PAPER

Mimicking Lombard Effect: An Analysis and Reconstruction[**]

Thuan Van NGO[†a)], **Rieko KUBO**[†*b)], ***Nonmembers*, *and*** **Masato AKAGI**[†c)], ***Member***

SUMMARY Lombard speech is produced in noisy environments due to the Lombard effect and is intelligible in adverse environments. To adaptively control the intelligibility of transmitted speech for public announcement systems, in this study, we focus on *perceptually* mimicking Lombard speech under backgrounds with varying noise levels. Other approaches map corresponding neutral 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, we utilize a rule-based method that automatically generates rules and flexibly controls features with any change of noise level. Specifically, we conduct a feature tendency analysis and propose a continuous rule generation model 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 neutral 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 our 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 and a state-of-the-art method. Intelligibility and naturalness are comparable with some feature modifications.

key words: *Lombard speech, perceptual mimicking, rule-based methods*

1. Introduction

In the public announcements provided in train stations, airports, and factories, the presence of noise often smears transmitted speech and makes it difficult for listeners to understand it. Two possible approaches to solving this problem are reducing noise and making the speech itself more intelligible. It is impractical to reduce the noise due to the complex structures of the facilities and the costly devices involved. On the other hand, it should be possible to increase the intelligibility of transmitted speech by applying the properties of Lombard speech. Lombard speech is a kind of intelligible speech that is produced in noisy environments due to the Lombard effect [1]. It has been shown that Lombard speech is more intelligible than neutral speech produced in quiet conditions [2] because the listeners are released from both energetic and informational masking [3]. Thus, by making the transmitted speech *perceivable* as Lombard speech, i.e., by *perceptually* mimicking Lombard speech, listeners are expected to be able to capture the transmitted information.

State-of-the-art methods that mimic Lombard speech are based on statistical Bayesian GMM (BGMM) [4] or DNN techniques [5]. In these methods, speech features or waveforms are automatically mapped from neutral 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 [6], [7], these state-of-the-art methods, especially DNN-based methods, require an extremely huge dataset to train, thus rendering them impractical.

Other methods have been based on acoustical analysis and acoustic feature modification by rules. By using these rule-based methods, we can get some insight into how Lombard speech is really produced and synthesize it systematically. In these methods, feature control is more robust to change yet requires detailed analyses and advanced modification techniques. Acoustical analyses [8], [9] have mainly revealed that, compared with neutral speech, the distinctive acoustic features of Lombard speech include increased duration, increased fundamental frequency (f_0), flattened spectral tilt, and increased vocal intensity [10]. In addition, Ngo *et al.* [6] 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.* [11], [12] and Rottschaefer *et al.* [13] make acoustical rules to directly modify neutral 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

Manuscript received September 27, 2019.
Manuscript revised January 17, 2020.
Manuscript publicized February 13, 2020.

[†]The authors are with Graduate School of Advanced Science and Technology, JAIST, Nomi-shi, 923–1292 Japan.
[*]Presently, with Universal Communication Research Institute, NICT, Kyoto-fu, 619–0289 Japan.

[**]Perceptually mimicking Lombard speech.

a) E-mail: vanthuanngo@jaist.ac.jp
b) E-mail: rkubo@nict.go.jp
c) E-mail: akagi@jaist.ac.jp

DOI: 10.1587/transinf.2019EDP7260

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.

In this study, to improve upon the analysis for feature tendencies in previous studies, we *adequately performed in-depth analyses* on the acoustical features of neutral and Lombard speech produced in backgrounds with various noise levels. Then, on the basis of the analyzed feature tendencies, we designed *a continuous rule generation model* of acoustic features to precisely estimate the effects of noise. This model is 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, we applied a coarticulation model [14] and a modified restricted temporal decomposition (MRTD) [15] with spectral-GMM [16] to our synthesis and modification methods. 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. Methodology

We used the analysis and synthesis methodology outlined in Fig. 1 to convert neutral speech to Lombard speech fit with varying noise levels. Our approach is a combination of STRAIGHT [17], a coarticulation model, MRTD [15], and our 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. 2), which is adopted from Nghia *et al.*'s study [14], 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.

