Evaluation of the Lombard effect model on synthesizing Lombard speech in varying noise level environments with limited data

Thuan Van Ngo*, Rieko Kubo* and Masato Akagi*
* Japan Advanced Institute of Science and Technology, Japan
E-mail: vanthuanngo@jaist.ac.jp, rkubo@jaist.ac.jp, akagi@jaist.ac.jp

***Abstract*—Lombard speech is intelligible speech produced by humans in noises. In this study, we focus on mimicking Lombard speech from natural neutral speech under backgrounds with varying noise levels to increase its intelligibility in these noises. Other approaches map corresponding speech features from the neutral speech to Lombard speech, which can only apply for an individual noise level, and cannot reveal feature tendencies. Instead, we implement a Lombard effect model to continuously estimate feature values with varying noise levels. The techniques, which are based on coarticulation, a source-filter model with MRTD and spectral-GMM, are used to easily modify features of the neutral speech to obtain their tendencies. Finally, these features are synthesized by STRAIGHT vocoder to obtain Lombard speech. The mimicking quality is evaluated in subjective listening experiments on similarity, naturalness, and intelligibility. The evaluation results show that the proposed method could convert neutral speech into Lombard speech in varying noise levels, which obtains comparable results with the state-of-the-art method.**

I. Introduction

Lombard speech [1] is intelligible speech produced in noisy environments. Lu and Cooke and colleagues [2], [3] reported that the distinctive acoustic features of Lombard speech included increased duration, increased f_0, and flattened spectral tilt, compared with the neutral speech (uttered in quiet environments). By manipulating these acoustic features, a mimicking Lombard speech [4], [5] can be synthesized. However, when the levels of noise are varying, mimicking Lombard speech has been still challenging. The state-of-the-art methods based on Bayesian GMM (BGMM) [6] or DNN techniques [7] would require a huge dataset to train to deal with such multiple noise levels. Rottschaefer *et al.* [8] proposed an online Lombard-adaptation in incremental speech synthesis to present and evaluate Lombard speech when its model parameters are updated continuously. The system achieved good results in adapting voice intensity and spectral emphasis (likewise, amplified speech) but failed with other features. It might be because of the lack in a detailed analysis and the simplicity of their proposed adaptation model. Sequentially, by analysis, Ngo *et al.* [9] found that these distinctive features of flattened spectral tilt, increased power envelope (or raises in modulation spectrum in specific frequencies), increased f_0, increased F_1, increased vowel duration are varying with increasing noise levels. It suggested a possibility to model and control these features with varying noise levels.

TABLE I: Acoustic feature groups and their parameters used to mimic Lombard speech under varying noise levels

Feature group	Parameter
Spectral tilt	Increased c_0, decreased c_1 and c_2
f_0	Increased f_0 mean and f_0 range
Power envelope	Increased consonant-to-vowel ratio and average power, positive correlation with f_0
Formants	Increased F_1, F_2, F_3, F_4, and decreased the vocal tract length correlated with the increase in f_0
Duration	Increased vowel duration

Then, we proposed a Lombard effect model among noise levels for parameter values of acoustic features. The synthesis and modification method based on coarticulation and source-filter models and modified-restricted-temporal decomposition can easily control the features with varying noise levels.

For evaluation, we compare our method with BGMM-based methods and Lombard speech produced in some typical noise levels of 66, 72, 78, 84 dB. To show the similarity with Lombard speech, we carried out experiments in noise-free conditions on the Lombard speech dataset (all mimicking speech vs. Lombard speech). To examine naturalness and intelligibility among mimicking speech in general (various mimicked features and datasets), we carried out the experiments among the mimicking speech on a different dataset (ATR dataset) in a background of pink noise.

II. Method

To mimic Lombard speech with varying noise levels, we describe modified acoustic features, the proposed Lombard model and modification synthesis techniques as follows.

A. *Modified acoustic features*

The features were mainly reported by Ngo *et al.* [9] and others with cepstral coefficients, f_0 range, F_2, F_3, F_4, the length of vocal tract, and average power envelope. These feature groups and their parameter changes with increasing noise levels are summarized as in Table I.

