), however, there was no significant interaction between f_0 and noise levels $p = 0.254 > 0.05$. Therefore, it could be seen that c_2 was the most significant factor. To study the significant differences among c_2 variants, pairwise comparisons with Bonferroni adjustment was carried out for these c_2 variations. The results showed that there were significant difference from $C2Normal$ for $C2No5$ ($p = 0.001$) and for $C2No6$ ($p = 0.001$), while there was no significant difference between $C2No5$ and $C2No6$ ($p = 1.000$).

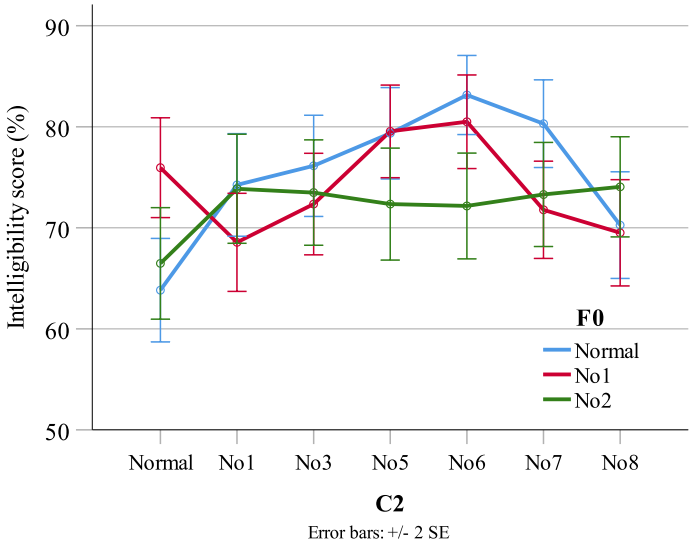
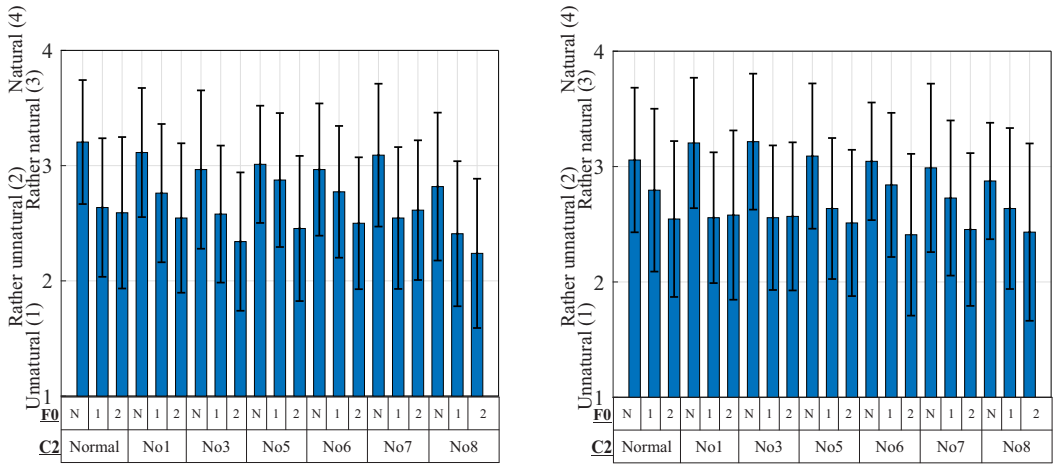


Figure 5-10: Intelligibility scores (percentage of correctly answered morae in a word) of the synthesized speech differed by **joint variations** between c_2 and f_0 in the presence of pink noise *pooled over 84 and 87 dB noise levels*.

To select a better variant between $C2No5$ and $C2No6$. The intelligibility scores averaged on all noise levels for each c_2 variant corresponding f_0 variants were shown in Fig. 5-10. As can be seen, the highest score was at $C2No6$ and $F0Normal$, i.e., the most contributing variant to intelligibility is increasing the spectral region between 2-6 kHz by 13 dB (Fig. 5-7) (this spectral region by c_2 was also mentioned in the previous chapter as a typical increase in the spectra of Lombard speech) with no modification of f_0. Under Fs = 16 kHz, c_2 was used as a representation of this mid-high frequency region. Increasing spectra 2-6 kHz region also corresponded to a plateau affecting the spectral regions between 2-6 kHz, which was highly produced by moving piriform fossa upward. f_0 had not much contribution to increasing intelligibility differently from the previous result on f_0 (Chapt. 3) might be because the speech was normalized out. At the same time, they were left as itself after modification of features in these experiments of the previous chapter.

- Perceptual test of naturalness

  Figure 5-11 shows the results of the intelligibility test for 21 speech types at the multiple levels of the pink noise. When f_0 was modified, the naturalness of synthesized speech became worst. The modification of c_2 did not make naturalness reduce much. This result meant that the modification on the spectral region made by c_2 was robust to



(a) 84 dB (b) 87 dB

Figure 5-11: Naturalness scores of the synthesized speech differed by **joint variations** between c_2 and f_0 in the presence of pink noise at 84 and 87 dB noise levels. The N, 1, and 2 in the row f_0 label corresponded to Normal, No1 and No2 of f_0 as defined in the f_0 variations.

naturalness.

5.1.4 Summary

In this section, it was to study the effects of varying effective acoustic features on the intelligibility and naturalness of speech at multiple levels of the pink noise. The results indicated that varying c_2 got the most contribution to increase intelligibility. The increases by a variation of c_2 indicated that the most effective variation was the increase in the spectrum by the plateau between 2-6 kHz by 13 dB. This increase might come from the upward movement of piriform fossa during the Lombard effect. This result was by the search space of feature variations at multiple noise levels of the pink noise only and produced by this study, however, what were the essential properties in the search space of more variable noise and SNRs and in comparison with other studies with their effective features? This question was answered in the next section.

5.2 Effective features in consideration with other studies

In the previous section, the most effective acoustic feature to vary with noise had been identified, i.e., the increase in the spectrum by the plateau between 2-6 kHz by 13 dB. Recently there were also other methods reported their effective acoustic features to control in noise. This section aimed to identify significant features to control speech to increase its intelligibility and natural-

ness in noise from these methods by expanding the search space to more SNRs and spectral-varied noise. A concept for the identification was proposed basing on modulation spectrum and modulation transfer function concepts and other related perceptual models. Promising features were extracted by analyzing relations of modulation spectra of the differently enhanced speech smeared by environments for intelligibility and naturalness. Tentative features are determined by combining the extracted features with supplement features from other models. The benefits for intelligibility and naturalness of these features by listening tests were examined. Finally, significantly effective features were identified.

5.2.1 Theory of effective feature extraction concept

Two typical ways to enhance the intelligibility of the presented speech are by controlling acoustic features with or without models.

Acoustic features for speech intelligibility can be directly controlled by spectral shaping, intensity amplifiers, and equalizers. Spectral shaping modifies speech spectra by increasing spectral regions for intelligibility (e.g., 1-4 kHz [16] or 2-6 kHz [88], and formants [16, 88]). Intensity amplifiers increase speech intensity by adjusting the gain of speech. Equalizers are audio processors that use a combination of different filters to alter the balance of frequencies in an audio signal [89, 90]. Thus, it can be used to boost important frequency regions for intelligibility. These important acoustic features to control were mainly extracted from clear speech [91] and Lombard speech [92]. Increasing the frequency regions between 1-4 kHz or 2-6 kHz is to decrease spectral tilt, modification on formants is to increase formant frequencies and amplitudes. Increasing spectra in different frequency regions also have different effects on the perception of related factors of the naturalness of speech. For example, decreasing spectral regions below 1 kHz reduces fullness, while increasing the spectral region above 1 kHz extends brightness [93]. Increasing speech intensity to increase vocal strength as in Lombard speech. Equalizers can combat feedback to imitate the production mechanism in Lombard speech. However, it is unable for those methods to control acoustic features to changing phenomena of environments appropriately.

Models such as perceptual models and room acoustic models can be used to estimate the equivalent amount of environment phenomena such as the signal-to-noise ratio (SNR) and noise level needed to compensate for degradation in intelligibility. Speech intelligibility has been

improved in noisy environments [1–4] by optimizing the index of the perceptual model used for intelligibility measurement such as SII [5], STI [6], and HEGP [7]. Further analyses of the speech after index optimization indicated that increasing the spectrum above 1 kHz increases intelligibility [4]. Another perceptual model that uses dynamic range compression (DRC) [16] has been used to reduce the speech amplitude on the basis of an input-output energy curve. The DRC model emphasizes loudness in the voice onsets and offsets and in the stops and nasals, thereby increasing intelligibility. Xu et al. [94] showed that an intensity range around peak amplitude yielded better intelligibility performance under noisy conditions than others. This finding indicates that compressing the speech amplitude into the regions of the peak amplitude might be useful in increasing speech intelligibility, which is in line with the effects of DRC model. Besides, the DRC model makes speech signals degraded, especially in naturalness. A method based on a room acoustic model uses the modulation transfer function (MTF) to control the speech modulation spectrum (MS) and has demonstrated a more systematic and explicit derivation to enhance speech intelligibility against environmental smears.

In the MTF concept, which was proposed by Houtgast and Steeneken [59, 61], the modulation reduction of the envelopes of an output signal are imitated from the input signal during transmission in a room. In the MS concept, the speech MS is produced by spectral analysis of the temporal amplitude envelope of the frequency spectra. The dominant MS component of continuous speech lies between modulation frequencies of 1 and 16 Hz, with a peak around 4 Hz [59, 63, 64]. Some recent studies [95, 96] reported that better intelligibility is obtained when increasing the MS indexes as high as in Lombard speech. In the other words, the higher the MS index in these modulation frequencies, the better the intelligibility. That is, speech is intelligibly presented if its MS resists smearing of the MTF by the environments. If a smeared MS (SMS) is given by

$$SMS = MS \times MTF \tag{5.1}$$

where MS is the MS of the original speech, then an optimally resistant MS (RMS) can be calculated using

$$RMS_{optimal} = MS \times MTF^{-1}. \tag{5.2}$$

If $RMS_{optimal}$ is presented in an adverse environment with such an MTF, the MS of the speech reaching the listeners should be MS as $MS = RMS_{optimal} \times MTF$, which has the original intelligible MS. Several studies tried to estimate MTF^{-1} [64]. However, directly obtaining MTF^{-1} is complicated, especially in realistic environments where backgrounds are diverse and varying

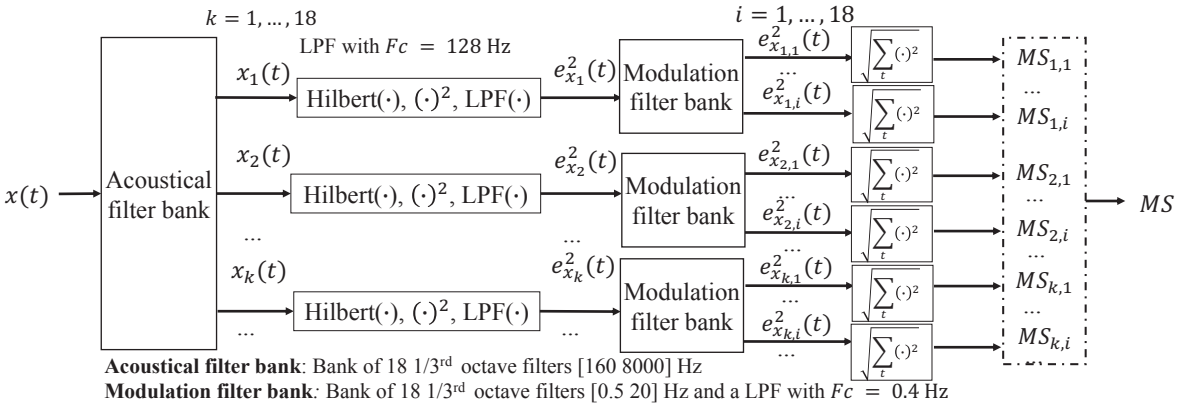


Figure 5-12: Modulation filtering

because it requires estimating MTF. This estimation is unsuited to realistic environments.

The present study aimed to extract significant features to modify MS of the original speech by analyzing relations of SMS for intelligibility and naturalness. The detail about SMS and this proposed method is given in the next sections.

5.2.2 Formation of smeared modulation spectrum

This section explains the methods used to calculate MS, RMS, MTF, and SMS.

Numerous methods have been reported for the extraction of speech MS [97, 98]. A study of modern psychophysical models of temporal processing indicated that temporal amplitude envelope is processed by a modulation filter bank [99]. As was used in Unoki and Zhu's study [100] and Zhu et al.'s study [99], a modulation filtering technique was used to extract the MS/RMS of speech. The filtering was done using an acoustic filter bank concatenated with a modulation filter bank. The former was a bank of 18 filters: $1/3^{rd}$ octave band-pass filters with bandwidths of 160-8000 Hz, which followed the SII specifications. The latter was also a bank of 18 filters: an LP filter with cutoff frequency $Fc = 0.4$ Hz and 17 $1/3^{rd}$ octave band-pass filters with bandwidths of 0.5-20 Hz. Houtgast et al. [61] also used 0.5-20 Hz band-pass filters to extract the speech power envelope spectrum. Our extracted MS/RMS thus contained 0 Hz modulation and showed as frequency features. The time features were above 0 Hz up until a modulation frequency of 20 Hz.

It was thus needed to discuss the capability in a representation of both local and global time-frequency features in the MS/RMS and relations of articulatory-acoustic features and the MS./RMS features. As $1/3^{rd}$ octave bands were used, the bandwidths were suitable to capture

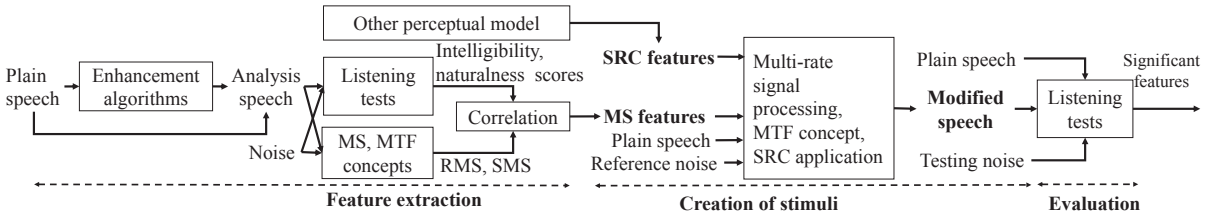


Figure 5-13: Implementation of theory of effective feature extraction model

both global frequency features, which might be from changes in spectral tilt and large frequency regions, and local features, which might be from changes in formants or narrow frequency regions. Also, as indicated in the calculation of STOI (short-time objective intelligibility) [101], to have a proper modulation frequency resolution, an appropriate length of the analysis time frame could be 300-500 ms. The modification in power envelope and some other temporal features could be presented in time features, which are in MS/RMS regions above 0 Hz until a modulation frequency of 20 Hz.

The MTF is fully determined mathematically for stationary noise by the signal-to-noise ratio [61, 62]. For each band-limited acoustic frequency, i.e, f_a, the MTF is independent of the modulation frequency, i.e., f_m and defined as

$$m_N(f_{a_k}, f_m) = \frac{1}{1 + 10^{\frac{-SNR_{f_{a_k}}}{10}}} \tag{5.3}$$

where $SNR_{f_{a_k}} = 10 \log 10\left(\overline{e^2_{x_k}} / \overline{e^2_{n_k}}\right)$. x_k and n_k were filtered speech and noise at the k^{th} f_a and e^2 was power envelope. It could be seen that the spectral properties of stationary noise or non-stationary noise were adequately concerned due to the calculation of the SNRs within the band-limited signals like this.

