RGC Ref.: N_CUHK404/15 NSFC Ref. : 61531166002
(please insert ref. above)

The Research Grants Council of Hong Kong
NSFC/RGC Joint Research Scheme
Joint Completion Report

(Please attach a copy of the completion report submitted to the NSFC by the Mainland researcher)

Part A: The Project and Investigator(s)

1. Project Title

Interactive Attribute Mining and Animated Speech Synthesis for Web-based Spoken Dialog Interactions
面向互聯網口語對話的交互屬性挖掘與特色語音生成的研究

2. Investigator(s) and Academic Department/Units Involved

	Hong Kong Team	Mainland Team
Name of Principal Investigator *(with title)*	Helen Mei-Ling Meng	Zhiyong Wu
Post	Professor and former Chairman	Associate Professor
Unit / Department / Institution	Department of Systems Engineering & Engineering Management, Human-Computer Communications Laboratory, The Chinese University of Hong Kong	Department of Computer Science and Technology, Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University
Contact Information	Phone: +852.3943.8327 Email: hmmeng@se.cuhk.edu.hk	Phone: +86.0755.26036870 Email: zywu@sz.tsinghua.edu.cn
Co-investigator(s) *(with title and institution)*	N.A.	N.A.

3. Project Duration

	Original	Revised	Date of RGC/ Institution Approval *(must be quoted)*
Project Start date	1 January 2016		
Project Completion date	31 Dec., 2019		
Duration *(in month)*	48		
Deadline for Submission of Completion Report	31 Dec., 2020		

Part B: The Completion Report

5. Project Objectives

5.1 Objectives as per original application

1. To extract natural speech attributes and social attributes that characterize Web-based spoken dialog interaction by applying data mining techniques to large-scale Web corpora of text and speech;
2. To develop a correlation model between the natural speech attributes and social attributes in order to generate different animated styles for interactive speech synthesis;
3. To develop a deep neural network (DNN) based framework for animated

speech synthesis that incorporate natural speech attributes and social attributes.

5.2 Revised Objectives N.A.

Date of approval from the RGC: _____

Reasons for the change: _____

_____

1.
2.
3. ....

6. Research Outcome

Major findings and research outcome
(maximum 1 page; please make reference to Part C where necessary)

Objective 1: Speech attributes and social attributes extraction

To address the first objective, we tried to investigate the interaction characteristics in realistic human-computer conversational dialogues. Understanding the user's intention is clearly critical for generating a spoken response with attributes that are appropriate in the dialog interaction. In [1,2], we investigated the user's intention and emotive state changes in a dialogue interaction. We defined the Intention Prominence (IP) as the semantic combination of focus by text and emphasis by speech, and proposed a multi-task deep learning framework to predict IP. Specifically, the adopted long short-term memory (LSTM) structure is use to model long short-term contextual dependencies to detect focus and emphasis, and the tasks for focus and emphasis detection utilize multi-task learning (MTL) to reinforce the performance of each other. We then employed Bayesian networks (BN) to incorporate multiple features (namely, focus, emphasis, and location reflecting the users' dialect conventions) to predict the IP based on feature correlations. Experiments based on a dataset of real-world Sogou Voice Assistant illustrate that our approach can outperform other approaches. Furthermore, experimental results based on a public emotive database and a real-world interaction database reflect the effectiveness of the proposed framework in predicting user emotive state changes.

Objective 2: Correlation model between the natural speech attributes and social attributes

We explored various architectures for modeling the correlations: (A) Attention-based structures that enable neural networks to automatically determine the weights in different parts of a neural representation in temporal and spatial dimensions, which has been demonstrated to be effective in many research areas [10,12-14]. We found that by employing multi-head self-attention, the network can model the inner dependencies between elements with different positions in the learned supra-segmental feature sequence, which enhances the importing of emotion-salient information and improves the performance of emotion recognition [12, 13]. Similar performance improvement is also obtained in the tasks of emphasis detection [10] and prosodic structure prediction [14]. (B) Capsule Networks (CapNets): we propose a novel architecture for SER based on CapsNets, which can consider spatial information that are important for emotion recognition, and provide an effective pooling method for constructing the utterance-level feature representation in emotion recognition [17]. To further enable the CapsNets to consider temporal information, we propose to introduce recurrent connections to the routing algorithm in between capsule layers. Experimental results show that the proposed CapNet-based system outperforms the baseline on both weighted accuracy and unweighted accuracy, which also demonstrates the effectiveness of CapsNets in SER; (C) Structure output layer (SOL): The SOL is a structure designed to model the correlation between the multiple output tasks. We have successfully leveraged it to model the dependency between the output of prosodic boundary prediction and part-of-speech (POS) tagging and achieve satisfied performance.

