Supplementary for "V2C: Visual Voice Cloning"

Qi Chen[1]* Mingkui Tan[2] Yuankai Qi[1] Jiaqiu Zhou[2,3] Yuanqing Li[2,3]† Qi Wu[1]†
[1]University of Adelaide [2]South China University of Technology [3]Pazhou Lab
{qi.chen04, qi.wu01}@adelaide.edu.au, qykshr@gmail.com
{auyqli, mingkuitan}@scut.edu.cn, mszjq@mail.scut.edu.cn

We organise the supplementary materials as follows.

- In Section A, we provide a detailed analysis of the proposed V2C-Animation dataset, *e.g.*, word cloud, distribution of utterance length, *etc*.
- In Section B, we pose more visualised pitch tricks of the samples from the V2C-Animation dataset and the related datasets.
- In Section C, we exhibit more visual results of the mel-spectrogram derived from our V2C-Net with comparisons against baseline and ground truth.
- In Section D, we report the implementation details.
- In Section E, we depict more details of the vocoder.
- In Section F, we analyse the limits of MCD-DTW.

A. More Analysis of V2C-Animation Dataset

A.1. Word Cloud and Count

In Figure 1, we visualise the texts/subtitles of our V2C-Animation dataset as Venn-style word cloud [2], where the size of each word refers to the harmonic mean of its count.

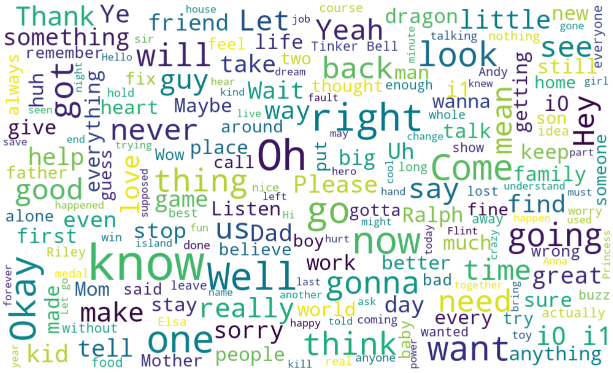


Figure 1. Word cloud of the texts on our V2C-Animation dataset.

Besides, we also provide the top 30 words on our V2C-Animation dataset along with their counts in Figure 2. More results (top 100) are in the following:
('know', 437), ('oh', 305), ('right', 255), ('one', 254), ('now', 250), ('well', 250), ('go', 233), ('okay', 217), ('come', 210), ('want', 201), ('look', 196), ('got', 181), ('going', 173), ('think', 167), ('will', 165), ('thing', 163), ('gonna', 163), ('need', 159), ('see', 155), ('back', 153), ('never', 151), ('us', 147), ('time', 141), ('say', 139), ('hey', 138), ('mean', 137), ('let', 137), ('good', 135), ('yeah', 131), ('guy', 128), ('really', 124), ('make', 124), ('thank', 124), ('little', 112), ('way', 108), ('love', 108), ('ye', 108), ('find', 104), ('help', 97), ('tell', 96), ('wait', 95), ('take', 93), ('kid', 92), ('please', 91), ('sorry', 88), ('something', 87), ('great', 87), ('dad', 87), ('friend', 84), ('day', 82), ('game', 80), ('stop', 75), ('even', 75), ('Uh', 74), ('big', 67), ('work', 66), ('Ralph', 66), ('much', 62), ('give', 62), ('first', 61), ('everything', 60), ('new', 59), ('still', 58), ('life', 58), ('keep', 58), ('dragon', 58), ('family', 57), ('sure', 56), ('made', 56), ('talk', 55), ('world', 53), ('place', 53), ('heart', 53), ('every', 53), ('maybe', 53), ('stay', 52), ('wanna', 51), ('better', 51), ('people', 50), ('huh', 50), ('anything', 50), ('getting', 49), ('thought', 48), ('man', 48), ('mom', 48), ('listen', 48), ('guess', 47), ('fine', 47), ('around', 47), ('gotta', 46), ('believe', 46), ('two', 45), ('someone', 45), ('home', 45), ('call', 45), ('boy', 45), ('son', 44), ('put', 43), ('fix', 43), ('always', 43)

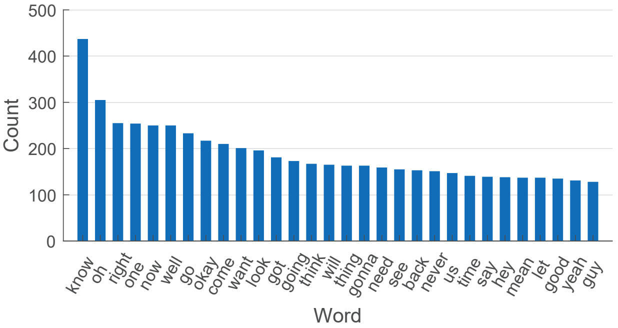


Figure 2. Top 30 words on V2C-Animation along with the counts.