From Eqs. 5.1 and 5.3, SMS can be calculated using

$$SMS(f_a, f_m) = MS(f_a, f_m) \times m_N(f_a, f_m). \tag{5.4}$$

SMS at 0 Hz modulation shows the effect of noise on the frequency features. The SMS at over 0-20 Hz modulation shows the effect of the environment on the time features.

5.2.3 Implementation of the theory for identifying effective features

As shown in Fig. 5-13 the proposed concept is based on three steps:

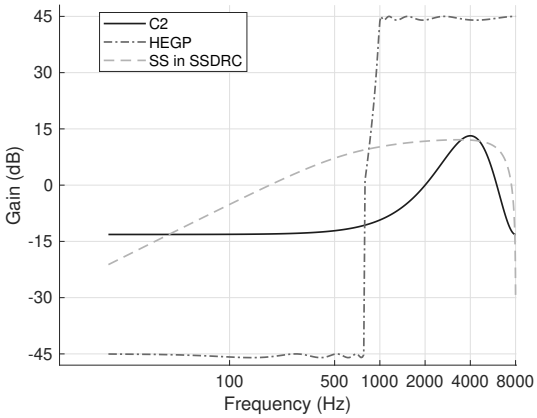


Figure 5-14: Investigated spectral shaping methods on analyzed speech

- Feature extraction: MS features and SRC features were extracted based on relations of SMS for intelligibility and naturalness and other perceptual models.
- Creation of stimuli: Modified speech was synthesized from plain speech by modifying extracted features by using a multi-rate signal processing technique, the MTF concept, and an SRC application. Then noise was added to create stimuli for listening tests.
- Evaluation: Listening tests for intelligibility and naturalness were conducted with the created stimuli to finally identify significant features.

Feature extraction

As shown in the feature extraction portion of Fig. 5-13, several speech enhancement methods were collected and applied to increase intelligibility, then the properties of the enhanced speech were investigated . Each method was used to mainly modify different acoustic and modulation frequency regions. The resulting intelligibility and naturalness scores differed. Then, significant acoustic and modulation frequency regions to modify the MS more were identified by using correlation. Finally, due to accompaniment with high modulation frequency regions in the MS, static range compression (SRC) was incorporated to fulfill the extraction feature set, as described below.

a. Differently enhanced speech

Three basic spectral shaping methods from different deriving hypotheses were used to enhance speech from plain speech (Fig. 5-14).

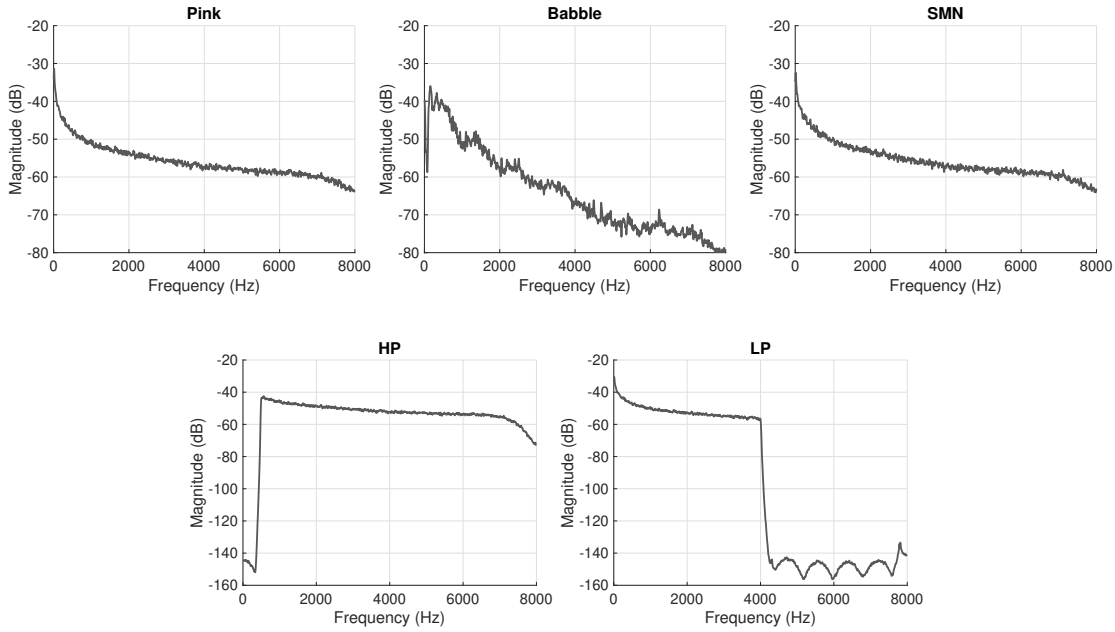


Figure 5-15: Long-term average spectra of the noise maskers used in the experiment for analyzed speech and in the creation of stimuli for evaluation of significantly effective features

The first method from the result from the previous section, which was called **C2**, was increasing spectral regions with a plateau from 2-6 kHz about 13 dB and decreasing spectra at other frequencies by 15 dB. The 13 and 15 dB were tunning values to best increase speech intelligibility. This shaping was the compensated spectrum derived from the 2^{nd} order cepstral coefficient

The second method, which was called **HEGP**, simulated the optimal spectral shaping in Tang et al. [4]: increasing spectral regions above 1 kHz by 45 dB, decreasing regions below 1 kHz by 45 dB.

The last method was the static spectral shaping in SSDRC [16], called SS. The SS was increasing spectra from 1-4 kHz by 12 dB, decreasing spectra by 6 dB/oct for below 0.5 kHz, and pre-emphasis. In addition to SS, formant sharpening (FS) and the DRC in SSDRC might well contribute to intelligibility. Thus, they were accumulated with SS and investigated as SSFS and SSDRC. The synthesizing method in SSDRC was inherited to apply those shaping modifications.

In total, there were six types of analysis speech: plain, C2, HEGP, SS, SSFS, and SSDRC. These speech were evaluated intelligibility and naturalness by performing a listening test experiment (Figs. 5-16 and 5-17, where Pearson correlation between intelligibility and natural-

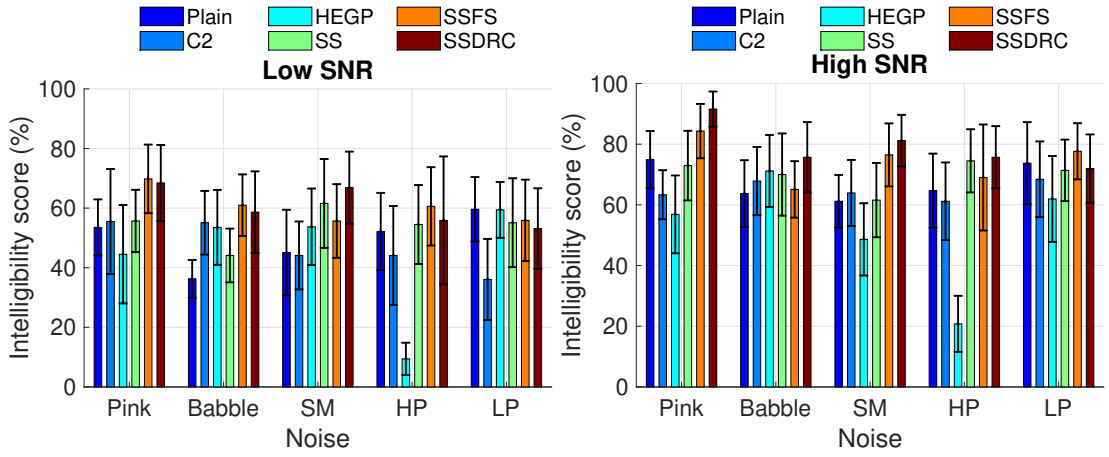


Figure 5-16: Intelligibility scores (percentage of correctly identified mora in a word) of analyzed speech in the presence of making noise at low and high SNRs.

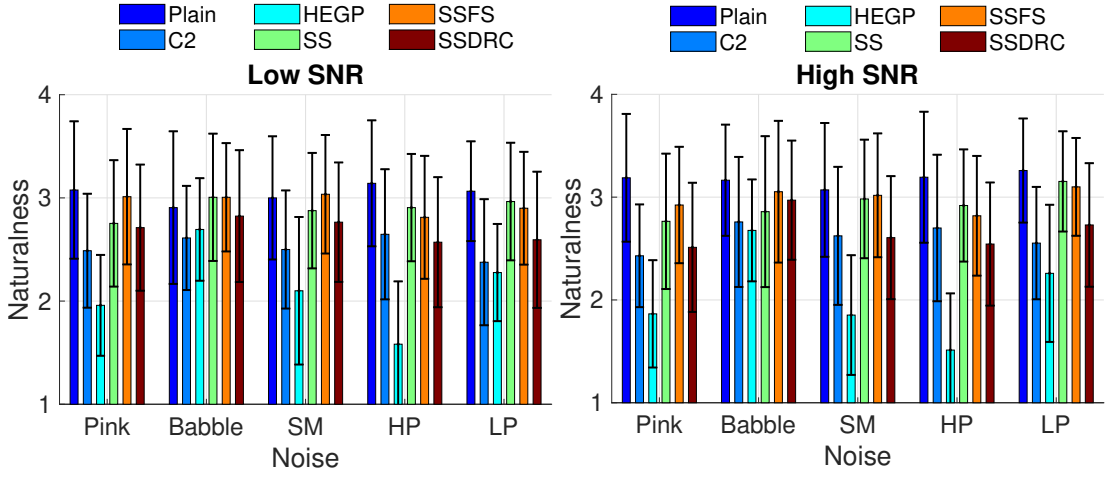


Figure 5-17: Naturalness scores of analyzed speech in the presence of making noise at low and high SNRs.

ness score was 0.61, which was only quite highly correlated. This correlation thus indicated that both intelligibility and naturalness scores still needed to be investigated in further analyses). We adopted the experiment design from the study of Tang et al. [4]. Speech materials were from the male speech in A-set of ATR dataset [87] with 600 three-mora words with an average duration about 350-450 ms, which was a suitable length for the MS extraction. Five noise maskers (Fig. 5-15, low and high SNR levels) used in the experiment were: Pink noise [74] (-9.5 dB of low SNR and -12 dB of high SNR), babble noise [75] (-12 and -9 dB), SM i.e., speech modulated noise created by multiplying the pink noise with the envelope of the babble noise (-10.5 and -7.5 dB), HP noise i.e., high-pass noise created by high-pass filtering the pink noise with cutoff frequency 0.5 kHz (-12 and -9 dB), and LP noise

i.e., low-pass noise created by low-pass filtering the pink noise with cutoff frequency 4 kHz (-13 and -10 dB). The SNR levels used in the tests were estimated to obtain at least 33% of correct-mora recognition. Thirty-six thousand stimuli (600 words $\times$ 6 speech types $\times$ 5 noises $\times$ 2 SNRs) re-sampled at 44,100 Hz and normalized to their average root mean square were involved in the experiment. Seventeen native Japanese (14 males and three females) aged mean 23.5 and standard deviation 1.7 with no report of hearing problems joined the tests and evaluated all speech types in all noise at two SNRs at 80 dB sound pressure level (SPL) in several sections. Within one section, one listener listened to 60 unique words at a noise type $\times$ an SNR level. The listeners did the intelligibility test and the naturalness test in sequence as follows.

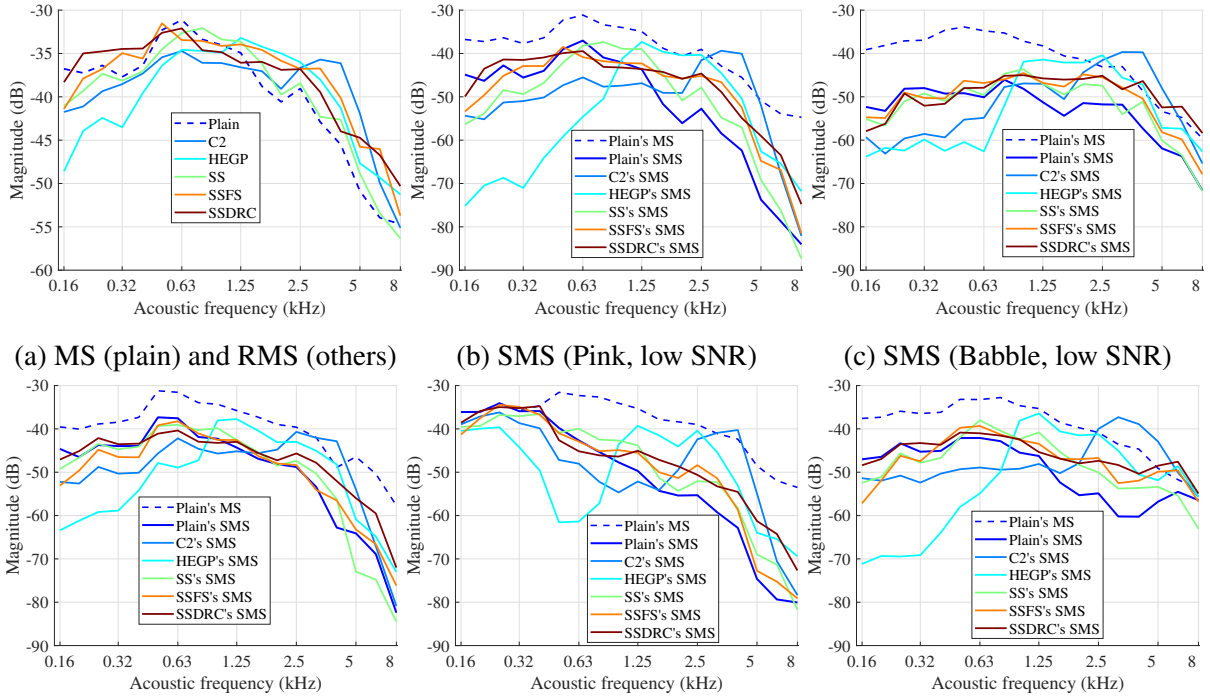
- Intelligibility: During this task, the stimulus was played only one time. The listeners were asked to write down the word they heard by using a keyboard. They clicked on the next button to continue.
- Naturalness: During this task, the stimulus could be played again. The listeners were asked to evaluate their feeling of naturalness (human voices) in four scales (1: unnatural, 2: rather unnatural, 3: rather natural, 4: natural) by clicking on the correspondent buttons. The next stimulus would be played immediately after that.
- Configuration: We conducted the experiment in a sound-proof room with a high-quality headphone (STAX SL51-2216) connected with a desktop computer via an amplifier (STAX SRM-1/MK-2). We used an amplifier to set an exact noise level of 80 dB SPL for the test, which was measured by a calibrated sound level meter (a hand-held analyzer type 2250 Bruel. & Kjar).

The clean speech (before added noise) and noise in the tests were used to calculate MS, RMS, MTF and SMS.

b. Features

We decomposed the enhancements on the basis of the relationships between RMS and intelligibility and between SMS and intelligibility to unfold their essential characteristics in “MS features." SRC was a supplemented feature for the MS features.

