2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Real-Time Noise Estimation for Lombard-Effect Speech Synthesis in Human–Avatar Dialogue Systems

Yuto Ishikawa*, Osamu Take*, Tomohiko Nakamura†, Norihiro Takamune*,
Yuki Saito*, Shinnosuke Takamichi*‡, and Hiroshi Saruwatari*
* The University of Tokyo, Tokyo, Japan
† The National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan
‡ Keio University, Kanagawa, Japan

***Abstract*—To facilitate smooth communication between a human (client) and an avatar/robot, it is essential to generate avatar/robot speech that reduces the listening effort for the client in noisy acoustic environments. In this paper, we propose a framework for the following two purposes by leveraging the estimated noise using our previously proposed real-time blind speech extraction method: (i) real-time adjustment so that the power of the avatar/robot speech relative to that of the estimated noise is constant and sufficiently large, and (ii) mimicking the Lombard effect by simply applying filters to the avatar/robot speech. The proposed framework can also smoothly connect speeches generated from voice conversion (avatar) and text-to-speech (robot) techniques. Subjective evaluations demonstrated that the proposed framework achieves natural synthesis in noisy environments.**

I. Introduction

Systems that allow human-like avatars and robots to interact with humans (clients) have been researched and developed for decades [1]. One essential component for realizing such systems is an autonomous dialogue system for robots [2]. By utilizing advances in technologies such as automatic speech recognition (ASR) [3], dialogue control [4], response generation [5], and text-to-speech (TTS) [6], robots can respond to clients' queries flexibly. These advances in automatic dialogue are expected to reduce human resources by replacing roles that only humans can take conventionally. However, in practical applications, irregular cases that an autonomous robot cannot deal with may exist (e.g., speech recognition errors and out-of-domain dialogue scenes). In such cases, the robot can be switched to an avatar, and a human operator can take over the response to handle client's questions. In this way, by having an automatic dialogue in normal cases and by having a human operator intervention only when necessary, a single operator can concurrently control multiple avatars. As a result, the number of avatars that a single operator can control increases, leading to a reduction in human resources. Additionally, in this case, the consistency between the avatar and robot speakers can be maintained by converting the operator's voice to the robot's speaker voice using a voice conversion (VC) technique [7]. On the other hand, since the power of the voice output by VC is correlated with that of the input voice spoken by the operator, there can be a large difference in power between the automatic speech (TTS) and the converted speech of the operator (VC). As a result, clients may feel discomfort owing to the discontinuity of speech that they hear. In the following, the robot's voice generated by a TTS technique, the voice converted from the operator's speech using a VC technique, and the avatar's/robot's voice that the clients hear are called *TTS voice*, *VC voice*, and *synthesized voice*, respectively. In this paper, we address a practical situation where an avatar/robot interacts with a single client in real-world noisy environments.

In such cases, the background noise makes a client's and synthesized voices obscure. Consequently, the client and the avatar's operator have difficulty listening to each other's speech, and the accuracy of ASR for the robot is also degraded. To extract the client's speech from the audio signals observed by a microphone array embedded in the avatar/robot, we previously proposed a real-time speech extraction framework [8]. This is expected to make it easier for the operator to understand the client's speech and improve the accuracy of ASR in noisy environments. On the other hand, since noise information is discarded through speech extraction, the avatar's operator and the robot cannot consider the background noise in the client's environment and cannot always generate a voice that the client can hear easily in noisy environments. As a result, the synthesized voice remains hidden in noise, which hinders smooth communication.

In this paper, we propose a framework for generating natural and audible speech for the client. In this framework, we utilize the property of a linear demixing filter, that is, it can accurately separate noise signals in a situation where a single directional target speaker exists in diffuse noise [9]. By utilizing this property, we estimate the client's background noise in real time using the multichannel speech extraction method proposed in [8] and adjust the power of the synthesized voice to be audible with less listening effort according to the estimated noise level. This adjustment results in the continuity of the power of the synthesized voice when switching between the TTS and VC

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011 (for framework development) and Tateisi Science and Technology Foundation (for practical experiment).

voices. Additionally, to further improve the intelligibility of the synthesized voice for the client in noisy environments, we partly simulate the Lombard effect [10], [11], which is the involuntary change in human speech features in noisy environments, on the synthesized voice. Subjective evaluations showed that the synthesized voice generated by our proposed framework sounds natural in noisy environments compared with that generated without processing or by simply applying a constant gain.

