# Comic-Guided Speech Synthesis

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

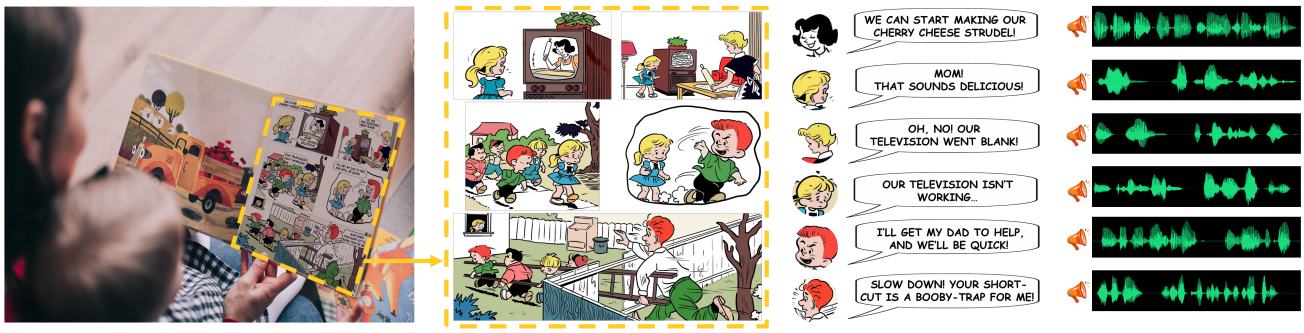


Fig. 1. Comic in, speech out: our approach automatically narrates the input comic page of "Clubhouse Rascals" (in the public domain) by synthesizing realistic speeches for each character. The synthesized speeches carry the identity properties (gender and age) and emotion conditions (*e.g.*, happy, sad, angry) of the characters inferred from the comic input.

We introduce a novel approach for synthesizing realistic speeches for comics. Using a comic page as input, our approach synthesizes speeches for each comic character following the reading flow. It adopts a cascading strategy to synthesize speeches in two stages: Comic Visual Analysis and Comic Speech Synthesis. In the first stage, the input comic page is analyzed to identify the gender and age of the characters, as well as texts each character speaks and corresponding emotion. Guided by this analysis, in the second stage, our approach synthesizes realistic speeches for each character, which are consistent with the visual observations. Our experiments show that the proposed approach can synthesize realistic and lively speeches for different types of comics. Perceptual studies performed on the synthesis results of multiple sample comics validate the efficacy of our approach.

CCS Concepts: • **Computing methodologies** → **Computer graphics**; **Image processing**; **Perception**.

Additional Key Words and Phrases: comics, speech synthesis, deep learning

**ACM Reference Format:**
Yujia Wang, Wenguan Wang, Wei Liang, and Lap-Fai Yu. 2019. Comic-Guided Speech Synthesis. *ACM Trans. Graph.* 38, 6, Article 187 (November 2019), 14 pages. https://doi.org/10.1145/3355089.3356487

* 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

## 1 INTRODUCTION

Comics, which use a combination of drawings and text balloons for storytelling, are a highly popular media for entertainment, advertising, and education worldwide. In addition to comics published in comic books and magazines, which are typically created by professional artists, there is a growing, vast volume of webcomics created by amateur artists. With the advancement of new media technologies and devices, the format of comics has also begun evolving. For example, nowadays, people can easily read and share comics through e-readers, such as Amazon Kindle and Sony Reader. There are also new augmented reality (AR) applications that bring traditional comic books to life by playing sounds and animations overlaid on top of comic images via a mobile phone or an AR headset [Lai et al. 2015].

To heighten the enjoyment brought by comics, a recent trend is to create audio comics akin to audiobooks, which have gained widespread popularity. To create audio comics, current practices [Rieman 2016] commonly resort to hiring professional narrators to read comic books page by page. Depending on the comic characters to be narrated, a professional narrator with a matching voice is hired. If the comics involve multiple characters with different personal attributes, such as age, gender and personality, several professional narrators may be hired to collaborate, each playing one or more characters.

During the narration process, the narrators need to pay close attention to the emotions of the comic 'speaker' by taking into account the facial expressions and texts associated with the character. To create an engaging, high-quality narration, the narrators need to control the tone and emotion of their speeches. Delivering such a skilled narration could be a daunting task, requiring significant expertise and hands-on experience. Narrating a comic book usually

involves hiring several professional narrators to read the book several hours a day, over multiple weeks. Consequently, audio comics are laborious, time-consuming, and costly to produce.

We propose a computational approach to facilitate and automate the creation of audio comics. Fig. 1 illustrates our approach. Given a comic page containing different characters and texts, our approach automatically narrates the characters with voices consistent with the visual content, driven by a visual analysis of the comic. The analysis results can be explicitly expressed in terms of the identity, gender, age, and emotion of different comic characters, identified across different panels. Our approach can potentially be applied to enable the scalable production of audio comics.

Synthesizing realistic speeches for comic characters presents a technical challenge because of the non-trivial mapping from the input vision and text domain to the output audio domain. To overcome this challenge, we devise a cascaded comic-guided speech synthesis approach comprising a *Comic Visual Analysis* stage and a *Comic Speech Synthesis* stage. In the visual analysis stage, our approach detects different comic elements (*i.e.*, panels, balloons, tails, texts, and characters) based on which it infers the identity and personal attributes (gender and age) of the speaking characters in each panel, as well as their emotion when speaking each text. In the speech synthesis stage, guided by the visual analysis results, our approach prompts each character to speak with a tone of voice that matches his/her personal attributes, as well as with an emotion consistent with the visual and textual context.

The proposed approach gives rise to a number of practical applications. For example, by narrating existing comic books with realistic speeches, old comic books can be revived as they are paired with new reading experiences. Additionally, narrating comic books in different languages (*e.g.*, narrating a manga with English) could also create a new means for foreign language and culture learning.

The main contributions of this paper are the following:

- We introduce a novel topic area of synthesizing speeches guided by the visual content of comics.
- We advise a computational approach for synthesizing speech, driven by comic visual analysis. The synthesized speeches take into account the inferred personal attributes and emotion states of the comic characters.
- We demonstrate the proposed approach for different practical applications and validate its effectiveness through perceptual studies.

## 2 RELATED WORK

### 2.1 Audiobook Narration

Stimulated by new forms of media, the audiobook market is growing rapidly. A lot of efforts have been dedicated to converting popular comics, e-books, newspapers, and teaching material into audiobooks.

Narrating an audiobook is a performance like putting on a play. It requires significant human effort. First, the producer has to come up with the voices for each character. Typically, to create an audiobook, narrators need to read collaboratively for several hours a day, over multiple weeks. The cost of professional narrators is high; narrating an audiobook usually requires several hundred dollars per hour. In contrast, if a layman wants to produce an audiobook, he/she may need to spend a significant amount of time on speech recording and post-production of the recorded audio. Please refer to Have and Pedersen [2013] for details on the recent development and widespread adoption of audiobooks, and the common processes used in industry for producing them.

To overcome the challenges of creating audio comics, we explore the possibility of devising a computational approach for synthesizing realistic speeches based on the visual content of comics.

### 2.2 Face Perception and Voice

Human faces, as well as human voices, reveal lots of information to a perceiver, *e.g.*, gender, age, and emotion [Belin et al. 2011; Bruce and Young 1986; Lass et al. 1976; McAleer et al. 2014]. Many social interactions involve inferring information from both the faces and voices of people [Campanella and Belin 2007; Kamachi et al. 2003].

In addition to observing facial expressions, people perceive emotions through speech content and voice [Trampe et al. 2015]. This process enables us to anticipate other people's behaviours and reactions. In the same way, understanding the emotions of a character in a comic enables the readers to understand and anticipate their actions and thoughts [Miall 1989].

In computer vision, natural language processing, and speech processing, the attributes (gender, age, and emotion) recognition problem is widely studied [Qi et al. 2018; Wang et al. 2018a,c]. Many deep learning methods have been proposed for facial attributes recognition [Liu et al. 2015; Rudd et al. 2016], voice attributes recognition [Fayek et al. 2017; Wang and Tashev 2017], and text emotion recognition [Kratzwald et al. 2018].

Inspired by these studies, we incorporate deep learning techniques to automatically infer the gender, age, and emotion of characters by performing visual analysis on comic pages, which guides the synthesis of realistic speeches for the characters.

### 2.3 Speech Synthesis and Processing

*Audiovisual Consistency.* Computer graphics researchers have worked on enhancing the audiovisual realism of virtual characters' motion. For example, different techniques have been proposed to synthesize high-fidelity facial animation based on input text or speech by synchronizing lip motion and speeches [Hu et al. 2017; Suwajanakorn et al. 2017], or by mixing and matching body motion, face motion and speeches for avatars [Ondřej et al. 2016]. Research progress in modelling and animating digital faces and avatars [Cao et al. 2016; Hu et al. 2017; Mcdonnell et al. 2009] has greatly facilitated the development of human-computer interfaces and virtual world applications which feature natural interactions.

*Speech Synthesis.* Speech synthesis refers to synthesizing human-like speeches, such as natural audio narration [Finkelstein et al. 2017]. We review relevant technologies of Text-to-Speech (TTS).

TTS synthesis provides a complete, end-to-end account of the speech synthesis process for converting a piece of normal language text into a speech. Several techniques have gained popularity in this field, such as concatenative synthesis [Hunt and Black 1996]; and parametric speech synthesis like WaveNet [van den Oord et al. 2016], SampleRNN [Mehri et al. 2017], Char2Wav [Sotelo et al. 2017], and

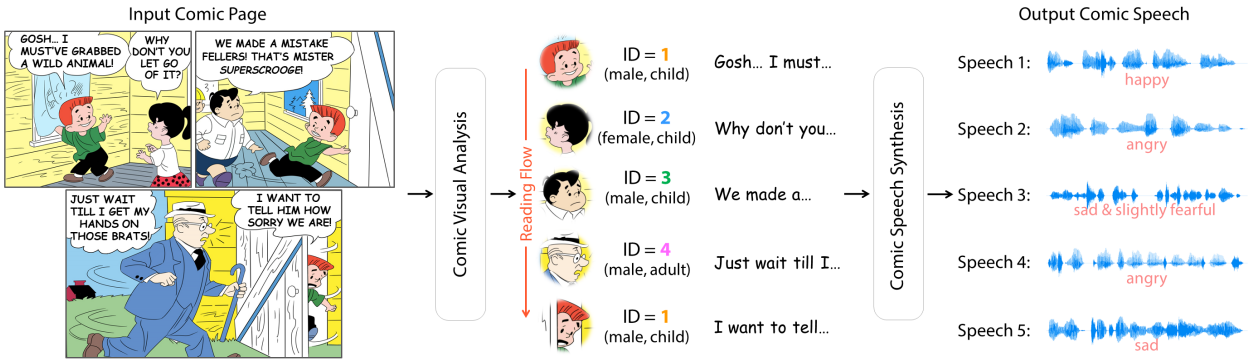


Fig. 2. Overview of our proposed approach for comic-guided speech synthesis. It consists of two major steps: *Comic Visual Analysis* and *Comic Speech Synthesis*. In the first step, our approach analyzes the input comic page of "Clubhouse Rascals" (in the public domain) to obtain information on each character (*e.g.*, identity, gender, age, and speech content). Guided by such information, our approach synthesizes realistic speeches for each character in the second step, following the reading flow.