(i) MS features as frequency features



(a) MS (plain) and RMS (others) (b) SMS (Pink, low SNR) (c) SMS (Babble, low SNR)

(d) SMS (SM, high SNR) (e) SMS (HP, high SNR) (f) SMS (LP, high SNR)

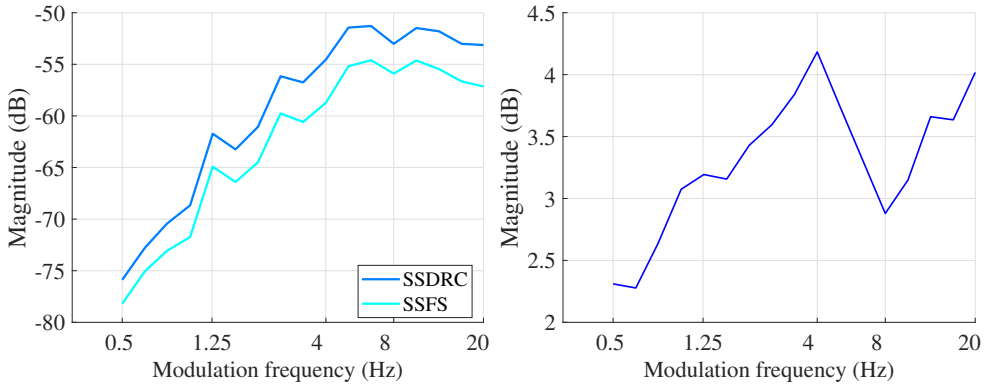
Figure 5-18: MS, RMS and SMS at 0 Hz modulation of analyzed speech in the presence of noise at some low and high SNRs.

The SMS at 0 Hz modulation showed the effect of noise on the frequency features Figure 5-18 shows the RMS and SMS at 0 Hz modulation of the analysis speech with different levels of intelligibility. It could be seen that MS (Fig. 5-18a) only reflected the increased modification made by the spectral shaping. However, SMS presented about what was increased comparing to the SMS of plain speech to reach to the MS of plain speech that might contribute to intelligibility shown in Fig. 5-16. The intelligibility was better for SS and its family than C2, HEGP, and plain speech. This better intelligibility might be because the SMS indexes of the frequency regions around 500 Hz, 1.25 kHz, 2.5-3 kHz, and 5-6 kHz were increased (as clearly shown in Fig. 5-18f). These frequencies seemed to relate to vowel formants (F_1 and F_2) and consonant bursts. The heavily decreased SMS for around 500 Hz might cause decreasing the intelligibility and naturalness of HEGP.

(ii) MS features as time features

The RMS obtained using DRC, which modified the time features, was the difference between the RMS of SSDRC and SSFS for 0-20 Hz modulation.

As shown in Figs. 5-16 and 5-17, SSDRC got higher intelligibility and lower naturalness than SSFS. This could be because DRC increased RMS at two regions: around 4 Hz and



(a) (b)

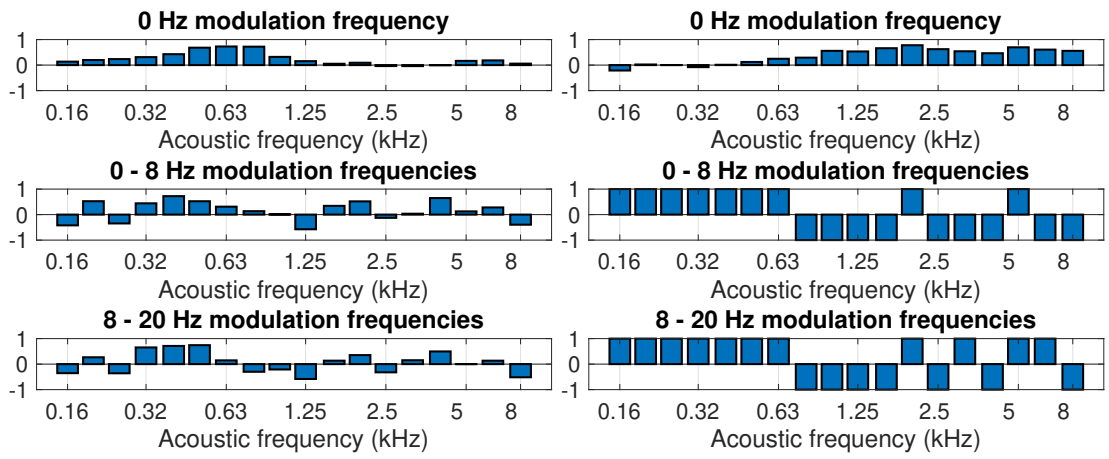
Figure 5-19: (a) RMS of SSDRC and SSFS and (b) their difference (RMS by DRC) over 0-20 Hz modulation in the acoustic spectrum of 5 kHz in the presence of SM noise at high SNR.

above 8 Hz modulation (Figs. 5-19a and 5-19b). As shown in Fig. 5-19b, this RMS had two peaks, one at around 4 Hz and one at around 20 Hz. These time features were coincident with the typical peak at 4 Hz of speech MS for speech intelligibility and above 8 Hz for speech prosody.

In particular, MS between 4 and 16 Hz modulation were reported to contribute to speech intelligibility most [61, 102, 103]. Also, Hermansky et al. [63] observed the dominated components between 2 Hz and 8 Hz modulation of continuous and uninterrupted speech. The components around 5-8 Hz modulation reflected rates of producing syllables across languages [104], which is an essential unit for speech perception. These frequencies were also hypothesized to be prominent in the cortical part of the human brain for syllabic segmentation with a time interval of 50-500 ms, i.e., 2-20 Hz modulation [105]. Furthermore, Zhu et al. [99, 106] argued that the MS between 8 Hz and 16 Hz modulation related to speaker individuality and emotions of speech. Therefore, speech intelligibility could be improved, and speech prosody could be affected, i.e., affecting naturalness by increasing MS at these two modulation frequency regions.

(iii) Correlation of MS features with intelligibility and naturalness

The Pearson correlations between SMS and RMS for each acoustic frequency band over three modulation frequency bands and the intelligibility/naturalness scores were calculated. Each acoustic frequency band for SMS at 0 Hz modulation was used in the calculation. Given that the environment was noise only, the effects of the environment on SMS for 0-20 Hz modulation were equal; it was thus needed to consider the RMS for 0-20 Hz, i.e., the



(a) In the stationary noise (b) In the babble noise

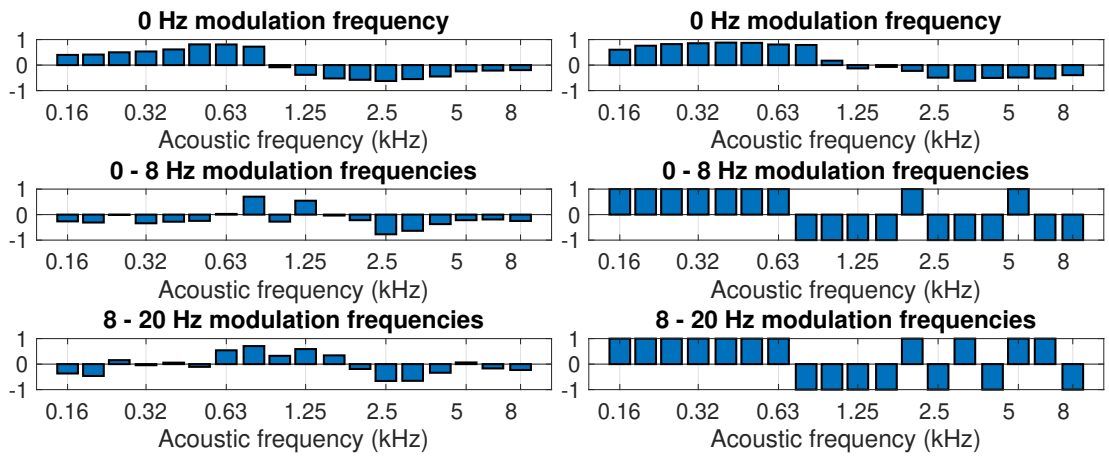
Figure 5-20: Pearson correlation between SMS and RMS for each acoustic frequency band (0, 0-8, 8-20 Hz modulation bands) and intelligibility scores for analyzed speech in noise in Fig. 5-16.

RMS by DRC. Also, due to the two peaks, each acoustic frequency band of RMS for the modulation frequency bands of 0-8 and 8-20 Hz of DRC was used in the calculation.

Figure 5-20 show correlations between SMS and RMS for each acoustic frequency band (0, 0-8, 8-20 Hz modulation bands) and intelligibility scores for analyzed speech in noise in Fig. 5-16. In the stationary noise of Pink, SM, HP, and LP noise (Fig. 5-20a), it was shown that some frequency regions below 500 Hz still had high correlation with the intelligibility scores. It seemed suitable to choose 250 Hz to include F1 for vowels /i/ and /u/. Thus the highly correlated frequencies with intelligibility were *relatively* 250/500-2250 Hz, 4.5-6.5 kHz, and around 4 Hz and above 8 Hz modulation in the acoustic spectra of 300-750, 1250-2250 Hz and 4.5-6.5 kHz.

In the non-stationary noise of the babble noise (Fig. 5-20b), it seemed that the highly correlated frequency regions with intelligibility could be started from 500 Hz or 1 kHz and continued expanding the entire higher frequency regions. This correlation meant that the frequency features could also be increased in MS above 500 Hz or above 1 kHz. These features seemed to be coincident with the arguments from previous studies [4], whether to increase the spectrum from 500 Hz or 1 kHz. Also, it could be seen that the time features in this babble noise were relatively as same as the time features in the stationary noise.

Figure 5-21 shows correlations between SMS and RMS for each acoustic frequency band (0, 0-8, 8-20 Hz modulation bands) and naturalness scores for analyzed speech in noise in Fig.



(a) In the stationary noise (b) In the babble noise

Figure 5-21: Pearson correlation between SMS and RMS for each acoustic frequency band (0, 0-8, 8-20 Hz modulation bands) and naturalness scores for analyzed speech in noise in Fig. 5-17.

5-17. In both stationary and non-stationary noises, in the SMS at 0 Hz modulation band, highly correlated frequencies with naturalness was the acoustical bands below 1 kHz. The increases in the MS at higher acoustical bands than 1 kHz harmed naturalness. However, it might be important to increase MS at these regions for speech intelligibility. Thus, they were still considered increasing with particular concerns with naturalness in the evaluation. RMS for acoustical bands around 500, 1250-2250 Hz, and 4.5-6.5 kHz (0-8, 8-20 Hz modulation bands) seemed highly correlated with naturalness.

The effects on the time features for naturalness were relatively the same as these for intelligibility. These shared features might come from a similar pattern between intelligibility and naturalness scores (as shown in Figs. 5-16 and 5-17). Therefore, the time-frequency features extracted by the correlation with intelligibility scores were used as the final features. Furthermore, the time features seemed to depend on the frequency features due to RMS for only specific acoustic frequency bands (0-8, 8-20 Hz modulation bands) positively correlating with intelligibility/naturalness scores.

(iv) Static range compression

The extracted MS seemed to present most of the wide-slowly changed time features due to its 0-20 Hz modulation. However, it was realized that the narrow-quickly equalization of the speech amplitude envelope by the SRC as one of essential local time features, which might be only appeared in the MS at higher modulation frequencies than 20 Hz, was out of

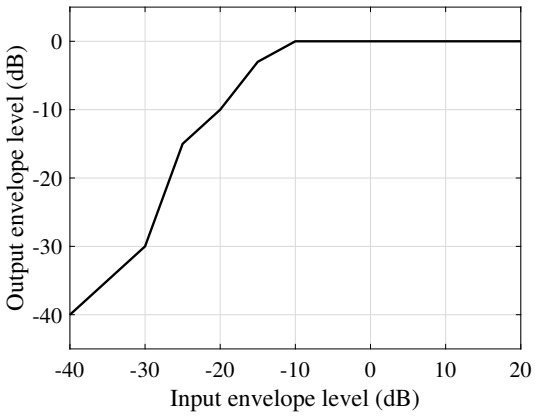


Figure 5-22: IOEC curve, redrawn by a linear interpolation from Fig. 1 in Zorila et al.'s study [16].

the scope of these MS features. Therefore, in this study, the SRC was ultimately inherited as a time feature to investigate. Figure 5-22 shows the IOEC curve, which had been used in SSDRC, was reused in this study. It shows that the input envelope was compressed in a range of −40 - 20 dB. The range of −20-20 dB of the input envelope was further reduced into the range of −10-0 dB to produce the output envelope.

From the explanation above, the tentative MS and SRC features were extracted as follows. The frequency features could be sparsely increased in the MS around 250/500-2250 Hz (acoustic frequency (AF) region 1) and 4.5 - 6.5 kHz (AF region 2) acoustic frequencies or continuously increasing in the MS above 500 Hz or above 1 kHz. The time features could be increased in MS around 2-6 Hz (modulation frequency (MF) region 1) and 8-20 Hz (MF region 2) modulation in the acoustic spectra of 300-750, 1250-2250 Hz and 4.5-6.5 kHz and the SRC characterized by the IOEC curve in Fig. 5-22.

Creation of stimuli

As shown in Fig. 5-13, to synthesize MS-modified speech, an analysis-synthesis method (see the following description of a. Modification of modulation spectrum) based on a multi-rate signal processing technique [107] was developed to modify MS of plain speech by amplifying acoustical and modulation bands as described in the extracted MS features. The amplified values were estimated from MTF of reference noise by Eq. 5.3, which were limited in ranges to preserve the voice quality of the plain speech. Basically, the MTF was averaged within specific regions of the AF or MF features and taken inverse. If the resulting value was still within

limited ranges, it was directly used as the amplified amount of these spectral/modulation spectral regions, otherwise the limitation was used. The AF regions 1 and 2 of the sparse frequency features and the frequency regions of the continuous frequency features were all empirically limited to 10-15 dB, i.e., a mostly flat response for a fair comparison of intelligibility. We increased both MF regions 1 and 2 of the time features by 4 dB (as the peak value from Fig. 5-19b). Modification of speech amplitude envelope by SRC (see the following description of b. Application of static range compression) was applied in the same way as in SSDRC. We created the stimuli for the evaluation by two steps:

(1) Plain speech was from the male speech in A and C sets of ATR dataset with 600 unique three-mora words (200 words from the C set and 400 from the A set). The plain speech of each word is one variant. The extracted features of the plain speech of each word were used to generate the remaining 19 variants of each word by either individually or jointly changing/applying MS feature as frequency features (AF features), SRC (SRC features), and MS features as time features (MF features) to a setting as the extracted features mentioned above. Table 5.1 gives an overview of the plain speech settings and the MS and SRC-modified speech settings for the three features. We synthesized all speech items as 16-bit mono signals with a sampling frequency of 16,000 Hz.

(2) To create stimuli for evaluation by listening tests, each speech variant was re-sampled at 44,100 Hz and added noise (as Fig. 5-15 with pink noise, babble noise, SM noise, LP noise, and HP noise) to have targeted global low and high SNRs for each noise type. The SNRs followed the same specifications as in the listening experiment with the analyzed speech. The stimuli were superimposed with noise had a total length of 1 s with the target words embedded in the middle. Afterward, all stimuli were normalized to their average root-mean-square to obtain the final stimuli.

a. Modification of modulation spectrum

To imitate the MS analysis, enabling reconstruction, a multi-rate signal processing technique was used for the MS analysis, modification, and synthesis steps (Fig. 5-23a). An acoustical analysis bank was used to filter plain speech into band-limited signals, and then the power envelopes of the band-limited signals were extracted. Next, to avoid modifying non-speech segments (modifying them might cause noise), VAD was used to mask the speech-absent portions of these power envelopes with a silence threshold of 0.005 on the

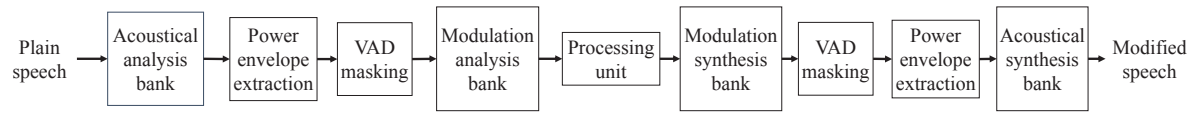
Table 5.1: Plain speech settings and MS and SRC-modified speech settings of three examined features used for multirate-signal processing synthesis.

