Speech Intelligibility Enhancement in Noisy Environments via Voice Conversion with Glimpse Proportion Measure

Taiho Takeuchi* and Yosuke Tatekura†
* Graduate School of Science and Technology, Shizuoka University, Shizuoka, Japan
E-mail: t-take@spalab.eng.shizuoka.ac.jp
† Graduate School of Integrated Science and Technology, Shizuoka University, Shizuoka, Japan
E-mail: tatekura.yosuke@shizuoka.ac.jp Tel/Fax: +81-53-478-1139

***Abstract*—In this study, we propose a new method for improving speech intelligibility in a public space such as station concourse that focuses on the difference in ease of listening due to the individuality of a speaker's voice. The proposed method evaluates the speech intelligibility caused by voice characteristic difference, and it converts the input voice to improve intelligibility measure by using voice-quality control based on the Gaussian mixture model. In the proposed method, glimpse proportion measure, which can estimate independent listening position, is used as the evaluation index of speech intelligibility. Speech that was converted by using the proposed method had larger energy than the background noise in some frequency bands. Results of subjective evaluation reveal that the proposed method can improve the speech intelligibility under a noisy environment.**

I. Introduction

For information transmission in a public space such as a station concourse, a voice announcement by using a loudspeaker is very effective. However, such speech's intelligibility decreases due to the masking caused by constant noises. Speech intelligibility should be improved because low intelligibility of an announcement voice in public spaces destroys smooth information transmission.

This study aims to improve speech intelligibility in additive noise. Valentini-Botinhao *et al*. proposed an improvement method of synthesized speech by using a text-to-speech (TTS) system in noisy environments [1]. In this method, the intelligibility in a noisy environment is estimated using a glimpse proportion (GP) measure [2] and is improved by converting the cepstrum as an acoustic feature. Given that using human speech is more intuitive to announce in public spaces than to convert text, speech generation based on the TTS system cannot satisfy this demand. However, immediacy is required in case of emergency announcement in public spaces. In such a case, TTS-based speech synthesis must able to prepare various announcement patterns in advance, or to input text according to circumstances. Meanwhile, if a speaker's voice can be converted directly, then advance preparation is unnecessary and usability is good.

The method of converting Lombard voice using the Gaussian mixture model (GMM) was investigated to directly improve speech intelligibility. The Lombard effect is defined as the involuntary raising of voice when speaking in a noisy environment. The intelligibility of a Lombard voice can be improved in comparison with normal voices. Hence, speech intelligibility can be improved through GMM training by using the Lombard voice as the training data. However, because the Lombard voice is often tense, it is not suitable for announcements in public areas. In addition, at least 50 utterances are necessary for the GMM training for stable voice conversion, but forcing this to all the people who will announce is unrealistic.

Based on the above perspective, we propose a speech intelligibility improvement method based on voice conversion that does not rely on Lombard effect. Given that the proposed method is voice conversion focused on voice individuality, the intelligibility of each speaker is evaluated using an evaluation index, and the input speech is converted on the basis of the evaluation scale. Therefore, converting speech obtained as normal utterance into speech that is easy to hear is possible.

II. Voice Quality Control

A voice conversion method based on GMM [3] is an excellent technique to convert voice characteristics, and it is also applied to, for example, high-quality conversion method based on trajectory estimation [4]. As an application for the voice conversion method, the voice quality control method based on multiple-regression Gaussian mixture model (MR-GMM) [5] has been proposed. The voice quality control method based on MR-GMM [5] converts an input voice into a target voice quality, and it correlates an intuitive parameter representing the voice quality with nonintuitive parameter of GMM. Recently, a voice conversion method by using a deep neural network (DNN) was proposed [6]. However, an established method of voice quality control does not exit. In this study, we use the MR-GMM method for voice quality control and not the DNN method. The outline of the voice quality control method in the following section is in the sequence of a basic voice conversion method by GMM, followed by one-to-many eigenvoice conversion method, and finally by the voice quality control method.

