人間とアバターとの対話システムにおける拡散性雑音下リアルタイム推定雑音を用いたLombard効果模擬音声合成のための検討*

◎石川 悠人，武 伯寒（東大），中村 友彦（産総研），高宗 典玄，
齋藤 佑樹（東大），高道 慎之介（東大/慶大），猿渡 洋（東大）

1 はじめに

近年，人間を模したアバターやロボットが対話をおこなうシステムの研究や実用化が進んでいる [1]．ロボットの対話システムに関しては，ユーザの問いかけに対して事前に決められた応答を一定のルールで出力するのみではなく，自動音声認識（automated speech recognition: ASR）や対話制御，テキスト音声合成（text-to-speech: TTS）といった技術の発展によりユーザの問いかけに対して自動かつ柔軟に対応することも可能となりつつある．従来は人間がユーザの問いかけに直接対応していたが，これらの技術によりロボットが人間の代わりを務めることで人的資源の削減が期待される．しかし，実環境での音声認識誤りやドメイン外対話シーンなど，実際の応用時には対応しきれないケースが生じうる．こうした状況では，人間が対応を引き継ぐことでユーザの問いかけに臨機応変に対処できる．このように平時には自動で対話を行い，必要が生じた場面でのみ人間が介入することで，1人の操縦者に対して複数のアバターを割り当てることも可能となる．またこの際，操縦者の音声を音声変換（voice conversion: VC）によってロボット音声話者に変換することでユーザに提示される音声の一貫性を保つことができる．その一方でVCが出力する音声のパワーは操縦者が発する入力音声のパワーに相関するため，自動発話（TTS）と操縦者による発話（VC）の間でパワーが大きく異なりうる．その結果，ユーザ側が聞く音声が断続的となり，ユーザに大きな違和感を与えるという課題が残る．以下では自動発話によるロボット音声をTTS音声，操縦者による発話にVCを適用した音声をVC音声と区別し，ユーザが実際に聞く音声をまとめて合成音声と呼ぶ．

本稿では，このような対話システムを搭載したアバターやロボットの運用シーンとして周囲に環境雑音が存在する実環境下での人間との対話を考える．この場合，ユーザもアバターの操縦者も環境雑音によって互いの発話が聞き取りにくくなり，またロボットの音声認識精度も下がる．ロボットや操縦者が聞く音からユーザの音声をリアルタイムに抽出する手法として，我々はリアルタイム音声強調フレームワークを提案してきた [2]．これにより雑音環境下において，操縦者がユーザの発話を聞き取りやすくしたり，ASRの精度を上げることが可能となる．その一方で雑音情報が切り捨てられるため，ロボットや操縦者はユーザの背景雑音状況を知ることができず，そうした雑音を考慮した合成音声を生成できなくなる．その結果，ユーザが聞く合成音声は雑音に埋もれたままになってしまい，円滑なコミュニケーションの妨げとなる恐れがある．

そこで本稿では，雑音下の対話においてアバターやロボットがユーザに向けて話すシーンを想定し，ユーザにとって聞き取りやすい音声を生成するフレームワークを提案する．[2] によって背景雑音をリアルタイムに推定し，これに対して合成音声が一定かつ十分大きなパワー比となるように調整する．さらに雑音下でより聞き取りやすくなるように，人間が雑音下で円滑に発話を伝達させるために発話の特徴量を変化させるLombard効果 [3] を合成音声上に模擬する．実験では，雑音環境下において単純に一定のゲインをかけ続ける手法と提案フレームワークを比較し，主観評価によって提案フレームワークが雑音環境下でより自然に聞こえることを確認する．

2 関連研究

2.1 リアルタイム音声強調フレームワーク [2]

我々は以前，独立低ランク行列分析（independent low-rank matrix analysis: ILRMA）[4] とランク制約付き空間共分散行列推定法（rank-constrained spatial covariance matrix estimation: RCSCME）[5] に基づく拡散性雑音下音声強調手法をリアルタイム化するフレームワークを提案した [2]．ILRMAは観測信号を短時間Fourier変換（short-time Fourier transform: STFT）し時間周波数領域で信号を扱う．マイクロホン数が音源数以上である優決定条件で各音源に対する時不変な分離フィルタを推定し，各音源信号に分離する．しかし，拡散性雑音下に単一方向性話者が存在するような状況にILRMAを適用すると，目的音声に対応する分離信号に同じ方位から到来する雑音成分が残留する．その一方で，雑音に対応する分離信号において高精度で目的音声を除去することができる [6]．RCSCMEは，ILRMAのこれらの性質を用いて高い精度で音声強調をおこなう手法である．また，RCSCMEはILRMAで推定されたパラメータ