Feature	Setting for plain speech	Settings for MS and SRC-modified speech
AF features	**N.A** (No modification in AF regions)	**A250s**[1], **A500s**[2], **A500c**[3] and **A1000c**[4] (all increased within the range 10-15 dB)
SRC features	**N.S** (No SRC applied)	**S** (SRC applied)
MF features	**N.M** (No modification in MF regions)	**M** (increased[5] both the MF region 1 of 2-6 Hz and the MF region 2 of 8-20 Hz by 4 dB)

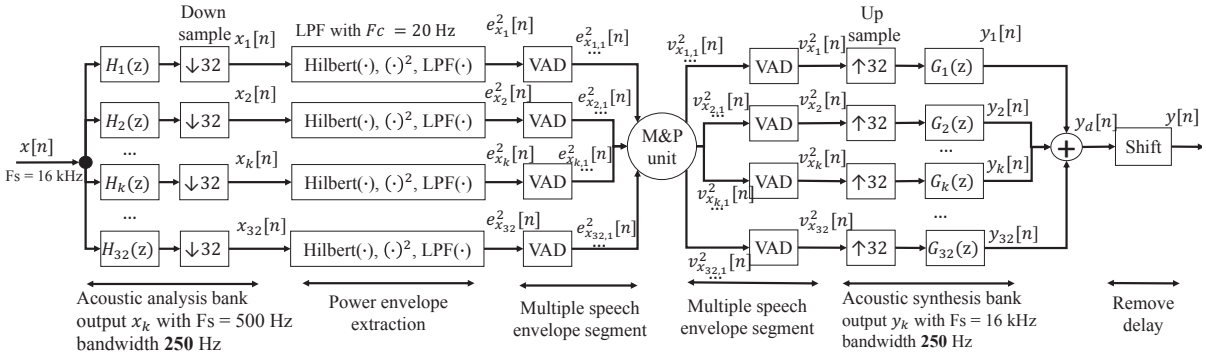
[1] A250s (AF250 sparse) includes two spare AF regions **250 Hz**-2250 Hz and 4.5-6.5 kHz
[2] A500s (AF500 sparse) includes two spare AF regions **500 Hz**-2250 Hz and 4.5-6.5 kHz
[3] A500c (AF500 continuous) includes only one continuous AF region of *500-8000 Hz*
[4] A1000c (AF1000 continuous) includes only one continuous AF region of *1000-8000 Hz*
[5] The increases were only in the acoustic spectra of 300-750, 1250-2250 Hz and 4.5-6.5 kHz.

max-value-normalized power envelopes. A modulation analysis bank was then applied to each speech-segment power envelope. Afterward, a processing unit with gain control amplified specific acoustical bands and modulation bands (as mentioned in settings in Table 5.1). It sequentially applied gains to the acoustic frequency regions (all 0-20 Hz modulation regions) and the modulation frequency regions (0-20 Hz, excluding 0 Hz). Finally, to obtain the modified speech, the reconstruction was processed in inverse order from the analysis with modulation synthesis bank, VAD, and acoustical synthesis bank. The processes are illustrated in Figs 5-23b and 5-23c. Further information for designing the acoustical analysis/synthesis banks and the modulation analysis/synthesis banks was described as follows.

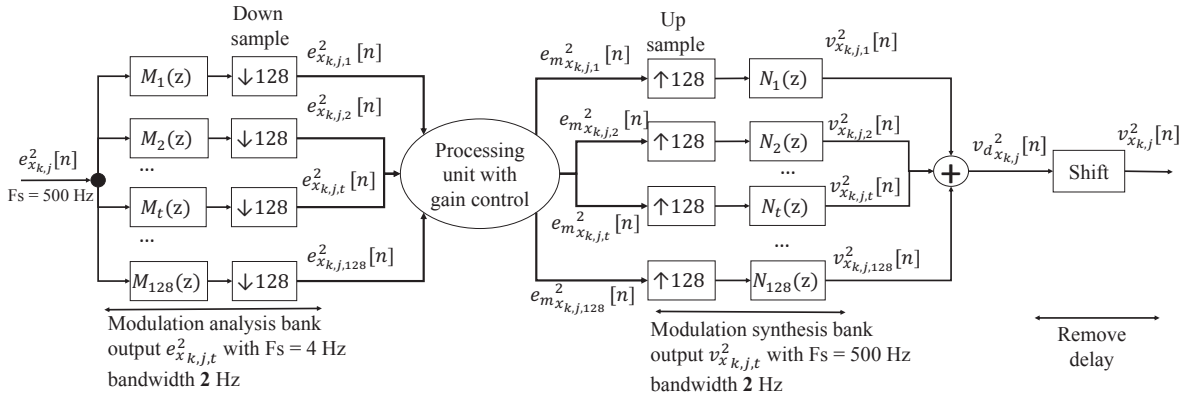
We designed equal-bandwidth filter banks with a perfect reconstruction of signals using a tree-based model [107, Chapter 12]. Starting from the basis of the analysis/synthesis orthogonal bank with two FIR filters (implemented by the function *firpr2chfb* in Matlab), expanding to deeper levels, i.e., five levels was to construct 32 (i.e., 2^5) filters for each acoustical analysis/synthesis bank. Also, expanding seven levels was to create 128 (i.e., 2^7) filters for each modulation analysis/synthesis bank. Thirty-two filters or bands were chosen because the sampling frequency of plain speech was at 16 kHz, and it was to obtain a bandwidth of 250 Hz on each acoustical band. This bandwidth was suitable to represent the extracted



(a) Block diagram of process for converting plain speech into MS-modified speech using multi-rate signal processing technique.



(b) Acoustic analysis and synthesis banks, power envelope extraction, and power envelope masking with voice activity detection (VAD)



(c) Modulation analysis and synthesis banks (**M&P unit**)

Figure 5-23: Conversion of plain speech into MS-modified speech using multi-rate signal processing technique.

MS features as frequency features. Due to the bandwidth of 250 Hz of the acoustical band, the modulation analysis/synthesis banks thus needed 128 filters to obtain a bandwidth of 2 Hz for each modulation band to enable us to modify the extracted MS features as time features. These banks of acoustical and modulation analyses/syntheses can obtain a perfect reconstruction of signals with only some delay. The delay was removed by shifting the reconstructed signal.

b. Application of static range compression

As was used in Zorila et al.'s study [16], SRC was applied in the final step after obtaining the modification of other features. We extracted the absolute amplitude envelope of the

plain speech (if no other features were modified) or the MS-modified speech of each word. The envelope extraction was proceeded by applying Hilbert transform and moving-average smoothing over 15 ms (15 ms was empirically tested as the best value for not making noise). Then, the amplitude, normalized by 0.3 times of the maximum amplitude value in the range of −40-20 dB only, was taken as the input envelope to IOEC curve (Fig. 5-22) to obtain the output envelope. The ratio between the output envelope and the input envelope of each word was calculated and defined as gain. By multiplying this gain with the plain speech or the MS-modified speech of each word, finally the SRC-modified speech of each word was obtained.

Evaluation

As shown in Fig. 5-13, the evaluation was about doing listening tests. It had two tasks. The first task was listening tests for intelligibility. Then the second task was the test for naturalness. One hundred twenty thousand stimuli from 20 speech variants differed with AF, SRC, MF features of 600 words added with five noise maskers at 2 SNR levels were involved in the evaluation. Seventeen native Japanese, including 13 men and four women with an average age of 25.2 years and a standard deviation of 3.1 years, participated in the tests. All participants gave informed consent and reported no hearing problems. Each participant evaluated all 20 speech variants in all five noise types at 2 SNR levels. The participant evaluated five speech variants at a time. Thus, each participant performed four times (separated by the minimum two-hour break) of the two tasks in two consecutive sessions.

In each time, for the evaluation of the intelligibility of the speech variants (first task), each participant went through 5 parts corresponded to 5 noise types (Pink noise, LP noise, HP noise, SM noise, and babble noise in sequence) of low and high SNR levels. In each part, the participant listened to 80 stimuli (80 unique words for five variants x 2 SNR levels of a noise type) in random order using high-quality headphones (STAX SL51-2216) connected with a desktop computer via an amplifier (STAX SRM-1/MK-2) in a soundproof room. The volume was adjusted to make all stimuli played at 80 dB SPL, which was measured by a calibrated sound level meter (a hand-held analyzer type 2250 Bruel. & Kjar). After each stimulus was played, the participants had to type a three-mora word by using a keyword. If they could not clearly understand the spoken word, they were allowed to type a random three-mora word. After typing

an answer to the current stimulus, they clicked on the Next button, and the next stimulus was automatically played (repetitions were *not possible*). After listening to every 20 stimuli, they could take a short break within one minute. After finishing the current part, they moved to the next one within a 2-minute break. The session took about 55 minutes.

After a short break, the participants started the second session for the second task. In this task, each participant went through the same five parts by using the same equipment as in the first session. Also, in each part, the participant listened to 80 stimuli in random order. After each stimulus was played, the participants were asked to rate the naturalness among four options, 1 - unnatural, 2 - rather unnatural, 3 - rather natural, or 4 - natural, by clicking on one of four buttons with the respective labels. After choosing the answer to the current stimulus, the next stimulus was automatically played (repetitions were *possible*). Also, after listening to every 20 stimuli, the participants took a short break of 1 minute. After finishing the current part, they moved to the next part within a 1-minute break. The whole session lasted about 30 minutes.

5.2.4 Results and discussion

Perceptual test of intelligibility

Figures 5-24 and 5-25 show the results of the intelligibility tests of 20 speech variants of the plain speech and the MS and SRC-modified speech in the presence of Pink noise, babble noise, SM noise, HP noise and LP noise at low and high SNR levels respectively. In general, it showed that A500s, A500c, A1000c, and S seemed to perform better than A250s and M. The A500s, A500c, A1000s features combined with both M and S had a better performance than each individual or a mutual combination of two features. Furthermore, S sometimes made a drastic increase in intelligibility when acting individually or combined with others. Also, M combined with A500s seemed to show a more increasing effect on intelligibility in the babble noise and the SM noise. On the other hand, A250s seemed not a suitable modification because its increasing scores on intelligibility were quite small compared to others. Moreover, A250s often got decreases in intelligibility in comparison to plain speech.

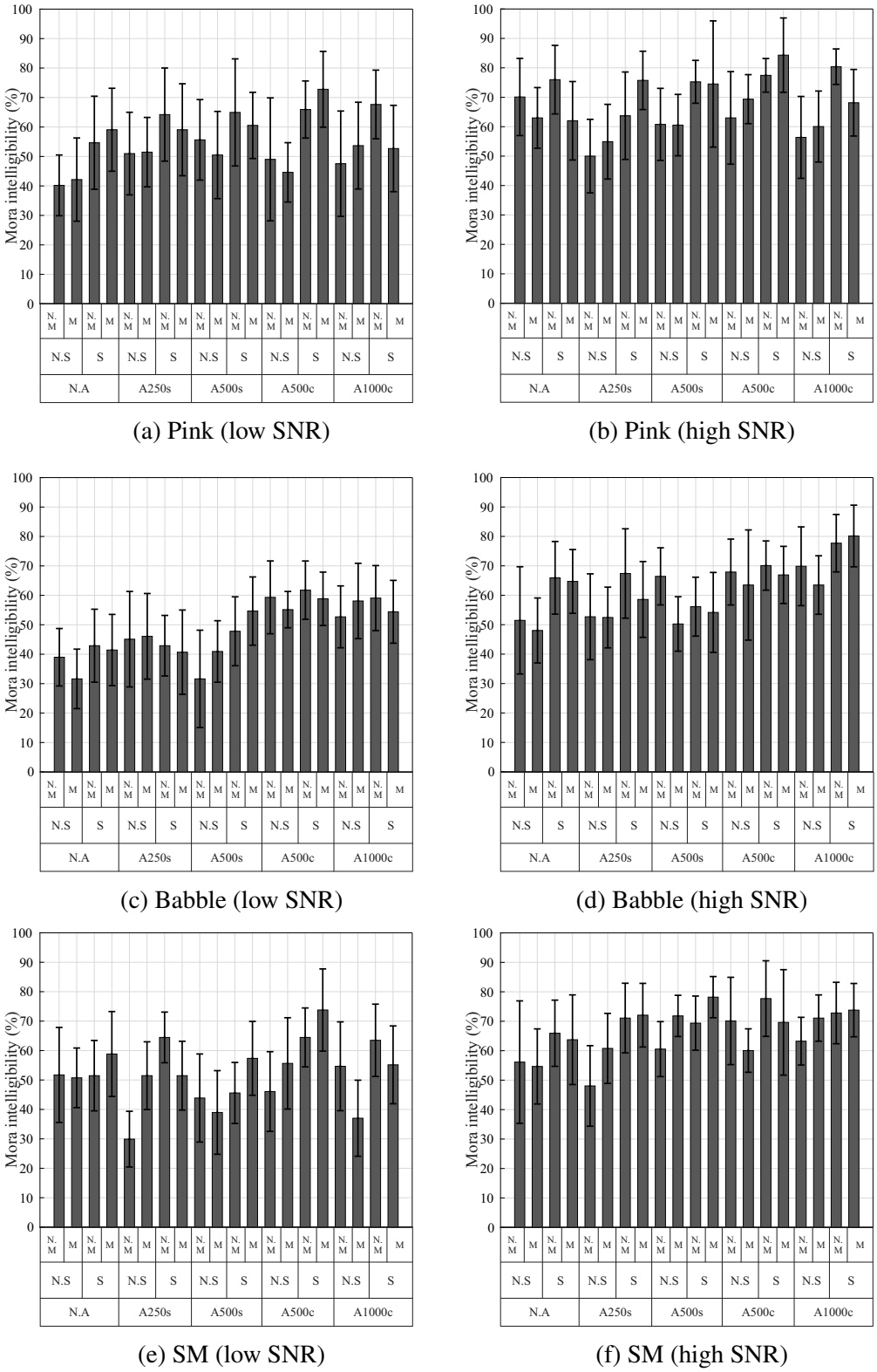
To study the effect of these features on the intelligibility score and their robustness against the environments in more detail, a five-way repeated ANOVA measure was conducted. The AF features, SRC features, MF features, noise types, and SNR levels were the five factors. The dependent factor was the intelligibility score. In particular, AF features had five levels (N.A,

A250s, A500s, A500c, A1000c). SRC features had two levels (N.S and S). MF features had two levels (N.M and M). Also, noise types had five levels (Pink, Babble, SM, HP, and LP). And, SNR levels had two levels (low and high SNRs).