Tacotron [Shen et al. 2018; Wang et al. 2017]. In addition, conditioning TTS approaches on emotions leads to more realistic synthesized speeches [Lee et al. 2017; Li et al. 2018]. Such conditioning can be achieved by prosody transfer [Skerry-Ryan et al. 2018; Wang et al. 2018b] or voice conversion [Mohammadi and Kain 2017].

*Speech Processing.* Audio processing tools such as Adobe Audition allow experts to edit waveforms via a timeline interface. Such general-purpose tools typically do not have a specific interface for speech synthesis, as they are unaware of the linguistic properties of the audio signal. Research-oriented tools such as Praat [Boersma 2002], STRAIGHT [Kawahara et al. 2008], and WORLD [Morise et al. 2016] allow a user to manipulate the phonetic and linguistic aspects of a speech, where audio is presented synchronously with other properties such as phonemes. Such tools are commonly used to achieve more complex tasks like speech synthesis.

One aspect of speech synthesis and processing that has received little attention, is associating the synthesized speeches with the input text as well as the facial attributes observed on the speaker's image. Achieving such an association, which is the main goal of our approach, is useful for a wide range of image-to-speech applications, such as digital comics, educational readings, and augmented reality.

### 2.4 Computational Tools for Comics Production

Computer graphics researchers have invented computational tools to make comic production accessible to general users. A number of techniques have been developed to facilitate the comic production process [Augereau et al. 2018; Dunst et al. 2018], such as manga face detection [Chu and Li 2019; Stricker et al. 2018], colourization [Qu et al. 2006], screening [Qu et al. 2008], layout generation [Cao et al. 2012], elements composition [Cao et al. 2014], structural lines cleanup and extraction [Li et al. 2017; Simo-Serra et al. 2016], and hand-colored cartoon animation [Dvorožnák et al. 2018]. Complementary to the existing works, we propose a novel approach for synthesizing realistic speeches for narrating comic books.

## 3 APPROACH OVERVIEW

Fig. 2 shows an overview of our approach. Given a digitized comic page as input, our approach aims to synthesize speeches consistent with the visual observations to narrate the page in a lively manner.

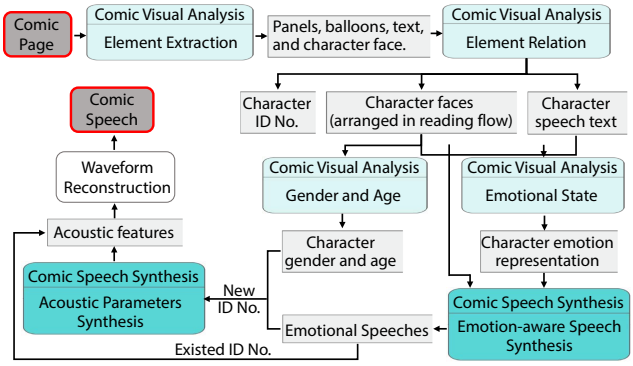


Fig. 3. Dataflow diagram of our proposed approach. The input and final output are highlighted in red rectangles. Essential modules and intermediate results are in green and gray colors, respectively.

Our approach proceeds in two phases. Fig. 3 illustrates the detailed dataflow. In the *Comic Visual Analysis* phase, our approach first performs visual analysis on the comic page to extract basic comic elements (*i.e.*, panels, speech balloons, text, and characters). It rearranges the extracted elements in the narrative order, associating the speech balloons with the comic characters. It also recognizes the identities and personal attributes (gender and age) of the characters, as well as inferring their emotional states in speaking each piece of text. Such an analysis leads to a comprehensive understanding of the content of the input comic. Our analysis also incorporates domain knowledge of comics. For instance, based on the knowledge that speech balloons, the containers of speech texts, typically have a tail pointing at their respective speakers [Rigaud et al. 2015], our approach associates speech balloons with the corresponding speaking characters.

To sound realistic, the synthesized speeches need to convey both the identity (*i.e.*, gender, age) and emotional states of the speakers. In the *Comic Speech Synthesis* phase, our approach synthesizes realistic comic speeches in two steps. First, for each piece of text in the balloon, our approach synthesizes an expressive speech that delivers the emotion (*e.g.*, happy, sad) inferred from the textual content and corresponding character's face. To maintain a consistent identity

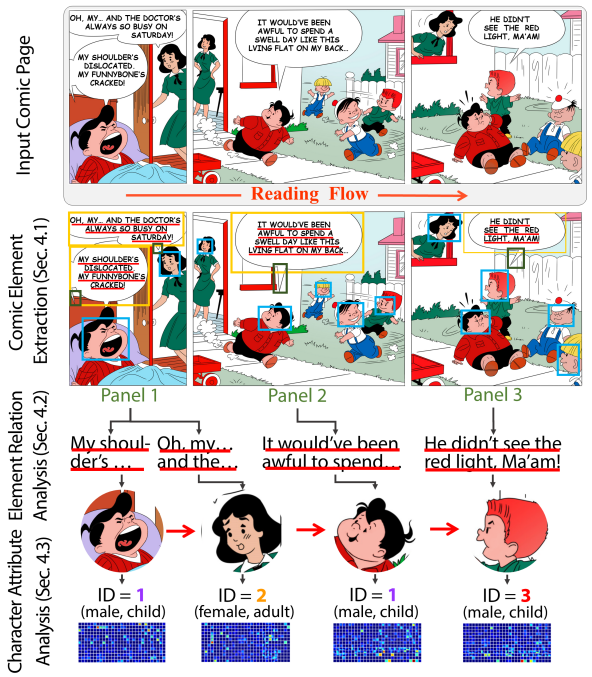


Fig. 4. *Comic Visual Analysis* for an input comic page from "Clubhouse Rascals" (in the public domain). Based on the relation analysis between comic elements (*i.e.*, panels, characters, speech balloons and tails, and text), we further infer multiple attributes for each character (*i.e.*, identity, gender, age, and emotion representation).

and speaking style for the same character appearing across different comic panels, our approach uniformly optimizes the acoustic parameters using all the speeches of that character (according to his/her personal attributes obtained in the *Comic Visual Analysis* phase). By applying our approach, the comic page is automatically narrated with realistic speeches consistent with the visual content.

## 4 COMIC VISUAL ANALYSIS

### 4.1 Comic Element Extraction

Our approach begins by pre-processing the comic page to extract comic elements with the following steps, using off-the-shelf techniques, as shown in Fig. 4:

(1) Apply the method proposed by Pang et al. [2014] to extract **panels** from the input comic page.
(2) Apply a balloon contour analysis method [Rigaud et al. 2015] to detect **speech balloons** and their **tail tips**, which are usually represented by a discontinuity on the balloons' contours;
(3) Apply the methods of [Ma et al. 2018; Shi et al. 2018] to extract and recognize **texts** from each speech balloon;
(4) Use an annotated *Manga109* dataset [Matsui et al. 2017; Ogawa et al. 2018] to fine-tune the deep network, MaskRCNN [Abdulla 2017], to detect **character faces** in each panel.

We conducted experiments to evaluate each step in *Comic Element Extraction*, *i.e.*, panel extraction, speech balloon detection, text extraction and recognition, and character face detection.

Table 1. Accuracies of different comic character face detection approaches, each of which is tested on the same comic panels.

| CNN Architecture | Gender |
|---|---|
| Nguyen *et al.* [2017] | 81.4% |
| Qin *et al.* [2017] | 90.4% |
| Chu *et al.* [2017] | 89.9% |
| Our Approach | **91.2%** |

*Dataset.* We adopted two different datasets: Manga109 [Matsui et al. 2017; Ogawa et al. 2018], which consists of 109 comic books with around 21.1k pages drawn by professional manga artists, for evaluating the performance of panel extraction, speech balloon detection, and character face detection; and Synth90k [Gupta et al. 2016], for evaluating the performance of text extraction and recognition. Synth90k contains 9 million images generated from a set of 90k common English words, which are rendered onto natural images with random transformations and effects. Every image is annotated with a groundtruth word.

*Panel Extraction.* The method proposed by Pang *et al.* [2014] comprises three steps, *i.e.*, panel block generation, panel block splitting and panel shape extraction. We test the approach with 1k randomly selected comic pages from Manga109. The panel is deemed to be successfully segmented if the overlap ratio of the detected region and ground-truth is over 90%. We achieve average accuracy of 92.5%.

*Speech Balloon Detection.* The approach [Rigaud et al. 2015] processes low-level information without relying on the text extractor's performance, thus avoiding error propagation. We evaluate the approach with 1k closed speech balloons (balloons with a fully connected outline), 94.7% of which are successfully detected.

*Text Extraction and Recognition.* We assess the performance of the trained text extraction model [Ma et al. 2018] and text recognition model [Shi et al. 2018] on Synth90k dataset. For text extraction, we obtain a precision of 94.3%, recall of 88.7%, and F-measure of 91.2%. The average accuracy of text recognition for different word lengths is 93.6% on the randomly selected 1k images.

*Character Face Detection.* If the overlap ratio of the detected region and ground-truth is over 80%, the face is considered successfully detected. The fine-tuned MaskRCNN [Abdulla 2017] achieves a precision of 91.2% on 1k randomly selected comic panels from Manga109. The experimental results of our method compared with those of deep neural networks are given in Table 1. Our fine-tuned network achieves the best performance compared to other deep learning approaches.