The MRTD [15] decomposes/interpolates acoustic features into/from temporal information (event functions) and

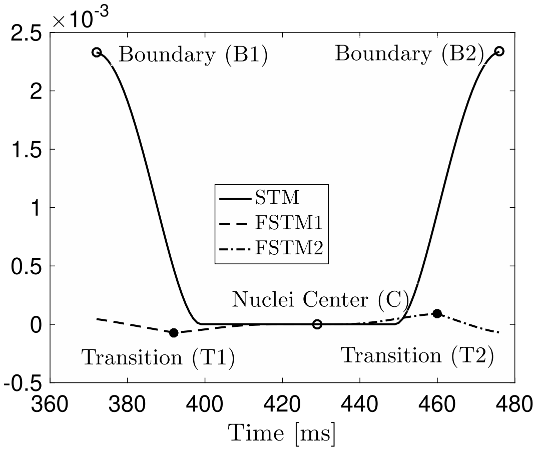


Fig. 2 Locations to extract event targets in temporal decomposition, based on coarticulation model. STM indicates the spectral transition rate of a phoneme. FSTM1 and FSTM2 are respectively derivatives of STM on the first and second halves. These rates represent both spectral dynamics and static spectral targets. The spectral dynamics, known to be context-sensitive, contain a lot of phonetic information of speech, which is crucial for speech intelligibility [18]. The static spectral targets, which contain linguistic-phonetic and non-paralinguistic information of speech, are important for speech intelligibility and naturalness [18].

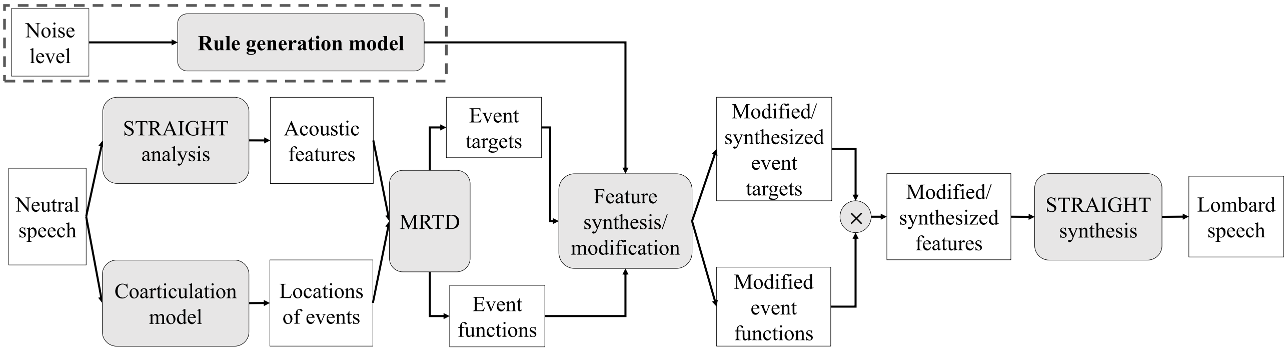


Fig. 1 Outline of our analysis/synthesis methods.

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, we extend the model to include other features. To represent co-articulation better, we use the event functions of a spectral feature to interpolate all features. To modify/synthesize acoustic features, we modify or synthesize the event targets of each feature according to the feature characteristics and the event functions of the spectral feature and then multiply them together. In this way, we can also obtain an interpolation of temporal information for modified acoustic features to maintain naturalness.

The most important element here is the *rule generation model* (see Sect. 3.2.5) for adaptability, which takes a noise level as input to give corresponding values of acoustic features as output. We constructed this model after our acoustical analysis of neutral and Lombard speech. With these components, our 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.

3. Feature Analysis and Rule Generation for Synthesis/Modification

3.1 Speech Dataset

Recorded speech by two speakers (one male, one female) in the pink noise levels of $-\infty$ (neutral 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 [19]. Three familiarity-controlled word lists, F0W7 [20] (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 our analyses.

3.2 Feature Analyses

We used the same dataset described above in the acoustical analysis in our previous study [6]. In that study, we 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 our previous studies [6] were carried out.

