様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 ３０ 日現在

機関番号：６２６１５
研究種目：研究活動スタート支援
研究期間：2019～2020
課題番号：１９Ｋ２４３７３

研究課題名（和文）Can we reduce misperceptions of emotional content of speech in the noisy environments?

研究課題名（英文）Can we reduce misperceptions of emotional content of speech in the noisy environments?

研究代表者

Ｚｈａｏ　Ｙｉ（Zhao, Yi）

国立情報学研究所・コンテンツ科学研究系・特任研究員

研究者番号：１０８４３１６２

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：私たちは主に、騒がしい環境でのスピーチの感情的な内容の誤解を減らすために調査しました。VQ-VAEベースの音声波形は、通常、不適切な韻律構造を持っていることがわかりました。 したがって、VQ-VAEに重要な拡張機能を導入しました音素と同時にF0関連の超分節情報を学習するため。会議論文を発表しました。クリーンな環境での感情的なスピーチを、VQVAEの下でロンバード効果のある感情的なスピーチに変換しようとしました。 私たちも持っていますデコードされた音声の感情的な了解度を改善するために、さまざまな敵対的ネットワークを調査しました。

研究成果の学術的意義や社会的意義
この作品は、騒がしい環境での感情表現を強化することにより、悪条件での人間のコミュニケーション効率を向上させます。 また、特定の話者に対して、ノイズに強い適切な感情的なスピーチを生成することもできます。

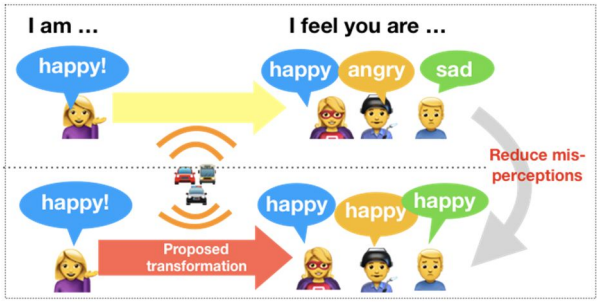
研究成果の概要（英文）：Under the real-life condition, people often need to express their emotions with appropriate speech in the noisy environments. In the past year, we mainly explored to reduce misperceptions of the emotional content of speech in the noisy environments. We found that VQ-VAE-based speech waveforms typically have inappropriate prosodic structure. Thus we introduced an important extension to VQ-VAE for learning F0-related suprasegmental information simultaneously along with phoneme features. We have published a conference paper on this work. We have tried to convert the emotional speech in the clean environment to the emotional speech with Lombard effect under the VQVAE. We have also investigated various adversarial networks to improve the emotional intelligibility of the decoded speech.

研究分野：voice conversion

キーワード：VQVAE　emotional enhancement　neural networks　voice conversion　Lombard speech　Adversarial network

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景



Humans usually adjust their way of talking in noisy environments involuntarily for effective communication. This adaptation is known as the Lombard effect [1]. The regular changes between normal and Lombard speech include not only loudness but also other acoustic features, such as prolonging the duration of their speech and increasing the pitch. By modifying the speech with the Lombard effect, people can improve intelligibility of their speech signal in the noisy environments. Such modification is referred as speech intelligibility enhancement [1].

Emotional speech in noise shows complex variations. The emotional speech uttered by male speakers is more confusable than that of female speakers in general. Older listeners are less good at recognizing the emotion contained in speech signals than young listeners [2]. Although emotional speech in noise shows the typical characteristics of the Lombard effect, the changes are complex. Emotional categories and Lombard effect mutually impact on the speech signal, which makes emotional speech in noise more easily confused.

Under the real-life condition, people often need to express their emotions with appropriate speech in the noisy environments. Although a lot of studies have been carried out on enhancing the speech intelligibility under the noisy environments, none of them considers the interaction of emotional categories and the Lombard effect in the noisy environments at the same time. Due to the complex variations of emotional speech under the noisy condition, traditional enhancement methods are no longer applicable to the emotional speech in noise. This leads to the key scientific question of this project: can we reduce misperceptions of the emotional content of speech in the noisy environments? However, this question is too broad and abstract to answer, it is better to be decomposed into several specific consecutive problems.