*Lombard-Effect-Mimicking Processing for Speech Synthesis Using Real-Time Diffuse Noise Estimation in Human-Avatar Dialogue Systems. by Yuto Ishikawa, Osamu Take (The University of Tokyo), Tomohiko Nakamura (The National Institute of Advanced Industrial Science and Technology (AIST)), Norihiro Takamune, Yuki Saito (The University of Tokyo), Shinnosuke Takamichi (The University of Tokyo/Keio University), Hiroshi Saruwatari (The University of Tokyo)

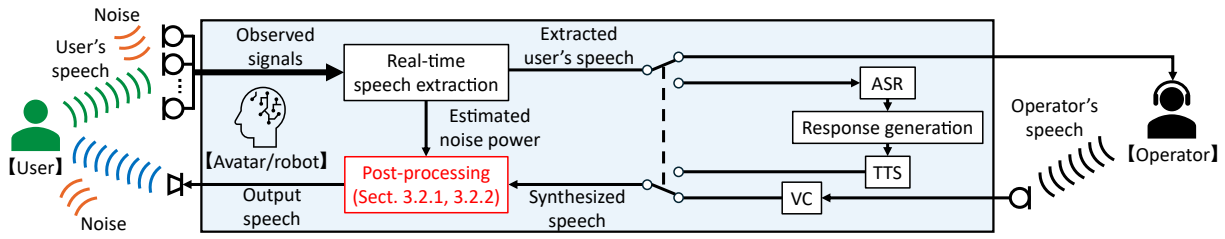

Fig. 1　Schematic of proposed framework for dialogue system embedded in avatar/robot.

を活用することで推定すべきパラメータ数を削減している．さらに [5] によって効率的な推定アルゴリズムが提案されており，非常に高速な実行が可能である．その一方で ILRMA は多数の行列演算を要するため計算コストが高い．そこで，RCSCME が要求する ILRMA の出力が時不変な分離フィルタのみであることと，RCSCME は非常に高速に動作し STFT のシフト長以内に実行可能であることに着目する．[2] ではこれらの事実から，ブロックワイズバッチアルゴリズムを導入し ILRMA と RCSCME を並列に動作させるフレームワークを提案した．これにより，STFT のシフト長間隔で RCSCME を実行し，強調音声を出力できる．一方で ILRMA はシフト長以内に実行することが困難なため，ある程度長い時間間隔で分離フィルタを推定し出力する．さらに [2] では，ILRMA に空間事前情報からなる正則化を導入することで分離性能を向上できることも報告されている．

2.2　Lombard 効果 [3]

雑音環境下において，人間は対話相手に円滑にかつ明瞭な発話を伝達するために無意識に発話の特徴量を変化させることが知られており，これは Lombard 効果と呼ばれている [3]．Lombard 効果が発現した発話は Lombard 発話と呼ばれ，発話音声のパワーやスペクトル特性，基本周波数などの特徴量が変化することが報告されている [7]．特に [7] によれば，非常にうるさい雑音環境における発話のスペクトルごとのエネルギーは 200 Hz 近傍の低周波数帯や 5000 Hz 程度の高周波数帯では静かな環境と比べて平均 5 dB ほど増加する一方で，500–4000 Hz の周波数帯では平均 15 dB ほど増加することや，騒がしい環境ではスペクトル傾斜が増大することが報告されている．

3　提案フレームワーク

3.1　動機

雑音環境下で合成音声を聞き取りやすくする方法として，合成音声に対して適当なゲインをかける方法が考えられる．しかし雑音は定常的ではないため，適当に定めたゲインでは変わらず雑音に埋もれたり，却って合成音声が過大な音量になり，ユーザに不快感をもたらす可能性がある．したがって，雑音の強さに対して適応的にゲインを変える手法が望ましい．

