Comic-Guided Speech Generation (Supplementary Material)

YUJIA WANG, Beijing Institute of Technology
WENGUAN WANG, Inception Institute of Artificial Intelligence & Beijing Institute of Technology
WEI LIANG*, Beijing Institute of Technology
LAP-FAI YU, George Mason University

In this supplementary, we provide: **1)** detailed character identity inference evaluations of four different comic pages in *Comic Element Relations Analysis*; **2)** quantitative comparison results on different CNN architectures for comic character personal attributes (*i.e.*, gender and age) recognition in *Comic Character Attribute Analysis*; **3)** an ablation study of the number of retrieved reference speeches used; **4)** quantitative comparison results on different speech recognition models used in *Synthesizing Acoustic Parameters for Characters*; **5)** detailed statistical results supplemented for *Comparison* in perceptual studies.

ACM Reference Format:
Yujia Wang, Wenguan Wang, Wei Liang, and Lap-Fai Yu. 2019. Comic-Guided Speech Generation (Supplementary Material). *ACM Trans. Graph.* 38, 6, Article 187 (November 2019), 3 pages. https://doi.org/10.1145/3355089.3356487

1 EVALUATION OF IDENTITY ANALYSIS IN *COMIC ELEMENT RELATIONS ANALYSIS*

We conducted experiments of comic character identity inference on four different comic pages. These pages cover different themes and genres, including American hero comics, Japanese detective manga, real-life comics, and educational comics for child. We randomly selected 2k images in total, containing 4.8k character faces. In addition, the numbers of faces of different characters are unbalanced.

The performance of character ID inference (clustering) is measured in accuracy computed from a confusion matrix (as shown in Fig. 1), which is derived from the match between the cluster labels of all comic character faces and ground truth identities. The average accuracies (%) of the four comics are: (a) 90.40 ± 4.13, (b) 90.60 ± 3.21, (c) 92.80 ± 4.15, and (d) 89.60 ± 4.04 .

2 CNN ARCHITECTURE COMPARISON IN *COMIC CHARACTER ATTRIBUTE ANALYSIS*

In this section we compare the performance of our model used in *Comic Character Attribute Analysis* with different popular backbone CNN architectures, *i.e.*, AlexNet, VggNet, and ResNet-152. To this end, we modify these networks as two-branch architecture,

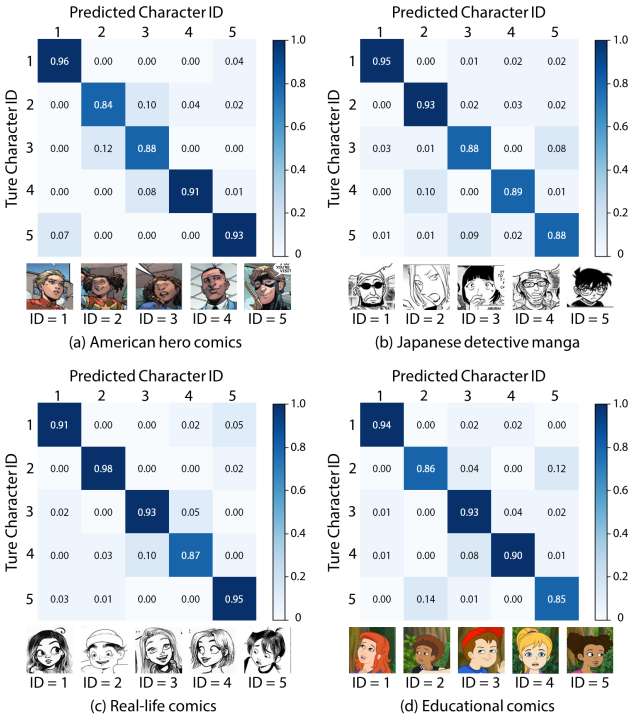

Fig. 1. Confusion matrices of comic character identity analysis. The characters are from four different comic pages covering different theme and genres ((a), ©Jim Mccann, Diego Olortegui, In-Hyuk Lee / Marvel Comics Inc.; (b), ©Gosho Aoyama / VIZ Media LLC (English); (c), ©Cassandra Calin; (d), ©Joanna Cole, Bruce Degen / Netflix Inc.).

Table 1. Testing accuracies of different CNN architectures for comic visual identity recognition. In each cell, the number refers to the testing accuracy on the corresponding dataset.

CNN Architecture	Gender	Age
AlexNet	80.21%	73.36%
VGGNet	87.54%	83.69%
ResNet-152	99.02%	95.33%
Ours	**99.23%**	**98.46%**

which is the same as our model, and apply standard cross-entropy classification losses to train the networks.