### 4.2 Comic Element Relations Analysis

Based on the extracted comic elements, our approach analyzes the relations between them. It deduces the chronological order of the panels by considering the reading flow. It also associates each speech balloon with its speaking character. As depicted in Fig. 4, our approach deduces which character is speaking for each piece of text detected; it also deduces whether the child in Panel 1 and 3 are the same child (ID No. is 2).

*4.2.1 Reading Flow.* Based on the comic design principles [Petersen 2011], our approach determines the chronological orders of the

detected panels, as well as of the speech balloons and characters that are present in each panel. By default, our approach uses the reading flow convention of American comics, which are read from left to right and top to bottom. If needed, an alternative reading flow convention, such as that of Japanese manga, which are read from right to left and from top to bottom, can be applied instead.

*4.2.2 Speaking Character.* Our approach then associates the detected speech balloons in each panel with their speaking characters. A speaking character is defined as the character pointed by the tail tip (extremity) of a speech balloon that contains a dialogue [Varnum and Gibbons 2007]. The tail direction is given by the line fitting algorithm. Akin to [Rigaud et al. 2015], we define the tail direction as the direction of the vector which points from the closest point of the fitted straight line to the tail root towards the tail tip (as shown in Fig. 5). Subsequently, we associate each detected speech balloon with a speaking character, who is taken as the first character encountered along the speech balloon's tail direction.

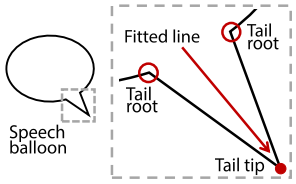

Fig. 5. Illustration of tail direction.

As shown in Fig. 4, by analyzing the comic element relations, our approach identifies all speaking characters and their corresponding speech text on the input comic page, following the reading flow.

*4.2.3 Identity Analysis.* It is common in comics that a character appears in multiple panels. Inferring the identity of each speaking character is essential to ensure that the voice of a character is consistent across different panels. In order to handle the large facial appearance variations of a character caused by different poses and facial expressions, we first use a deep metric learning (DML) approach to learn a facial feature representation. Then, the facial features are clustered in an unsupervised way. Faces, clustered in the same category, are regarded as coming from one character.

*Dataset.* We leverage a large-scale dataset of different themes manga images, *Manga109* [Matsui et al. 2017; Ogawa et al. 2018], which consists of 109 comic books of around 21.1k pages drawn by professional manga artists. In total, there are about 68k panels and 96k comic character faces. Some of the comic pages could pose great challenges for character identification. For example, some pages contain many characters showing non-frontal or occluded faces. Some pages may show a dark scene presented by different screentones or colors.

*Feature Representation.* In order to extract effective features from comic characters' faces, we fine-tune the GoogLeNet using the DML approach. To guide the training process, we constructed a training set $\{(I_a, I_b, l)\}$, where $l$ is the identity label of the paired face images (*i.e.*, $I_a$ and $I_b$):

$$l(I_a, I_b) = \begin{cases} 1 & I_a \text{ and } I_b \text{ refer to the same identity,} \\ -1 & I_a \text{ and } I_b \text{ refer to different identities.} \end{cases} \quad (1)$$

We leverage the contrastive loss [Hadsell et al. 2006] to capture the relationship between pairwise data, *i.e.*, similar or dissimilar.

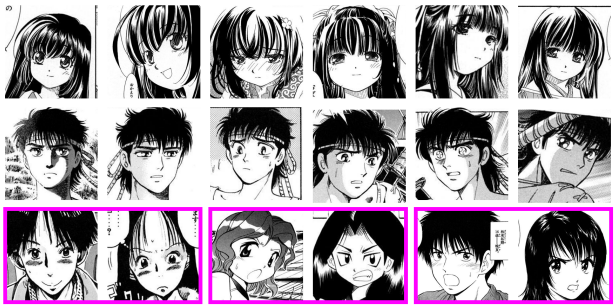
Fig. 6. Results of ID clustering for characters in *Manga109* [Matsui et al. 2017; Ogawa et al. 2018]. Each of the first two rows shows face images from the same character that are clustered successfully by our method. The bottom row shows some failure cases, where each pink rectangle contains a pair of face images with similar appearance which are incorrectly clustered. ©Sakurano Minene, Shimazaki Yuzuru, Taka Tsukasa, Ki Takashi, Shirai sanjirou, Inokuma Shinobu

The training process learns the network weights $\theta$ by minimizing a loss function, which is defined as:

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{i=1}^{P} l\mathcal{L}_1 + (1-l)\mathcal{L}_2, \\ \mathcal{L}_1 &= ||G_\theta(I_a)^{(i)} - G_\theta(I_b)^{(i)}||, \\ \mathcal{L}_2 &= \max(0, \rho - ||G_\theta(I_a)^{(i)} - G_\theta(I_b)^{(i)}||). \end{aligned} \quad (2)$$

$G_\theta(I)$ is the mapped features of an input face $I$. $\rho > 0$ is a constant and set as 2. By minimizing the loss function $\mathcal{L}(\theta)$, the distance between the mapped features of $I_a$ and $I_b$ is driven by $\mathcal{L}_1$ to be small if $I_a$ and $I_b$ correspond to the same character, and is driven by $\mathcal{L}_2$ to be large vice versa.

After the training process, we replace the final fully connected layer with a $L_1$ normalization layer to obtain deep feature vectors.

*Inferring Character ID..* Each comic character face is represented as a 1,024-d feature vector and all feature vectors are clustered by an affinity propagation (AP) approach [Frey and Dueck 2007] using the Euclidean distance. Specifically, we set the examplar preference as the average similarity values among all pairs of sequences and use $\lambda = 0.5$ as the damping factor.

*Evaluation.* We conducted experiments on different comic pages, such as American hero comics, Japanese detective manga, etc. We randomly selected 50 chapters from each corresponding comic book, which contain around 400 pages, 2k panels, and 4.8k character faces.

The performance of character ID inference is measured in terms of accuracy computed from a confusion matrix, which is derived from the match between the cluster labels of all comic character faces and ground truth identities. The average accuracy and standard deviation over the four comics are 90.67% ± 1.22%. Please see the supplementary material for more details about the confusion matrices and the accuracy of each comic.

Fig. 6 shows some identity analysis results of the comic characters in *Manga109*, where each row except for the bottom one denotes a cluster. It can be observed that each cluster (each row) covers faces of the same character with different head poses, expressions, occlusions, and scene conditions.

### 4.3 Comic Character Attribute Analysis

Next, our approach infers the gender, age, and emotional states of each speaking character, so as to enable it to narrate the character with a realistic tone of voice and proper emotional states, which are important for lively and engaging speeches, as demonstrated in [Belin et al. 2004].

*4.3.1 Gender and Age Analysis.* The gender and age of a speaker are directly relevant to the structure of his/her vocal system, such as the vocal cords, and hence are related to the speaker's voice [Ghazanfar and Rendall 2008]. Therefore, to synthesize realistic comic speeches, we need the speeches to convey the characters'gender and age. To analyze a comic character's gender and age, we fine-tune GoogLeNet with two branches, one for gender classification and one for age classification. For a character that appears multiple times, his/her gender and age are voted on the inferred results for all faces of the character so as to improve the inference accuracy and mitigate the influence of varying head poses and occlusions.

*Dataset.* To train the network, we further annotate *Manga109* [Matsui et al. 2017; Ogawa et al. 2018] with gender (*female* and *male*) and age (*child*, *adult*, and *senior*). Following random selection, we split the dataset into a separate training set (19k comic images) and testing set (2.1k comic images).

*Learning.* For the purpose of training, the comic character face images in the training set are uniformly resized to $256 \times 256$ and fed to the network. Standard cross-entropy classification losses are applied to the two branches. The network is trained using the Adam optimizer with a batch size of 128 samples and a momentum of 0.9. The learning rate is set as 0.01 and the weight decay is set as 0.0005.

*Evaluation.* We test the performance of our model on the test set of *Manga109*. For testing on single panel, the average classification accuracy is 99.23% over all the gender categories, and 98.46% over all the age categories.

For comic pages, the average classification accuracy is 99.47% over gender categories, and 99.30% over age categories. In a comic page, if a character has several identity proposals, *i.e.*, the same character appears multiple times across different panels, the final category is voted on by all proposals. We can observe an accuracy improvement compared to testing on single images. We also compare the performance of our model against different famous backbone architectures, such as AlexNet, VggNet, and ResNet-152, and find that the GoogLeNet-based model achieves the best performance. Please refer to the supplementary material for more quantitative comparison results and implementation details.

*4.3.2 Emotional State Analysis.* Comics typically convey characters' emotions via their facial expressions and the speech texts. Analyzing the emotional states of characters helps synthesize vivid speeches, enriching reader's auditory experience.

To capture high-quality emotion information, we apply two top-performing, deep learning-based emotion recognition models, which work on images [Rudd et al. 2016] and texts [Felbo et al. 2017], respectively. We further annotated the face images of the *Manga109* dataset with emotions (*neutral*, *happy*, *sad*, *angry*, and *fearful*), to fine-tune the image emotion recognition model. The features extracted from their penultimate fully-connected layers are used as the visual and textual emotion representations, respectively. The visual emotion feature is a 1,024-d vector and textual emotion feature is a 2,304-d vector. They are concatenated to form a 3,328-d visual-textual emotion feature, which will be used in our emotion-aware comic speech synthesis (Sec. 5).

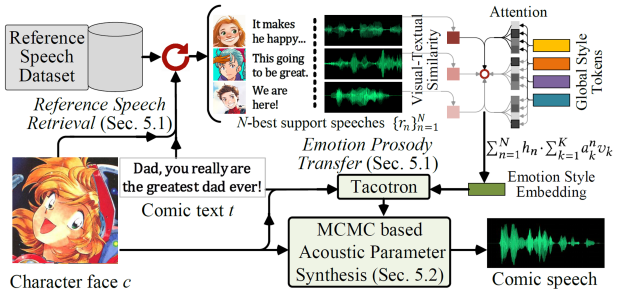


Fig. 7. Illustration of our *Comic Speech Synthesis* stage (Sec. 5), which first synthesizes expressive comic speeches through an emotion prosody transfer process, and then refines all the speeches of a same character through synthesizing acoustic parameters. ©Miyone Shi

## 5 COMIC SPEECH SYNTHESIS

With the personal attributes and emotional states recognized from the comic, next our approach synthesizes realistic and emotional speeches so as to yield realistic speaking styles for the characters. As demonstrated in [Belin et al. 2004], the human voice is the carrier of speech, but also an "auditory face" that conveys important identity and affective information (*i.e.*, emotional states). The affective information carried by speech is conveyed by affective prosody [Dimos et al. 2015], *i.e.*, qualities of a speech including stress, speed, loudness, and voice quality [Banse and Scherer 1996]. The identity information of voices is directly influenced by the personal attributes (*i.e.*, gender and age) of the speaker.