２．研究の目的

Our proposed idea is aimed at reducing misunderstanding of emotional content of speech produced under the noisy condition. We had three main goals that all relate to the overall goal of improving the emotional intelligibility under noisy environment.

We found that the emotional speech produced in the noisy environments could cause more confusion to listeners' judgment of emotional content compared with the emotional speech produced in quiet environments. We also observed that the emotional speech uttered by well-trained speakers resulted in much less confusion than that of less-trained speakers, especially in the noisy environment. From these phenomena, we come up with the first question: (1) how the well-trained speakers modify their emotional speech when they are in the noisy environments? Can we learn these modifications? Also, we are wondering: (2) can we apply the modifications learnt from well-trained speakers to less-trained speakers, to make the less trained speakers'

emotional speech in noise less confusable? Further, we hope to extend our study with the third question: (3) can we enhance emotion of speech for any given speaker in the noisy environments?

３．研究の方法

In this research, we will investigate emotional speech transformation, with the goal of enhancing the emotional content of speech under the noisy condition. To achieve the purpose, we consider the following strategies.

3.1 Investigating the appropriate model for emotional conversion and enhancement. Building a voice conversion model could conduct a mapping from general emotional speech (e.g. emotional speech produced in quiet) to emotional speech robust to noise using the well-trained speakers' data. We hope to learn the modifications made by well-trained speakers for emotional speech robust to noise. We consider training a speaker independent speech transformation neural network with speaker and emotional embeddings to learn the modifications in a supervised approach. The target of this network will be the emotional speech less confusable in the noisy environments.

3.2. Improving the speech quality under the voice conversion framework.

3.3 Improving the speaker similarity under the voice conversion framework.

3.4 Semi-supervised adaptation for the less-trained speakers based on the trained model.

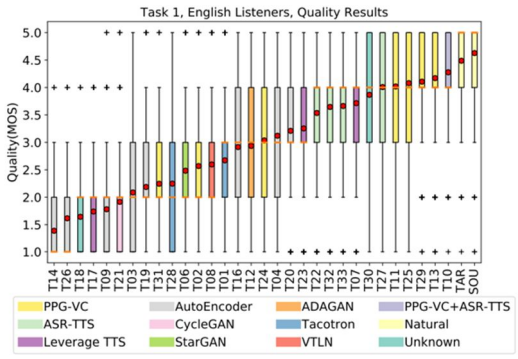
To make sufficient use of the recorded data, we will select the less-confusable speech of the less-trained speakers according to listeners' judgments and use selected data for supervised adaptation. High-confusable speech data can be used for unsupervised adaptation. We will embed x-vectors to the speech transformation model to control the latent factors such as speaker identity and speaking style for unsupervised modeling.

We have employed online crowd-sourcing listening tests to evaluate our experimental results.

４．研究成果

4.1 **We organized the voice conversion challenge 2020 and investigated advantages/disadvantages of the state-of-art systems for voice conversion (Published in Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020).**

We have finished analyzing the performance of state-of-art voice conversion systems, and found that although most of the best performing systems used PPG entirely or partially, encoder-decoder based systems could achieve better speaker similarity while speech quality was degraded.



4.2 **Built a voice conversion model under VQ-VAE framework and improved prosody from learned F0 codebook representations for speech waveform reconstruction under the VQ-VAE framework. (This work has been published at Interspeech2020.)**