Objective 3: Deep neural network (DNN) based framework for animated speech synthesis

Conventional statistical parametric speech synthesis (SPSS) methods generates frame-level acoustic features in two separately optimized steps, i.e. duration prediction and acoustic feature generation. This incorporates the uboptimal conditional independence assumption. In [19], we proposed to apply the sequence-to-sequence (seq2seq) structure attention-based recurrent generator (ARG) with Gaussian Tolerance (GT) for SPSS for joint optimization, and avoid error propagation during generation by means of GT. We also introduced residual error embedding extracted from the spoken exemplar [26], which can be automatically learned without manual emotion annotations, and efficiently generated for rapid adaptation to a target emotion using only a single adaptation utterance.

To further enhance the synthesis model's ability to synthesize various speaking styles, i.e. declarative, interrogative, etc., we used feature-based adaptation to fine-tune the top layers of the neural synthesis model. The style features are extracted from (i) a bottleneck DNN trained with target-style data, and (ii) a novel cross-style residual feature regression DNN.

Potential for further development of the research and the proposed course of action *(maximum half a page)*

For future development, it will be good to collect and annotate a corpus that can support the joint analysis of the methods used above, but this will be a costly endeavor. The corpus should contain enough instantiations of the correlation between speech attributes and social attributes. Also, the size of the corpus should be large enough to support the training of state-of-the-art neural architectures. Since the annotation of animated styles is difficult and ambiguous, it is promising to leverage the generative adversarial network (GAN) technologies to learn an automatic criterion to teach the speech synthesis model. For the correlation model between attributes, the enhancement of interpretability is still required. Understanding the functionality of each component in the networks will be useful for improving the controllability in the extracted feature representations. Although we have initial experimental results in interpreting the impact of manipulating each dimension of the extracted attribute embedding on the expressiveness of the final synthetic speech, there is still much room for improvement. Currently, our project is focused on audio and text data, but integration with video and vision data for interaction modeling will be interesting as well.

7. The Layman's Summary

(describe in layman's language the nature, significance and value of the research project, in no more than 200 words)

This project aims to develop a neural framework to synthesize animated speech that suits the ever-proliferating speech-based information services, such as voice search, personal smart phone assistants and web chatbots or social assistants. The framework consists of three integral parts: (i) attribute extraction; (ii) attribute encoding and (iii) speech synthesis. The speech and social attributes, e.g. prosody, intention and emotion, etc., are either extracted from a provided spoken exemplar, or inferred from the interaction context. The synthesis model generates animated speech utterances that contain the corresponding attributes in the extracted feature representations. We utilize data mining and deep learning techniques to build the models for the three parts. Our framework achieves good performance in modeling how the attributes govern the acoustic realizations in the interaction speech data. This is demonstrated in both public academic corpora and real-word datasets from smart phone assistant products. The project provides a promising solution to modeling web-based dialogue interaction that can serve as a baseline for future research.

Part C: Research Output

8. Peer-reviewed journal publication(s) arising directly from this research project

(Please attach a copy of each publication and/or the letter of acceptance if not yet submitted in the previous progress report(s). All listed publications must acknowledge RGC's funding support by quoting the specific grant reference.)