Inspired by these studies, our algorithm is designed to have two steps. First, it maps the balloons into speeches that carry the emotional states conveyed by the corresponding comic character's faces and texts (Sec. 5.1). Then, for each character, it synthesizes voice tone parameters based on his/her personal attributes. Such parameters are used to tune the emotional speeches to match the personal attributes of the character (Sec. 5.2).

### 5.1 Emotion-aware Comic Speech Synthesis

The emotion of a speech is strongly associated with several affective acoustic factors (*e.g.*, intonation, flow, and stress) and is difficult to model. Inspired by the studies of prosody transfer [Wang et al. 2018b], our comic speech synthesizer mimics the process of voice actors emotionally narrating comics, using a template learning framework. As depicted in Fig. 7, it proceeds by two main steps: searching for several reference speeches that carry the target emotion; and refining the emotion of a speech through conditioning on an emotion prosody transfer model using the reference speeches.

*5.1.1 Reference Speech Retrieval.* Through the use of a reference speech set that contains speeches covering a variety of affective

acoustic factors, our approach is capable of synthesizing emotionally-rich and expressive speeches.

*Dataset.* We created a *Reference Speech Dataset* from audio comics downloaded from the Internet, which contains 300 combinations of character faces, texts, and speeches. The dataset covers a variety of emotions (*e.g.*, neutral, happy, sad, angry, and fearful). Each speech lasts for at least 5 seconds. To extract emotion from images and texts in this dataset, we use the same method as in Sec. 4.3.

*Similarity Metric.* Given a comic character $c$ and text $t$ from a corresponding speech balloon, we retrieve $N$-best reference speeches $\{r_i\}_{i=1}^{N}$ from the *Reference Speech Dataset*. The similarity between $(c, t)$ and a reference speech $r_i$ is computed using the mean squared distance over their 3,328-d visual-textual emotion features (Sec. 4.3); the shorter the distance, the larger the similarity.

The *Reference Speech Dataset* serves as an intermediate representation of the large range of affective acoustic factors. In our implementation, we empirically set $N = 3$. Please refer to the supplementary material for examples of the retrieved speeches and an ablation study on the retrieved reference speech quantity $N$.

*5.1.2 Emotion Prosody Transfer.* As shown in Fig. 7, given the comic input $(c, t)$, our comic speech synthesizer seeks to synthesize an expressive speech through leveraging the emotional speaking styles of the retrieved reference speeches $\{r_i\}_{i=1}^{N}$. We achieve this by modifying an advanced emotion prosody transfer model called global style tokens (GST) [Wang et al. 2018b].

GST is devised upon a popular text-to-speech model, Tacotron [Shen et al. 2018; Wang et al. 2017], which predicts mel spectrograms directly from grapheme inputs. It extends the Tacotron with global style tokens, a bank of $K$ randomly initialized speaking style vectors $\{v_k\}_{k=1}^{K}$. An attention-based prosody learning strategy is used to model the style representation of a reference speech as the weighted sum of the global style tokens: $\sum_{k=1}^{K} a_k v_k$, where the combination weights $\{a_k \in [0, 1]\}_{k=1}^{K}$ are the learnable attentions.

Given a text input, Tacotron is conditioned on the reference style embedding and the comic face to synthesize the corresponding acoustic representations (spectrogram frames). The spectrogram frames are then converted to waveforms by a vocoder, *i.e.*, WaveRNN [Kalchbrenner et al. 2018]. GST is trained with the reconstruction loss from Tacotron without any other explicit emotion or prosodic labels. Please refer to [Wang et al. 2018b] for more details.

We modify the inference mode of the GST model according to our specific task settings:

(1) In order to leverage the comic's visual information, we embed multi-modal hints (comic character face images and text from the corresponding speech balloons), during reference speech retrieval. Such multi-modal embedding carries a comprehensive representation of the emotions conveyed by the input comic, which efficiently captures the dynamic range of affective acoustic factors while inheriting the advantage of GST of not requiring any prosodic annotations (*e.g.*, intonation, stress).
(2) Instead of relying on only one reference speech to model the diverse speaking styles, as in GST, our approach employs a reference speech set $\{r_i\}_{i=1}^{N}$ containing the $N$-best emotion-matched speeches. Considering several reference speeches is important for enriching the emotion expressiveness and variety of acoustic representations.
(3) For each reference speech $r_i$, we compute its importance $h_i \in [0, 1]$ using its similarity to the comic input $(c, t)$ within the 3,328-d visual-textual emotion space, normalized over the similarities of all the retrieved reference speeches $\{r_i\}_{i=1}^{N}$. Then the overall emotion style embedding of the reference speech set $\{r_i\}_{i=1}^{N}$ is computed as a weighted sum: $\sum_{i=1}^{N} h_i \cdot \sum_{k=1}^{K} a_k^i v_k$, where $\sum_{i=1}^{N} h_i = 1$. Finally, the computed overall emotion style embedding and the comic image are fed to Tacotron to produce an emotional speech.

Following the above, our approach synthesizes a speech for each piece of text that reflects the emotional states of the speaking character, as inferred from the comic page. In the next step, for each character, a realistic tone of voice is synthesized in accordance with the personal attributes of the character. This tone of voice is used to refine all the speeches of the character to deliver a consistent speaking style. Note that the character may have different inferred emotions when speaking the content of different speech balloons, but it should have only one tone of voice across different speech balloons to maintain a consistent identity.

## 5.2 Synthesizing Acoustic Parameters for Characters

In our approach, the tone of voice is represented by two basic acoustic features: fundamental frequency $f_0$ and formant frequency $f_f$. These two frequencies have been proven to be strongly associated with the vocal gender and age [Ghazanfar and Rendall 2008; Ou and Mak 2017]. We optimize these frequencies such that the synthesized tone of voice matches the character in terms of his/her inferred personal attributes (*i.e.*, gender and age). Note that a character may appear multiple times in comics; our approach identifies repeated characters (Sec. 4.3) and only synthesizes one pair of $f_0$ and $f_f$.

Specifically, for each character $c$, our approach optimizes the fundamental frequency $f_0$ and formant frequency $f_f$ according to the personal attributes inferred for the character. This is achieved by minimizing the following cost function:

$$\mathrm{C}(f_0, f_f, c) = \mathrm{C}_{\mathrm{iden}}\Big(\mathcal{V}(f_0, f_f), c\Big) + \lambda \mathrm{C}_{\mathrm{auth}}(\mathcal{V}\Big(f_0, f_f\Big)\Big). \quad (3)$$

Here $\mathrm{C}_{\mathrm{iden}}(\cdot)$ is a visual-acoustic identity cost that penalizes the incompatibility between the voice tone $\mathcal{V}(f_0, f_f)$, parameterized by $f_0$ and $f_f$, and the input character $c$ over the gender and age. $\mathrm{C}_{\mathrm{auth}}(\cdot)$ is a speech-authenticity cost that penalizes the distortion of a speech spoken with the voice tone. $\mathcal{V}(\cdot)$ represents the formant Vocoder [Morris and Clements 2002], which reconstructs the voice tone wavelet from $f_0$ and $f_f$. We fix the trade-off coefficient $\lambda$ to be 1 for all of our experiments, unless otherwise specified.

*5.2.1 Speech Identity and Authenticity Recognition.* Before describing the details of the cost terms in Equation (3), we introduce a *Speech Identity and Authenticity Recognition* model based on which the cost terms are defined.

*Recognition Model:* This recognition model is designed to have three branches. The first two are for recognizing the gender and age attributes of speeches and the last one is for justifying the speech-quality, *i.e.*, the speech being real or distorted. Inspired by previous

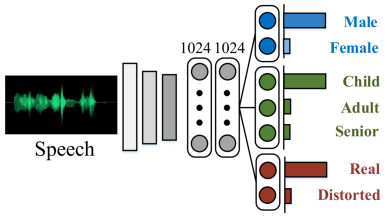


Fig. 8. Network architecture of the *Speech Identity and Authenticity Recognition*, which has three branches for speech gender recognition, speech age recognition, and speech authenticity recognition.

speech recognition models [Engel et al. 2017; Gabbay et al. 2018], the bottom of our model has five cascaded convolution layers (with 64, 64, 128, 128, and 128 output sizes). Batch Normalization and Leaky-ReLU activation are applied for each convolution layer. The final output feature is fed into two consecutive fully-connected layers, each of which has 1,024 neurons. Then, three fully-connected layer based classifiers are added. The three branched classification layers have sizes of 2, 3, and 2, respectively, and are used to predict the gender (*female* and *male*), age (*child*, *adult*, and *senior*), and level of distortion of the input speech, as shown in Fig. 8.

*Training Data.* To train our model, we collected a *Speech Corpus Dataset*, which has 13.7k real and 4.5k distorted speech clips obtained from different speech corpus [Veaux et al. 2016] and the Internet. Each of the speeches lasts at least 2 seconds. The real speech clips are annotated with gender and age labels, similar to the ones used in *Manga109*. We randomly selected 1.7k real speeches and 0.5k distorted speeches for testing; the rest were used for training.

*Learning.* Each training speech was first converted to a spectrogram format. The two speech-identity recognition branches were only trained with the 12k real speeches with identity labels, while the last speech-authenticity classification branch was trained with all the 16k training data.

*Evaluation.* We assessed the performance of our model on the test set of our *Speech Corpus Dataset*. The average classification accuracy of the gender and age attributes is 92.45% and 88.76%, respectively. The speech-authenticity classifier achieves an accuracy of 89.00%. Please refer to the supplementary material for more experimental details.

Based on our *Speech Identity and Authenticity Recognition* model, we elaborate the costs $\mathrm{C}_{\mathrm{iden}}(\cdot)$ and $\mathrm{C}_{\mathrm{auth}}(\cdot)$ defined in Equation (3).