II. Related works

A. Real-time speech extraction framework

In [8], we previously proposed the real-time speech extraction framework based on independent low-rank matrix analysis (ILRMA) [12] and rank-constrained spatial covariance matrix estimation (RCSCME) [13]. In ILRMA, the short-time Fourier transform (STFT) is utilized to process the observed audio signals in the time-frequency domain. Let us define the STFTs of the observed, source, and separated signals as $\boldsymbol{x}_{ij} \in \mathbb{C}^M$, $\boldsymbol{s}_{ij} \in \mathbb{C}^N$, and $\boldsymbol{y}_{ij} \in \mathbb{C}^N$, respectively. Here, $i \in \{1, ..., I\}$, $j \in \{1, ..., J\}$, $m \in \{1, ..., M\}$, and $n \in \{1, ..., N\}$ are the indices of the frequency bins, time frames, microphones, and sources, respectively. If each source is a point source and the reverberation time is sufficiently shorter than the STFT window length, the observed signal is approximately represented as $\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij}$. Here, $\boldsymbol{A}_i = (\boldsymbol{a}_{i1}, ..., \boldsymbol{a}_{iN}) \in \mathbb{C}^{M \times N}$ is the mixing matrix and $\boldsymbol{a}_{in}$ is the steering vector for the nth source to the microphone array. In ILRMA, under a determined condition ($M = N$), we estimate the time-invariant demixing matrix $\boldsymbol{W}_i \in \mathbb{C}^{N \times M}$ and subsequently separate the observed signals into each source signal as $\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}$. Here, to resolve the scale ambiguity in each frequency bin of the separated signals, the frequency-wise scale of each separated signal is aligned with that of the observed signal at the reference microphone channel using projection back (PB) [14].

However, when we apply ILRMA to a mixture that consists of a directional target speech and diffuse noise, one separated signal corresponding to the target speech includes diffuse noise components [15]. On the other hand, the other separated signals contain only diffuse noise and suppress the target speech with high accuracy [9]. RCSCME utilizes these properties of the separated signals and efficiently estimates the parameters for a multichannel Wiener filter. RCSCME uses parameters estimated by ILRMA to reduce the number of to-be-estimated parameters and can achieve rapid processing. However, it is difficult to extend the speech extraction method consisting of ILRMA and RCSCME for real-time processing in a straightforward manner because the update algorithm for ILRMA is computationally costly owing to numerous matrix operations. Here, we exploit two facts: 1) the ILRMA output required in RCSCME is only the time-invariant demixing matrix and 2) RCSCME itself is computationally efficient and can be executed within the shift length of the STFT. On the basis of these facts, in [8], a framework that performs ILRMA and RCSCME in parallel by introducing the blockwise batch algorithm [16], [17] has been proposed. This framework can execute RCSCME at the shift length intervals and output the extracted target speech with low latency (less than 100 ms, excluding input/output latency). On the other hand, since it is difficult to perform ILRMA within the shift length, it estimates and outputs the demixing matrix at longer time intervals.

Since ILRMA is a fully blind method, it is necessary to select the channel containing the target speech from the separated signals without any prior information. However, by introducing a regularizer based on spatial prior information into ILRMA, we can induce the separated signal at a specific channel to contain the target speech. As one such regularizer, in [8], ILRMA with null-based spatial regularization (NSR-ILRMA) has been proposed. The spatial regularization term is calculated using a steering vector for the target speech, $\hat{\boldsymbol{a}}_i$, derived from prior information about the client's position against the avatar/robot. The cost function of NSR-ILRMA is expressed as

$$\begin{aligned}\mathcal{T} =& \frac{1}{J} \sum_{i,j,n} \left(\frac{|\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2}{r_{ijn}} + \log r_{ijn} \right) - \sum_{i} \log |\det \boldsymbol{W}_i|^2 \\ &+ \sum_{i,n} \mu_{in} (1 - \delta_{nn^{(\mathrm{t})}}) |\boldsymbol{w}_{in^{(\mathrm{t})}} \hat{\boldsymbol{a}}_i|^2 + \mathrm{const.}, \qquad (1)\end{aligned}$$