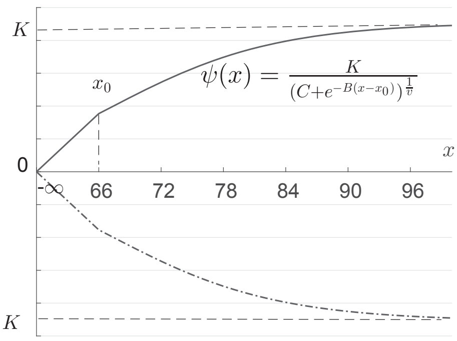

Fig. 1: Lombard effect model of acoustical parameter values ψ in log scale depending on the noise level x. K indicates the upper or lower limit, to which the saturation approximates, x_0 indicates the noise level, at which the drastic change to Lombard speech occurs.

B. Lombard effect model

We updated our previous model [10] based on the model reported by Hodgson *et al.* [11], in which the Lombard effect represents the relationship between the constitutional factors of environments with noise levels. Our model represents the relationship between acoustical parameter values and noise levels. It was estimated with a drastic change around 66 dB [11] and a saturation started from 90 dB as shown in Fig. 1 and Eq 1.

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

By applying this model, for each acoustical parameter, a model function was estimated by non-linear least square fit (lsqcurvefit in Matlab) with initial values of $(K, C, B, v) = (K_0, 1, B_0, 1)$. K_0 = maximum of the estimated values if the changing tendency of the values with noise levels was realized to be increased, and vice versa i.e. the minimum of that values. B_0 was set equal to the linear slope estimating these values. The lower and upper bound are $(-\infty, 0, 0, 0)$, $(\infty, \infty, \infty, \infty)$ respectively with a step size of 10^{-6}. The root mean square errors of the fitting were about 1.3 dB, 0.1 dB, and 0.1 dB for c_0, c_1, and c_2 of spectral tilt respectively, 0.04 dB for power envelope, 1 Hz for all F_1, F_2 and f_0, and 1ms for duration. The errors were small to compare with variations of acoustical parameter values among noise levels.

C. Modification synthesis techniques

In an addition to STRAIGHT [12], we used the following techniques to extract and modify features.

1) Modified restricted temporal decomposition: MRTD [13] is mainly applied to spectral feature to decompose and interpolate temporal information and spectral parameters at some specific time locations. The spectra are decomposed into event targets (which are spectral parameters) and event functions (which are temporal information used in the interpolation). We extended the decomposition and interpolation to other features. To represent coarticulation better, we used the same event function of the spectral feature to interpolate all features. Modifications were carried on event targets of each feature, excepting the scaling of duration was done by modifying the event functions.

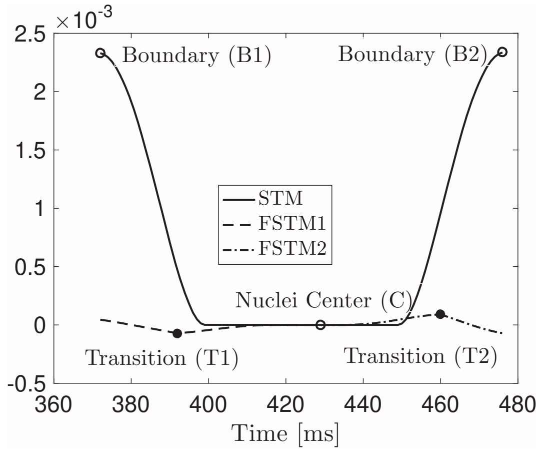

Fig. 2: Locations to extract event targets in temporal decomposition, based on coarticulation model. STM indicates the spectral transition rate of the phoneme. FSTM1 and FSTM2 are respectively the derivatives of STM on the first and second halve.

2) Coarticulation model for MRTD: The coarticulation effect of two consecutive phonemes is critical in perceiving natural sound. Therefore, a model of this effect (Figure 2) become important as well. According to Nghia *et al.* [14], in a phoneme, it has five locations: two boundaries, two transitions, and a nuclei center to represent the coarticulated transition regions. The nuclei center is the minimal point of the spectral transition rate (STM) of the phoneme. The transitions are respectively minimal and maximal points of the derivatives of each half of the STM. Those points were locations to extract event targets in temporal decomposition for modeling the phonemes and modifying features to Lombard speech.