*5.2.2 Visual-Acoustic Identity Cost.* $\mathrm{C}_{\mathrm{iden}}(\cdot)$ drives the synthesized voice tone $\mathcal{V}(f_0, f_f)$ to be consistent with the inferred identity of the character $c$. It is defined as:

$$\mathrm{C}_{\mathrm{iden}}(\cdot) = \mathrm{C}_{\mathrm{gen}}\big(\mathcal{V}(f_0, f_f), l^c_{\mathrm{gen}}\big) + \mathrm{C}_{\mathrm{age}}\big(\mathcal{V}(f_0, f_f), l^c_{\mathrm{age}}\big), \quad (4)$$

where the cost terms $\mathrm{C}_{\mathrm{gen}}(\cdot)$ and $\mathrm{C}_{\mathrm{age}}(\cdot)$ measure the inconsistency between the voice tone and the character's gender and age, respectively. $l^c_{\mathrm{gen}}$ and $l^c_{\mathrm{age}}$ are the gender and age attributes of character $c$ obtained from Sec. 4.3. The two cost terms $\mathrm{C}_{\mathrm{gen}}(\cdot)$ and $\mathrm{C}_{\mathrm{age}}(\cdot)$ are computed as:

$$\begin{aligned} \mathrm{C}_{\mathrm{gen}}(\cdot) &= \exp\Big(-p_{l^c_{\mathrm{gen}}}\big(\mathcal{V}(f_0, f_f)\big)\Big), \\ \mathrm{C}_{\mathrm{age}}(\cdot) &= \exp\Big(-p_{l^c_{\mathrm{age}}}\big(\mathcal{V}(f_0, f_f)\big)\Big). \end{aligned} \quad (5)$$

In $\mathrm{C}_{\mathrm{gen}}(\cdot)$, $p_{l^c_{\mathrm{gen}}}(\cdot) \in [0, 1]$ is the classification score of the gender label $l^c_{\mathrm{gen}}$. This score is computed by the *Speech Identity and Authenticity Recognition* model, which indicates the probability that the voice tone $\mathcal{V}(f_0, f_f)$ belongs to gender $l^c_{\mathrm{gen}}$. We define $\mathrm{C}_{\mathrm{age}}(\cdot)$ similarly.

*5.2.3 Speech Authenticity Cost.* In addition to constraining the identity consistency between the speech and the character through $\mathrm{C}_{\mathrm{iden}}(\cdot)$, the speech is also desired to be natural and realistic. To achieve this, a speech-authenticity cost $\mathrm{C}_{\mathrm{auth}}(\cdot)$ is employed to enhance the authenticity of a speech spoken with the synthesized voice tone. This cost uses the speech distortion classifier of our *Speech Identity and Authenticity Recognition* model to evaluate the quality of the speech spoken with the synthesized voice tone $\mathcal{V}(f_0, f_f)$:

$$\mathrm{C}_{\mathrm{auth}}(\cdot) = \exp\Big(-p_{\mathrm{real}}\big(\mathcal{V}(f_0, f_f)\big)\Big), \quad (6)$$

where $p_{\mathrm{real}}(\cdot) \in [0, 1]$ refers to the probability of the speech spoken with the synthesized voice tone $\mathcal{V}(f_0, f_f)$ being realistic.

*5.2.4 Optimization.* For the cost function $\mathrm{C}(\cdot)$ in Equation (3), we adopt a Markov chain Monte Carlo (MCMC) optimization to explore the space of possible intrinsic acoustic parameters extensively. For each character, our approach uses his/her first speech synthesized in Sec. 5.1 to initialize $f_0$ and $f_f$. Then it optimizes $f_0$ and $f_f$ by performing the following two steps iteratively:

- Propose a $(f_0', f_f')$ by modifying $f_0$ or $f_f$ with an equal probability. The move to modify $f_0$ is defined as: $f_0' = f_0 + \Delta f_0$, where $\Delta f_0$ is sampled from a Gaussian distribution whose mean is zero and variance is $0.1 f_0$. The move to modify a $f_f$ is formulated as similarly: $f_f' = f_f + \Delta f_f$, where $\Delta f_f$ is also sampled from a Gaussian distribution whose mean is zero and variance is $0.02 f_f$.
- Accept or reject the proposed $(f_0', f_f')$ based on the Metropolis-Hastings's acceptance probability [Hastings 1970]:

$$\mathcal{A} = \min\Big\{1, \frac{\mathrm{C}(f_0, f_f, c)}{\mathrm{C}(f_0', f_f', c)}\Big\}, \quad (7)$$

where $\mathrm{C}(\cdot)$ is the cost function defined in Equation (3).

The optimization is terminated if the absolute change in cost $\mathrm{C}(\cdot)$ is less than 5% over the past 20 iterations. Then, we obtain the optimized voice tone parameters $f_0^*$ and $f_f^*$ for character $c$.

Finally, similar to [Morris and Clements 2002], we use the two optimized voice tone parameters $f_0^*$ and $f_f^*$ to refine all the synthesized speeches of character $c$, such that each synthesized speech carries the voice tone as well as the emotion style of $c$. By applying the emotion prosody transfer described in Sec. 5.1 and the optimized voice tone parameters, our approach synthesizes a realistic speech consistent with the personal attributes (age, gender) of the character as well as his/her emotion as inferred from the comic for speaking the speech bubble content.

## 6 RESULTS AND DISCUSSIONS

We demonstrate the efficacy of our approach and present several useful applications. As our results consist of synthesized speeches, we encourage readers to watch the accompanying video to view

and hear the full results. Our approach was implemented on an Intel Core i7-5930K machine running at an NVIDIA TITAN GPU with 12GB graphics card memory. Our approach, including *Comic Visual Analysis* and *Comic Speech Synthesis*, took about 350 milliseconds for each synthesis. All synthesized speech clips had a sample rate of 22050Hz and a bit depth of 16 bits.

### 6.1 Audio Comics

We applied our approach for synthesizing speeches to four comic pages. These pages cover different color, themes, and styles, including American hero comics ("Captain Marvel"), Japanese manga ("Detective Conan"), Western fairy tale comics ("Briar Rose"), and Educational comics for children ("The Magic School Bus Rides Again"). Please refer to supplementary materials for the synthesized results, which were also used in our perceptual study (Sec. 7).

When applying our approach to a comic with a specified theme, the user can synthesize speeches consistent with this theme. Such comic speech synthesis can be achieved by using reference speeches that match the theme. Take the comic speech synthesis for "Captain Marvel" as an example. We can limit the retrieval to comic character faces and text in hero comics, *e.g.*, Batman, Arrow, etc. Currently, we have incorporated four different comic themes into our speech synthesis system, but our analysis and synthesis framework is flexible for incorporating additional themes during speech synthesis.

In addition, we demonstrate how our approach can be applied to tackle several common scenarios from *Manga109* [Matsui et al. 2017; Ogawa et al. 2018] in comic narration:

(1) **Different Emotions**: As shown in Fig. 9 (a), the same speech text can be spoken with different emotions as inferred from the different facial expressions.
(2) **Different Degrees of the Same Emotion**: As shown in Fig. 9 (b), the same speech text can be spoken in different degrees of the same emotion. In this case, the three characters speak the same speech text with different degrees of anger as inferred from their facial expressions.
(3) **Combined Emotions**: As shown in Fig. 9 (c), speech texts can be spoken with different combined emotions as inferred from the characters' faces and their speech texts. For example, a speech text can be spoken with fearfulness and sadness, or fearfulness and anger.

To handle these scenarios, our approach represents the target emotion as a latent variable. It matches the visual-textual emotion feature inferred from the character's face and speech text with that of the references in the *Reference Speech Dataset* in order to retrieve the three nearest reference speeches for emotion transfer (see Sec. 5.1). Then it conditions the emotion prosody transfer model on the reference speeches to synthesize a speech consistent with the target emotion. We performed a perceptual study (Sec. 7) to evaluate people's perception of the emotion of these synthesized speeches, which we find to be consistent with the target emotion inferred from the visual content.

### 6.2 Other Applications

Our comic-guided speech synthesis approach leads to a number of interesting digital comic applications.

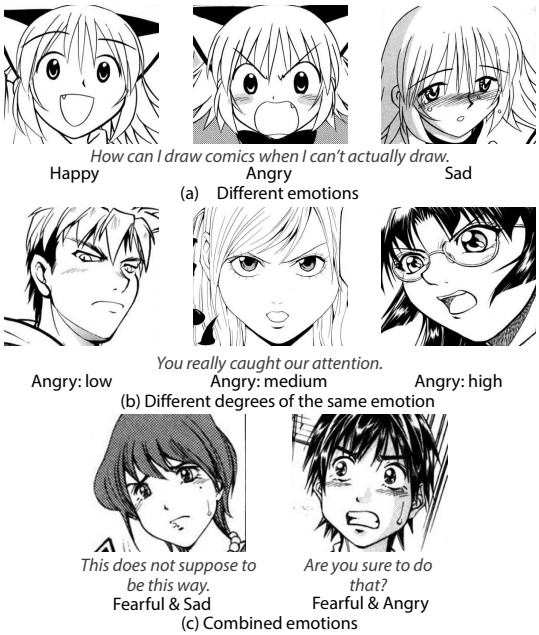


Fig. 9. (a) Same speech text spoken with different emotions. (b) Same speech text spoken with different degrees of the same emotion (angry). (c) Speech texts spoken with combined emotions inferred from faces and texts. ©Yagami Ken, Takeyama Yusuke, Ishioka Shoei, Deguchi Ryusei

*AR Comic Book Reading.* Read through an augmented reality device such as a HoloLens headset, comic books can be made even more appealing and engaging by showing animated 3D characters and models. Fig. 10(a) depicts an example. Our approach can make such comic books audible, which further enriches the reading experience. In the example, our approach is applied to narrate the comic character with a male, child voice, which talks about the Lincoln Memorial in a vivid manner with emotion matching the facial expression and speech text.

*Game Character Narration.* Many games use comic-like characters for storytelling and explaining game settings. For example, a role-playing game usually involves multiple characters who speak a large amount of text to deliver the story. Our approach can be applied for narrating such game characters as well. Fig. 10(b) shows an example, where a female adult speech is synthesized for the game character. The emotion of the synthesized speech changes with the emotion inferred from the character's face and the game texts. Our approach can potentially be applied to automate the narration of a game which involves a lot of different characters and texts, saving much effort and time in game production.