where $r_{ijn} > 0$ denotes the time-varying variance of the separated signal y_{ijn}, $\mathrm{const.}$ represents the term independent of $\boldsymbol{w}_{in}$ and r_{ijn}, and $\cdot^{\mathsf{H}}$ is the Hermitian transpose. Here, $n^{(\mathrm{t})}$ is a prespecified channel index corresponding to the target speech. This regularizer induces $\boldsymbol{w}_{in}$ $(n \neq n^{(\mathrm{t})})$ to form the null in the direction of the target speech, and thus, $\boldsymbol{w}_{in^{(\mathrm{t})}}$ seems to correspond to the target speech. The process of minimizing (1) is detailed in [8]. In this case, we can choose $n^{(\mathrm{t})}$ for the channel selection.

B. Lombard effect

The Lombard effect is the involuntary modification of the characteristics of human speech produced in noisy environments to enhance its audibility and intelligibility for listeners [10]. It has been reported that the Lombard speech, which refers to speech produced under the Lombard effect, exhibits changes in various characteristics, such as its power, spectral characteristics, and fundamental frequency [11]. In [11], the spectra of speech produced in a markedly noisy environment were compared with those produced in a quiet environment. This comparison showed an average increase in the power of approximately 5 dB in the low-frequency band around 200 Hz and in the high-frequency band above 5000 Hz, whereas an average increase in the power of approximately 15 dB is observed in the frequency band around 500 to 4000 Hz. It is also reported that relative to quiet environments, the mean vocal level (the power of speech) increased by approximately 14.5 dB.

III. Proposed method

A. Motivation

A simple approach to improving audibility in noisy environments is by applying a constant gain to the synthesized

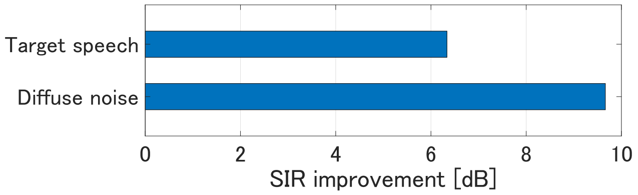

Fig. 1. SIR improvements of the target speech (above) and diffuse noise (below) estimated by proposed method with offline NSR-ILRMA.

voice. However, since noise is unstationary, a fixed gain can be an inappropriate setting. For example, if the gain is small, the synthesized voice can remain inaudible in very noisy environments, and conversely, if the gain is large, the synthesized voice can become uncomfortably loud in quiet environments. Therefore, we must adaptively adjust the gain to the noise intensity in real time.

To address this issue, we propose a method to estimate the power of time-varying noise in real time and adjust that of the synthesized voice to ensure that the power of the synthesized voice is sufficiently higher than that of noise at a constant rate. This approach is expected to generate an audible voice suitable for various noisy environments. Moreover, even in a special case where TTS and VC voices switch, this processing makes the power of the synthesized voice almost constant as long as the noise power does not change abruptly, enabling a smooth connection of the synthesized voice. If the observed signals are directly used to calculate the noise power without any processing, the client's speech might be erroneously considered as background noise. Therefore, the power of synthesized voice increases in correlation with not only the power of the background noise but also that of the client's speech. As a result, even if the environment is consistently quiet, the synthesized voice will become louder when the client speaks loudly. Therefore, a method is required to suppress the client's speech and estimate only the noise power in real time.

B. Proposed real-time noise estimation procedure

In this section, we aim to remove the target speech and estimate the noise accurately in real time. First, we focus on the characteristics of source separation methods based on the linear demixing filter, such as ILRMA. In [9], it is reported that the linear demixing filter can accurately separate noise signals in a situation where a single directional target speaker exists in diffuse noise. In this situation, the observed signal $\boldsymbol{x}_{ij}$ can be represented as $\boldsymbol{x}_{ij} = \tilde{\boldsymbol{a}}_i \tilde{s}_{ij} + \boldsymbol{u}_{ij}$, where $\tilde{\boldsymbol{a}}_i$ and $\tilde{s}_{ij}$ denote the steering vector and the source signal of the target speech, respectively, and $\boldsymbol{u}_{ij}$ denotes the source image of diffuse noise. By performing the real-time speech extraction method with NSR-ILRMA, we can obtain the linear demixing matrix $\boldsymbol{W}_i$ and the index corresponding to the target speech $n^{(\mathrm{t})}$ in real time. Here, the scale of $\boldsymbol{W}_i$ is modified by PB as