3.2.1 Spectral Envelope

The spectral envelope at three locations (transitions and nuclei center) was taken from the spectral sequence. 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

$$\log|X(\omega)| = T(\omega) + \log\frac{B(\omega)}{A(\omega)}. \tag{1}$$

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

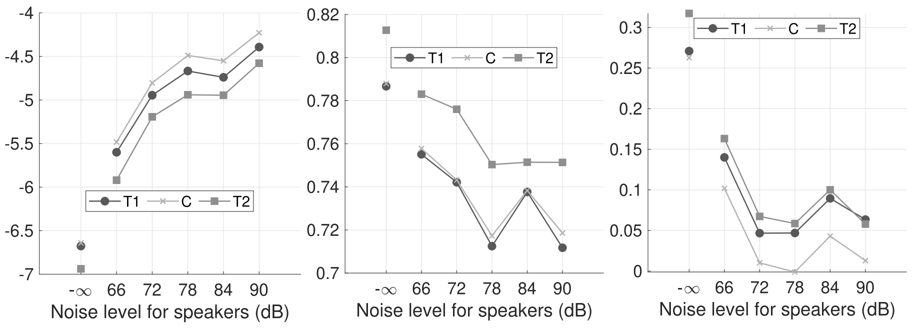
$$T(\omega) = c_0 + 2c_1\cos(\omega) + 2c_2\cos(2\omega), \tag{2}$$

$$c_n = \frac{1}{N}\sum_m \log|X(\omega)|\cos(n\omega), \tag{3}$$

where $\omega = \frac{2\pi k}{2(N-1)}$; $k = 0, 1, \ldots, N-1$.
Figure 5 shows the analysis results of these cepstral coefficients at transitions and the nuclei center of vowels.
- Formants: The formant frequencies were estimated by KARMA [21]. In our previous study [6], we observed shifting in F_1 and F_2 to higher frequency regions. With an assumption of the appearances of the plateau in the spectral tilt, we skipped the pattern of formant amplitudes. Further analyses on F_3 and F_4 were also conducted. Figure 6 shows the analysis results of these four formants.
- Vocal tract length (VT length): To the best of our 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 [22], [23]. Figure 7 shows the analysis results of this relationship.

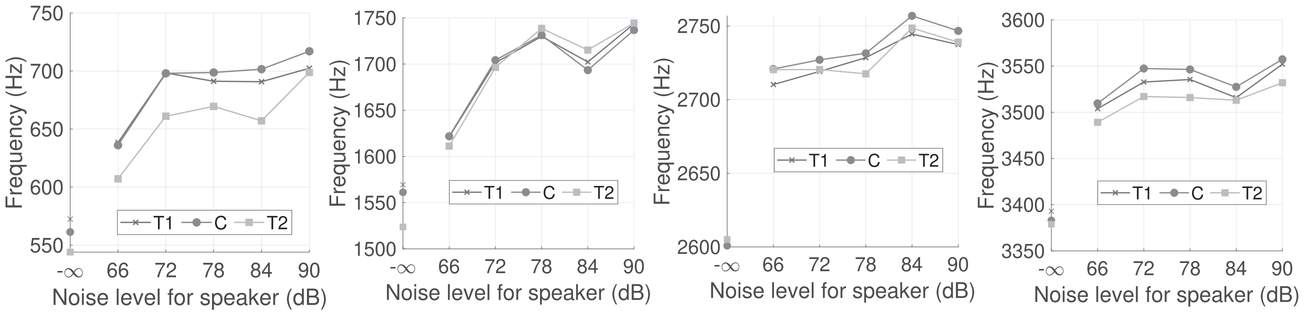
3.2.2 Fundamental Frequency (f_0)

In our previous study [6], f_0 mean showed a clear correlation with increasing noise levels. In this study, the dynamic



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

Fig. 5 Analysis results of cepstral coefficients of vowels. T1, T2 are the transitions. C is the nuclei center.



(a) F_1 (b) F_2 (c) F_3 (d) F_4

Fig. 6 Analysis results of formants. T1, T2 are the transitions. C is the nuclei center.

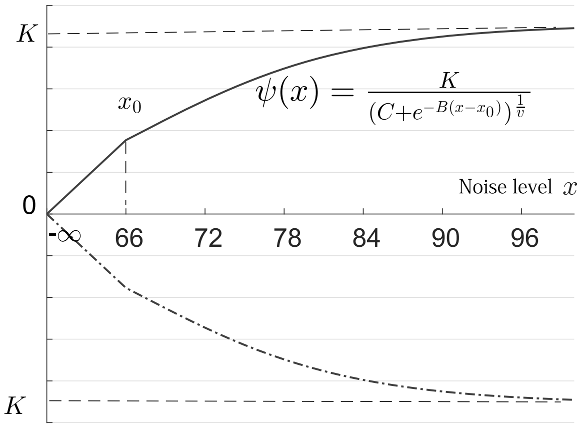


Fig. 3 Rule generation model of acoustical parameter values ψ in log scale depending on the noise level x. K indicates the upper or lower limit to which the saturation approximates. x_0 indicates the noise level at which drastic change to Lombard speech occurs.