In detail, there were significant main effects for four of the factors, that is, SRC features [$F(1, 16) = 187.11$, $p < 0.001$], AF features [$F(4, 13) = 59.71$, $p < 0.001$], noise types [$F(4, 13) = 31.79$, $p < 0.001$], and SNR levels [$F(1, 16) = 3040.69$, $p < 0.001$]. Across all noise types and SNR levels, on average, the SRC-modified speech increased 6.76 % intelligibility score more than the non-SRC-modified speech and the AF-modified speech increased 5.29 % more than the non-AF-modified speech. The MF features (i.e., increased in MS 2-6 Hz and 8-20 Hz modulation regions) have no significant main effect [$F(1, 16) = 1.19$, $p = 0.291 > 0.05$]. The effect size was strongest for SNR levels ($\eta^2 = 0.995$), followed by AF features ($\eta^2 = 0.948$), SRC features ($\eta^2 = 0.921$), noise types ($\eta^2 = 0.907$), and MF features ($\eta^2 = 0.069$). In addition, among features there was a multiple significant interaction between factors, namely, between AF features and MF features [$F(4, 13) = 6.23$, $p = 0.005$, $\eta^2 = 0.657$]. however, there were no significant multiple interaction between AF features and SRC features ($p = 0.083$). There was no multiple significant interaction between AF features and noise types ($p = 0.089 > 0.05$) or between AF features combined with SRC features and noise types ($p = 0.572 > 0.05$), while there were multiple significant interactions between SRC features and noise types ($p < 0.001$) and between MF features and noise types ($p = 0.014$). In interaction with SNR levels, there was no significant interaction for SRC features, while there were significant interactions for AF features and MF features. In interaction with noise types along with SNR levels, there were no significant interactions for AF features, SRC features, and AF features combined with SRC features and MF features. At the same time, there was a significant interaction for MF features. With larger effect sizes than noise types and no significant interactions with noise types and SNR levels, it indicated that the AF features and SRC features seemed to be able to robust with noise types and SNR levels.

Figure 5-26 shows the intelligibility for each AF feature averaged over all other factors. It indicates that A500c seemed to get the highest score. To study the significant differences among AF features of Normal A, A250s, A500s, A500c, and A1000c, a pairwise comparison was conducted on the basis of a posthoc test with Bonferroni correction. It was shown that comparing to Normal A (N.A), there were significant differences for A500s ($p = 0.003$), A500c

($p < 0.001$), and A1000c ($p < 0.001$), while there was no significant difference for A250s ($p = 1.000$). Also, across all noise types and SNR levels, on average, the speech by modifying A500s, A500c, and A1000c increased 6.72 % intelligibility score more than the speech without modifying A500s, A500c and A1000c. Also, the difference between A500s and A500c and the difference between A500s and A1000c were significant ($p < 0.001$). There was no significant difference between A500c and A1000c ($p = 1.000$). Also, across all noise types and SNR levels, on average, the speech by modifying A500c and A1000c increased 8.59 % intelligibility score more than the speech without modifying A500c and A1000c.



(a) Pink (low SNR)

(b) Pink (high SNR)

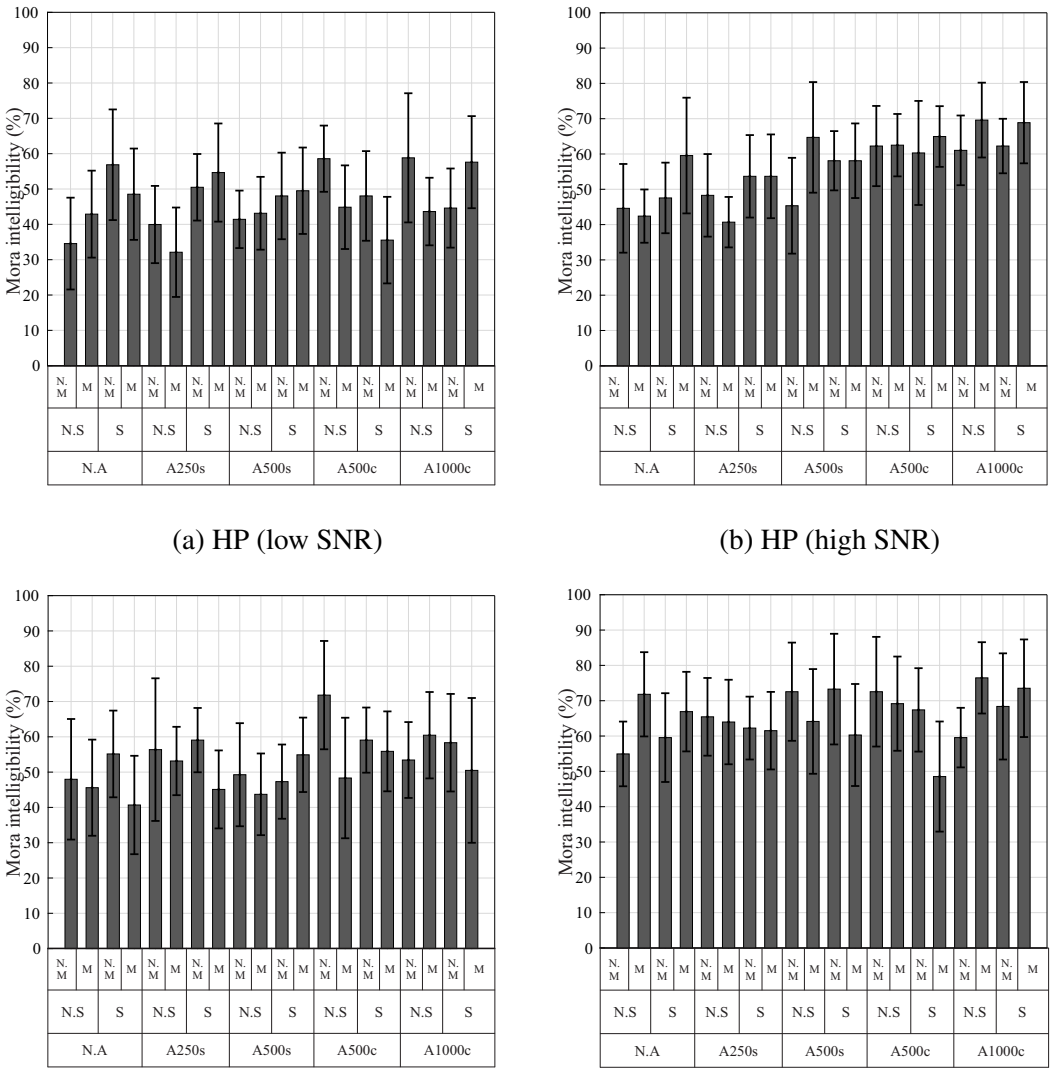
(c) Babble (low SNR)

(d) Babble (high SNR)

(e) SM (low SNR)

(f) SM (high SNR)

Figure 5-24: Intelligibility scores (percentage of correctly answered mora in a word) of plain speech and the MS and SRC-modified speech in the presence of Pink noise, babble noise, and SM noise, at low and high SNRs. The labels indicate feature combinations, where "N." stands for the neutral setting of a feature (as in plain speech), "A" stands for the MS feature setting as frequency features as in Table 3.1, "S" stands for SRC, and "M" for the MS feature setting as time features.



(a) HP (low SNR)

(b) HP (high SNR)

(c) LP (low SNR)

(d) LP (high SNR)

Figure 5-25: Intelligibility scores (percentage of correctly answered mora in a word) of plain speech and the MS and SRC-modified speech in the presence of HP noise and LP noise at low and high SNRs. The labels indicate feature combinations, where "N." stands for the neutral setting of a feature (as in plain speech), "A" stands for the MS feature setting as frequency features as in Table. 3.1, "S" stands for SRC, and "M" for the MS feature setting as time features.

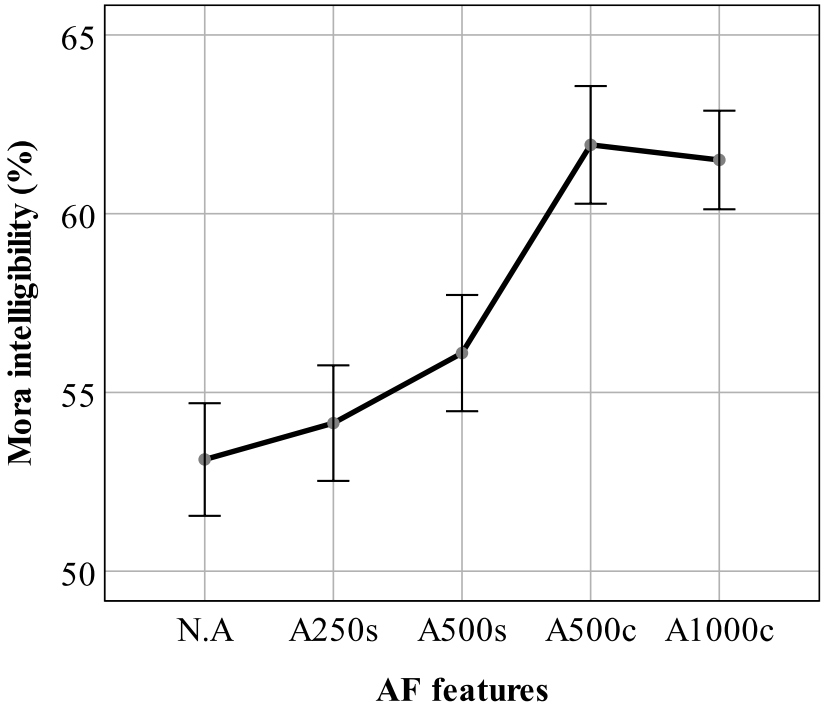


Figure 5-26: Intelligibility scores (percentage of correctly answered mora in a word) of the AF-modified speech for each AF feature, averaged over all the other factors (MF features, SRC features, noise types, and SNR levels)

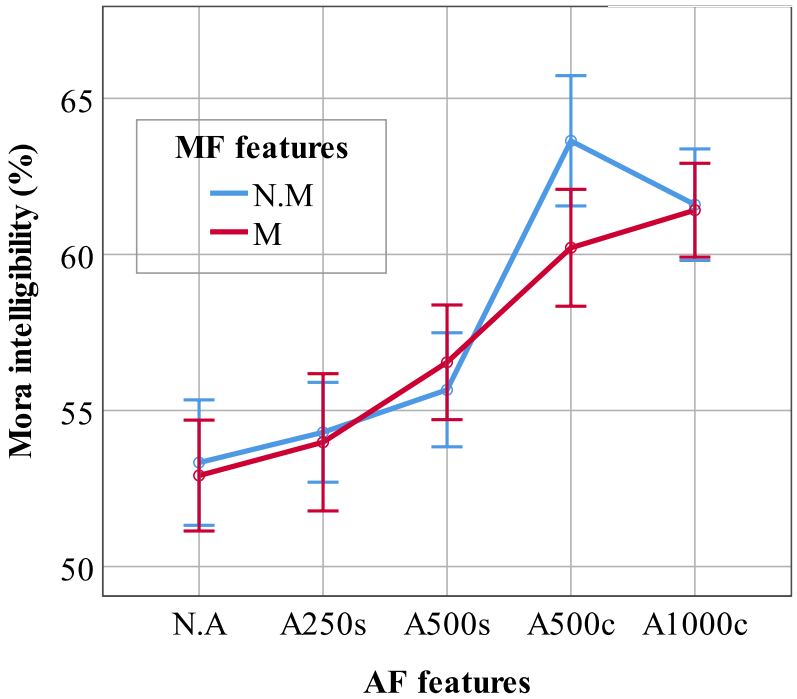


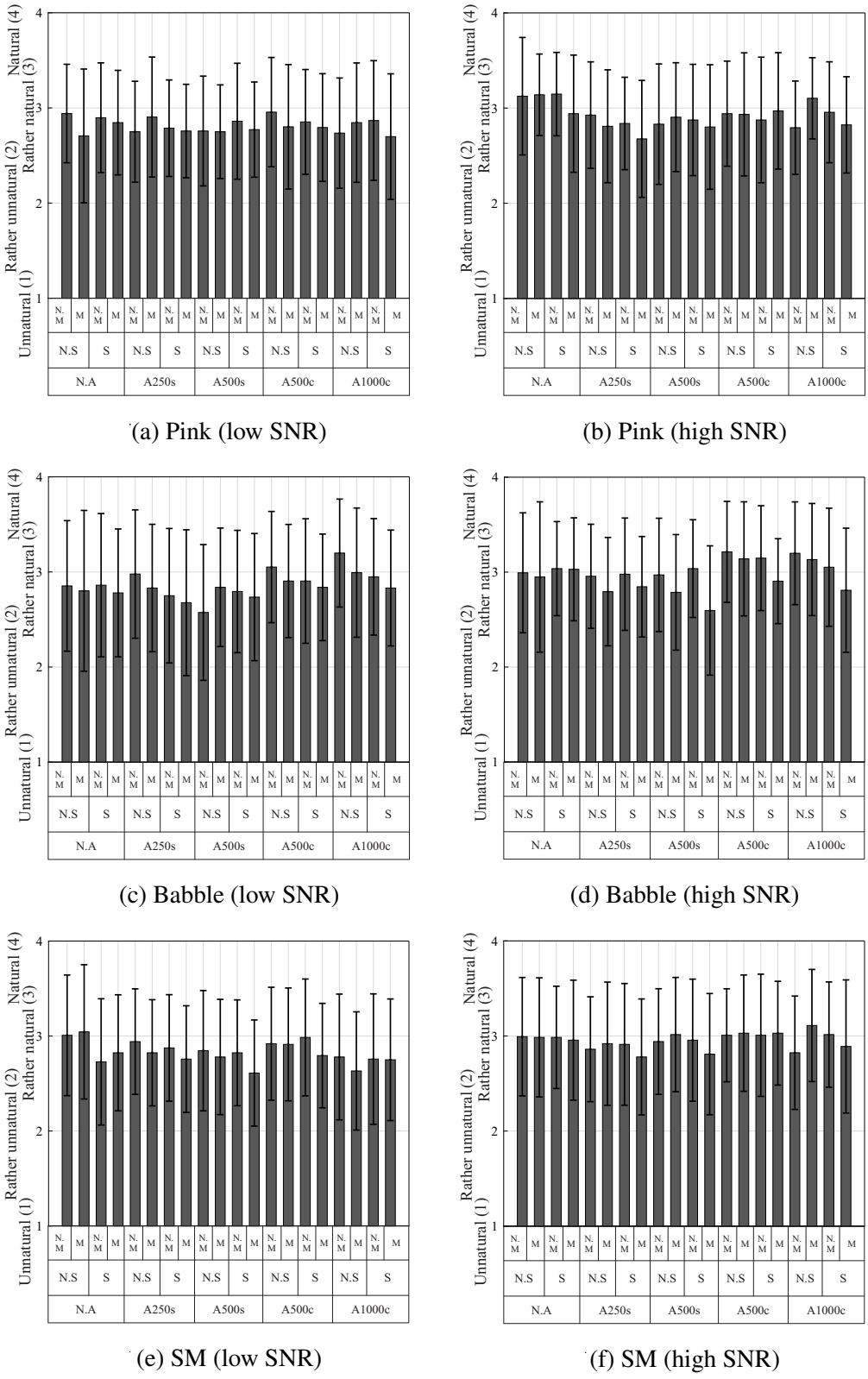
Figure 5-27: Intelligibility scores (percentage of correctly answered mora in a word) of the AF and MF modified speech for each AF feature combined with each MF feature, averaged over all the other factors (SRC features, noise types, and SNR levels)

In summary, it was found that the change of AF features from normal to A500c and A1000c improved the intelligibility of the words most. This improvement was independently from the type of background noise but dependently on SNR levels of noise. Given that the effect of mid-high frequency regions, this finding answers the question in the previous study of [4] that increasing frequency regions from 500 Hz or increasing frequency regions from 1000 Hz is equivalent. The dependence on SNR levels for AF features might be because the limited range of increasing AF regions made this modification less adaptive with different SNR levels. Still, for comparison among features, it was necessary.