In the voice conversion method based on GMM, the acoustic feature of the source and target speakers is modeled in a one-to-one correspondence. In such conversion, a dynamic feature [4], which is considered to improve the quality of converted speech, is extensively used. Let $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ be the acoustic feature vectors of the source and target speakers in frame t, respectively. When the dynamic feature is represented by Δ, the acoustic feature of the source and target speakers are $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$, $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$, where $\top$ denotes the transposition of the vector. By introducing a joint vector $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$ that is aligned by the dynamic time warping to ensure that $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ correspond to each frame, the joint probability density is modeled by GMM as follows:

$$
\begin{aligned}
P(\boldsymbol{Z}_t \mid \boldsymbol{\lambda}^{(Z)}) &= \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\mu}_m^{(Z)}, \boldsymbol{\Sigma}_m^{(ZZ)}), \\
\boldsymbol{\mu}_m^{(Z)} &= \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \\
\boldsymbol{\Sigma}_m^{(ZZ)} &= \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix},
\end{aligned} \tag{1}
$$

where $\mathcal{N}(\ \cdot\ ; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, M is the total number of mixture components, α_m is the weight of the m-th mixture component, and $\boldsymbol{\lambda}^{(Z)}$ is a parameter set of GMM, which is estimated through the EM algorithm by using a parallel utterance pair. The converted feature sequence $\hat{\boldsymbol{y}}$ is determined by the maximum likelihood estimation by using the estimated parameter set $\boldsymbol{\lambda}^{(Z)}$ as

$$
\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}), \tag{2}
$$

where $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t, \ldots, \boldsymbol{X}_T]$ with the number of frames of the source feature T and $\hat{\boldsymbol{y}}$ and $\boldsymbol{Y}$ are also expressed similarly.

In one-to-many voice conversion based on the eigenvoice Gaussian mixture model (EV-GMM), the relationship between the source and arbitrary target speakers is trained using the utterance pair comprising a single source and many pre-stored target speakers [7]. The model representing the relationship with the specific speaker s via the EV-GMM is shown by controlling the output mean vector in (1) using the weight vector $\boldsymbol{w}^{(s)}$ expressed as follows:

$$
\boldsymbol{\mu}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{B}_m^{(Y)} \boldsymbol{w}^{(s)} + \boldsymbol{b}_m^{(Y)} \end{bmatrix}, \tag{3}
$$

where $\boldsymbol{B}_m^{(Y)}$ is the eigenvector and $\boldsymbol{b}_m^{(Y)}$ is the bias vector. These two vectors can be obtained by applying the principal component analysis of the output mean vector $\boldsymbol{\mu}^{(Y)}(s)$, which can be expressed as

$$
\begin{aligned}
\boldsymbol{\mu}^{(Y)}(s) &\simeq \boldsymbol{B}^{(Y)} \boldsymbol{w}^{(s)} + \boldsymbol{b}^{(Y)}, \\
\boldsymbol{B}^{(Y)} &= \left[{\boldsymbol{B}_1^{(Y)}}^\top, \ldots, {\boldsymbol{B}_m^{(Y)}}^\top, \ldots, {\boldsymbol{B}_M^{(Y)}}^\top \right]^\top, \\
\boldsymbol{b}^{(Y)} &= \left[{\boldsymbol{b}_1^{(Y)}}^\top, \ldots, {\boldsymbol{b}_m^{(Y)}}^\top, \ldots, {\boldsymbol{b}_M^{(Y)}}^\top \right]^\top.
\end{aligned} \tag{4}
$$

When converting a specific output speaker by using the EV-GMM, performing an unsupervised estimation of model parameters by estimating the weight vector $\boldsymbol{w}^{(s)}$ only from the target speaker's acoustic feature without using the parallel data is possible [8]. However, directly controlling the required voice quality of the voice conversion using the EV-GMM is difficult because the relationship between the elements of the weight vector $\boldsymbol{w}^{(s)}$ and the intuitive voice quality is unknown.