We trained these networks on Manga109 [Ogawa et al. 2018], respectively. In total, there are about 68k panels and 96k comic character faces. We further annotated each character face with character personal attributes: gender (*female* and *male*) and age (*child*, *adult*, and *senior*). We randomly split the dataset of comic character faces into training (76k images) and testing set (20k images). Moreover, these networks were trained using Adam optimizer with a batch

* Corresponding author: Wei Liang (liangwei@bit.edu.cn).
Authors' addresses: Yujia Wang, Beijing Institute of Technology, wangyujia@bit.edu.cn; Wenguan Wang, Inception Institute of Artificial Intelligence & Beijing Institute of Technology, wenguanwang.ai@gmail.com; Wei Liang, Beijing Institute of Technology, liangwei@bit.edu.cn; Lap-Fai Yu, George Mason University, craigyu@gmu.edu.

0730-0301/2019/11-ART187 $15.00
https://doi.org/10.1145/3355089.3356487

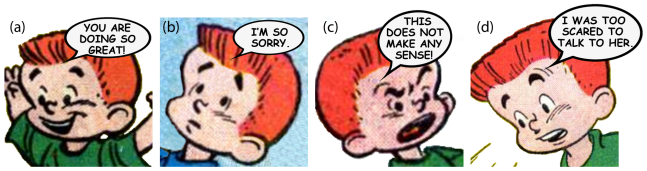

Fig. 2. The same comic character from "Clubhouse Rascals" (in the public domain) used in the ablation study of the number of retrieved reference speeches used, showing (a) a happy face; (b) a sad face; (c) an angry face; and (d) a fearful face, respectively.

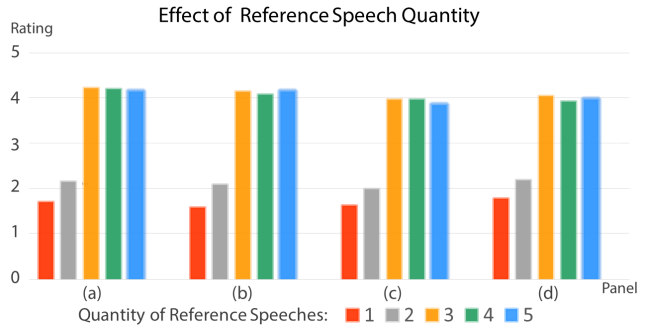

Fig. 3. The performance of our method in the test of different numbers of reference speeches used in *Emotion-aware Comic Speech Synthesis*.

size of 128 examples and a momentum of 0.9. The learning rate was set as 0.01 and the weight decay was 0.0005.

Table. 1 presents the recognition accuracy of each network on the testing set. It shows that for both gender and age recognition tasks, our model performs better than other CNN architectures, achieving the accuracy of 99.23% on gender recognition and 98.46% on age recognition.

3 ABLATION STUDY OF THE NUMBER OF RETRIEVED REFERENCE SPEECHES

In the *Emotion-aware Comic Speech Synthesis* process, our approach automatically finds 3-best support reference speeches to be an intermediate representation of the large range of affective acoustic factors. In this section, we conduct an ablation study to investigate the influence of different numbers of retrieved reference speeches used on the quality of the synthesis results and to verify the adequacy of using three retrieved reference speeches.

Procedure. We selected 4 comic panels with the same character from "Clubhouse Rascals" (in the public domain), showing different emotional faces (as shown in Fig. 2) to conduct the experiment. Each comic penal and the corresponding speeches were displayed to the recruited 30 participants, who were required to rate the emotion consistency between the comic (combined facial expression and speech text) and speech. Ratings range from 1 to 5, with 1 representing the inconsistent that speech does not show the consistency emotion with the comic page, while 5 representing the very consistent. Note that, the order of each paired comic panel and speech was displayed randomly, *i.e.*, participants will view the panel for rating and then listen to the speech for rating, or vice versa.

Table 2. Ablation study on comics (shown in Fig. 2). The test is to investigate how the quality of the synthesized speeches is influenced by the number of retrieved reference speeches used.

Quantity	Scores			
	a	b	c	d
1 Ref.	1.73 ± 0.33	1.60 ± 0.26	1.62 ± 0.30	1.81 ± 0.44
2 Ref.	2.23 ± 0.67	2.08 ± 0.66	1.99 ± 0.57	2.25 ± 0.71
3 Ref.	**4.20 ± 1.03**	**4.18 ± 1.06**	**3.99 ± 0.94**	**4.05 ± 0.85**
4 Ref.	4.18 ± 0.88	4.07 ± 0.81	3.99 ± 0.83	3.95 ± 0.80
5 Ref.	4.10 ± 0.80	4.19 ± 0.81	3.85 ± 0.84	4.02 ± 0.86

Table 3. Testing accuracies of different methods for speech identity and authenticity recognition. In each cell, the number refers to the testing accuracy on the corresponding dataset.