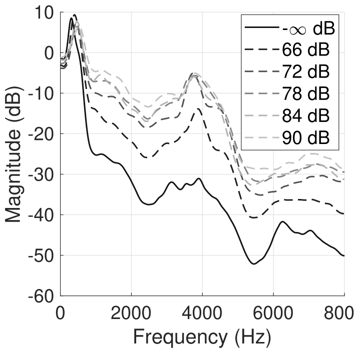


Fig. 4 Spectral envelope of neutral (−∞ dB, solid line) and Lombard speech (66–90 dB, dashed lines) at the nuclei center (C) of vowels. A plateau between 2–6 kHz appears in Lombard speech.

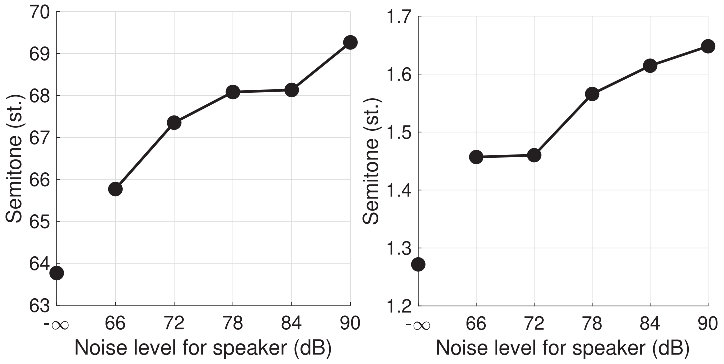
range of f_0 was also investigated. Figure 8 shows the analysis results of these acoustic features of f_0.

3.2.3 Power Envelope

In our previous study [6], we 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. Our 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. 9. Also, a correlation between the power envelope and the f_0 contour was estimated.

3.2.4 Duration

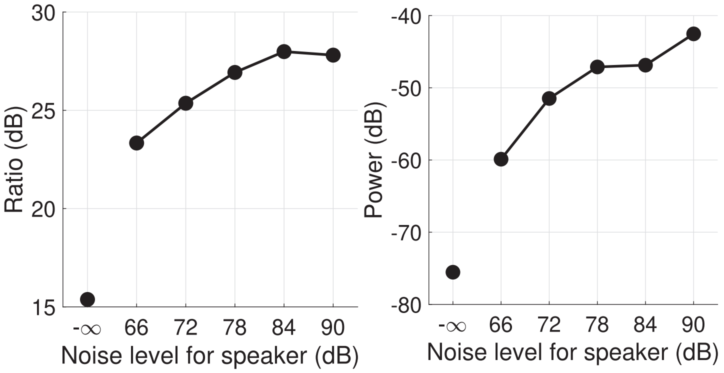
The same as in our previous study [6], we extracted vowel duration, which is shown in Fig. 10.



(a) f_0 mean

(b) f_0 range

Fig. 8 Analysis results of f_0.



(a) Consonant-to-vowel ratio

(b) Average power

Fig. 9 Analysis results of power envelope.

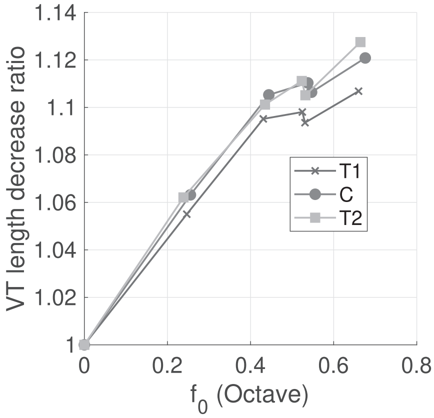


Fig. 7 Analysis results of vocal tract length with f_0. T1, T2 are the transitions. C is the nuclei center.