Ohta *et al*. proposed a voice quality control method based on MR-GMM [5], which is an extension of EV-GMM, as a method to flexibly control voice individuality. In this method, a voice quality control vector $\boldsymbol{w}_e$ with elements directly corresponding to parameters expressing an intuitive voice quality (e.g. age, gender, and physique) is estimated. The voice quality control vector $\boldsymbol{w}_e$ is estimated via the speaker adaptation training by using the parameter obtained by the EV-GMM training. When training the parameter $\boldsymbol{\lambda}^{(\mathrm{MR})}$ of the MR-GMM, the voice quality vector $\boldsymbol{w}_e$ is fixed to express each pre-stored target speaker and estimate the other parameters, which is expressed as

$$
\hat{\boldsymbol{\lambda}}^{(\mathrm{MR})} = \underset{\boldsymbol{\lambda}^{(\mathrm{MR})}}{\operatorname{argmax}} \prod_{s=1}^{S} \prod_{t=1}^{T_s} P\left(\boldsymbol{Z}_t^{(s)} \mid \boldsymbol{\lambda}^{(\mathrm{MR})}, \boldsymbol{w}_e^{(s)} \right), \tag{5}
$$

where S denotes the total number of pre-stored target speakers, T_s indicates the total number of frame of the speaker s, and $\boldsymbol{w}_e^{(s)}$ denotes the voice quality vector of the speaker s.

The converted feature sequence is obtained using the maximum likelihood estimation. In this study, the output mean vector $\boldsymbol{E}_{m,t}^{(Y)}$ of m-th conditional probability density distribution is obtained as follows:

$$
\boldsymbol{E}_{m,t}^{(Y)} = \boldsymbol{B}_m^{(Y)} \boldsymbol{w}_e + \boldsymbol{b}_m^{(Y)} + \boldsymbol{\Sigma}_m^{(YX)} {\boldsymbol{\Sigma}_m^{(XX)}}^{-1} (\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}). \tag{6}
$$

By setting the voice quality control vector $\boldsymbol{w}_e$ to the required parameter, we can convert the speech with an arbitrary voice quality.

III. Glimpse Proportion

The GP measure [2] is an index for speech intelligibility by focusing on speech component that is not masked by noise when recognizing speech in noisy environment. Speech intelligibility estimation through GP measure is obtained by comparing the speech energy and noise in time-frequency domain considering human auditory characteristics. In [1], an approximate value of a spectro-temporal excitation pattern (STEP) is used.

The approximated GP measure $I^{(\mathrm{GP})}$ is expressed as

$$I^{(\mathrm{GP})} = \frac{100}{N_f N_t} \sum_{t=1}^{N_t} \sum_{f=1}^{N_f} \mathcal{L}(y_{t,f}^{(\mathrm{sp})} - y_{t,f}^{(\mathrm{ns})}), \tag{7}$$

where $y_{t,f}^{(\mathrm{sp})}$ and $y_{t,f}^{(\mathrm{ns})}$ are the approximated STEP representation for speech and noise, respectively, at analysis frame t and frequency channel f. Futher, N_t and N_f are the total number of time frames and frequency channels, respectively, and $\mathcal{L}(\cdot)$ indicates the logistic sigmoid function.

The approximated STEP is calculated using a Gammatone filter. The Gammatone filter is extensively known to approximate the human auditory characteristics; hence, expressing a physical process wherein we can recognize a speech is possible. The approximated STEP value is calculated as follows:

$$y_{t,f}^{(\cdot)} = \frac{1}{N} (\boldsymbol{G}_f \boldsymbol{h}_t \textcircled{N} \boldsymbol{G}_f \boldsymbol{h}_t)^\top \boldsymbol{V} \boldsymbol{b}, \tag{8}$$

where Ⓝ denotes the circular convolution operator with the dimension N, $\boldsymbol{h}_t = [|H_t(\omega_1)|, \ldots, |H_t(\omega_N)|]^\top$ denotes the amplitude spectrum of windowed speech signal at the analysis frame t, $\boldsymbol{G}_f = \mathrm{diag}\,[g_{f,1}, \ldots, g_{f,N}]$ indicates the $N \times N$ diagonal matrix whose diagonal contains the Gammatone filter frequency response at the frequency f. In addition, $\boldsymbol{V} = \mathrm{diag}\,[v_1, \ldots, v_N]$ denotes $N \times N$ diagonal matrix whose diagonal contains the smoothing filter frequency response and $\boldsymbol{b} = [b_1, \ldots, b_N]$ is $N \times 1$ vector containing the coefficients of the averaging filter. Moreover, the center frequency of the Gammatone filter bank is equally arranged on the equivalent rectangular bandwidth (ERB) scale. The ERB is a scale expressing the frequency resolution of the auditory characteristics by a logarithmic scale. Hence, reflecting the frequency resolution of human by using a filter equally arranged on the ERB scale is possible.