Furthermore, it was also found the change by SRC to be highly effective at increasing the intelligibility in all types of noise along with SNR levels, just followed after A500c and A1000c. Though A500s had lower performance than A500c, it could also be considered useful features to combine with M features (as shown in Fig. 5-27, only A500s additively increased intelligibility with M). This lower performance is also in line with the finding in [18], where applying the continuous frequency response was better than using the sparse frequency response to increase speech intelligibility. The MF features did not affect individually. This result might be because the testing environments were noise when it is more common for testing in reverberation with MF features. However, their combination with AF features got an effect. This accumulative effect is also in line with the synergistic contribution between time and frequency features found in the previous studies by [18] and [17]. However, although A500c or A1000c was shown the best intelligibility overall in noise, it seemed not right when combining them with the time feature M (see shown in Fig. 5-27). This result might lead to another assumption. That is, frequency features and time features only support each other if they are in a suitable pair. Therefore, the mission to identify an appropriate time feature for these frequency features has still remained. And, it is possible to find them out when investigating in noisy reverberant conditions.

Also, it can be seen that not only 0-20 Hz modulation regions but also the higher modulation regions than 20 Hz, which was reflected by SRC, contributed much to the intelligibility of speech in noise.

Perceptual test of naturalness



(a) Pink (low SNR)

(b) Pink (high SNR)

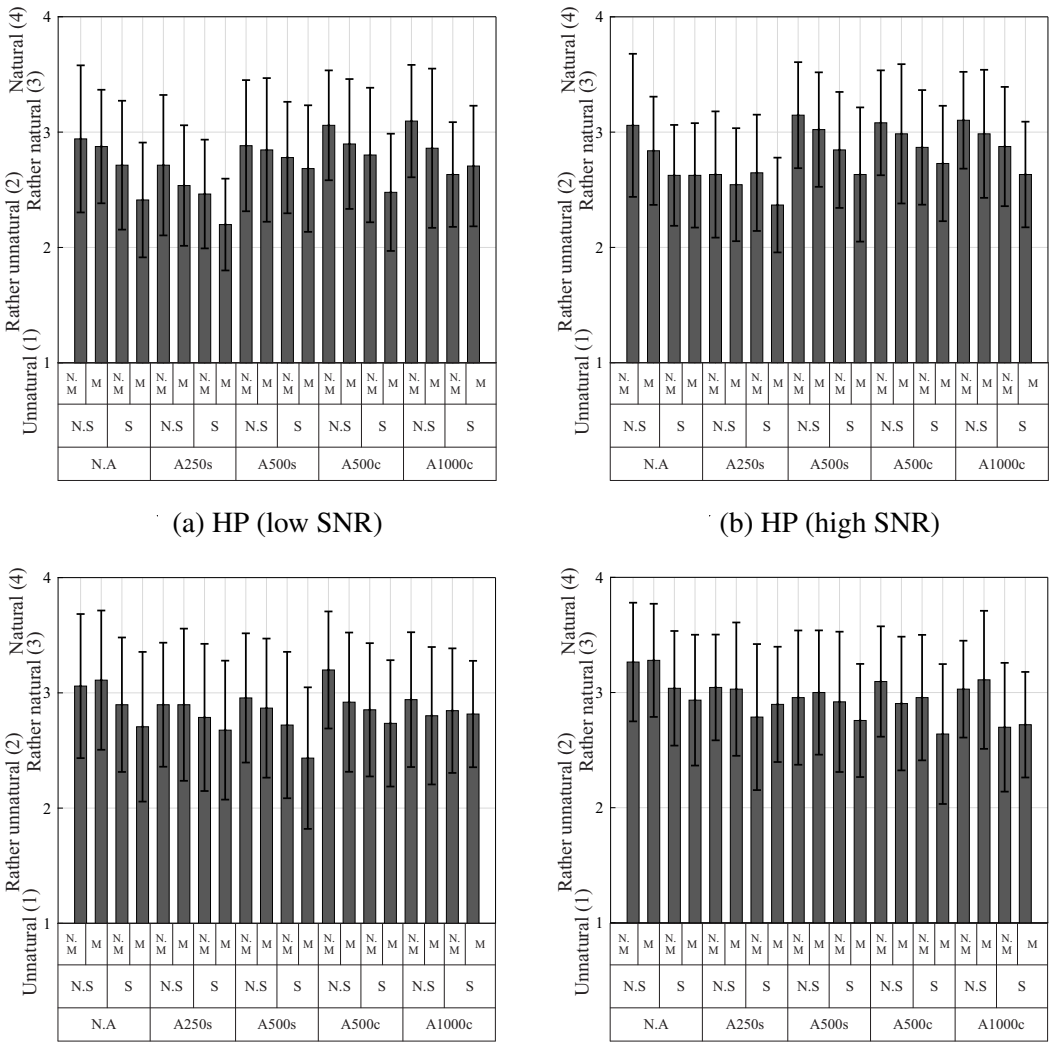
(c) Babble (low SNR)

(d) Babble (high SNR)

(e) SM (low SNR)

(f) SM (high SNR)

Figure 5-28: Naturalness scores of plain speech and MS and SRC-modified speech in the presence of of Pink noise, babble noise, and SM noise at low and high SNRs. The labels indicate feature combinations, where "N." stands for the neutral setting of a feature (as in plain speech), "A" stands for the MS feature setting as frequency features as in Table. 3.1, "S" stands for SRC, and "M" for the MS feature setting as time features.



(a) HP (low SNR)

(b) HP (high SNR)

(c) LP (low SNR)

(d) LP (high SNR)

Figure 5-29: Naturalness scores of plain speech and MS and SRC-modified speech in the presence of HP noise and LP noise at low and high SNRs. The labels indicate feature combinations, where "N." stands for the neutral setting of a feature (as in plain speech), "A" stands for the MS feature setting as frequency features as in Table. 3.1, "S" stands for SRC, and "M" for the MS feature setting as time features.

Figures 5-28 and 5-29 show the results of the naturalness tests of 20 speech variants of plain and MS and SRC-modified speech in the presence of Pink noise, babble noise, SM noise, HP noise and LP noise at low and high SNR levels respectively. In general, it showed that most of the speech variant was rather natural. This result seemed reasonable due to the features that had been extracted by considering the positive correlation with naturalness. Perhaps, the variants by A250s or by SRC as modified features slightly reduced the naturalness of speech comparing to others. This reduction was in line with previous studies about SRC. A250s increased low-frequency regions around 250 Hz, and this might be similar to proximity effect [93]. This effect makes speech sound closer to the ear, thus affected naturalness.

5.2.5 Summary

This section presented the method for identify effective features from various enhancement methods.

Smeared modulation spectra of the speech signals by both stationary and non-stationary noisy environments were derived on the basis of the modulation spectrum and modulation transfer function concepts. Correlated acoustic, and modulation frequencies in the modulation spectrum with intelligibility and naturalness under noise were extracted from differently enhanced speech by analyzing relations of the smeared modulation spectra for intelligibility and naturalness scores from listening tests and considered as modulation spectral features. Concerning the other model supporting the proposed concept, the static range compression was incorporated as an additional feature to investigate. The extracted features then included: MS features as frequency features or AF features, MS features as time features or MF features, and SRC features. We then synthesized the modified speech from a plain speech by controlling these features. Listening tests of intelligibility and naturalness for the modified speech and plain speech were conducted under various noisy conditions. As a result, significant features to increase intelligibility and preserve naturalness were identified. In other words, the proposed concept, obtaining significant characteristics from different enhancement methods by relations of smeared modulation spectra for intelligibility and naturalness, is valid. That is, increasing AF features from 500 Hz or 1 kHz improved intelligibility most, followed by SRC. MF features (i.e., increasing 2-6 Hz and 8-20 Hz modulation regions in the acoustic spectra 300-750, 1250-2250 Hz, and 4.5-6.5 kHz) individually had no significant contribution to the intelligibility of speech in

noise. However, when combined with the AF features (in particular, sparse AF feature, i.e., A500s with an effect of accumulated increasing intelligibility), they played a synergistic effect. It also indicated that the higher modulation frequency regions (somewhere above 20 Hz modulation) should be modified. Modifying these features still preserved the naturalness of speech in noise. Concerning the A500s plus M, for frequency features, A500s seemed to boost spectral amplitude at more specific important frequency regions of vowel formant frequency F_1 and F_2 (around 500-2000) and consonant bursts or the piriform fossa (around 4500-6500 Hz). The other enhanced methods aimed at some of them: C2 increased the amplitude F_3, piriform fossa/consonant burst; SS increased amplitudes of formants F_2 and F_3; HEGP increased F_2, F_3, piriform fossa/consonants bursts. It seems that A500s collected most of the specific properties of others, introduced one new (F_1), and filtered out one (F_3). For time features, M was the effect of DRC, perhaps highly capable of the dynamic state of DRC. The increases in modulation in specific spectra might be also related to vowel formant regions, consonant bursts, and piriform fossa. The only limitation was in the representation of time-frequency features by MS. Due to strictly following the modulation frequency regions between 0 and 20 Hz, this prevented MS from being able to represent the quick modulation existed in higher modulation frequency regions, which was also found to relate to intelligibility in Ngo et al. [28]. Therefore, the static range compression had to be inherited as a quick fix to resolve this problem. In future work, it will be to explore these higher modulation frequency components in MS.

5.3 Effective features and their final mission to exceeding the intelligibility and the naturalness of Lombard speech

Lombard speech was studied to extract its effective features and the most effective one was increasing the spectrum by a plateau between 2-6 kHz about 13 dB. However, this effective feature was varied among many advanced enhancement methods. Thus the concept of extracting effective features among the differently enhanced speech synthesized by these enhancements was proposed. As reported in Cooke et al.'s study [56], SSDRC speech obtained better intelligibility of Lombard speech. To confirm the mission of exceeding the intelligibility of Lombard speech, so far, an experiment had been conducted to compare the speech modified by the effective features with SSDRC.

The testing environment was noisy reverberation. Because only A500s additively increased intelligibility with M, the features A500s and M were used to synthesize the MS modified speech in this comparison, so-called MS500 speech. Because the environment contained reverberation, MTF for reverberation was then defined as $m_R(f_a, f_m)$ using the modulation filtering technique for given a delivered room impulse response (RIR). Then, MTF in noisy reverberant conditions was then calculated by Eq. (5.5) as

$$m(f_{a_k}, f_m) = m_N(f_a, f_m) \times m_R(f_a, f_m) \tag{5.5}$$

The amplified values for the synthesis of MS500 speech were then estimated from MTF of reference noise and RIR by Eq. (5.5), which were empirically limited in ranges to preserve voice quality from plain speech and obtained the best settings for increasing intelligibility. That was, the AF regions 1 and 2 were limited in 5-15 dB and 15-25 dB, respectively, i.e., an incremental response for intelligibility. Both MF regions 1 and 2 were limited in 6-10 dB.

Speech materials were drawn from the male speech in A-set of ATR dataset [87] with 120 three-mora words, and used as plain speech. The clean speech was filtered by far RIR [taken from Hurricane Challenge 2.0 (HC 2.0) [24]] and then added the babble noise (in HC 2.0 dataset) to have −4 dB SNR. Four hundred eighty stimuli (120 words x 4 methods x 1 noisy reverberant condition) were involved in this experiment. Nine native Japanese (1 female and eight males) aged mean 25.1 and standard deviation 1.9 with no report of hearing problems participated in the experiment. Speech types included: plain speech, MS500, and SSDRC.

Figure 5-30 shows that MS500 obtained a comparable result to SSDRC. And as a result of the previous section, naturalness of plain speech was still preserved for MS500. Thus, it can be concluded that the effective features could help to exceed the intelligibility of Lombard speech and preserve the naturalness of Lombard speech.

The ways to control effective features were also discussed a lot in previous studies. It was confused between a flat frequency response or an incremental response. In this study, the incremental response showed its effectiveness.

Also, A500c or A1000c had been shown as the best intelligibility overall in noise. However, it seemed not right when combining them with the time feature M, which might lead to another assumption. That is, frequency features and time features only support each other if they are in a suitable pair. Therefore, the mission to identify an appropriate time feature for these frequency

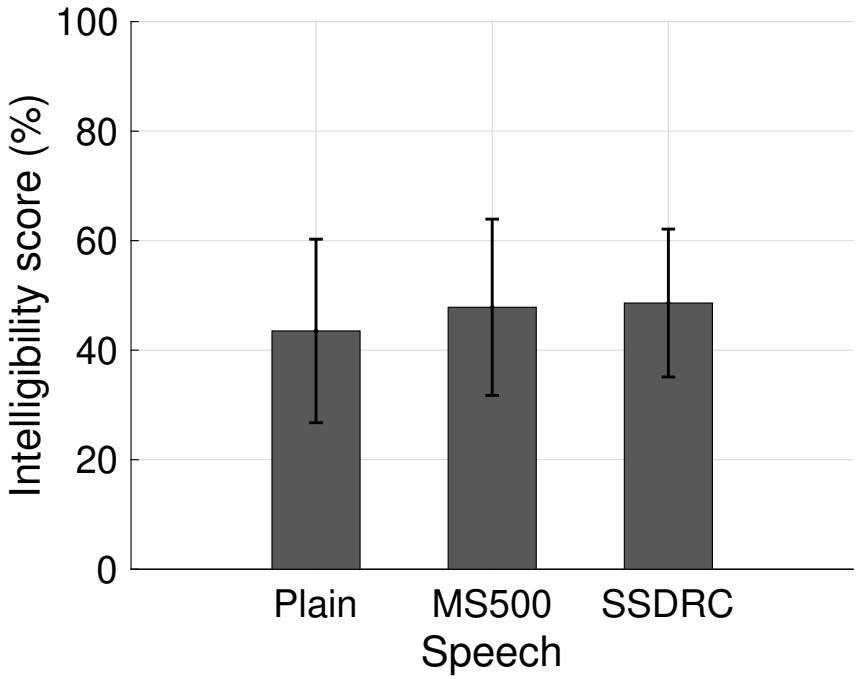


Figure 5-30: Intelligibility score of plain speech, MS500 speech, and SSDRC speech in noisy reverberant conditions

features has still remained. And, it is possible to find them out when investigating in noisy reverberant conditions.

5.4 Discussion on applying effective features with varying noise levels and SNRs

In general, applying correct Lombard speech/mimicking Lombard speech to a noisy environment at the same noise level as it was produced is a proper way to obtain better intelligibility and less annoyance for listeners (Kubo et al. [14]). Besides, applying effective features can also be systematically handled by the modification of MS according to SNRs, as was described in Sec. 5.2. There might be two ways to handling for both noise levels and SNRs.

(1) Apply them separately: If the produced mimicking Lombard speech still reaches a suitable SNR for listeners (not so small) to understand speech content, it should be to go with the mimicking Lombard speech to avoid annoyance by loud voices. In the case of the produced mimicking Lombard speech cannot reach a suitable SNR for listeners (so small), the modification of effective featured by the modification of MS should be used. This method is simple and easy to apply.

(2) Apply them in one consistent way: The synthesis system needs to be constructed taken two inputs of a noise level and an SNR to modify the effective features under the constraint by the rule generation model. It is to preserve the characteristics of Lombard speech and obtain excellent intelligibility and naturalness of the effective features. Further investigation should be done to make that system efficiently.

Chapter 6

Application to improving speech intelligibility under noisy reverberant conditions

In this section, it was to apply the effective features to increase the intelligibility of speech in noisy and reverberant environments. In reverberation, time features could play its effect. As was indicated in the previous chapter, only A500s, i.e., increasing spectral regions at 500-2250 and 4500-6500 Hz got a accumulated increasing effect with M features (i.e., increasing modulation spectra at 2-6 Hz and 8-20 Hz modulation in the spectra of 300 - 750, 1250 - 2250, and 4500 - 6500 Hz). Therefore, this feature combination was applied for the application. MS modified speech was synthesized by modifying these features. Perceptual experiments were performed to evaluate the intelligibility of MS modified speech in various noisy reverberant conditions and languages. The results indicated that the proposed method in most conditions, excepting German could increase speech intelligibility from plain speech about $5 - 20\%$.