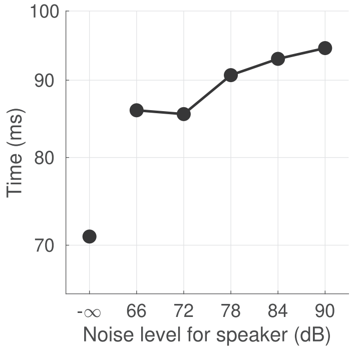


Fig. 10 Analysis results of vowel duration.

3.2.5 Rule Generation Model

The analysis results are summarized in Table 1. For all the analyzed acoustic features (Figs. 4 to 10), 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 neutral 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 our rule generation model.

Therefore, on the basis of the model reported by Hodgson *et al.* [24], in which the Lombard effect repre-

Table 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. 5)
Formants	Increased F_1, F_2, F_3, F_4 (see Fig. 6)
Vocal tract length	Decreased vocal tract length (see Fig. 7)
f_0	Increased f_0 mean and f_0 range (see Fig. 8)
Power envelope	Increased consonant-to-vowel ratio and average power (see Fig. 9), and a positive correlation of 0.47 with f_0
Duration	Increased vowel duration (see Fig. 10)

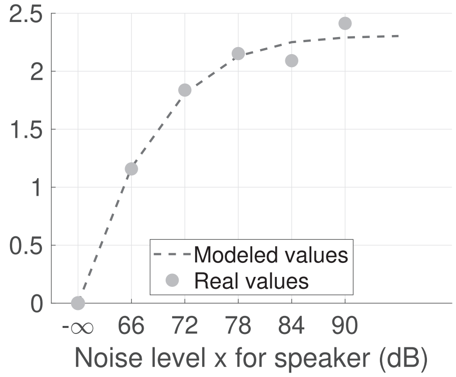


Fig. 11 Rule generation model for c_0 at the nuclei center of vowels, shown in Fig. 5 (a), *relative to neutral speech*.

sents the relationship between the constitutional factors of environments with noise levels, we propose a rule generation model. This model represents the relationship between acoustical parameter values and noise levels. It is estimated to have a drastic change around 66 dB [24] and a saturation starting from 90 dB, as shown in Fig. 3 and Eq. (4).

$$\psi(x) = \frac{K}{(C + e^{-B(x-x_0)})^{1/v}} \tag{4}$$

By applying this model, for each acoustic feature, relative to neutral 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 neutral 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. 11 represents the modeling for c_0 at the nuclei center (C) of vowels, shown in Fig. 5 (a), *relative to neutral speech*, with the estimated coefficients $K = 9.97$, $C = 2.22$, $B = 0.19$, $x_0 = 66$, and $v = 0.55$.

3.3 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.

3.3.1 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 [16]. 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.

3.3.2 Fundamental Frequency (f_0)

To synthesize f_0 contour, f_0 contour was parameterized and controlled by the Fujisaki model [25]. 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.

3.3.3 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 [26], [27]. After that, the modified power envelope at event-target locations was taken as the modified event target of the power envelope.

3.3.4 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 [28], we modeled two halves of an EF (left and right halves, separated by the location of event function) by Eqs. (5) and (6). Expected scales in vowel duration were obtained by controlling N and s.

$$EF_L(t) = \left| 1 - \frac{2}{1 + e^{s \frac{t}{t - t_N}}} \right| \quad (5)$$

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

$$EF_R(t) = \left| 1 - \frac{2}{1 + e^{s \frac{t}{t - t_1}}} \right| \quad (6)$$

where $t = N - 1, N - 2, \ldots, 0$ 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 our synthesizer to produce Lombard speech.

4. Experiments

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

4.1 Experiments for Similarity

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

4.1.1 Setup

- Speech material: Speech material was drawn from recorded speech (both Lombard speech produced at 66, 72, 78, and 84 dB noise levels and neutral speech) [6]. We used 105 Japanese words (4-mora) spoken by one male and one female.
- Speech types: We used two BGMM-based methods [4]: GlottalDNN-based (called **Glottal_BGMM**) and STRAIGHT-based (called **STRAIGHT_BGMM**) synthesis. 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, we synthesized two more types by our 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: We used 12 native Japanese: nine males and three females aged 23 to 25 years (mean: 24) with no history 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 × 4 pair types × 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 neutral speech.