The Latest Status of Publications				Author(s) *(**bold** the authors belonging to the project teams and denote the corresponding author with an asterisk*)*	Title and Journal / Book *(with the volume, pages and other necessary publishing details specified)*	Submitted to RGC *(indicate the year ending of the relevant progress report)*	Attached to this report *(Yes or No)*	Acknowledged the support of this Joint Research Scheme *(Yes or No)*	Accessible from the institutional repository *(Yes or No)*
Year of publication	Year of Acceptance *(For paper accepted but not yet published)*	Under Review	Under Preparation *(optional)*						
		√ (under second review)		**Xixin Wu**, Yuewen Cao, Hui Lu, Songxiang Liu, Shiyin Kang, **Zhiyong Wu**, Xunying Liu, **Helen Meng**	Exemplar-based Emotive Speech Synthesis, **IEEE**/ACM Transactions on Audio, Speech, and Language Processing	No	No	Yes	No
		√ (under second review)		**Sougxiang Liu,** Yuewen Cao, Disong Wang, **Xixin Wu,** Xunying Liu and **Helen Meng**	Any-to-Many Voice Conversion with Location-Relative Sequence-to-Sequence Modeling, **IEEE**/ACM Transactions on Audio, Speech, and Language Processing	No	No	Yes	No

9. Recognized international conference(s) in which paper(s) related to this research project was/were delivered *(Please attach a copy of each delivered paper. All listed papers must acknowledge RGC's funding support by quoting the specific grant reference.)*

#	Month/Year/ Place	Title	Conference Name	Submitted to RGC *(indicate the year ending of the relevant progress report)*	Attached to this report *(Yes or No)*	Acknowledged the support of this Joint Research Scheme *(Yes or No)*	Accessible from the institutional repository *(Yes or No)*
1	05/2019/Brighton	End-to-end Code-switched TTS with Mix of Monolingual Recordings	ICASSP	No	Yes	Yes	Yes
2	05/2019/Brighton	Quasi-fully Convolutional Neural Network with Variational Inference for Speech Synthesis	ICASSP	No	Yes	Yes	Yes
3	05/2019/Brighton	Dilated Residual Network with Multi-head Self-attention for Speech Emotion Recognition	ICASSP	No	Yes	Yes	Yes
4	05/2019/Brighton	Learning Discriminative Features from Spectrograms Using Center Loss for Speech Emotion Recognition	ICASSP	No	Yes	Yes	Yes
5	05/2019/Brighton	A Compact Framework for Voice Conversion Using Wavenet Conditioned on Phonetic Posteriorgrams	ICASSP	No	Yes	Yes	Yes
6	05/2019/Brighton	Speech Emotion Recognition Using Capsule Networks	ICASSP	No	Yes	Yes	Yes
7	08/2019/Macau	Towards Discriminative Representation Learning for Speech Emotion Recognition	IJCAI	No	Yes	Yes	Yes

8	09/2019/Graz	Knowledge-based Linguistic Encoding for End-to-End Mandarin Text-to-Speech Synthesis	INTERSPEECH	No	Yes	Yes	Yes
9	09/2019/Graz	Disambiguation of Chinese Polyphones in an End-to-End Framework with Semantic Features Extracted by Pre-trained BERT	INTERSPEECH	No	Yes	Yes	Yes
10	09/2019/Graz	One-shot Voice Conversion with Global Speaker Embeddings	INTERSPEECH	No	Yes	Yes	Yes
11	11/2019/Lanzhou	Learning Contextual Representation with Convolution Bank and Multi-head Self-attention for Speech Emphasis Detection	APSIPA	No	Yes	Yes	Yes
12	11/2019/Lanzhou	Prosodic Structure Prediction using Deep Self-attention Neural Network	APSIPA	No	Yes	Yes	Yes
13	04/2018/Calgary	Feature based Adaptation for Speaking Style Synthesis	ICASSP	No	Yes	Yes	Yes
14	04/2018/Calgary	Emphatic Speech Generation with Conditional Input Layer and Bidirectional LSTMs for Expressive Speech Synthesis	ICASSP	No	Yes	Yes	Yes
15	05/2018/Beijing	Emphatic Speech Synthesis and Control based on Characteristic Transferring in End-to-End Speech Synthesis	ACII Asia	No	Yes	Yes	Yes
16	09/2018/Hyderabad	Emotion Recognition from Variable-Length Speech Segments using Deep Learning on Spectrograms	INTERSPEECH	No	Yes	Yes	Yes
17	10/2018/Seoul	Inferring User Emotive State Changes in Realistic Human-Computer Conversational Dialogs	ACM Multimedia	No	Yes	Yes	Yes
18	11/2018/Taipei	Speech Super Resolution Using Parallel WaveNet	ISCSLP	No	Yes	Yes	Yes