6.1 Dataset

The speech material was provided by HC 2.0 in German and Spanish (100 sentences each) and English (90 sentences) as recorded by native male speakers and was used as plain speech.

Table 6.1: SNR (decibels, dB) under various conditions used in HC 2.0 listening tests.

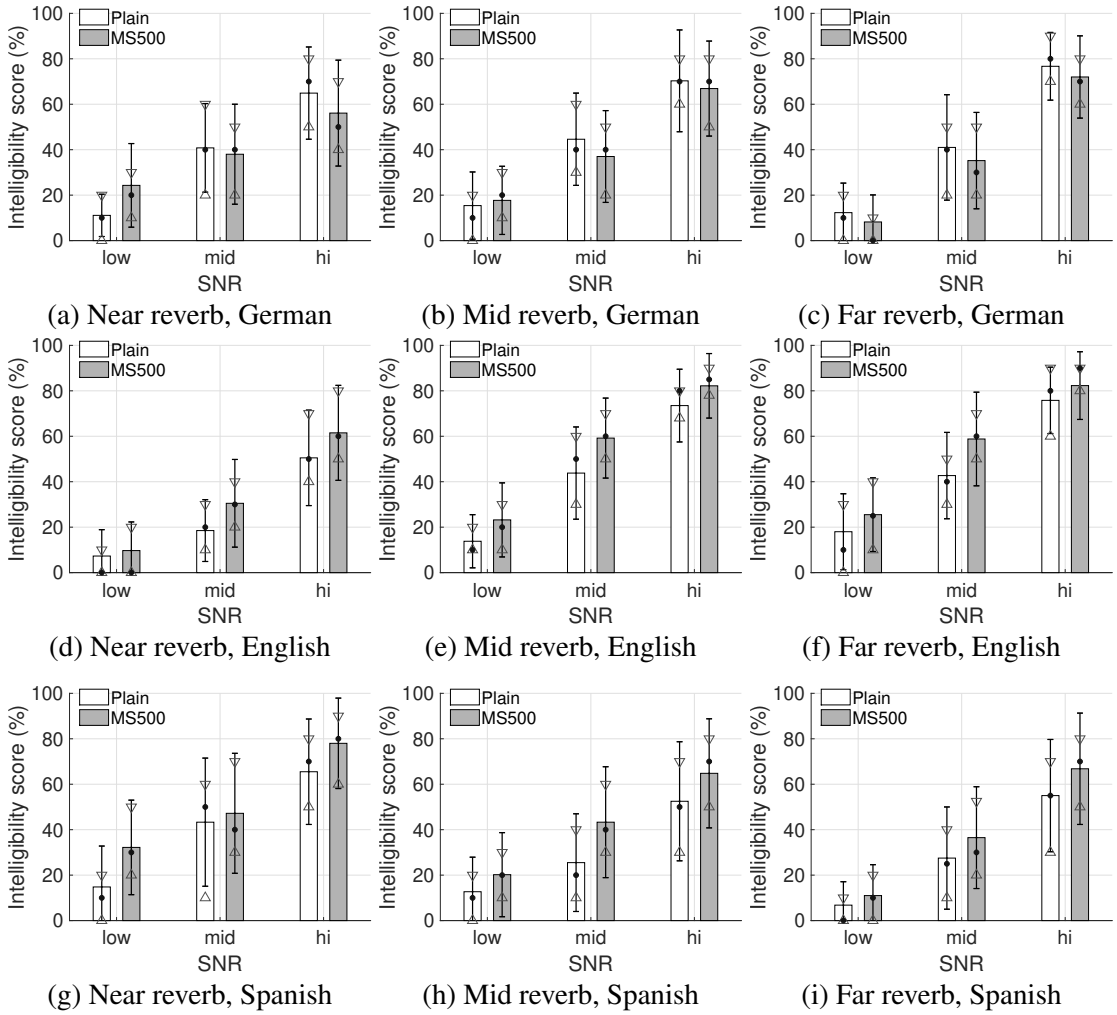
Reverberation	SNR	German	English	Spanish
near (1 m)	low	−15.0	−13.0	−17.0
	mid	−12.5	−8.5	−14.5
	high	−10.0	−4.0	−11.5
mid (2.5 m)	low	−13.0	−11.0	−17.0
	mid	−10.0	−5.0	−14.0
	high	−7.0	−1.0	−11.0
far (4 m)	low	−13.0	−10.0	−18.0
	mid	−9.0	−4.0	−14.0
	high	−5.0	2.0	−10.0

6.2 Modified speech

The same technique as described in the previous chapter was used to modify A500s and M features on the MS of plain speech to synthesize MS modified speech, which was previously called the MS500 speech.

6.3 Perception experiment

Two speech types were used: the plain speech and the MS500 speech. The evaluation was performed by HC 2.0 with about 180 listeners. They created stimuli for the experiment using the MS500 speech and their plain speech, babble noise, and the RIR. The clean speech was filtered using the RIR, and then the noise was added to obtain the targeted global SNR. Table 6.1 shows the evaluated SNRs and reverberation conditions in terms of the distance between the loudspeaker and the listener.



(a) Near reverb, German (b) Mid reverb, German (c) Far reverb, German

(d) Near reverb, English (e) Mid reverb, English (f) Far reverb, English

(g) Near reverb, Spanish (h) Mid reverb, Spanish (i) Far reverb, Spanish

Figure 6-1: Intelligibility scores (percentage of correctly identified words) of plain and MS-modified speech under 3 SNR noise levels (low, mid, hi) × 3 reverberation (reverb) conditions (near, mid, far) × 3 languages (German, English, Spanish). Bars indicate mean and standard deviation. Triangles, inverse triangles, and circles indicate first quartile, third quartile, and median respectively.

6.4 Results and discussion

The results of the listening tests are shown in Figure 6-1. They indicate that the MS500 speech had better intelligibility than the plain speech under all noisy reverberant conditions for English and Spanish (5-20%). The higher first and third quartiles for MS500 further support the substantial improvement obtained. However, intelligibility was not improved in other cases for the German dataset. One possible reason could be that the German dataset contained highly intelligible speech, leaving little room for improvement. Further, German words often contain plosive consonants, and modifying their MS is a delicate operation.

6.5 Summary

This section presented an application of effective features for improving the intelligibility of speech under noisy reverberant conditions. As a result, the applied features could effectively increase intelligibility for the modified speech.

Chapter 7

Conclusion

In this chapter, first, all of the work of this study is summarized. Contributions are then discussed. Finally, future work is introduced.

7.1 Summary

The purpose of this research was to improve the intelligibility and naturalness of speech in noise using conversion rules inspired by Lombard effect. It came up with two sub-goals: (I) obtaining feature understanding and control which contributes to the intelligibility of Lombard speech under **noise-level-varying and various noise**, which was achieved and (II) identifying and applying the **effective feature control methods** for exceeding the intelligibility and naturalness of Lombard speech, which was almost achieved. Thus, the investigation had three steps to cover the search space of features, noise levels, feature variations, SNRs, and spectral-varied noise for finding features and applied them:

1. Mimicking Lombard speech by controlling articulatory and acoustic features from one to multiple noise levels was achieved,

2. Effective features for the intelligibility and naturalness of speech in noise with various SNRs and spectral types to exceed intelligibility and naturalness of Lombard speech were achieved,

3. Application to improve the intelligibility of speech under noisy reverberant conditions was almost achieved.

For the first step: The goal for this step was to understand features and obtain control of features contributing to the intelligibility of Lombard speech under various noise with varying noise levels. The sub-goals were then to understand about articulatory-acoustic features for intelligibility and naturalness of Lombard speech in noise and obtain the control of acoustic features to mimic Lombard speech under backgrounds with varying noise levels.

Because previous studies have been unclear about the contribution of articulatory features to the intelligibility of Lombard speech in noise, analysis-by-synthesis methods based on an articulatory synthesis to mimic Lombard speech at a noise level with the most extent of its articulatory features were used. Consequently, the contribution of articulatory-acoustic features to the intelligibility of speech in the pink noise and the babble noise was obtained. Expanding to multiple noise levels, mimicking Lombard speech under varying noise levels had been a problem. A rule-based method with a rule-generation model and acoustical control were proposed to mimic variations of Lombard speech under background with varying noise levels of the pink noise. The mimicking speech to each noise level was equivalent with Lombard speech and the other mimicking speech by other state-of-the-art methods.

In this step, these subgoals were obtained. That is, the contributive articulatory-acoustic features to the intelligibility and naturalness of speech in noise were thoroughly understood and identified, which include spectral tilt, formants, and f_0. Also, the independent control of acoustic features was achieved. This achievement was the basis to investigate further effective features to increase speech intelligibility and naturalness in the next step.

For the second step: The goal was to exceed the intelligibility and naturalness of Lombard speech. The first subgoal was to obtain the best variation of acoustic features for intelligibility in multiple levels of noise, identified as effective features. Then, it was also to consider other effective features from other studies to extract the most significant properties from all of them (including the effective feature from the first subgoal) under various SNRs and noise types and identified the final effective features. The exceeding of the intelligibility of Lombard speech with the effective features was discussed. Each sub-goal was obtained with the following findings:

The best variation of acoustic features to increase the intelligibility and preserve the naturalness of speech in varying noise:

(1) Increases in spectra with a plateau between 2-6 kHz about 13 dB.

Effective features for all kinds of noise from all investigated studies (including the effective feature of increasing in spectra with a plateau between 2-6 kHz about 13 dB):

(1) Frequency features: Increases in spectral regions from 500 or 1000 kHz, effective increased controls were about 15-25 dB, which was also the most effective feature.

(2) Time features: Static range compression as the second effective feature could contribute to the intelligibility and naturalness of speech in noise.

(3) Accumulated contribution to speech intelligibility of the sparse frequency features (i.e., increases in spectra at 500-2250 and 4500-6500 Hz) with the time features (i.e., increases in modulation spectra at 2-6 and 8-20 Hz in the spectra of 300-750, 1250-2250, and 4500-6500 Hz). These features was evaluated to be able to exceed the intelligibility of Lombard speech and preserve the naturalness of Lombard speech.

In addition, basing on the modulation transfer function and modulation spectrum concepts in relationship with listening tests, a concept to extract effective features from differently enhancement methods was proposed.

For the third step: The goal was to increase the intelligibility of speech in various noisy reverberant conditions by applying the effective features from the second step.

The application to increase the intelligibility of speech under noisy and reverberant conditions has been successful:

(1) The speech by modifying the sparse frequency features, i.e., increases in spectra at 500-2250 and 4500-6500 Hz) and the time features (i.e., increases in modulation spectra at 2-6 and 8-20 Hz in the spectra of 300-750, 1250-2250, and 4500-6500 Hz) could increase the intelligibility for the speech in most of the noisy reverberant conditions and languages.

7.2 Contributions

The main contribution is that the results of this research can help us to understand the production of Lombard speech from the viewpoints of articulatory features to acoustic features, contributing to the intelligibility and naturalness of speech in noise. Also, it was about from a type of Lombard speech to varying types with noise levels of Lombard speech with the rule-generation model. Especially, the effective features were identified for all kinds of noise, which is important for increasing the intelligibility of speech in realistic situations.

The second contribution of this research is that the results can enlighten the fields of speech

enhancements, objective intelligibility measurements, voice conversion, and synthesis. It provides critical basic information for the areas of speech enhancement engineering.

The third contribution of this research is that the synthesis method and rule-generation model that can control acoustic features independently and synthesize Lombard speech under varying noise levels were proposed. Furthermore, the concept for extracting effective features from different enhancement methods was also described.

7.3 Future work

This study focused on the investigation of Japanese, effectively applied to English and Spanish. However, the effective features were not good for German. So, enhancements of intelligibility and naturalness among languages can be different. Thus, it will be to try to generalize the present research with more languages. It was realized that the contribution of time and frequency features should be synergistic. However, in the present study, one pair of these features was just found out. The mission to identify the remaining time features for the most effective frequency features mentioned above is still open. In other words, a way of time features to interact with frequency features well is still unknown. Therefore, the study should go further to investigate in noisy reverberant conditions to find out the optimal time-frequency feature combination.

Also, it will be to try to improve the modeling of the modulation spectrum to capture the significant characteristics of static range compression into it.

Finally, it is also to perform more evaluation for proving exceeding Lombard speech of the effective features.

Bibliography

[1] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations," in *EUSIPCO*, IEEE, 2010, pp. 1919–1923.

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

[3] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," *Computer Speech & Language*, vol. 28, no. 4, pp. 858–872, 2014.

[4] Y. Tang and M. Cooke, "Learning static spectral weightings for speech intelligibility enhancement in noise," *Computer Speech & Language*, vol. 49, pp. 1–16, 2018.

[5] A. ANSI, "S3. 5-1997, methods for the calculation of the speech intelligibility index," *New York: American National Standards Institute*, vol. 19, pp. 90–119, 1997.

[6] P. CODE, "Sound system equipment–part 16: Objective rating of speech intelligibility by speech transmission index," 2003.

[7] Y. Tang, M. Cooke, *et al.*, "Glimpse-based metrics for predicting speech intelligibility in additive noise conditions.," in *INTERSPEECH*, 2016, pp. 2488–2492.

[8] J. C. Krause and L. D. Braida, "Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2165–2172, 2002.

[9] ——, "Acoustic properties of naturally produced clear speech at normal speaking rates," *The Journal of the Acoustical Society of America*, vol. 115, no. 1, pp. 362–378, 2004.

[10] E. Lombard, "Le signe de l'élévation de la voix," *Annales des Maladies de L'Oreille et du Larynx*, vol. 37, pp. 101–119, 1911.

[11] J. J. Dreher and J. O'Neill, "Effects of ambient noise on speaker intelligibility for words and phrases," *The Journal of the Acoustical Society of America*, vol. 29, no. 12, pp. 1320–1323, 1957.

[12] A. L. Pittman and T. L. Wiley, "Recognition of speech produced in noise," *Journal of Speech, Language, and Hearing Research*, 2001.

[13] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3261–3275, 2008.

[14] R. Kubo and M. Akagi, "Effects of speaker's and listener's acoustic environments on speech intelligibility and annoyance," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Institute of Noise Control Engineering, vol. 253, 2016, pp. 3366–3371.

[15] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, 1988.

[16] T.-C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *INTERSPEECH*, 2012.

[17] N. Chennupati, S. R. Kadiri, and B. Yegnanarayana, "Spectral and temporal manipulations of SFF envelopes for enhancement of speech intelligibility in noise," *Computer Speech & Language*, vol. 54, pp. 86–105, 2019.

[18] M. Cooke, V. Aubanel, and M. L. G. Lecumberri, "Combining spectral and temporal modification techniques for speech intelligibility enhancement," *Computer Speech & Language*, vol. 55, pp. 26–39, 2019.

[19] P. Birkholz, *Vocaltractlab [software]*, version 2.2, 2017.

[20] ——, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLoS ONE*, vol. 8, no. 4, e60603, 2013.

[21] P. T. Nghia, L. C. Mai, and M. Akagi, "Improving the naturalness of concatenative vietnamese speech synthesis under limited data conditions," *Journal of Computer Science and Cybernetics*, vol. 31, no. 1, pp. 1–16, 2015.

[22] P. C. Nguyen, T. Ochi, and M. Akagi, "Modified restricted temporal decomposition and its application to low rate speech coding," *IEICE T. INF. SYST.*, vol. 86, no. 3, pp. 397–405, 2003.