3) Source filter model with cepstrum-based spectral tilt and spectral-GMM-based vocal tract spectrum: This model precisely decomposes and modifies spectral tilt and formants. The tilt was estimated by a smooth cepstrum, which is represented by three first coefficients c_0, c_1, and c_2. The vocal tract spectrum was divided into two parts: the positive (peaks) and negative (dips) components after subtracting c_0, c_1 and c_2. They were further modeled by spectral-GMM [15]. The modification of F_1, F_2, F_3, F_4 (the formant frequencies were estimated by using KARMA [16]) and the length of the vocal tract were done on the positive component, while the negative one was preserved.

4) Fujisaki model to control f_0: f_0 was parameterized and controlled by Fujisaki model [17], [18]. In the model, f_0 baseline Fb, amplitude of accent commands (Aa) were increased, the amplitude of phase commands were varied to obtain the target f_0 mean and range by non-linear optimization.

5) Target prediction model to control power envelope: Power envelope was parameterized by the second order damping modeling, in which the parameter *target* was used to control power envelope portions to expected powers. The *target* was extracted using target prediction model [19], [20].

In short, duration was controlled by event functions. Spectral tilt was modified by cepstral coeffients. Formants were modified by spectral-GMM. f_0 was modified by using Fujisaki model. Power envelope was controlled by the target of the second order damping model. After all features were modified, they were used to synthesize the mimicking Lombard speech by STRAIGHT.

III. Listening experiments

To evaluate our models with any noise levels, there are two main experiments: similarity, and intelligibility and naturalness.

A. Experiments of similarity

The purpose of this experiment was to compare our model with BGMM in a mean of resembling Lombard speech.

1) Speech material: Speech material was drawn from the recorded speech (both Lombard speech produced at 66, 72, 78, 84 dB noise levels and neutral speech) [21]. 105 Japanese words (4-mora) of a male and a female were taken.

2) Speech types: BGMM-based methods had two types: Glottal vocoder-based (called **GlottalBGMM**) and STRAIGHT-based (called **STRAIGHTBGMM**) synthesis. In both, the modified features were spectral tilt, f_0, duration, and power envelope. In addition, we synthesized two more types: **ProposedF0Tilt** and **ProposedF0TiltFormant**. The former's modified features were *spectral tilt*, f_0, duration, and power envelope. The latter's modified features were *spectral tilt*, f_0, *Formants*, duration, and power envelope. In total, it had four mimicking types: GottalBGMM, STRAIGHTBGMM, ProposedF0Tilt, and ProposeF0TiltFormant.

3) Listeners: Twelve native Japanese including 9 males and 3 females from 23 to 25 years old (a mean of 24) with no report of hearing problems.

4) Procedure: The complete set was 105 words in both Lombard speech and four mimicking types mentioned above produced at 4 noise levels: 66, 72, 78, 84 dB noise levels. A stimulus was a pair of concatenated the mimicking speech and Lombard speech with the same content. There were 1680 stimuli in total (105 words x 4 pair types x 4 noise levels). Each listener was assigned 64 pairs at a specific noise level using balanced design. Each pair type/noise level was listened by the same number of listeners. The listeners were asked to evaluate how the mimicking speech resemble Lombard speech in a five scale (1: none, 2: little, 3: moderately, 4: much, 5: very much) by clicking the correspondent buttons.

The experiment was 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 for the test measured by a sound level meter (hand-held analyzer type 2250 Bruel. & Kjar), which had been calibrated. Before carrying the experiment, listeners were familiarized with Lombard speech by listening Lombard and neutral speech.