*Talking Head Narration.* 3D character creation tools, such as CrazyTalk and 3D face synthesis [Lang et al. 2019], allow users to convert a 2D comic figure into a 3D "talking head" that can be animated and instructed to read a piece of text, for applications such as virtual reality, animation, and video conferencing. By synthesizing speeches that match the personal attributes inferred from the input 2D comic figure and the input text, our approach can potentially be applied for automatically narrating a talking head generated by such tools without the need of any user-provided sound recording. Fig. 10


(a) AR Comic Book Reading (b) Game Character Narration (c) Talking Head Narration

Fig. 10. Applications driven by our approach: (a) a virtual character shown via a HoloLens introduces the content of a book with synthesized speech; (b) a character in a role-playing game is automatically narrated by our approach; (c) a talking head speaks with a synthesized speech that matches her attributes. In each example, the left shows the application scenario. The right shows the character face, text, and the corresponding synthesized speech.

(c) shows a talking head automatically narrated by our approach based on the input 2D face and text.

## 7 PERCEPTUAL STUDY

We evaluated the effectiveness of our approach and the acoustic fidelity of the results through perceptual studies. First, we evaluated the consistency between comics and synthesized speeches. Second, we compared our approach with two other approaches. Finally, the quality of the synthesized speeches was evaluated.

*Participants.* We recruited 66 participants whose age ranged from 18 to 40. They were divided into two groups randomly. The first group contained 36 participants, who took part in the consistency evaluation and the speech quality evaluation. The second group contained 30 participants, who took part in the comparison evaluation. All participants reported normal or corrected-to-normal vision with no colour-blindness and normal hearing. None of them had expertise in speech synthesis.

*Procedure.* A website was set up to conduct the perceptual study. The participants were seated 35 cm in front of a screen (with 1440× 900 resolution). Auditory input was provided by a pair of Logitech G430 gaming headphones with 7.1 channel surround sound output. Before each study, the participants were given a task description and encouraged to ask any questions. Each participant was allowed to view the comic page and to play the speech clip an unlimited number of times. Please refer to the supplementary materials for the detailed description, all comic pages, and the corresponding synthesized speeches.

### 7.1 Comic and Speech Consistency Evaluation

Since we apply gender, age, and emotion analyzed from the comic page to guide the speech synthesis, we conducted three consistency evaluations between the comic pages and the synthesized speeches to validate our approach: (1) *Emotion Consistency Evaluation*, validating whether the synthesized speeches reflected the emotions conveyed by the comic character faces and speech texts; (2) *Gender and Age Consistency Evaluation*, validating whether the synthesized speeches successfully carried the personal attributes (*i.e.*, gender and age) conveyed by comic characters; and (3) *Overall Consistency Evaluation*, validating whether the comics and the synthesized speeches were consistent in terms of overall visual and hearing criteria.

Comic pages, covering different colors, styles, and themes, were used to conduct the perceptual studies, *i.e.*, American hero comics ("Infinity Countdown : Captain Marvel #1"[1]), Japanese detective comics ("Detective Conan - Chapter 1025"[2]), Western fairy tale comics ("Briar Rose"[3]), and educational comics ("The Magic School Bus"[4]). The pages including 34 panels, 20 different characters, and 37 groups of speaking characters, speech text, and the corresponding synthesized speech. In the study, pairs of comic pages and the corresponding narrations were randomly selected to display to the participants. Each pair was rated by 9 participants. The participants were asked to rate each comic panel according to the instructions. Note that, the display order of each paired panel and speech was random, *i.e.*, participants may have seen the panel for rating first and then heard the speech for rating, or vice versa.

Table 2. Statistics of rating differences between the ratings of comics and speeches in terms of emotion. The last column shows the percentage of participants whose rating differences are within 1 for one emotion.

| Emotion | Mean | SD | Diff. within 1 |
|---|---|---|---|
| Neutral | 0.80 | 1.09 | 78.43% |
| Happy | 0.39 | 0.77 | 90.36% |
| Sad | 0.45 | 0.83 | 88.83% |
| Angry | 0.42 | 0.82 | 89.09% |
| Fearful | 0.35 | 0.76 | 92.64% |

#### 7.1.1 *Emotion Consistency Evaluation.*

The main goal of this evaluation was to test whether the synthesized speeches accurately express the comic characters' emotion states.

Given each panel of the comic page and the corresponding synthesized speech, the participants were asked to rate them in terms of their emotion perception. We provided 5 common emotions for participants to rate (*i.e.*, neutral, happy, sad, angry, and fearful). It is worth noting that the participants needed to rate all emotion categories so that we could evaluate whether the synthesized speeches could reflect both single emotions ( *e.g.*, very angry) and combined emotions ( *e.g.*, slightly happy and slightly fearful) conveyed by the character's face and the speech text. The rating was between 1 and 5, in which 1 indicated the comic or speech did not convey a certain emotion, and 5 indicated it expressed the emotion most strongly. Furthermore, the overall emotion consistency (also varying from 1 to 5) between the comic and the corresponding speech was rated as well, where 1 meant very inconsistent and 5 meant very consistent.

For each participant's ratings, we first calculated an average rating difference across the 5 emotion categories between the comic emotion ratings and the speech emotion ratings on each comic panel. The smaller the difference, the more likely both the comic and the

[1]https://www.marvel.com/comics/issue/67346/infinity_countdown_captain_marvel_2018_1

[2]https://www.detectiveconanworld.com/

[3]https://disney.fandom.com/wiki/Disney_ Princess_(comic_ book)

[4]https://www.netflix.com/title/80108373

speech belong to the same emotion category. Then we analyzed the correlation between all participants' average rating differences and their overall comics-speech emotion consistency ratings to verify the relation between them.

We conducted a Bivariate (Pearson) correlation analysis and obtained a negative correlation between all participants' average rating differences ($M = 0.48, SD = 0.47$) and their overall comics-speech emotion consistency ratings ($M = 4.01, SD = 0.88$), with $r = -.62$, $p < .05$. The results support that the lower the average rating difference was, the higher the emotion consistency rating.

To further demonstrate that the synthesized speeches effectively expressed different emotion states conveyed by comics, we analyzed the distribution of the emotion differences between the comic emotion ratings and the speech emotion ratings on each panel. In Table 2, we show the average rating difference and the corresponding standard deviation on each emotion. We observe that the emotion 'fearful' has the lowest mean and standard deviation, whereas the emotion 'neutral' has the highest mean and standard deviation.

To investigate the distribution of the rating differences, we considered the percentage of participants whose rating differences (within 1) between the comics and the speeches in terms of one emotion, in Table 2. A higher percentage means that more participants thought the paired panel and speech expressed the same emotional state. Four out of five emotions have more than 89% participants whose rating differences were within 1. Only the emotion 'neutral' had a relatively lower percentage of 78.43%. According to feedback from the participants, most felt that judging the neutral emotion on speech was more difficult than on comics. This explains the diversities in judgment for the neutral emotion.

The results suggest that our approach makes the most of the emotion information obtained from the visual analysis to guide the speech generation so that the synthesized speeches can carry the corresponding emotions.

*7.1.2 Gender and Age Consistency Evaluation.* The main goal of this evaluation was to test whether the personal attributes, *i.e.*, gender and age, conveyed by the synthesized speeches match the ones carried by the character, hence, to evaluate the optimization process in the acoustic parameters synthesis (Sec. 5.2).

The participants were required to rate the personal attributes for each comic character and the corresponding speech. Due to different drawing styles, some character faces were hard to distinguish in terms of gender and age. To maintain a unified and objective rating standard for different comics, we set the rating with ordinal variables. The gender rating varied from 1 to 5, in which 1 represented male and 5 represented female. The age rating also ranged from 1 to 5, in which 1 represented a child, 2 to 4 represented an adult, and 5 represented a senior. We showed users examples of each category to help them intuitively understand the meaning of the variables. Moreover, participants were asked to rate the overall consistency (varying from 1 to 5) of the comics and the paired speeches regarding gender and age respectively, in which 1 indicated very inconsistent and 5 indicated very consistent.

Similar to the emotion consistency evaluation, we calculated the rating differences between the gender/age ratings of the comics and speech. Then, we conducted the Bivariate (Pearson) Correlation analysis between the gender rating differences ($M = 0.50, SD = 1.14$) and the overall gender consistency ratings ($M = 4.33, SD = 0.92$). There was a negative correlation between them with $r = -.40, p < .05$. A negative correlation was also observed between the age rating differences ($M = 0.70, SD = 0.94$) and the overall age consistency ratings ($M = 4.07, SD = 0.90$), $r = -.33, p < .05$.

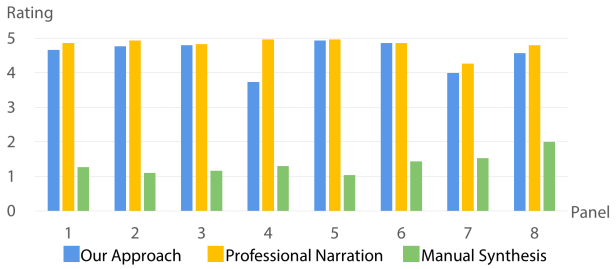


Fig. 11. Statistics of average consistency rating of our approach, of *Professional Narration* approach, and of *Manual Synthesis* approach on 8 panels. The results of our approach and *Professional Narration* approach are comparable across 8 panels, and both achieved much higher average ratings than the *Manual Synthesis* approach.

93.37% and 90.60% of the participants had rating differences between the comics and synthesized speeches for gender and age within 1. The results suggest reasonable performance for the synthesized speech with respect to conveying personal attributes.

*7.1.3 Overall Consistency Evaluation.* The goal was to evaluate the participants' overall perception about the consistency between the input comic page and the output narrations. Participants were asked to rate the overall consistency for the given pair of panels and speeches. The rating ranged from 1 to 5, in which 1 meant they were very inconsistent and 5 meant they were very consistent.

We received positive feedback from the participants regarding the consistency ($M = 3.76, SD = 0.91$). Among the results, the American hero comics achieved the highest consistency rating on average ($M = 4.13, SD = 0.72$), whereas the Japanese detective manga achieved the lowest rating on average ($M = 3.57, SD = 0.85$). According to the participants' feedback on the comic page of Japanese detective manga, we observed that some participants gave high ratings on the consistency of gender, age and emotion between the comics and the paired speeches, but gave relatively lower ratings on the overall consistency. They explained that the generated speech did not meet their expectations in other aspects, *e.g.*, accents. The results give us some interesting insights about considering more attributes to narrate comics in the future.