IV. Voice Conversion with GP

To improve speech audibility in a noisy environment, we propose a speech intelligibility improvement method by using the voice quality control based on MR-GMM with GP measure. Fig. 1 shows the framework of the proposed method. In this voice quality control method, the model is trained to correlate the evaluation index given for each pre-stored speaker. In [5], although the voice quality can be manually manipulated by associating the subjective evaluation value with the voice of the speaker for training, performimg the original voice quality control by arbitrarily setting the corresponding parameter is also possible. Therefore, we associate the parameter of the target voice quality to the speech intelligibility to control the speech intelligibility. In this study, the GP measure is estimated for each speaker, and the voice quality is modeled corresponding to the intelligibility by using the MR-GMM training. The GP measure for each speaker is calculated for each utterance with respect to the target noise, and the average is regarded as the intelligibility of the speaker.

The speech intelligibility index (SII) [9] is extensively regarded as highly correlated measure with speech intelligibility.

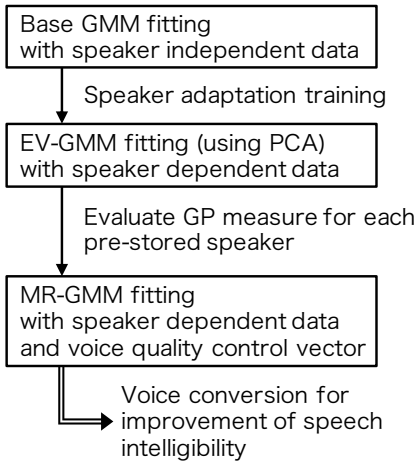

Fig. 1. Framework of the proposed method. In the proposed method, the MR-GMM training is performed to evaluate speech intelligibility via the GP measure and to control the speech quality for speech intelligibility. By performing voice quality control based on the intelligibility of the GMM due to the differences in speaker individuality, we generate a speech that is easy to hear even in a noisy environment.

However, SII is a measure for estimating speech intelligibility at a specific sound receiving position. Hence, the GP measure, which is not dependent on the sound receiving position, is more effective than SII because this study assumes that an unspecified number of listeners in a public area are targeted.

In the conversion step, the element of the voice quality vector $\boldsymbol{w}_e$ corresponding to the intelligibility is set large. Given that the naturalness of the obtained speech depends on the voice conversion method, the speech distortion of the converted speech is expectedly similar to the rule based method. In addition, given that the model can be constructed without using the Lombard speech in the proposed method, the elimination of data cost per one speaker is a possible advantage.

Given that the voice quality control based on the MR-GMM assigns the voice quality by $\boldsymbol{w}_e$, the proposed method can convert the voice quality while simultaneously fixing it, as well as the intelligibility. Particularly, by introducing the weight vector $\boldsymbol{w}^{(s)}$ estimated by using the EV-GMM into voice quality control vector $\boldsymbol{w}_e$, implicitly holding the voice characteristic of the source speaker is possible. In the proposed method, not only intelligibility but also the gender and age are explicitly regarded as an element of the voice quality control vector. In the conversion step, the gender and age of the source speaker are provided.