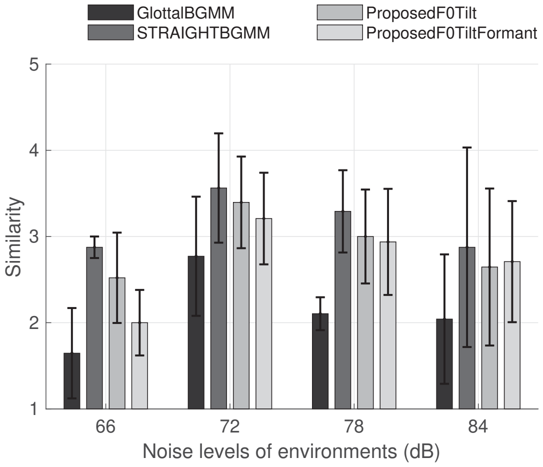

Fig. 3: Similarity of the mimicking speech. The bar and error values indicate the mean and standard deviation among listeners. The values of similarity mean 1: none, 2: little, 3: moderately, 4: much, 5: very much similar to Lombard speech

5) Results and discussion: Figure 3 shows the results of similarity to Lombard of the mimicking speech. Throughout all noise levels, the similarity scores decreased by STRAIGHT-BGMM, ProposedF0Tilt, ProposedF0TiltFormant, and Glottal-BGMM repectively. ProposedF0Tilt seemed comparable with STRAIGHTBGMM. It could be seen that the Lombard effect model could help to obtain a similar result with the statistical methods. The results shows that proposed model could correctly represent Lombard speech with varying noise levels.

B. Experiments of naturalness and intelligibility

The purpose of this experiment was to evaluate the intelligibility and the naturalness of the mimicking speech by our model compared with BGMM-based methods when different set of features are modified. This might reveal some clues to improve intelligibility and naturalness for the speech in noise. The experiment was carried out in the other dataset without Lombard speech. This also proved the generality of our proposed model.

1) Speech material: Speech material was drawn from the ATR dataset. 384 words (3-mora) of neutral speech of six different speakers (3 males, 3 females)

2) Speech types: We used four types: **ProposedTilt**, **ProposedF0Tilt**, **ProposedTiltFormant**, and **ProposedF0TiltFormant**. ProposedTilt's modified features were *spectral tilt*, duration and and power envelope. ProposedF0Tilt's modified features were *spectral tilt*, f_0, duration and power envelope. ProposedTiltFormants's modified features were *spectral tilt*, *Formants*, duration and power envelope Lastly, ProposedF0TiltFormants's modified features were f_0, *spectral tilt*, *Formants*, duration and power envelope. We chose STRAIGHTBGMM as a reference due to the same vocoder. In total, it had 5 types: STRAIGHTBGMM, ProposedTilt, ProposedF0Tilt, ProposeTiltFormant, and ProposeF0TiltFormant.

3) Listeners: Seven native Japanese including 5 males and 2 females from 22 to 25 years old (a mean of 23.57) with no report of hearing problems.

4) Maskers: Pink noise [22] at 4 noise levels: 66, 72, 78, 84 dB had been used, thus there were 4 maskers.

5) Procedure: The complete set was 7680 stimuli (384 words x 5 speech types x 4 noise level maskers). Within a test of intelligibility or naturalness, 60 unique words were assigned to one listener at each noise level. Each listener listened to all 4 noise levels in an increasing order. They did the intelligibility test and naturalness test in sequence.

- 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 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 the correspondent buttons. The next stimulus would be played immediately after that.

6) Results and discussion:

- Intelligibility
  Figure 4 shows the results that only with the modification of spectral tilt, our method obtained a comparable result with STRAIGHTBGMM. With the modification of the other feature sets, it obtained lower intelligibility. However, throughout all noise levels, the scores are varied in a similar way. This might be due to some interactions among features rather than the proposed model. Therefore, the proposed model still well contributed to this intelligible adaption with varying noise levels, which represented Lombard speech.
- Naturalness
  Figure 5 shows the results among different feature sets. Our method with the modification of spectral tilt showed a comparable result with STRAIGHTBGMM. With the modification of the other feature sets, it obtained lower naturalness. It could be explained by effects of the modification of some parameter features rather than the proposed model. For an example, the modified f_0 range might cause wrong pitch accents, thus reduced the naturalness. Therefore, it could be seen that the proposed model still worked in this evaluation.