To further verify the effectiveness of our approach, *i.e.*, comic speech perception is affected by our considered factors (*e.g.*, gender, age, and emotion) besides voice preference and priors of the voice for faces, we computed Bivariate (Pearson) correlation coefficients between the ratings of overall consistency and gender, age, and emotion consistency, respectively. We obtain positive correlations: gender ($r = .37, p < .05$); age ($r = .57, p < .05$); emotion ($r = .51, p < .05$). The results suggest that increases in the consistency ratings of emotion, gender, and age correlate with increases in the overall consistency. This supports that our adopted strategy, *i.e.*, narrating comics by considering the consistency between the comics and the speeches in terms of emotion expression and personal attributes, is reasonable.

## 7.2 Comparison

We next evaluated how well our approach narrates comics through comparing it to *Professional Narration* and *Manual Synthesis*. Our approach and the other two approaches were applied to narrate a comic page of "The Magic School Bus Rides Again - Season 2, Episode 10 - Tim and the Talking Trees"[5], which contains 8 panels and 6 characters. Please refer to the supplementary material for the comic page used and the synthesized speeches.

Professional Narration: To realistically imitate the application scenarios, we extracted the audio track from the publicly released animation clips of "The Magic School Bus Rides Again" as the compared speeches. The comic characters were narrated by six professional narrators who were recruited by the animation producer.

Manual Synthesis: The speeches were manually synthesized by three professional audio editors who majored in broadcasting for about four years. They all received training experience in narrating books and editing audio clips. They used the state-of-the-art text-to-speech system (*i.e.*, Google Cloud Text-to-Speech) and audio editing tool (*i.e.*, Cool Edit and Adobe Audition) to narrate the comic page.

In the perceptual study, the 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 to very inconsistent to the very consistent.

*Outcome and Analysis.* The statistics of the average consistency rating on each panel are shown in Fig. 11. For the overall average consistency across 8 panels, the *Professional Narration* obtained the highest rating ($M = 4.81, SD = 0.44$), followed by our approach ($M = 4.54, SD = 0.63$) and *Manual Synthesis* ($M = 1.35, SD = 0.64$). Please refer to the supplementary material for all participants' ratings and detailed mean and standard deviation results.

In Fig. 11, comparison results on each panel show that the consistency ratings of our approach and *Professional Narration* are mostly comparable. The ratings of our approach are lower than those of *Professional Narration* by about 0.7% (on panel 3 and 5) to 6% (on panel 7), except for the results of panel 4. For panel 4, our approach had a rating of 3.73% while *Professional Narration* had a rating of 4.97%, which led to statistical significance ($F_{[1,479]} = 30.02, p < .05$) of One-Way ANOVA test between the results of the two approaches on the entire comic page.

We further conducted a One-Way ANOVA test on the results of each individual panel with one independent variable, *i.e.*, the character in each panel was treated independently and the corresponding results generated by different narration approaches were used to carry out the test. The detailed results shown in Table 3 support that our approach and *Professional Narration* were comparable with regard to the consistency ratings on most of the panels. 7 out of 8 panels (panel 4 being the exception) did not show any statistical significance between the ratings of our approach and the ones of *Professional Narration*. The reason for the lower rating in panel 4 arises from the emotional state analysis. Although we fine-tuned the top-performing emotion recognition model on comic data, the emotion recognition task itself is still very challenging. So in some cases, *e.g.*, panel 4, the performance of the emotion state analysis affects the performance of the emotion prosody transfer, which may affect the participants' perception about the consistency between the comics and the synthesized speeches. We believe that by improving the emotion analysis, our approach will perform better.

Table 3. One-Way ANOVA test results on overall consistency scores of each independent panel between our approach and *Professional Narration*, and *Manual Synthesis*, respectively.

| Panels | Ours and Professional | Ours and Manual |
|---|---|---|
| 1 | $F_{[1,29]} = 3.43, \quad p = 0.07 > .05$ | $F_{[1,29]} = 398.21, \quad p < .05$ |
| 2 | $F_{[1,29]} = 1.66, \quad p = 0.20 > .05$ | $F_{[1,29]} = 717.71, \quad p < .05$ |
| 3 | $F_{[1,29]} = 0.15, \quad p = 0.87 > .05$ | $F_{[1,29]} = 632.43, \quad p < .05$ |
| 4 | $F_{[1,29]} = 50.68, \quad p < .05$ | $F_{[1,29]} = 153.58, \quad p < .05$ |
| 5 | $F_{[1,29]} = 0.18, \quad p = 0.84 > .05$ | $F_{[1,29]} = 2298.70, \quad p < .05$ |
| 6 | $F_{[1,29]} = 0.08, \quad p = 0.93 > .05$ | $F_{[1,29]} = 267.86, \quad p < .05$ |
| 7 | $F_{[1,29]} = 1.17, \quad p = 0.32 > .05$ | $F_{[1,29]} = 86.44, \quad p < .05$ |
| 8 | $F_{[1,29]} = 2.02, \quad p = 0.14 > .05$ | $F_{[1,29]} = 103.85, \quad p < .05$ |

On the other hand, the difference between *Manual Synthesis* and our approach ($F_{[1,479]} = 3059.58, p < .05$), and between *Manual Synthesis* and *Professional Narration* ($F_{[1,479]} = 4782.08, p < .05$) were statistically significant. We also conducted a One-Way ANOVA test between the consistency ratings of our approach and of *Manual Synthesis* on each panel. The results shown in Table 3 indicate a statistically significant difference at the $\alpha = 0.05$ significance level between the two approaches on all panels. Since comics usually carry complex character attributes, manually tuning speeches to match the visual observations on a comic page can be challenging.

## 7.3 Mean Opinion Score Test

The quality of the synthesized speeches is another factor which affects users' perception on the consistency between the comics and the speeches. Therefore, we conducted a Mean Opinion Score (MOS) test for each synthesized speech to evaluate the speech quality. The participants were asked to rate the quality of the speeches using 1 to 5: 1 = bad (very annoying), 2 = poor (annoying), 3 = fair (slightly annoying), 4 = good (perceptible but not annoying) and 5 = excellent (imperceptible, almost real). We obtained an MOS of $4.04 \pm 0.71$ , which corresponds to around "Good" in terms of speech quality, showing that the quality of the synthesized speeches is acceptable.

## 8 CONCLUSION

We propose a novel computational approach for automatically narrating comics. Guided by the visual analysis results of a comic page, our approach synthesizes speeches that match comic characters' personal attributes as well as their emotional states as inferred from their faces and speech content. We demonstrated that our approach can be used for narrating different types of comics, and for synthesizing speeches with different degrees and combinations of emotions. In addition to automating audio comic production, our approach could also enable other interesting applications such as AR comic book reading, game character narration and talking head narration. Perceptual studies also find the synthesized speeches of reasonable quality and consistent with the comic's visual content.

[5]https://www.netflix.com/title/80108373

*Limitations.* Comics with fantasy themes featuring non-human characters are common. For example, a comic character could be a cat (*e.g.*, Garfield), a dog (*e.g.*, Snoopy), or a mouse (*e.g.*, Mickey). Due to the difficulty of learning effective and general representations of such non-human characters by computer vision techniques, we only tested our approach for narrating human-like characters. Based on our approach, the user might circumvent the narration of non-human characters by explicitly specifying personal attributes and reference speeches which carry emotions deemed appropriate for the non-human characters.

To handle failures of each step in *Comic visual analysis*, we set corresponding validation criteria. Take the comic character identification as an example, we computed the normalized Euclidean distance (in [0,1]) of a detected face with each cluster center. The face was assigned to the nearest cluster center by default. However, if the smallest normalized distance of a face was longer than 0.1, this could imply confusing identification (*e.g.*, caused by occlusion), and the user could be promoted to manually assign the ID. The advancement of the relevant visual analysis techniques, such as panel extraction and text extraction, will improve the visual analysis performance and reduce the cost of user engagement.

Although the comic character identification and a speech authenticity cost are designed to ensure cross-panel consistency for the acoustic features of the same character and enhance the realism of the synthesized speeches, the speech quality heavily relies on the TTS technique (Tacotron) used in the comic speech synthesis step. The advancement of TTS techniques, the availability of large-scale datasets for training, and the size and quality of *Reference Speech Dataset* will help synthesize more realistic comic speeches.

To verify the effectiveness of our approach, we designed perceptual studies to compare our results with those of other approaches (*i.e.* professional narration and manual synthesis) and conducted corresponding One-Way ANOVA tests. We further tested the results of each panel, aiming to validate the performance of our approach on each independent comic character shown in each panel. For overall differences between the approaches, the multiple groups One-Way ANOVA tests will facilitate such analysis and avoid the high chance of getting a significant result of multiple One-Way ANOVA tests with two groups. Furthermore, since only two comic characters in the comic page appeared twice, we did not test on the results of the same character across different panels generated by different approaches, *i.e.*, two independent variables (comic characters and narration approaches). Testing more comic pages with more recurring characters, considering more results generated by different approaches, and performing Two-Way ANOVA test will help us analyze our approach more comprehensively.

*Future Work.* Our current approach considers gender, age, and emotion parameters for synthesizing speeches for comic characters. It would be interesting to consider additional factors such as personality and intention of the characters (*e.g.*, head pose [Wang et al. 2019]), scene context, and comic story, to yield more diverse and realistic speech synthesis results in future work. In our experiments, we tested our approach for synthesizing speeches based on the visual analysis of a comic page. For future extension, it would be helpful to comprehensively test the practicality of our approach for the scalable production of audio comics by using the content of a fully digitalized comic book as input. Another challenging but interesting avenue for future research would be to extend our approach to automatically narrating cartoons and animations.

## ACKNOWLEDGMENTS

We would like to thank the *Manga109* dataset for exploring the possibility of devising a computational approach for synthesizing realistic speeches based on the visual content of comics. Lap-Fai Yu was supported by the National Science Foundation under award number 1565978 when working on this project. Wenguan Wang is supported by the CCF-Tencent Open Fund, and Zhijiang Lab's International Talent Fund for Young Professionals.

## REFERENCES

Waleed Abdulla. 2017. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. https://github.com/matterport/Mask_RCNN.

Olivier Augereau, Motoi Iwata, and Koichi Kise. 2018. A survey of comics research in computer science. *Journal of Imaging* 4, 7 (2018), 87.

Rainer Banse and Klaus R Scherer. 1996. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology* 70, 3 (1996), 614.

Pascal Belin, Patricia EG Bestelmeyer, Marianne Latinus, and Rebecca Watson. 2011. Understanding voice perception. *British Journal of Psychology* 102, 4 (2011), 711–725.

Pascal Belin, Shirley Fecteau, and Catherine Bedard. 2004. Thinking the voice: neural correlates of voice perception. *Trends in cognitive sciences* 8, 3 (2004), 129–135.