19	02/2017/San Francisco	Multi-task Deep Learning for User Intention Understanding in Speech Interaction Systems	AAAI	No	Yes	Yes	Yes
20	03/2017/New Orleans	Learning Cross-Lingual Knowledge with Multilingual BLSTM for Emphasis Detection with Limited Training Data	ICASSP	No	Yes	Yes	Yes
21	03/2017/New Orleans	Multi-Task Learning of Structured Output Layer Bidirectional LSTMs for Speech Synthesis	ICASSP	Yes	Yes	Yes	Yes
22	08/2017/Stockholm	Speech Emotion Recognition with Emotion-Pair based Framework Considering Emotion Distribution Information in Dimensional Emotion Space	INTERSPEECH	No	Yes	Yes	Yes
23	08/2017/Stockholm	Attention-based Recurrent Generator with Gaussian Tolerance for Statistical Parametric Speech Synthesis	ASMMC workshop on INTERSPEECH 2017	Yes	Yes	Yes	Yes
24	08/2017/Stockholm	Spectro-Temporal Modelling with Time-Frequency LSTM and Structured Output Layer for Voice Conversion	INTERSPEECH	Yes	Yes	Yes	Yes
25	08/2017/Stockholm	Multi-Task Learning for Prosodic Structure Generation using BLSTM RNN with Structured Output Layer	INTERSPEECH	No	Yes	Yes	Yes
26	03/2016/Shanghai	Question Detection from Acoustic Features using Recurrent Neural Network with Gated Recurrent Unit	ICASSP	Yes	Yes	Yes	Yes
27	03/2016/Shanghai	Learning Cross-lingual Information with Multilingual BLSTM for Speech Synthesis of Low-resource Languages	ICASSP	No	Yes	Yes	Yes
28	03/2016/Shanghai	Low Level Descriptors based DBLSTM Bottleneck Feature for Speech Driven Talking Avatar	ICASSP	No	Yes	Yes	Yes

29	06/2016/Seattle	Recognizing Stances in Mandarin Social Ideological Debates with Text and Acoustic Features	ICME	No	Yes	Yes	Yes
30	09/2016/San Francisco	Analysis on Gated Recurrent Unit based Question Detection Approach	INTERSPEECH	Yes	Yes	Yes	Yes
31	09/2016/San Francisco	Combining CNN and BLSTM to Extract Textual and Acoustic Features for Recognizing Stances in Mandarin Ideological Debate Competition	INTERSPEECH	No	Yes	Yes	Yes
32	09/2016/San Francisco	Expressive Speech Driven Talking Avatar Synthesis with DBLSTM using Limited Amount of Emotional Bimodal Data	INTERSPEECH	No	Yes	Yes	Yes
33	09/2016/San Francisco	Phoneme Embedding and its Application to Speech Driven Talking Avatar Synthesis	INTERSPEECH	No	Yes	Yes	Yes
34	10/2016/Tianjin	DBLSTM-based Multi-Task Learning for Pitch Transformation in Voice Conversion	ISCSLP	Yes	Yes	Yes	Yes

10. Student(s) trained *(Please attach a copy of the title page of the thesis.)*

Name	Degree registered for	Date of registration	Date of graduation
Lifa Sun	PhD	1 August 2013	16 November 2017
Xixin Wu	PhD	1 August 2015	30 December 2019
Songxiang Liu	PhD	1 August 2017	Not yet

11. Other impact N.A. *(e.g. award of patents or prizes, collaboration with other research institutions, technology transfer, etc.)*

N.A.

12. Statistics on Research Outputs *(Please ensure the summary statistics below are consistent with the information presented in other parts of this report.)*

	Peer-reviewed journal publications	Conference papers	Scholarly books, monographs and chapters	Patents awarded	Other research outputs (Please specify)
No. of outputs arising directly from this research project [or conference]	2 (under review)	34	N.A.	N.A.	N.A.