We found that VQ-VAEbased speech waveforms typically have inappropriate prosodic structure in the case of Japanese. This is probably because Japanese is a pitch-accented language, which means that pitch accents directly affect the meaning of words and the perceived naturalness of the speech. However, unlike tonal languages such as Mandarin, in which each syllable is coupled with a specific tone index, Japanese pitch accentual patterns (high or low) of each mora are affected by the information at a different linguistic layer from the syllable layer, such as adjacent words. Hence, we hypothesize that a successful VQ-VAE architecture needs to simultaneously extract not only representations corresponding to the segmental features, but also another set of representations corresponding to the supra-segemental features. Motivated by this, we propose an extension to VQ-VAE structure utilizing two encoders at the same time. One encoder uses a raw speech waveform for input exactly as the original VQ-VAE does. The other encoder uses F0 trajectory as input to separately learn the pitch patterns as well as other F0-related supra-segmental information. This model was trained using a loss function that jointly considers two types of VQ losses as well as the usual coarse-to-fine waveform losses used for training WaveRNN. Listening test results show that this simple yet effective extension significantly improves prosody and naturalness of reconstructed Japanese speech waveforms. The extended VQ-VAE structure was motivated by Japanese prosody, but it can be applied to speech in any language. Therefore, we also show results of the extended VQ-VAE using a Mandarin speech database for further analysis.

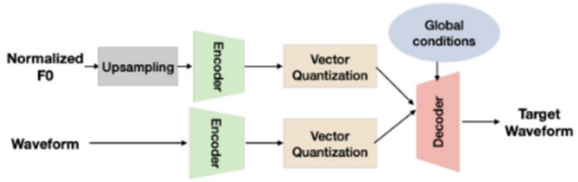


Figure 1. VQ-VAE based voice conversion framework with improved F0.

Table 1: *F0 RMSE errors and P563-based estimated MOS scores of each Japanese test speaker and their average. Natural speech MOS score is 4.2 on average. Original VQ-VAE doesn't have F0 encoder but the extended one does.*

Speaker	Gender	VQ-VAE	log F0 RMSE	P563 MOS
Speaker 1	M	Original	0.51	3.8
		Extended	0.30	4.2
Speaker 2	M	Original	0.46	3.6
		Extended	0.20	5.0
Speaker 3	M	Original	0.49	3.8
		Extended	0.27	3.9
Speaker 4	M	Original	0.43	3.7
		Extended	0.24	3.9
Speaker 5	F	Original	0.36	4.0
		Extended	0.25	4.5
Speaker 6	F	Original	0.30	3.7
		Extended	0.23	4.2
Speaker 7	F	Original	0.36	3.0
		Extended	0.27	3.5
Speaker 8	F	Original	0.41	4.1
		Extended	0.27	4.1
Average	M+F	Original	0.42	3.8
		Extended	0.26	4.2

4.3 **Improving the speaker similarity of converted speech under semi-supervised VQ-VAE paradiam. (This work has been published at ICASSP 2021)**

In this work, we present an end-to-end solution to disentangle sub-phone content and global-level speaker characteristics by building upon the original vector-quantized variational autoencoder (VQ-VAE) used in voice conversion. While VQ-VAE learns a type of sub-phone representation, the system cannot generalize well to unseen speakers or unseen content. We present our method to learn two different VQ codebooks at the same time while producing synthetic speech that is highly intelligible and with high speaker similarity in unseen conditions. We also demonstrate that the learned representations can be used in downstream tasks: phone recognition from VQ sub-phone codes, and speaker diarization from VQ speaker codes.

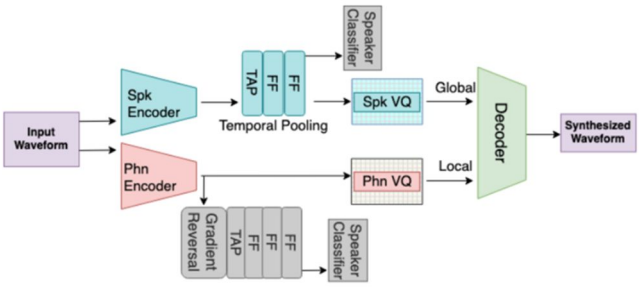


Figure 2. Speaker VQ encodes global conditions with temporal average pooling layer (TAP). Phone VQ encodes local conditions.

Table 2: Speaker diarization error (DER) scores on concatenated VCTK audio. (S: softmax, AS: angular-softmax).