P. Praat Boersma. 2002. A System for Doing Phonetics by Computer. *Glot International* 5, 9/10 (2002), 341–345.

Vicki Bruce and Andy Young. 1986. Understanding face recognition. *British Journal of Psychology* 77, 3 (1986), 305–327.

Salvatore Campanella and Pascal Belin. 2007. Integrating face and voice in person perception. *Trends in cognitive sciences* 11, 12 (2007), 535–543.

Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. 2016. Real-time facial animation with image-based dynamic avatars. *TOG* 35, 4 (2016), 126.

Ying Cao, Antoni B. Chan, and Rynson W. H. Lau. 2012. Automatic stylistic manga layout. *TOG* 31, 6 (2012), 1–10.

Ying Cao, Rynson W. H. Lau, and Antoni B. Chan. 2014. Look over here:attention-directing composition of manga elements. *TOG* 33, 4 (2014), 1–11.

Wei-Ta Chu and Wei-Wei Li. 2017. Manga facenet: Face detection in manga based on deep neural network. In *ICMR*. ACM, 412–415.

Wei-Ta Chu and Wei-Wei Li. 2019. Manga face detection based on deep neural networks fusing global and local information. *Pattern Recognition* 86 (2019), 62–72.

K Dimos, L Dick, and V Dellwo. 2015. Perception of levels of emotion in speech prosody. *The Scottish Consortium for ICPhS* (2015).

Alexander Dunst, Jochen Laubrock, and Janina Wildfeuer. 2018. *Empirical Comics Research: Digital, Multimodal, and Cognitive Methods*. Routledge.

Marek Dvorožnák, Wilmot Li, Vladimir G Kim, and Daniel Sỳkora. 2018. Toonsynth: example-based synthesis of hand-colored cartoon animations. *TOG* 37, 4 (2018), 167.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. 2017. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. In *ICML*. 1068–1077.

Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. 2017. Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Networks* 92 (2017), 60–68.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Conference on Emirical Methods in Natural Language Processing*.

Adam Finkelstein, Adam Finkelstein, Adam Finkelstein, Adam Finkelstein, and Adam Finkelstein. 2017. VoCo: text-based insertion and replacement in audio narration. *TOG* 36, 4 (2017), 96.

Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science* 315, 5814 (2007), 972–976.

Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. 2018. Visual Speech Enhancement. In *Interspeech*. 1170–1174.

Asif A Ghazanfar and Drew Rendall. 2008. Evolution of human vocal production. *Current Biology* 18, 11 (2008), R457–R460.

Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. 2016. Synthetic data for text localisation in natural images. In *CVPR*. 2315–2324.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*, Vol. 2. IEEE, 1735–1742.

W Keith Hastings. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 1 (1970), 97–109.

Iben Have and Birgitte Stougaard Pedersen. 2013. Sonic mediatization of the book: affordances of the audiobook. *MedieKultur: Journal of media and communication research* 29, 54 (2013), 18–p.

Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. 2017. Avatar digitization from a single image for real-time rendering. *TOG* 36, 6 (2017), 195.

Andrew J Hunt and Alan W Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *ICASSP*, Vol. 1. 373–376.

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. 2018. Efficient Neural Audio Synthesis. In *ICML*, Vol. 80. 2410–2419.

Miyuki Kamachi, Harold Hill, Karen Lander, and Eric Vatikiotis-Bateson. 2003. Putting the face to the voice': Matching identity across modality. *Current Biology* 13, 19 (2003), 1709–1714.

Hideki Kawahara, Masanori Morise, Toru Takahashi, Ryuichi Nisimura, Toshio Irino, and Hideki Banno. 2008. TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *ICASSP*. 3933–3936.

Bernhard Kratzwald, Suzana Ilić, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. 2018. Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems* 115 (2018), 24–35.

Andy SY Lai, Chris YK Wong, and Oscar CH Lo. 2015. Applying augmented reality technology to book publication business. In *International Conference on e-Business Engineering*. IEEE, 281–286.

Yining Lang, Wei Liang, Yujia Wang, and Lap-Fai Yu. 2019. 3d face synthesis driven by personality impression. In *AAAI*, Vol. 33. 1707–1714.

Norman J Lass, Karen R Hughes, Melanie D Bowyer, Lucille T Waters, and Victoria T Bourne. 1976. Speaker sex identification from voiced, whispered, and filtered isolated vowels. *The Journal of the Acoustical Society of America* 59, 3 (1976), 675–678.

Younggun Lee, Azam Rabiee, and Soo-Young Lee. 2017. Emotional End-to-End Neural Speech Synthesizer. In *NIPS Workshop*.

Chengze Li, Xueting Liu, and Tien-Tsin Wong. 2017. Deep extraction of manga structural lines. *TOG* 36, 4 (2017), 117.

Hao Li, Yongguo Kang, and Zhenyu Wang. 2018. EMPHASIS: An Emotional Phoneme-based Acoustic Model for Speech Synthesis System. *arXiv preprint arXiv:1806.09276* (2018).

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *ICCV*. 3730–3738.

Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. 2018. Arbitrary-oriented scene text detection via rotation proposals. *TMM* 20, 11 (2018), 3111–3122.

Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2017. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications* 76, 20 (2017), 21811–21838.

Phil McAleer, Alexander Todorov, and Pascal Belin. 2014. How do you say "Hello"? Personality impressions from brief novel voices. *PLOS ONE* 9, 3 (2014), e90779.

Rachel Mcdonnell, Cathy Ennis, Simon Dobbyn, and Carol O'Sullivan. 2009. Talking bodies:Sensitivity to desynchronization of conversations. *TAP* 6, 4 (2009), 1–8.

Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. 2017. SampleRNN: An unconditional end-to-end neural audio generation model. In *ICLR*.

David S. Miall. 1989. Beyond the schema given: Affective comprehension of literary narratives. 3, 1 (1989), 55–78.

Seyed Hamidreza Mohammadi and Alexander Kain. 2017. An overview of voice conversion systems. *Speech Communication* 88, 88 (2017), 65–82.

Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. 2016. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *Ieice Transactions on Information and Systems* 99, 7 (2016), 1877–1884.

R. W. Morris and M. A. Clements. 2002. Reconstruction of speech from whispers. *Medical Engineering & Physics* 24, 7 (2002), 515–520.

Nhu-Van Nguyen, Christophe Rigaud, and Jean-Christophe Burie. 2017. Comic characters detection using deep learning. In *IAPR international conference on document analysis and recognition*, Vol. 3. IEEE, 41–46.

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).

Jan Ondřej, Cathy Ennis, Niamh A Merriman, and Carol O'sullivan. 2016. FrankenFolk: Distinctiveness and attractiveness of voice and motion. *TAP* 13, 4 (2016), 20.

Dayi Ou and Cheuk Ming Mak. 2017. Optimization of natural frequencies of a plate structure by modifying boundary conditions. *The Journal of the Acoustical Society of America* 142, 1 (2017), EL56–EL62.

Xufang Pang, Ying Cao, Rynson WH Lau, and Antoni B Chan. 2014. A robust panel extraction method for manga. In *International Conference on Multimedia*. ACM, 1125–1128.

Robert S Petersen. 2011. *Comics, manga, and graphic novles: a history of graphic narratives*. ABC-CLIO.

Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. 2018. Learning human-object interactions by graph parsing neural networks. In *ECCV*. 401–417.

Xiaoran Qin, Yafeng Zhou, Zheqi He, Yongtao Wang, and Zhi Tang. 2017. A faster R-CNN based method for comic characters face detection. In *IAPR International Conference on Document Analysis and Recognition*, Vol. 1. IEEE, 1074–1080.

Yingge Qu, Wai Man Pang, Tien Tsin Wong, and Pheng Ann Heng. 2008. Richness-preserving manga screening. *TOG* 27, 5 (2008), 1–8.

Yingge Qu, Tien Tsin Wong, and Pheng Ann Heng. 2006. Manga colorization. *TOG* 25, 3 (2006), 1214–1220.

Richard Rieman. 2016. *The Author's Guide to Audiobook Creation*. Breckenridge Press.

Christophe Rigaud, Clément Guérin, Dimosthenis Karatzas, Jean-Christophe Burie, and Jean-Marc Ogier. 2015. Knowledge-driven understanding of images in comic books. *International Journal on Document Analysis and Recognition* 18, 3 (2015), 199–221.

Ethan M Rudd, Manuel Günther, and Terrance E Boult. 2016. Moon: A mixed objective optimization network for the recognition of facial attributes. In *ECCV*. 19–35.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *ICASSP*. IEEE, 4779–4783.

Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2018. Aster: An attentional scene text recognizer with flexible rectification. *TPAMI* (2018).

Edgar Simo-Serra, Satoshi Iizuka, Kazuma Sasaki, and Hiroshi Ishikawa. 2016. Learning to simplify: fully convolutional networks for rough sketch cleanup. *TOG* 35, 4 (2016), 121.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J Weiss, Rob Clark, and Rif A Saurous. 2018. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. In *ICML*. 4700–4709.

Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. 2017. Char2wav: End-to-end speech synthesis. In *International Conference on Learning Representations Workshop*.

Marco Stricker, Olivier Augereau, Koichi Kise, and Motoi Iwata. 2018. Facial Landmark Detection for Manga Images. *arXiv preprint arXiv:1811.03214* (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.

Debra Trampe, Jordi Quoidbach, and Maxime Taquet. 2015. Emotions in Everyday Life. *PLOS ONE* 10, 12 (2015).

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. In *9th ISCA Speech Synthesis Workshop*. 125–125.

Robin Varnum and Christina T Gibbons. 2007. *The language of comics: Word and image*. Univ, Press of Mississippi.

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

Wenguan Wang, Jianbing Shen, and Haibin Ling. 2018a. A deep network solution for attention and aesthetics aware photo cropping. *TPAMI* 41, 7 (2018), 1531–1544.

Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. 2018c. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *CVPR*. 4271–4280.

Yujia Wang, Wei Liang, Jianbing Shen, Yunde Jia, and Lap-Fai Yu. 2019. A deep Coarse-to-Fine network for head pose estimation from synthetic data. *Pattern Recognition* 94 (2019), 196–206.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards End-to-End Speech Synthesis. In *Proceedings of Interspeech*. 4006–4010.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. 2018b. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. In *ICML*.

Zhong-Qiu Wang and Ivan Tashev. 2017. Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. In *ICASSP*. IEEE, 5150–5154.