[23] B. Nguyen and M. Akagi, "A flexible spectral modification method based on temporal decomposition and gaussian mixture model," *Acoustical science and technology*, vol. 30, no. 3, pp. 170–179, 2009.

[24] J. Rennies-Hochmuth, M. Cooke, and C. Valentini-Botinhao, *The hurricane challenge*.

[25] A. R. Bradlow and J. A. Alexander, "Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners," *The Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2339–2349, 2007.

[26] J. C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, vol. 93, pp. 510–524, 1993.

[27] C. Davis, J. Kim, K. Grauwinkel, and H. Mixdorff, "Lombard speech: Auditory (a), visual (v) and av effects," in *Proceedings of the Third International Conference on Speech Prosody*, 2006, pp. 248–252.

[28] T. Van Ngo, R. Kubo, D. Morikawa, and M. Akagi, "Acoustical analyses of tendencies of intelligibility in lombard speech with different background noise levels," *Journal of Signal Processing*, vol. 21, no. 4, pp. 171–174, 2017.

[29] Y. Uemura, M. Morise, and T. Nishiura, "The lombard speech recognition based on the voice conversion towards neutral speech," *ICA2010, PaperID*, vol. 167, 2010.

[30] M. Garnier, L. Bailly, M. Dohen, P. Welby, and H. Loevenbruck, "An acoustic and articulatory study of lombard speech: Global effects on the utterance," in *Interspeech/ICSLP 2006*, Pittsburgh, United States, 2006, pp. 2246–2249.

[31] M. Garnier, "May speech modifications in noise contribute to enhance audio-visible cues to segment perception?" In *AVSP*, 2008, pp. 95–100.

[32] J. E. Huber and B. Chandrasekaran, "Effects of increasing sound pressure level on lip and jaw movement parameters and consistency in young adults," *J. Speech Language Hearing Res.*, vol. 49, pp. 1368–1379, 2006.

[33] J. Simko, S. Benus, and M. Vainio, "Hyperarticulation in lombard speech: Global coordination of the jaw, lips and the tongue," *The Journal of the Acoustical Society of America*, vol. 139, pp. 151–162, 2016.

[34] M. Garnier, L. Ménard, and G. Richard, "Effect of being seen on the production of visible speech cues. a pilot study on lombard speech," in *13th Annual Conference of the International Speech Communication Association (InterSpeech 2012)*, Portland, United States, 2012, pp. 611–614.

[35] J. Scobbie, J. Ma, and J. White, "The tongue and lips in lombard speech: A pilot study of vowel-space expansion," CASL, 2012.

[36] M. Garnier, L. Ménard, and B. Alexandre, "Hyper-articulation in lombard speech: An active communicative strategy to enhance visible speech cues?" *The Journal of the Acoustical Society of America*, vol. 144, pp. 1509–1074, 2018.

[37] K. N. Stevens, *Acoustic Phonetics*. The MIT Press, 2000.

[38] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Synthesis and perception of breathy, normal, and lombard speech in the presence of noise," *Computer Speech & Language*, vol. 28, no. 2, pp. 648–664, 2014.

[39] B. Langner and A. W. Black, "Improving the understandability of speech synthesis by modeling speech in noise," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, vol. 1, 2005, pp. I–265.

[40] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Evaluating speech intelligibility enhancement for hmm-based synthetic speech in noise," in *SAPA-SCALE Conference*, 2012.

[41] M. Cooke, V. Aubanel, and M. L. G. Lecumberri, "Combining spectral and temporal modification techniques for speech intelligibility enhancement," *Computer Speech & Language*, vol. 55, pp. 26–39, 2019.

[42] Y. Lu and M. Cooke, "The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, vol. 51, pp. 1253–1262, 2009.

[43] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "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 Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[44] M. Cooke, C. Mayo, and J. Villegas, "The contribution of durational and spectral changes to the lombard speech intelligibility benefit," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 874–883, 2014.

[45] P. Boersma and D. Weenink, *Praat: Doing phonetics by computer (version 5.1.13)*, 2009.

[46] M. Cooke and V. Aubanel, "Effects of linear and nonlinear speech rate changes on speech intelligibility in stationary and fluctuating maskers," *The Journal of the Acoustical Society of America*, vol. 141, pp. 4126–4135, 2017.

[47] A. R. López, S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, "Speaking style conversion from normal to lombard speech using a glottal vocoder and bayesian gmms.," in *Interspeech*, 2017, pp. 1363–1367.

[48] B. Bollepalli, M. Airaksinen, and P. Alku, "Lombard speech synthesis using long short-term memory recurrent neural networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 5505–5509.

[49] S. Matsumoto and M. Akagi, "Variation of formant amplitude and frequencies in vowel spectrum uttered under various noisy environments," in *2019 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP2019)*, Research Institute of Signal Processing, Japan, 2019.

[50] T. Nishigaki and M. Akagi, "Influence of auditory feedback on uttering vowel speech in noisy environment," in *Proc. 2020 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP 2020)*, Research Institute of Signal Processing Japan, 2020, pp. 303–306.

[51] C. Hotchkin and S. Parks, "The lombard effect and other noise-induced vocal modifications: Insight from mammalian communication systems," *Biological Reviews*, vol. 88, no. 4, pp. 809–824, 2013.

[52] J.-C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.*, vol. 93, no. 1, pp. 510–524, 1993.

[53] D. Y. Huang, S. Rahardja, and E. P. Ong, "Lombard effect mimicking," in *ISCA*, 2010.

[54] D. Y. Huang and E. P. Ong, "Lombard speech model for automatic enhancement of speech intelligibility over telephone channel," in *ICALIP*, IEEE, 2010, pp. 429–434.

[55] S. Rottschafer, H. Buschmeier, H. Welbergen, and S. Kopp, "Online lombard adaptation in incremental speech synthesis," in *ISCA*, 2015.

[56] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.

[57] N. Morales, Z. Tang, and D. Manocha, "Receiver placement for speech enhancement using sound propagation optimization," *Applied Acoustics*, vol. 155, pp. 53–62, 2019.

[58] M. Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.

[59] T. Houtgast and H. J. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acta Acustica United with Acustica*, vol. 28, no. 1, pp. 66–73, 1973.

[60] M. Long, *Architectural acoustics*. Elsevier, 2005.

[61] T. Houtgast and H. J. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *The Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1069–1077, 1985.

[62] M. Unoki, Y. Yamasaki, and M. Akagi, "MTF-based power envelope restoration in noisy reverberant environments," in *EUSIPCO*, IEEE, 2009, pp. 228–232.

[63] H. Hermansky, "Modulation spectrum in speech processing," in *Signal Analysis and Prediction*, Springer, 1998, pp. 395–406.

[64] A. Kusumoto, T. Arai, K. Kinoshita, N. Hodoshima, and N. Vaughan, "Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments," *Speech Communication*, vol. 45, no. 2, pp. 101–113, 2005.

[65] P. Birkholz, S. Drechsel, and S. Stone, "Perceptual optimization of an enhanced geometric vocal fold model for articulatory speech synthesis," in *Interspeech 2019*, Graz, Austria, 2019, pp. 3765–3769.

[66] P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube, "Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis," in *Interspeech 2011*, Florence, Italy, 2011, pp. 2681–2684.

[67] P. Birkholz and D. Jackèl, "Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system," in *Interspeech 2004-ICSLP*, Jeju, Korea, 2004, pp. 1125–1128.

[68] P. Birkholz and D. Pape, "How modeling entrance loss and flow separation in a two-mass model affects the oscillation and synthesis quality," *Speech Communication*, vol. 110, pp. 108–116, 2019.

[69] P. Birkholz, "Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets," in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, 2007, pp. 2865–2868.

[70] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

[71] P. Birkholz, L. Martin, Y. Xu, S. Scherbaum, and C. Neuschaefer-Rube, "Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis," *Computer Speech & Language*, vol. 41, pp. 116–127, 2017.

[72] I. R. Titze, *Principles of Voice Production*. Prentice Hall, 1994.

[73] S. Prom-on, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in mandarin and english as a process of target approximation," *JASA*, vol. 125, no. 1, pp. 405–424, 2009.

[74] Pink-Noise, *Various - audio test CD-1 - 91 test signals for home and laboratory use*, 1984.

[75] Babble-Noise, *Noisex*. NOISE-ROM-0, NATO: AC243/(Panel 3)/RSG10, 1990.

[76] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[77] T.-N. Phung, M. C. Luong, and M. Akagi, "An investigation on perceptual line spectral frequency (PLP-LSF) target stability against the vowel neutralization phenomenon," in *2011 3rd International Conference on Signal Acquisition and Processing (ICSAP 2011)*, Institute of Electrical and Electronics Engineers (IEEE), 2011.

[78] K. Kondo, S. Amano, Y. Suzuki, and S. Sakamoto, "Japanese speech dataset for familiarity-controlled spoken-word intelligibility test (fw07)," *NII Speech Resources Consortium*, 2007.

[79] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1732–1746, 2012.

[80] A. C. Lammert and S. S. Narayanan, "On short-time estimation of vocal tract length from formant frequencies," *PloS one*, vol. 10, no. 7, e0132193, 2015.

[81] P. F. Assmann and T. M. Nearey, "Relationship between fundamental and formant frequencies in voice preference," *J. Acoust. Soc. Am.*, vol. 122, no. 2, EL35–EL43, 2007.

[82] M. Hodgson, G. Steininger, and Z. Razavi, "Measurement and prediction of speech and noise levels and the lombard effect in eating establishments," *J. Acoust. Soc. Am.*, vol. 121, no. 4, pp. 2023–2033, 2007.

[83] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE, vol. 1, 2002, pp. I–509.

[84] M. Akagi and Y. Tohkura, "Spectrum target prediction model and its application to speech recognition," *Computer Speech & Language*, vol. 4, no. 4, pp. 325–344, 1990.

[85] Y. Xue, Y. Hamada, and M. Akagi, "Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space," *Speech Communication*, vol. 102, pp. 54–67, 2018.

[86] B. O. Bush and A. Kain, "Modeling coarticulation in continuous speech,"

[87] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.

[88] T. V. NGO, R. KUBO, and M. AKAGI, "Mimicking lombard effect: An analysis and reconstruction," *IEICE Transactions on Information and Systems*, vol. E103.D, no. 5, pp. 1108–1117, 2020.

[89] EQ, *Equalization (audio)*.

[90] N. Westerlund, M. Dahl, and I. Claesson, *Adaptive gain equalizer for speech enhancement*, 2002.

[91] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing ii: Acoustic characteristics of clear and conversational speech," *Journal of Speech, Language, and Hearing Research*, vol. 29, no. 4, pp. 434–446, 1986.

[92] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Mal. de L'Oreille et du Larynx*, pp. 101–119, 1911.

[93] A. Raake, "Speech quality of voip," *Assessment and Prediction*, 2006.

[94] D. Xu, F. Chen, F. Pan, and D. Zheng, "Factors affecting the intelligibility of high-intensity-level-based speech," *The Journal of the Acoustical Society of America*, vol. 146, no. 2, EL151–EL157, 2019.

[95] H. R. Bosker and M. Cooke, "Enhanced amplitude modulations contribute to the lombard intelligibility benefit: Evidence from the nijmegen corpus of lombard speech," *The Journal of the Acoustical Society of America*, 2020.

[96] J. H. Hansen, J. Lee, H. Ali, and J. N. Saba, "A speech perturbation strategy based on "lombard effect" for enhanced intelligibility for cochlear implant listeners," *The Journal of the Acoustical Society of America*, vol. 147, no. 3, pp. 1418–1428, 2020.

[97] A. Ivanov and X. Chen, "Modulation spectrum analysis for speaker personality trait recognition," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[98] L. Moro-Velázquez, J. A. Gómez-García, and J. I. Godino-Llorente, "Voice pathology detection using modulation spectrum-optimized metrics," *Frontiers in bioengineering and biotechnology*, vol. 4, p. 1, 2016.

[99] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech," *Acoustical Science and Technology*, vol. 39, no. 3, pp. 234–242, 2018.

[100] M. Unoki and Z. Zhu, "Relationship between contributions of temporal amplitude envelope of speech and modulation transfer function in room acoustics to perception of noise-vocoded speech," *Acoustical Science and Technology*, vol. 41, no. 1, pp. 233–244, 2020.

[101] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2010, pp. 4214–4217.

[102] K. K. Wójcicki and P. C. Loizou, "Channel selection in the modulation domain for improved speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2904–2913, 2012.

[103] S. Greenberg and T. Arai, "The relation between speech intelligibility and the complex modulation spectrum," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[104] F. Pellegrino, C. Coupé, and E. Marsico, "A cross-language perspective on speech information rate," *Language*, pp. 539–558, 2011.

[105] S. Greenberg, "On the origins of speech intelligibility in the real world," in *Robust Speech Recognition for Unknown Communication Channels*, 1997.

[106] Z. Zhu, Y. Nishino, R. Miyauchi, and M. Unoki, "Study on linguistic information and speaker individuality contained in temporal envelope of speech," *Acoustical Science and Technology*, vol. 37, no. 5, pp. 258–261, 2016.

[107] L. Milic, *Multirate Filtering for Digital Signal Processing: MATLAB Applications*. IGI Global, 2009.

Publications

Journal

[1] **Thuanvan Ngo**, Rieko Kubo, Daisuke Morikawa and Masato Akagi, "Acoustical Analyses of Tendencies of Intelligibility in Lombard Speech with Different Background Noise Levels," Journal of Signal Processing, vol. 21, no. 4, pp. 171–174, 2017.

[2] **Thuanvan Ngo**, Masato Akagi, and Peter Birkholz, "Effect of articulatory and acoustic features on the intelligibility of speech in noise: An articulatory synthesis study," Speech Communication, vol. 117, pp. 13–20, 2020.

[3] **Thuanvan Ngo**, Rieko Kubo, and Masato Akagi, "Mimicking Lombard effect: An analysis and reconstruction," IEICE Transactions on Information and Systems, vol. E103.D, no. 5, pp. 1108–1117, 2020.

International Conference

[4] **Thuanvan Ngo**, Rieko Kubo, Daisuke Morikawa and Masato Akagi, "Acoustical analyses of Lombard speech by different background noise levels for tendencies of intelligibility," 2017 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP'17), 2017, pp. 309–312.

[5] **Thuanvan Ngo**, Rieko Kubo, and Masato Akagi, "Evaluation of the Lombard effect model on synthesizing Lombard speech in varying noise level environments with limited data," In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 133-137.

Domestic Conference

[6] **Thuanvan Ngo**, Rieko Kubo, and Masato Akagi, "Intelligibility improving and naturalness preserving for evacuation speech in noisy environments," IEICE Technical Report, Engineering Acoustics, 2019.

[7] **Thuanvan Ngo**, Rieko Kubo, and Masato Akagi, "Improved quality and intelligibility of mimicking Lombard speech by source-filter and coarticulation model-based synthesis," ASJ Spring Meeting, 2019.

[8] **Thuanvan Ngo**, Rieko Kubo, and Masato Akagi, "Speaker-independent control model for mimicking Lombard speech uttered in background noises with various levels," ASJ Spring Meeting, September, 2018, pp. 1371-1374.

[9] **Thuanvan Ngo**, Rieko Kubo, and Masato Akagi, "Acoustical control method for increasing intelligibility based on Lombard speech uttered in background noises with various levels," ASJ Fall Meeting, 2018, pp. 313-316.

[10] **Thuanvan Ngo**, Rieko Kubo, and Masato Akagi, "Acoustical rules for mimicking Lombard speech produced in a various noise level background," IEICE Technical Report, Engineering Acoustics, vol. 117, no. 170, 2017.