$$\boldsymbol{W}_i \leftarrow \mathrm{diag}\big(\boldsymbol{e}_{m_{\mathrm{ref}}}^{\mathsf{H}} \boldsymbol{W}_i^{-1}\big) \boldsymbol{W}_i, \tag{2}$$

where $\mathrm{diag}(d_1, ..., d_N)$ denotes the operator that outputs an $N \times N$ diagonal matrix whose diagonal elements are given by the vector $(d_1, ..., d_N)$, m_{ref} is the index of the reference microphone channel, and $\boldsymbol{e}_m \in \mathbb{C}^M$ is a one-hot vector with 1 only in the mth element. By utilizing these parameters, we can define the estimated noise signal $\hat{u}_{ij}$ as

$$\hat{u}_{ij} = \sum_{n \neq n^{(\mathrm{t})}} \boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}. \tag{3}$$

Here, we focus on the facts that the demixing filters $\boldsymbol{w}_{in}$ $(n \neq n^{(\mathrm{t})})$ can form the null of the steering vector for the target speech $\tilde{\boldsymbol{a}}_i$ and then $\boldsymbol{w}_{in}^{\mathsf{H}} \tilde{\boldsymbol{a}}_i \approx 0$ $(n \neq n^{(\mathrm{t})})$, resulting in $\hat{u}_{ij} = \sum_{n \neq n^{(\mathrm{t})}} \boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{u}_{ij}$. We note that the noise power is reduced compared with the true value by the noise component leaking in the target speech direction, $\boldsymbol{w}_{in^{(\mathrm{t})}}^{\mathsf{H}} \boldsymbol{u}_{ij}$. We experimentally verified this characteristic, and Fig. 1 shows the source-to-interference ratio (SIR) [18] improvements for the target speech and the diffuse noise when offline NSR-ILRMA is applied to mixture signals consisting of a single directional target speech and diffuse noise. The experimental conditions for Fig. 1 are as follows: We used a 4.2-s-long female speech signal from the JSUT dataset [19] as a dry source. Then the dry source was convolved with the recorded impulse response and its signal length was adjusted to 5.12 s by padding with the zero value. The recording conditions for both the diffuse noise and impulse response and the settings for STFT were the same as those described in Section IV-A. The target speech was mixed with the diffuse noise so that the input signal-to-noise ratio (SNR) became 0 dB for the entire signal at a reference microphone channel except for the silent intervals. We applied NSR-ILRMA to the mixture signal, and an estimated noise signal was then obtained by calculating (3). As demonstrated in Fig. 1, we empirically confirmed that NSR-ILRMA can separate the background noise with higher accuracy than in the case of the target speech under the experimental conditions used in this study. This result is consistent with [9], [13]. Moreover, it is expected that we can accurately estimate the diffuse noise in real time using the real-time speech extraction method.

Using the real-time estimated noise, we attempt to generate audible speech by applying a time-varying gain to the synthesized voice so that the ratio between the power of the output speech and noise estimated in real time is constant. Furthermore, by imitating the spectral energy changes described in Section II-B, we simulated the Lombard effect on the synthesized voice. These modifications will enable the generation of more natural-sounding speech across diverse noisy environments.

C. Proposed framework

Fig. 2 illustrates our proposed framework. In the following, we describe the details of the processing part highlighted in red frame.

1) SNR modification using real-time noise estimation: First, we consider modifying the vocal level of the synthesized voice. For the client to hear the synthesized voice clearly in noisy environments, we should calculate the gain to achieve a specified output SNR. This requires obtaining the power of both the synthesized voice and noise.