V. Evaluation

A. Experimental conditions

To evaluate the effectiveness of the proposed method, we confirm the speech intelligibility via the mean opinion score (MOS) test. The conditions in the experiment are described as follows. We use a speech dataset of Japanese newspaper article

sentences (JNAS) [10] for the training and data evaluation. Thirty three people are slected to evaluate the speakers as the pre-stored speaker from the subset F of JNAS, which has 45 sentences. The speaker who has the lowest GP value for each gender in the subset F is chosen as the source speaker. Five sentences that are not used for the data training are employed as the evaluation data. The converted feature is the 1st-24th-order mel-cepstral coefficients obtained by compressing 1024 dimensional smoothed spectrum extracted via the WORLD [11]. The shift length is set to 5 ms. The sampling frequency is 16 kHz. The number of MR-GMM components is set to 32, and a full covariance matrix is utilized. When calculating the GP measure, the frequency band for the evaluation is set to 55 channels at 100–7500 Hz, the tap length of the moving average filter to use the temporal frame smoothing is set to 8 ms, the frame length is set to 30 ms, and the shift interval is 10 ms. The GP measure as the target at the conversion step is calculated for each utterances with regard to the background noise, and the average value is utilized as the speaker intelligibility. The noise used for the evaluation is that of a crowd recorded in station premises. The voice quality control vector is in the order of gender, age, and GP measure, and gender and age are fixed to the source speaker's one at the conversion. The elements of the voice quality control vector are normalized into the Z-score (i.e. zero mean and unit variance).

Subsequently, we describe the condition of the subjective evaluation. The listeners are nine students in their 20s with normal hearing. They evaluate the speech randomly presented by listening via headphones. The opinion score is set from 1 (Bad) to 5 (Excellent). The presented speeches are in a total of seven types with analysis-by-synthesis speeches by using the WORLD of the input speech The speeches are converted with the set target GP measure from 0 to 50 in ten increments for one sentence.

B. Results and discussion

Fig. 2 shows the averaged spectrum of the converted speech obtained by the proposed method when the target GP measure is set to 10 or 50. The figure also presents the averaged spectrum of the input speech and the background noise. When the GP measure is set to 10, we can determine that the averaged spectrum of the converted speech decreases compared with the input speech in the frequency band of more than 2 kHz. By comparison, when the GP measure is set to 50, we can identify that the averaged spectrum of the converted speech increases over the noise spectrum in the frequency band of more than 4 kHz. Particularly, the concentration of power exceeding the background noise occurs within 4–5 kHz, thereby indicating that this band is emphasized on average of the conversion. The concentration of power also occurs at approximately 2.5 kHz, although it does not exceed the background noise. At approximately 7 kHz, an excessive emphasis is applied to deviate from the general speech characteristics. However, given that the converted speech did not deviate audibly, it is implied that the model perform a unique processing to produce an intelligibility of a speech.

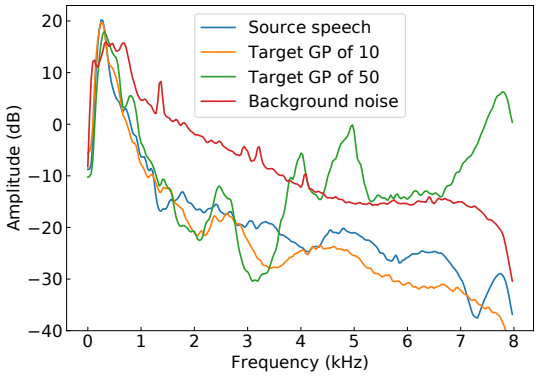

Fig. 2. Averaged spectrum of the source speech, background noise, and converted speeches with the target GP of 10 and 50.

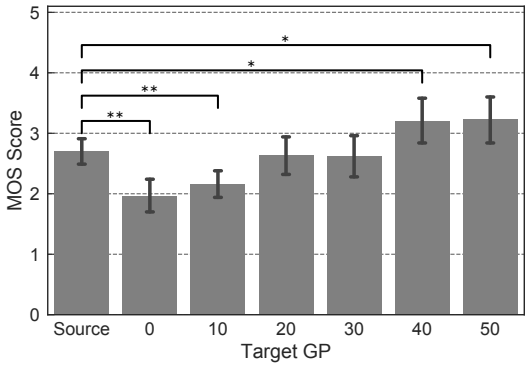

Fig. 3. Evaluation result of the MOS test for the female speaker.