IV. Conclusions

In this paper, we have presented the concept of the Lombard effect model with varying noise levels and its application with the modification-synthesis method. The method was based on coarticulation and source filter model and MRTD with spectral-GMM, which can easily control features with multiple noise levels. The results showed that our method can be comparable with the state-of-the-art method. The proposed model has correctly represented Lombard speech with varying noise levels. Specifically, at a fixed noise level, the state-of-the-art method could be better. When noise levels are continuous, it cannot adapt features to the noise levels. Otherwise, our model can interpolate Lombard speech with any noise levels. In order to obtain better intelligibility and naturalness, we aim to improve our modification methods in f0 contour and formants in future work. This Lombard effect model is expected to be used in an extrapolation model for an even better intelligible speech based on Lombard speech.

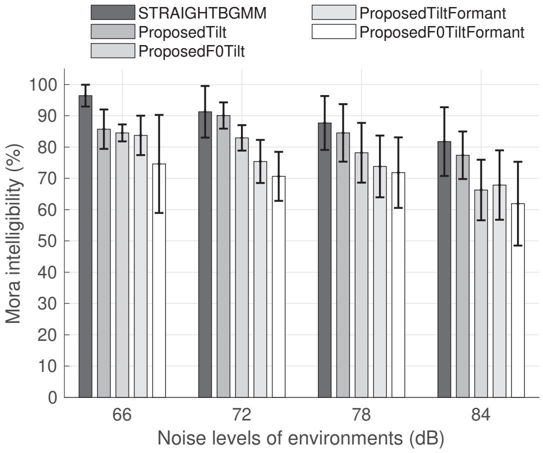

Fig. 4: Intelligibility of the speech when various features are mimicked, i.e. percentage of correctly answered mora in a word. The bar and error values indicate the mean and standard deviation of among participants.

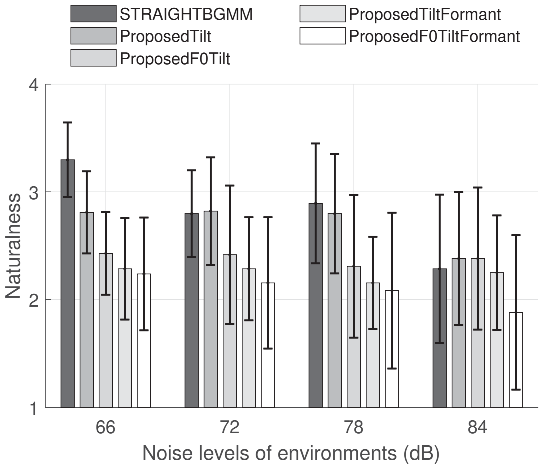

Fig. 5: Naturalness of the speech when various features are mimicked. The bar and error values indicate the mean and standard deviation among participants. The values of naturalness mean 1: unnatural, 2: rather unnatural, 3: rather natural, 4: natural.

V. Acknowledgements

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'e'le'vation de la voix," *Annales des Maladies de L'Oreille et du Larynx*, vol. 37, pp. 101–119, 1911.

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

[3] 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.

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

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

[6] 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*, pp. 1363–1367, 2017.

[7] 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)*, pp. 5505–5509, IEEE, 2017.

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

[9] T. V. Ngo, R. Kubo, D. Morikawa, and M. Akagi, "Acoustical analyses of tendencies of intelligibility in lombard speech with different background noise levels," *JSP*, vol. 21, pp. 171–174, 2017.

[10] T. V. Ngo, R. Kubo, and M. Akagi, "Acoustical rules for mimicking lombard speech produced in a various noise level background," in *Proceedings of the auditory research meeting*, vol. 47, pp. 475–480, 2017.

[11] 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.

[12] 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.

[13] 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.

[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] 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.

[16] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1732–1746, 2012.

[17] H. Fujisaki, S. Narusawa, and M. Maruno, "Pre-processing of fundamental frequency contours of speech for automatic parameter extraction," in *WCC 2000-ICSP 2000. 2000 5th International Conference on Signal Processing Proceedings. 16th World Computer Congress 2000*, vol. 2, pp. 722–725, IEEE, 2000.

[18] 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*, vol. 1, pp. I–509, IEEE, 2002.

[19] 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.

[20] Y. Xue and M. Akagi, "A study on applying target prediction model to parameterize power envelope of emotional speech," in *2016 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP'16)*, , 2016.

[21] 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*, vol. 253, pp. 3366–3371, Institute of Noise Control Engineering, 2016.

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