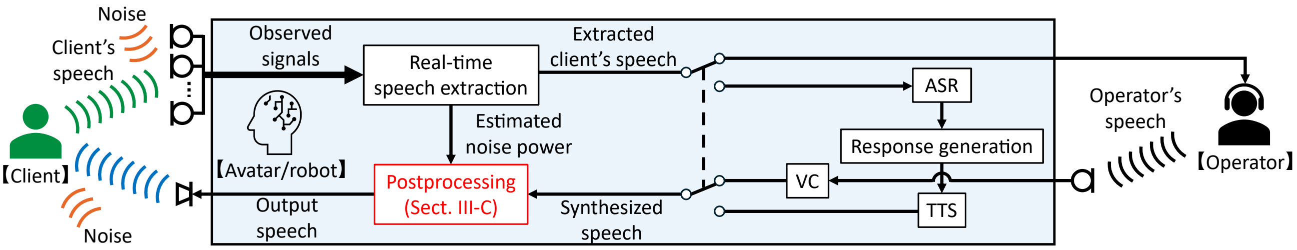

Fig. 2. Schematic of proposed framework for dialogue system embedded in avatar/robot.

For the power of the synthesized voice, sample speeches are prepared for both TTS and VC voices, and their powers are precalculated as references. That is, when a $T^{(l)}$-sample-long segment of TTS or VC voices $s_{\mathrm{sample}}^{(l)}[t]$ ($l \in \{\mathrm{TTS}, \mathrm{VC}\}$, $t = 1, \dots, T^{(l)}$) is given beforehand, we define the average power (the power per unit time) of the synthesized voice $S^{(l)}$ as

$$S^{(l)} = \frac{1}{T^{(l)}} \sum_{t=1}^{T^{(l)}} \left(s_{\mathrm{sample}}^{(l)}[t] \right)^2 . \tag{4}$$

On the other hand, for the noise power, to suppress the client's speech and adapt to the time-varying noisy environment, we use the noise signal estimated at the shift length intervals using the real-time speech extraction framework and (3). Here, to suppress the rapid changes in noise power, the exponential smoothing method is introduced and noise information history is also considered. The noise signal estimated by the real-time speech extraction framework is denoted as $u[t]$. This noise signal $u[t]$ is estimated at every shift length ΔT. For the integer k, the noise power estimate U_k for the time $t = (k-1)\Delta T + 1, \dots, k\Delta T$ is defined using the estimated noise signal at that time interval as

$$U_k = (1-\lambda) U_{k-1} + \frac{\lambda}{\Delta T} \sum_{t=(k-1)\Delta T+1}^{k\Delta T} (u[t])^2 , \tag{5}$$

where $\lambda \in (0, 1]$ denotes a forgetting factor and U_0 is set to 0.

Finally, when the output SNR is set to R, the time-varying gain $\alpha_k^{(l)}$ for the synthesized voice is calculated as

$$\alpha_k^{(l)} = 10^{\frac{R}{20}} \sqrt{\frac{U_k}{S^{(l)}}}. \tag{6}$$

2) Simple imitation of Lombard effect: Next, to imitate the increased energy in the relatively high-frequency band observed in the Lombard speech, we consider boosting the energy in the 500–4000 Hz range by 5–10 dB compared with the other frequency bands. In this paper, we apply a bandpass filter that allows the 500–4000 Hz frequency band to pass through and an all-pass filter with a similar phase response to the bandpass filter to the input TTS/VC voice in parallel. Subsequently, the filtered signals are summed. That is, if the transfer functions of the bandpass and the all-pass filters in the z-domain are represented as $h_{\mathrm{B}}(z)$ and $h_{\mathrm{A}}(z)$, respectively, the transfer function of our proposed filter is expressed as $h_{\mathrm{B}}(z) + h_{\mathrm{A}}(z)$.

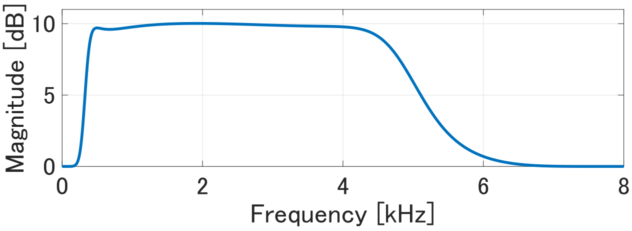

Fig. 3. Magnitude frequency response for proposed filter.