そこで解決策として，時間変化する雑音のパワーをリアルタイム推定し，それに対する合成音声の相対パワーが一定かつ十分大きくなるように調整する．この処理により，多様な雑音環境に適した聞き取りやすい合成音声を生成できると考えられる．また雑音状況が急変しない限りは合成音声のパワーをある程度一定に保てるようになるため，TTS 音声と VC 音声が切り替わるようなケースでも合成音声を滑らかに接続できる．もし単純に観測信号をそのまま雑音のパワーの計算に使った場合，ユーザの発話を背景雑音と判断してしまうために，背景雑音だけでなくユーザの発話のパワーにも相関して合成音声のパワーが増加する．そのため，背景雑音は一定に静かでもユーザが大きな声を発したときに合成音声のパワーも大きくなってしまう．したがって，リアルタイムにユーザ発話を分離して雑音を推定する手段が求められる．

本稿では ILRMA が雑音を高精度に推定できる [6] ことから，[2] によって雑音をリアルタイム推定し，推定雑音信号に対して一定のパワー比を保つようにリアルタイムに計算されたゲインを合成音声にかけることでユーザにとって聞き取りやすい音声を生成する．また，雑音下でより聞き取りやすい音声を生成するために，2.2 節で述べた Lombard 発話における変化を合成音声上に再現する．今回は [7] で示されたスペクトルエネルギーの変化を再現する．

3.2　提案フレームワーク

提案フレームワークの概略図を Fig. 1 に示す．以下では図中赤枠で示される処理部分の詳細を述べる．

3.2.1　リアルタイム推定雑音を用いた SNR 調整

まず合成音声のパワーを調整する方法を考える．あらゆる雑音下でユーザが聞き取りやすいように，指定した SNR となるようなゲインをかければよいが，そのためには合成音声と雑音のパワーを取得する必要がある．

合成音声のパワーについては，TTS/VC 音声ともにサンプル音声を用意し事前に計算しておく．すなわち，事前に標本数 T のサンプル音声 $s_{\mathrm{sample}}[t]$ $(t = 1, ..., T)$ が与えられたとき，合成音声のパワー S を以下で定義する．

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

一方で雑音のパワーは時間に応じて変化しうるた

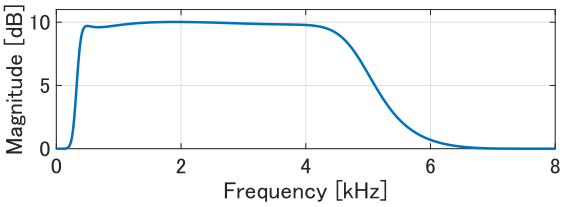

Fig. 2　Magnitude frequency response for proposed filter.

め，リアルタイム音声強調フレームワークによりシフト長間隔で推定された雑音を用いることで，雑音環境の変化に対応する．雑音のパワーの急激な変化を抑制するために，指数平滑法を導入し過去の雑音情報も考慮する．リアルタイム音声強調フレームワークにより，STFT のシフト長 ΔT ごとに推定される雑音信号を $n[t]$ とおく．ある自然数 k に対して時刻 $t = (k-1)\Delta T+1, ..., k\Delta T$ における雑音のパワーの推定値 N_k は，その時刻における推定雑音信号 $n[t]$ $(t = (k-1)\Delta T + 1, ..., k\Delta T)$ を用いて以下で定義する．

$$N_k = (1-\lambda)N_{k-1} + \frac{\lambda}{\Delta T}\sum_{t=(k-1)\Delta T+1}^{k\Delta T}(n[t])^2 \quad (2)$$

ただし，$\lambda \in [0, 1]$ は忘却係数を表し，$N_0 = 0$ とする．

以上より，出力の SNR を R としたとき，合成音声に対するゲイン α_k は

$$\alpha_k = 10^{\frac{R}{20}}\sqrt{\frac{N_k}{S}} \quad (3)$$

と計算される．

3.2.2　Lombard 効果の簡易的な再現

雑音下での Lombard 効果を再現するために，[7] を参考に 500–4000 Hz の周波数帯のエネルギーを他の周波数帯域と比べて 5–10 dB 程度増大させることを考える．今回は合成音声に 500–4000 Hz の周波数帯を通過させるバンドパスフィルタ（band-pass filter: BPF）と，作成した BPF の位相特性に近い位相特性を持ったオールパスフィルタを並列に適用し，足し合わせることで 500–4000 Hz の周波数帯のみを強調する．遅延を小さくするために BPF は 5 次の Butterworth 型を用い，オールパスフィルタは位相特性が BPF の位相特性に近づくように 5 次の infinite impulse response フィルタを用いた．また，足し合わせる前に BPF を適用した信号に一定のゲインをかけることで 500–4000 Hz の周波数帯のパワーが阻止帯域と比べて 10 dB 増加するように調整する．最終的なフィルタの振幅特性を Fig. 2 に示す．