		Condition				
Method		C1	C2	C3	C4	Avg
x-vector		24.3	44.6	27.4	46.7	35.8
VQ-VAE		–	–	–	–	–
+ Global VQ		44.4	39.1	44.7	39.6	42.0
+ Speaker label	S	32.4	32.2	31.0	33.1	32.2
	AS	34.6	35.9	36.4	35.9	35.7
+ Adversarial loss	S	**32.2**	**32.3**	**30.5**	**32.9**	**31.9**
	AS	37.2	35.6	36.1	35.2	36.0

4.4 We have used speaker embeddings which were extracted based on environmental and emotional categories for emotional enhancement and conversion under VQVAE framework. We have also involved adversarial training for this task. (Not published yet)

We have employed crowd-sourcing test to finish the experiment. Experimental results have shown that the proposed method could improve the less-trained speakers' performance under the noisy environment.

References:

[1]Zhao, Yi, et al. "Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion." *arXiv preprint arXiv:2008.12527* (2020).

[2]Zhao, Yi, et al. "Improved Prosody from Learned F0 Codebook Representations for VQ-VAE Speech Waveform Reconstruction}}." *Proc. Interspeech 2020* (2020): 4417-4421.

[3]Williams J, Zhao Y, Cooper E, Yamagishi J. Learning Disentangled Phone and Speaker Representations in a Semi-Supervised VQ-VAE Paradigm. InICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2021 Jun 6 (pp. 7053-7057). IEEE.

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　4件／うちオープンアクセス　4件）

１．著者名 Zhao Yi、Li Haoyu、Lai Cheng-I、Williams Jennifer、Cooper Erica、Yamagishi Junichi	４．巻 2020
２．論文標題 Improved Prosody from Learned F0 Codebook Representations for VQ-VAE Speech Waveform Reconstruction	５．発行年 2020年
３．雑誌名 Proc. Interspeech 2020	６．最初と最後の頁 4417--4421
掲載論文のDOI（デジタルオブジェクト識別子） 10.21437/Interspeech.2020-1615	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Zhao Yi, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhen-Hua Ling, Tomoki Toda	４．巻 2020
２．論文標題 Voice Conversion Challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion	５．発行年 2020年
３．雑誌名 Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020	６．最初と最後の頁 80--98
掲載論文のDOI（デジタルオブジェクト識別子） 10.21437/VCC_BC.2020-14	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Rohan Kumar Das, Tomi Kinnunen, Wen-Chin Huang, Zhen-Hua Ling, Junichi Yamagishi, Zhao Yi, Xiaohai Tian, Tomoki Toda	４．巻 2020
２．論文標題 Predictions of Subjective Ratings and Spoofing Assessments of Voice Conversion Challenge 2020 Submissions	５．発行年 2020年
３．雑誌名 Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020	６．最初と最後の頁 99--120
掲載論文のDOI（デジタルオブジェクト識別子） 10.21437/VCC_BC.2020-15	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Yi Zhao ; Xin Wang ; Lauri Juvela ; Junichi Yamagishi	４．巻 -
２．論文標題 Transferring Neural Speech Waveform Synthesizers to Musical Instrument Sounds Generation	５．発行年 2020年
３．雑誌名 ICASSP 2020	６．最初と最後の頁 6269 - 6273
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICASSP40776.2020	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計2件（うち招待講演　2件／うち国際学会　0件）

１．発表者名 Yi Zhao
２．発表標題 Modeling and evaluation methods in current voice conversion tasks
３．学会等名 言語処理学会第27回年次大会（招待講演）
４．発表年 2021年

１．発表者名 Yi Zhao
２．発表標題 Waveform loss-based acoustic modeling for text-to-speech synthesis and speech-to-musical sound transferring
３．学会等名 Seminar in National University of Singapore（招待講演）
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

Samples for emotional clean/noisy speech https://nii-yamagishilab.github.io/EmotionaLombardSpeech/ Samples for neural waveform vocoders https://nii-yamagishilab.github.io/samples-nsf/neural-music.html

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関