The bandpass filter $h_{\mathrm{B}}(z)$ is designed as follows: 5th-order Butterworth type, a passband of 500–4000 Hz, a passband ripple of 3 dB, stopband frequencies of 50 and 7000 Hz, and a stopband attenuation of 60 dB. The cutoff frequencies of this filter become about 376 and 4720 Hz. Then, the 5th-order infinite impulse response all-pass filter $h_{\mathrm{A}}(z)$ is designed by searching for parameters using a greedy algorithm to minimize the mean squared error between phase responses of this filter and the bandpass filter. Here, the gain of the bandpass filter relative to that of the all-pass filter is determined to increase the energy in the passband by 10 dB. The magnitude frequency response of the proposed filter $h_{\mathrm{B}}(z) + h_{\mathrm{A}}(z)$ is shown in Fig. 3.

Since this filtering affects the power of the entire output speech, the processing should be conducted in the following order: first, perform the filtering described in this section, and then perform the SNR modification described in Section III-C1.

IV. Experiments

In our experiments, we consider three application scenarios where a human and an avatar/robot are talking under a diffuse-noise condition: (i) listening to only the TTS voice, (ii) listening to only the VC voice, and (iii) listening to the TTS voice followed by the VC voice almost continuously with different powers. We evaluate the effectiveness of the proposed method by subjective evaluation.

A. Experimental conditions

Diffuse noise and impulse responses were recorded at the Ito International Research Center, The University of Tokyo. A circular microphone array with a radius of 3.25 cm and composed of four omnidirectional microphones was placed at a height of 1 m from the floor. During the diffuse noise recording, 10 participants were seated 2–4 meters apart from the microphone array and engaged in conversations with others

TABLE I

Preference scores on naturalness of output voice. Values enclosed in parentheses denote p-values of Student's t-test. **Bold** indicates significantly (p-value < 0.025) better scores.

Case	Noise condition	Number of evaluators	Baseline / Naive	Baseline / Proposed	Naive / Proposed
(i)	-5 dB	26	0.115 / **0.885** ($< 10^{-10}$)	0.154 / **0.846** ($< 10^{-10}$)	0.558 / 0.442 (9.70×10^{-2})
	0 dB	28	0.098 / **0.902** ($< 10^{-10}$)	0.089 / **0.911** ($< 10^{-10}$)	0.268 / **0.732** ($< 10^{-10}$)
	5 dB	25	0.070 / **0.930** ($< 10^{-10}$)	0.040 / **0.960** ($< 10^{-10}$)	0.100 / **0.900** ($< 10^{-10}$)
(ii)	-5 dB	25	0.110 / **0.890** ($< 10^{-10}$)	0.030 / **0.970** ($< 10^{-10}$)	0.090 / **0.910** ($< 10^{-10}$)
	0 dB	25	0.110 / **0.890** ($< 10^{-10}$)	0.060 / **0.940** ($< 10^{-10}$)	0.040 / **0.960** ($< 10^{-10}$)
	5 dB	25	0.050 / **0.950** ($< 10^{-10}$)	0.020 / **0.980** ($< 10^{-10}$)	0.020 / **0.980** ($< 10^{-10}$)
(iii)	-5 dB	24	0.083 / **0.917** ($< 10^{-10}$)	0.073 / **0.927** ($< 10^{-10}$)	0.094 / **0.906** ($< 10^{-10}$)
	0 dB	25	0.090 / **0.910** ($< 10^{-10}$)	0.030 / **0.970** ($< 10^{-10}$)	0.060 / **0.940** ($< 10^{-10}$)
	5 dB	26	0.058 / **0.942** ($< 10^{-10}$)	0.067 / **0.933** ($< 10^{-10}$)	0.038 / **0.962** ($< 10^{-10}$)

around them or read a text provided to them beforehand. Simultaneously, music was played from loudspeakers embedded in the ceiling. The impulse responses were recorded under the following conditions: the height of the target speaker was 1.1 m, the horizontal distance between the target speaker and the microphone array was 1 m, and the reverberation time T_{60} was around 750 ms.

Next, we describe the processes for synthesizing the TTS and VC voices. For both, the synthesized speech speaker was one female speaker from the SaSLaW corpus [20]. For the TTS voice, FastSpeech 2 [21] was used as the deep neural network model to output the mel-spectrogram of the speech, and the trained HiFi-GAN [22] was used as the vocoder to generate the speech waveform. The details of TTS models are based on [20]. Then, FastSpeech 2 was pretrained on the JSUT dataset [19] and fine-tuned on single-speaker recordings of SaSLaW. For the VC voice, RVC[1] trained on the synthesized speech speaker's data was used, and the input speaker was one female speaker from the JVS corpus [19].