Methods	Gender	Age	Authentic
SVM	75.37%	74.01%	78.25%
Our model	**92.45%**	**88.76%**	**89.00%**

Outcome and Analysis. As summarised in Fig. 3, by referring with multiple reference speech clips (baseline: one reference speech clip), we observe obvious performance improvement (*i.e.*, ratings at 3 reference speech clips) and eventual stabilization (*i.e.*, ratings at 4 and 5 reference speech clips). The trend is consistent across all comics. Based on such observation, we retrieve three reference speeches, whose mean rating is 2.62 times higher than the baseline. Table. 2 illustrates the detailed statistical results of Fig. 3.

4 MODEL COMPARISON IN *SYNTHESIZING ACOUSTIC PARAMETERS FOR CHARACTERS*

In this section we compare the performance of our model used in *Synthesizing Acoustic Parameters for Characters* for speech identity and authenticity recognition with traditional SVM classifiers. We created the Speech Corpus Dataset, which contains 13.7k real and 4.5k distorted speech clips from different speech corpus, such as [Veaux et al. 2016], and from the Internet. The dataset was used in this experiment to verify the effectiveness of our model used in *Speech Identity and Authenticity Recognition*. Each speech in the dataset lasts at least 2 seconds. The real speech clips were annotated with gender (*female* and *male*) and age (*child*, *adult*, and *senior*) labels. We randomly selected 1.7k real speeches and 0.5k distorted ones for testing, and the rest was used for training.

We first extracted MFCC coefficients from the speech clips in our collected Dataset, similar with [Suwajanakorn et al. 2017], and obtained a 13-d feature vector. Two SVMs for the gender and age recognition are trained on real speeches (12k) and one SVM for the speech-authentic classification is trained with all the 16k training speeches.

Table. 3 presents the results on the testing set. It shows that for all recognition tasks, our model performed better than the traditional SVM models.

5 COMPARISON

In this section, we include the evaluation results on how well our approach narrates comics through comparing to *Professional Narration* approach and *Manual Synthesis* approach. Our approach and the other two approaches were applied to narrate a comic page of

Table 4. Means and standard deviations of the three approaches compared, *i.e.*, our approach, *Professional Narration*, and *Manual Synthesis*.

Panels	Comparison Approaches					
	Our		Professional		Manual	
	M	SD	M	SD	M	SD
1	4.67	0.48	4.87	0.35	1.27	0.45
2	4.77	0.43	4.93	0.25	1.10	0.31
3	4.80	0.41	4.83	0.38	1.17	0.38
4	3.73	0.64	4.97	0.18	1.30	0.54
5	4.93	0.25	4.97	0.18	1.03	0.18
6	4.87	0.35	4.87	0.35	1.43	0.73
7	4.00	0.64	4.23	0.74	1.53	0.82
8	4.57	0.50	4.80	0.41	2.00	0.83

Fig. 4. Comic page ("The Magic School Bus Rides Again", ©Joanna Cole, Bruce Degen / Netflix Inc.) used for comparing the performance of different narrations in terms of different approach, *i.e.*, our approach, *Professional Narration*, and *Manual Synthesis*.

"The Magic School Bus Rides Again" shown in Fig. 4, which contains 8 panels and 6 characters. The recruited 30 participants were shown the comic page as well as the synthesized speeches panel by panel. The speeches were randomly selected from the results of our approach, *Professional Narration*, and *Manual Synthesis* so as to avoid bias. After that, the participants were asked to rate the consistency between the comic page and the synthesized speeches. The range of the rating was 1 to 5, referring "very inconsistent" to "very consistent".

Outcome and Analysis. The statistics of the average consistency score and standard deviations on each panel (1 to 8) is shown in Table 4. The *Professional Narration* received the highest score ($M = 4.85, SD = 0.39$) on average, followed by our approach ($M = 4.14, SD = 0.73$) and *Manual Synthesis* ($M = 1.25, SD = 0.51$).

The results of the comparison on each panel show that the consistency scores of our approach and of *Professional Narration* are mostly comparable. The scores of our approach were lower than the score of *Professional Narration* from 0.7% (on panel 3 and 5) to 6% (on panel 3), except for the scores on panel 4, on which the score of our approach was lower than the one of *Professional Narration* by about 24.8% (4.97 and 3.73).

REFERENCES

Toru Ogawa, Atsushi Otsubo, Rei Narita, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. Object Detection for Comics using Manga109 Annotations. *CoRR* abs/1803.08670 (2018).

Supasorn Suwajanakorn, Steven M. Seitz, Ira Kemelmacher-Shlizerman, Supasorn Suwajanakorn, Steven M. Seitz, Ira Kemelmacher-Shlizerman, Supasorn Suwajanakorn, Steven M. Seitz, Ira Kemelmacher-Shlizerman, and Supasorn Suwajanakorn. 2017. Synthesizing Obama: learning lip sync from audio. *TOG* 36, 4 (2017), 1–13.

Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. 2016. SUPERSEDED-CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit. (2016).