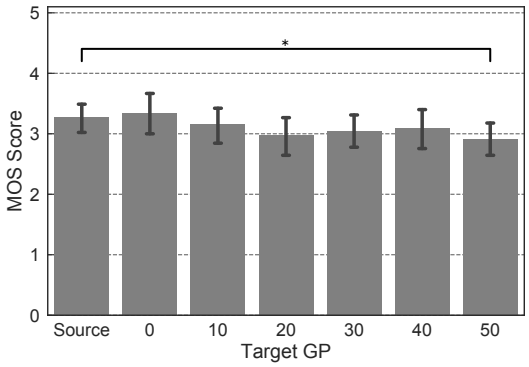

Fig. 4. Evaluation result of the MOS test for the male speaker.

Figs. 3 and 4 show the result of the MOS test for the female and male speakers, respectively, where the error bar indicates 95% confidence intervals and $*$ and $**$ indicate the significant difference $p < 0.05$ and $p < 0.01$, respectively. As shown in Fig. 3, the proposed method works effectively for the female speaker and significantly improves the intelligibility

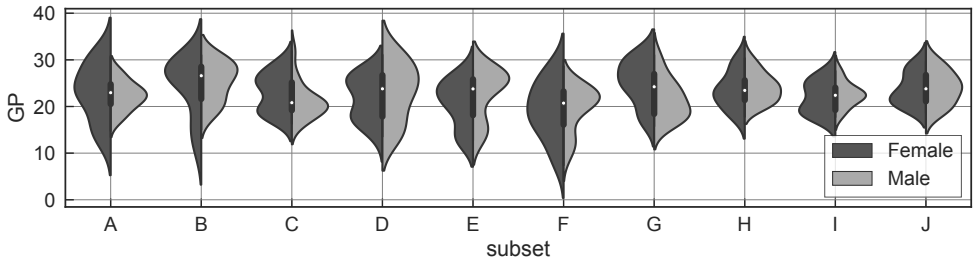

Fig. 5. Distribution of the GP measure for the JNAS dataset.

when the target GP measure is set to 40 or 50. Meanwhile, Fig. 4 reveals that the proposed method cannot improve the speech intelligibility. The reason for different in the result on improving the speech intelligibility is possible the distribution influence of the GP measure on the training data. Fig. 5 shows the distribution of the GP measure for each subset included in the JNAS dataset. Evidently, the distribution of the female speaker in the subset F is wide, whereas that of the male speaker is biased. Therefore, the speech intelligibility does not improve in the conversion because the intelligibility could cannot be sufficiently modeled in the training of the male speaker. This result suggests that the speech intelligibility of the training data should be extensively dispersed to effectively improve the speech intelligibility. In the future work, the voice quality control method based on the DNN should be implemented because the DNN-based method is considered to improve speech quality.

VI. Conclusion

To improve speech intelligibility of annoucing voice in public areas, this study proposed an improvement method based on voice conversion. Specifically, an intelligibility scale was introduced to the voice control method through MR-GMM, and models were established to cope with the intelligibility due to speaker's characteristic difference. In this study, by introducing the GP measure as the intelligibility evaluation index, we can convert a clear speech by not depending on the listening position. The subjective evaluation experiments showed that the proposed method works effectively under certain conditions. However, no effect was obtained when the training data with biased evaluation scale were used. Hence, addressing this problem is a future task. Although gender and age are explicitly specified in this experiment, the speaker's characteristic of the source speaker can be further emphasized by estimating a weight vector that determines the source speaker and incorporating it into the voice quality control vector. In addition, given that the DNN-based voice conversion method is considered to improve the speech quality, using this method for voice quality control is a future work.

References

[1] C. Valentini-Botinhao, R. Maia, J. Yamagishi, S. King, and H. Zen, "Cepstral analysis based on the glimpse proportion measure for improving the intelligibility of hmm-based synthetic speech in noise," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*., pp. 3997–4000.

[2] 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.

[3] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[4] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[5] K. Ohta, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Regression approaches to voice quality control based on one-to-many eigenvoice conversion," in *Proceedings 6th ICASSP SSW6*, pp. 101–106, 2007.

[6] Y. Saito, S. Takamichi and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, 2018.

[7] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on gaussian mixture model," *the 9th International Conference on Spoken Language Processing*, pp. 2446–2449, 2006.

[8] R. Kuhn, J. C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, 2000.

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

[10] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[11] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.