In this paper, we simulate a situation where the VC voice is quieter and more difficult to hear than the TTS voice in different noisy environments. To achieve this, the VC voices were adjusted so that the average power of each VC voice became -5 dB relative to that of each TTS voice. Here, the TTS voices used were adjusted beforehand so that their average power became constant. Then, we adjusted the noise signal so that the average power of the noise signal at the reference channel became -5, 0, and 5 dB relative to that of the TTS voice, thereby creating three types of noise condition. We note that, as a result, the SNRs of the TTS voice were 5, 0, and -5 dB, and the SNRs of the VC voice were 0, -5, and -10 dB, in the same order as above. Then, for experiments (i) and (ii), to obtain clean synthesized speech signals, two utterances were selected from either TTS or VC voices and concatenated with 5-s-long silent intervals in between. The silent interval before the first utterance was used to calculate the noise power sufficiently and the silent interval between the utterances was used to simulate a conversation between a human and an avatar/robot. Similarly, for experiment (iii), a clean synthesized speech signal was created by concatenating a 5-s-long silent interval, one TTS utterance, a 1-s-long silent interval, and one VC utterance in this order. The 1-s-long interval between the TTS and VC utterances simulates a scene where the operator intervenes suddenly, whereas the 5-s-long silent interval before the TTS utterances serves the same purpose as in (i) and (ii). In total, there were nine experimental conditions: Three experimental scenarios and three noise mixing conditions. For each experimental condition, 10 clean synthesized speech signals were created. The clean synthesized speech signals created in this way are referred to as *standard speech signals*.

Next, we describe how to generate the input signals for the real-time speech extraction framework in each noisy environment. To simulate client utterances in noisy environments, we used single-female-speaker utterances from the JSUT dataset [19] as client's dry sources. These utterances were within 5 s in length. After convolving the dry sources with the impulse responses, the resulting signals were mixed with the noise signals after power adjustment with TTS voices. In the mixing process, the average power of each convolved signal was adjusted to be 0 dB relative to the power of the noise, and the timing of each convolved signal was aligned with the 5-s-long silent intervals inserted before two TTS or VC voices in the experiments (i) and (ii), and before the first TTS voice in the experiment (iii). The sampling rate of all signals was set to 16 kHz.

For comparison, the following three methods were used.

- **Baseline**: the standard speech signals are output without any additional processing.
- **Naive**: signals applied at a constant gain of $+10$ dB to the standard speech signals are output.
- **Proposed**: signals adjusted by the proposed real-time framework with an output SNR of $R = 15$ dB for each noise power estimate are output.

For the real-time speech extraction framework, NSR-ILRMA was used in the ILRMA part described in [8], and the other conditions were also the same as those in [8]. STFT was performed using a 64-ms-long Hann window with a shift length of 32 ms. The sample speeches $s^{(\mathrm{TTS})}_{\mathrm{sample}}[t]$ and $s^{(\mathrm{VC})}_{\mathrm{sample}}[t]$

[1] https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI

used for SNR modification were generated by selecting one utterance from TTS and VC voices, respectively, and applying the filtering process described in Section III-C2. For the forgetting factor λ, we set it to 0.5 by experimentally choosing it from some values in $(0, 1]$.

For subjective evaluation, we generated evaluation speech signals to simulate listening under each noise condition by mixing the output speech signal for each compared method with the corresponding noise signal. Then, two of the three evaluation speech signals were selected as an evaluation pair, and AB preference tests (listening tests) were conducted for each pair to determine which evaluation speech signal sounded more natural in a noisy environment. Each evaluator listened to and evaluated a total of 12 evaluation pairs containing the same number of evaluation pairs for each combination of the compared method.