4.1.2 Results and Discussion

Figure 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 our mimicked speech could success-

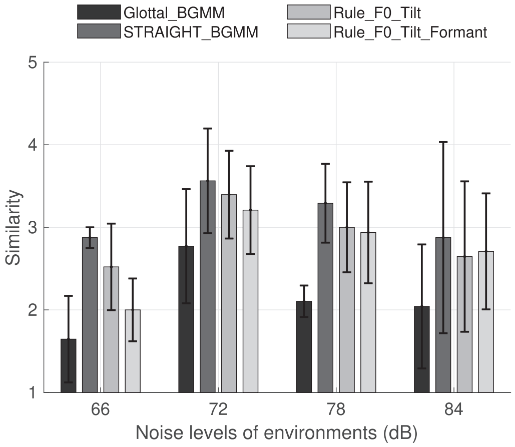


Fig. 12 Similarity of the mimicked speech. The bar and error values indicate the mean and standard deviation among listeners. The values of similarity mean 1: not at all, 2: a little, 3: moderately, 4: a lot, and 5: quite a lot similar to Lombard speech.

fully adapt to noise levels. This demonstrates that the proposed model can correctly represent Lombard speech with varying noise levels.

4.2 Experiments for Intelligibility and Naturalness

The purpose of this experiment was to evaluate the intelligibility and naturalness of the mimicked speech by our 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.

4.2.1 Setup

- Speech material: Speech material was drawn from the ATR dataset [29]. A total of 384 words (3-mora) of neutral speech from six different speakers (3 males, 3 females) was used.
- Speech types: Table 2 lists the five speech types we examined. Four types were synthesized using our 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: We used ten native Japanese: 8 males and 2 females aged 22 to 25 years (mean: 23.4) with no history of hearing problems.
- Maskers: Pink noise [30] 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 × 5 speech types × 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 were asked to write down the word they heard using a keyboard. They clicked 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.

Table 2 Speech types used 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

4.2.2 Results and Discussion

Figure 13 shows the results of speech intelligibility when various features were mimicked. We found that our method 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 14 shows the results of naturalness. Similar to the results of intelligibility, our method with the modifi-

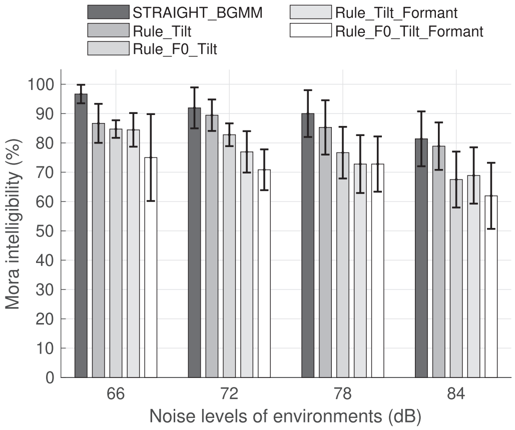


Fig. 13 Intelligibility of speech when various features are mimicked, i.e., percentage of correctly perceived mora in a word. The bar and error values indicate the mean and standard deviation among participants.

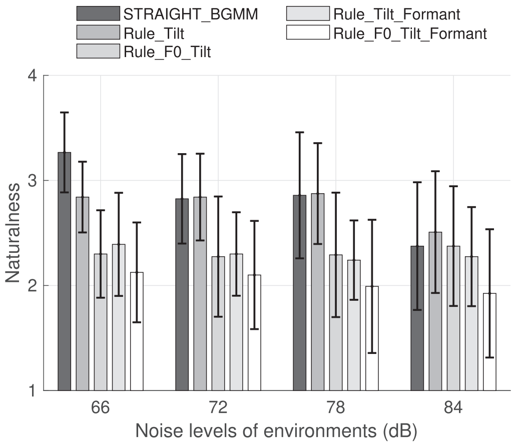


Fig. 14 Naturalness of speech when various features are mimicked. The bar and error values indicate the mean and standard deviation among participants. The values of naturalness are 1: unnatural, 2: rather unnatural, 3: rather natural, 4: natural.

cation of spectral tilt showed comparable naturalness with STRAIGHT_BGMM. With the modification of the other feature sets, it obtained lower naturalness.

The equivalent intelligibility and naturalness when spectral tilt was mimicked indicates that 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. We can find 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, we conclude that the proposed model performed adequately in all of the evaluations.