4　実験

今回は 1 節で述べた応用シーンを再現するため，(i) 雑音下で TTS 音声のみを聞く場合，(ii) 雑音下で VC 音声のみを聞く場合，及び（iii）雑音下で TTS 音声と VC 音声が切り替わる場合の 3 つのケースを再現した音声を作成し，主観評価によって提案フレームワークの有効性を確認する．

4.1　実験条件

拡散性雑音とユーザ音声のインパルス応答は東京大学伊藤国際学術センターにて録音した．収録には半径 3.25 cm の円上に 4 つの無指向性マイクを備えた円形マイクロホンアレイを，高さ 1 m に設置した．拡散性雑音を再現するために，10 人の協力者にマイクロホンアレイの周り 2 m 以上 4 m 以内の地点に座ってもらい，周りの人との会話や，事前に渡した読み上げ用文章を読んでもらった．同時に，天井に設置されたスピーカから音楽を再生した．収録された拡散性雑音の信号長は 476 s であった．インパルス応答はマイクロホンアレイから水平方向に 1 m 離れた地点で，高さ 1.1 m の位置に音源を配置し収録した．

次に合成音声の作成方法を述べる．合成音声の話者は TTS/VC 音声ともに SaSLaW コーパス [8] の女性話者 1 名とした．TTS 音声の作成には，音声のメルスペクトログラムを出力する深層ニューラルネットワークモデルとして FastSpeech 2 [9] を，音声波形を生成するボコーダとしては学習済の HiFi-GAN [10] を用いた．TTS モデル構築の実験条件は [8] に基づいて決定した．VC 音声の作成には，合成音声話者のデータで学習した RVC[1] を用い，入力話者は JVS コーパス [11] から女性話者 1 名とした．今回は VC 音声が TTS 音声と比べて小さく聞こえにくい状況を再現するために，TTS 音声の二乗和の時間平均を基準として VC 音声の二乗和の時間平均が –5 dB となるように調整した．ただし，使用する TTS 音声のパワーは一定となるように事前に調整した．その後，異なる大きさの雑音を再現するために，TTS 音声のパワーに対して雑音信号の参照チャネルにおける時間波形の二乗和の時間平均が –5，0，5 dB となるように調整し，3 種類の雑音状況を作成した．すなわち，TTS 音声については SNR がそれぞれ –5，0，5 dB となり，VC 音声については SNR がそれぞれ –10，–5，0 dB となる．また 1 節の応用シーンを再現するために，実験（i）及び（ii）ではそれぞれ TTS または VC 音声から 2 発話を選択し，5 s 間の無音区間と選択した音声を交互に結合することで，実験（iii）では 5 s 間の無音区間，TTS 音声，1 s 間の無音区間，VC 音声の順に結合することで各実験ケースに対してクリーンな合成音声信号を作成した．以上により，実験（i），(ii)，及び（iii）のそれぞれに対して 3 種類の雑音混合条件の合計 9 種類の実験条件を設定した．各実験条件に対してそれぞれ 10 種の合成音声信号を作成した．またユーザの発話を模擬し対話を再現するために，リアルタイム音声強調フレームワークへの入力

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

信号にはユーザ発話として JSUT データセット [11] から信号長 5 s 以内の単一女性話者音声を用いた．インパルス応答を畳み込んだ後，無音区間を除いた信号全体で二乗和の時間平均が雑音のパワーに対して 0 dB となるように調整し，実験（i）と（ii）では発話前と発話間の 2 か所の無音区間に，実験（iii）では発話前の 1 か所の無音区間に含まれるように雑音信号と混合した．合成音声信号，入力信号，及び雑音信号はいずれも標本化周波数を 16 kHz とした．