B. Subjective evaluation

For each experiment and each noise mixing condition, 24–28 evaluators were recruited and 96–112 responses were collected via the crowdsourcing platform "Lancers"[2]. The significant differences in preference scores were evaluated using Student's two-tailed t-test with a significance level of 5%. The results are shown in Table I. Under almost all experimental conditions, the scores were significantly higher for Naive than for Baseline and for Proposed than for Naive. We also consider the following to be the reason why there was no significant difference between Naive and Proposed only in the case (i) with a noise condition of -5 dB. First, under this condition, the ratio between the power of the TTS voice and noise signal was 5 dB. Therefore, the synthesized voice was adjusted so that the resulting SNR became 15 dB for Naive. On the other hand, Proposed also functions so that the average power of the synthesized voice became 15 dB relative to the noise power estimate under this experimental condition. However, since noise components remain in the separated signal corresponding to the target speech by ILRMA, the noise power estimate is slightly lower than the actual noise power. Then, in the gain calculation (6), U_k becomes smaller and the gain also becomes smaller, resulting in a slightly lower SNR than the ideal output SNR of $R = 15$ dB. Thus, Naive has slightly higher power than Proposed, making it easier and slightly more natural for the client to hear in a noisy environment. This results in both methods showing a competitive performance.

V. Conclusion

In this paper, for the dialogue system embedded in the avatar/robot, we proposed a real-time framework to modify the synthesized speech so that it is easy for clients to hear. We estimated the noise power in real time by utilizing the real-time speech extraction framework based on ILRMA and RCSCME. In the postprocessing part, we performed (i) power adjustment using estimated noise power in real time to make the synthesized speech audible and (ii) filter processing to partially imitate the Lombard effect. Subjective evaluation experiments confirmed that the proposed framework can generate a more natural speech in noisy environments than the simple approaches.

[2] https://www.lancers.jp/

References

[1] O. Mubin, M. I. Ahmad, S. Kaur, W. Shi, and A. Khan, "Social robots in public spaces: A meta-review," in *Proc. ICSR*, 2018, pp. 213–220.

[2] T. Kawahara, "Spoken dialogue system for a human-like conversational robot ERICA," in *Proc. IWSDS*, 2018, pp. 65–75.

[3] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. ASLP*, vol. 22, no. 4, pp. 745–777, 2014.

[4] H. Chen, X. Liu, D. Yin, and J. Tang, "A survey on dialogue systems: Recent advances and new frontiers," *SIGKDD Explor. Newsl.*, vol. 19, no. 2, pp. 25–35, 2017.

[5] C. Dong, Y. Li, H. Gong, *et al.*, "A survey of natural language generation," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–38, 2022.

[6] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv*, 2021.

[7] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. ASLP*, vol. 29, pp. 132–157, 2020.

[8] Y. Ishikawa, K. Konaka, T. Nakamura, N. Takamune, and H. Saruwatari, "Real-time speech extraction using spatially regularized independent low-rank matrix analysis and rank-constrained spatial covariance matrix estimation," in *Proc. HSCMA*, 2024.

[9] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Trans. ASLP*, vol. 17, no. 4, pp. 650–664, 2009.

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

[11] A. L. Pittman and T. L. Wiley, "Recognition of speech produced in noise," *JSLHR*, vol. 44, no. 3, pp. 487–496, 2001.

[12] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. ASLP*, vol. 24, pp. 1626–1641, 2016.

[13] Y. Kubo, N. Takamune, D. Kitamura, and H. Saruwatari, "Blind speech extraction based on rank-constrained spatial covariance matrix estimation with multivariate generalized Gaussian distribution," *IEEE/ACM Trans. ASLP*, vol. 28, pp. 1948–1963, 2020.

[14] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

[15] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures," *EURASIP JASP*, vol. 2003, no. 11, pp. 1–10, 2003.

[16] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction," *IEICE Trans. Fundamentals*, vol. E87-A, pp. 1941–1948, 2004.

[17] Y. Mori, H. Saruwatari, T. Takatani, *et al.*, "Blind separation of acoustic signals combining SIMO-model-based independent component analysis and binary masking," *EURASIP JASP*, vol. 2006, no. 034970, pp. 1–17, 2006.

[18] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] S. Takamichi, R. Sonobe, K. Mitsui, *et al.*, "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *AST*, vol. 41, no. 5, pp. 761–768, 2020.

[20] O. Take, S. Takamichi, K. Seki, Y. Bando, and H. Saruwatari, "SaSLaW: Dialogue speech corpus with audio-visual egocentric information toward environment-adaptive dialogue speech synthesis," in *Proc. INTERSPEECH*, 2024.

[21] Y. Ren, C. Hu, X. Tan, *et al.*, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, 2021.

[22] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020, pp. 17 022–17 033.