5. General Discussion

Under the rule generation model, we tried to generalize the rules in Lombard effect mimicking regarding noise levels. We generated the values of model parameters 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 our 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 [7]. To apply for these different directional changes of these features, we only need 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, our method can be controlled by the generated rules according to variation of noise levels. Thus, our 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, we can improve our proposed model 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.

6. Conclusion

In this study, we conducted analyses of Lombard speech and presented our rule generation model under backgrounds with varying noise levels for adaptively controlling the intelligibility of transmitted speech in public announcement systems. We described our modification-synthesis method, 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 our proposed mimicking model. Our 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, our proposed model can interpolate Lombard effect in any noise level. In order to obtain better intelligibility and naturalness, we aim to improve our 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.

Acknowledgments

This work was supported by SECOM Science and Technology foundation and JST-Mirai Program (Grant Number: JPMJMI18D1).

References

[1] 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.

[2] Y. Lu and M. Cooke, “Speech production modifications produced by competing talkers, babble, and stationary noise,” J. Acoust. Soc. Am., vol.124, no.5, pp.3261–3275, 2008.

[3] M. Cooke and Y. Lu, “Spectral and temporal changes to speech produced in the presence of energetic and informational maskers,” J. Acoust. Soc. Am., vol.128, no.4, pp.2059–2069, 2010.

[4] 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,” Interspeech, pp.1363–1367, 2017.

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

[6] T.V. 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.

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

[8] 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, no.12, pp.1253–1262, 2009.

[9] M. Cooke, C. Mayo, and J. Villegas, "The contribution of durational and spectral changes to the Lombard speech intelligibility benefit," J. Acoust. Soc. Am., vol.135, no.2, pp.874–883, 2014.

[10] 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.

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

[12] D.-Y. Huang and E.P. Ong, "Lombard speech model for automatic enhancement of speech intelligibility over telephone channel," 2010 International Conference on Audio, Language and Image Processing, pp.258–263, 2010.

[13] S. Rottschafer, H. Buschmeier, H. Welbergen, and S. Kopp, "Online Lombard adaptation in incremental speech synthesis," ISCA, pp.80–84, 2015.

[14] 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.

[15] P.C. Nguyen, T. Ochi, and M. Akagi, "Modified restricted temporal decomposition and its application to low rate speech coding," IEICE Trans. Inf. & Syst., vol.E86-D, no.3, pp.397–405, March 2003.

[16] B.P. 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.

[17] 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.3-4, pp.187–207, 1999.

[18] 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," 2011 3rd International Conference on Signal Acquisition and Processing (ICSAP 2011), Institute of Electrical and Electronics Engineers (IEEE), pp.512–514, 2011.

[19] T. Kondo, S. Amano, S. Sakamoto, and Y. Suzuki, "Development of familiarity-controlled word-lists (FW07)," IEICE Tech. Rep., vol.107, no.432, pp.43–48, 2008.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.I-509–I-512. 2002.

[26] 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.

[27] 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.

[28] B.O. Bush and A. Kain, "Modeling coarticulation in continuous speech," 15th Annual Conference of the International Speech Communication Association, INTERSPEECH 2014, pp.193–197, 2014.

[29] 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.

[30] Pink-Noise, "Various - audio test CD-1 - 91 test signals for home and laboratory use."

Thuan Van Ngo received the B.S. degree from Posts and Telecommunications Institute of Technology in 2015 and the M.S. degree in Information Science from Japan Advanced Institute of Science and Technology in 2017. Since 2017, he has been a Ph.D. student under the doctor research fellow program and a research assistant in Akagi Laboratory, Japan Advanced Institute of Science and Technology. His research areas include speech production and speech intelligibility in advert environments.

Rieko Kubo graduated from Osaka University in 1995. She received the Ph.D. degree in Information Science from Japan Advanced Institute of Science and Technology in 2015. Her research areas are speech perception and understanding. She specializes in aging effects on phoneme acquisition, and interactions between perception and production.

Masato Akagi received the B.E. from Nagoya Institute of Technology in 1979, and the M.E. and Ph.D. Eng. from the Tokyo Institute of Technology in 1981 and 1984. He joined the Electrical Communication Laboratories of Nippon Telegraph and Telephone Corporation (NTT) in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992 he has been on the faculty of the School of Information Science of JAIST and is now a full professor. His research interests include speech perception, modeling of speech perception mechanisms in human beings, and the signal processing of speech.