比較手法には一定のゲインをかけ続ける方法（Naive）と提案フレームワーク（Proposed）の 2 つを用いる．Naive では合成音声信号を常に +10 dB するようなゲインを用いた．Proposed では推定した雑音パワーに対して出力 SNR が $R = 15$ dB となるように合成音声信号をリアルタイムに調整した．リアルタイム音声強調フレームワークについて，ILRMA 部分は [2] より NSR-ILRMA を用い，他の条件も [2] と同じものを用いた．STFT は窓長 64 ms，シフト長 32 ms，Hann 窓でおこなった．合成音声のパワーを調整するためのサンプル音声 $s_{\mathrm{sample}}[t]$ には対応する TTS 及び VC 音声から 1 発話を選択し，同様のフィルタ処理を施した信号を用いた．忘却係数は $\lambda = 0.5$ とした．

主観評価の際は各手法の出力信号と雑音信号を混合することで，雑音環境での聴取を模擬した評価用音声を作成した．その後，各比較手法に対応する評価用音声 2 つを評価対とし，どちらの評価用音声が雑音環境下で自然で聞き取りやすいかについての AB テストをおこなった．各評価者は含む 4 つの評価対を聴き，評価した．

4.2 主観評価実験

AB テストの評価結果はランサーズ[2]を介したクラウドソーシングにより収集した．各実験ケース・各雑音混合条件ごとにクラウドワーカーの重複を許容して 24–28 名，合計 96–112 件の回答を収集した．評価結果から，各実験ごとにプリファレンススコアを算出した．プリファレンススコアの差に関する評価は有意水準 5%の Student の両側 t 検定によりおこなった．

結果を Tab. 1 に示す．合成音声信号に比べて雑音のパワーが小さい場合でのみ有意差が見られなかったが，それ以外のすべての実験条件で Naive と比べて Proposed はスコアが有意に高かった．これらの結果から，提案フレームワークにより雑音下で自然に聞き取りやすい音声が生成できることがわかった．

5 おわりに

本稿では，人間とアバターやロボットが対話する状況においてユーザがより聞き取りやすく，また違和感がないように合成音声をリアルタイムに調整するフレームワークを提案した．さらに，雑音環境下でより聞き取りやすくなるように Lombard 効果の一部をアバター音声上に模擬した．主観評価実験によって，提案フレームワークにより従来より自然な音声が生成できることを確認した．

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

	Noise	Naive	Proposed	p-value
(i)	−5 dB (26)	0.558	0.442	9.70×10^{-2}
	0 dB (28)	0.268	**0.732**	$< 10^{-10}$
	5 dB (25)	0.100	**0.900**	$< 10^{-10}$
(ii)	−5 dB (25)	0.090	**0.910**	$< 10^{-10}$
	0 dB (25)	0.040	**0.960**	$< 10^{-10}$
	5 dB (25)	0.020	**0.980**	$< 10^{-10}$
(iii)	−5 dB (24)	0.094	**0.906**	$< 10^{-10}$
	0 dB (25)	0.060	**0.940**	$< 10^{-10}$
	5 dB (26)	0.038	**0.962**	$< 10^{-10}$

Table 1 Preference scores on naturalness of output voices with p-value of Student's t-test. The values enclosed in parenthesis denotes the number of participants. **Bold** indicates significantly (p-value $<$ 0.025) better scores.

謝辞 本研究は，JST ムーンショット型研究開発事業 JPMJMS2011（フレームワークの設計），公益財団法人立石科学技術振興財団 2023 年度研究助成（S）（主観評価実験）の支援を受けたものです．

参考文献

[1] O. Mubin et al., "Social robots in public spaces: A meta-review," in *Proc. ICSR*, 2018, pp. 213–220.

[2] Y. Ishikawa et al., "Real-time speech extraction using spatially regularized independent low-rank matrix analysis and rank-constrained spatial covariance matrix estimation," in *Proc. HSCMA*, 2024.

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

[4] D. Kitamura et al., "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. ASLP*, vol. 24, pp. 1626–1641, 2016.

[5] Y. Kubo et al., "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.

[6] Y. Takahashi et al., "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Trans. ASLP*, vol. 17, no. 4, pp. 650–664, 2009.

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

[8] O. Take et al., "SaSLaW: Dialogue speech corpus with audio-visual egocentric information toward environment-adaptive dialogue speech synthesis," in *Proc. INTERSPEECH*, 2024.

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

[10] J. Kong et al., "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020.

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