*Partial of the work was performed when Qi Chen was an intern at Pazhou Lab.

†Corresponding authors.

A.2. Distribution of Emotion Labels

Following the categories of FER-2013 [4] (a dataset for human facial expression recognition), we divide the collected video/audio clips into 8 types (*i.e.*, 0: angry, 1: disgust, 2: fear, 3: happy, 4: neutral, 5: sad, 6: surprise, and 7: others). The number and distribution of each emotion label can be found in Table 1 and Figure 3, respectively.

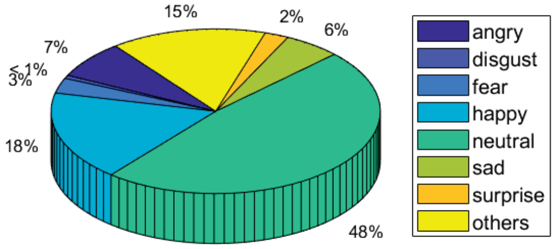


Figure 3. Distribution of emotion labels on V2C-Animation.

Emotion	angry	disgust	fear	happy
Count	756	64	305	1799
Emotion	neutral	sad	surprise	others
Count	4919	572	240	1562

Table 1. Counts of the emotion labels on V2V-Animation dataset.

A.3. Distribution of Utterance Length

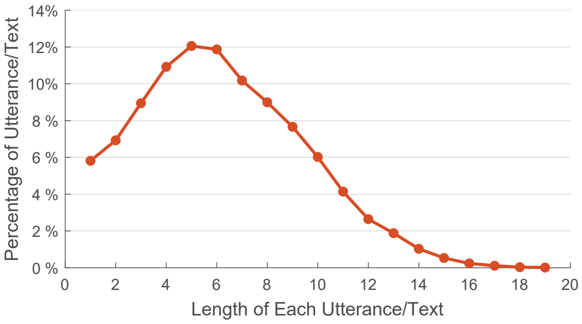


Figure 4. Distribution of utterance/text length.

Figure 4 exhibits the distribution of utterance/text length on V2C-Animation dataset, which shows that most utterance range from 3 to 8 words. Besides, we also list the number of utterance/text and their corresponding percentages in the following (format: *length, count, percentage*): (1, 594, 5.81%), (2, 708, 6.93%), (3, 914, 8.95%), (4, 1116, 10.92%), (5, 1232, 12.06%), (6, 1213, 11.87%), (7, 1040, 10.18%), (8, 919, 8.99%), (9, 783, 7.66%), (10, 615, 6.02%), (11, 423, 4.14%), (12, 270, 2.64%), (13, 192, 1.88%), (14, 105, 1.03%), (15, 54, 0.53%), (16, 24, 0.23%), (17, 11, 0.11%), (18, 3, 0.03%), (19, 1, 0.01%)

A.4. More Examples of Subtitle and Video Clip

We show several examples of how to crop movies based on a corresponding subtitle file. Here, we use an SRT type subtitle file. Besides the subtitles/texts, the SRT file also contains starting and ending time-stamps to ensure the subtitles match with video and audio. The sequential number of subtitle (*e.g.*, No. 726 and No. 1340 in Figure 5) indicates the index of each video clip. Based on the SRT file, we cut movie into a series of video clips via FFmpeg toolkit [11] (an automatic audio and video processing toolkit).

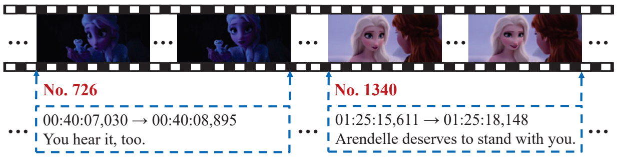


Figure 5. Examples of how to cut a movie into a series of video clips according to subtitle files. Note that the subtitle files contain both starting and ending time-stamps for each video clip.

A.5. Samples of Character's Emotion

Figure 6 shows some samples of the reference videos on V2C-Animation dataset with their corresponding emotions.

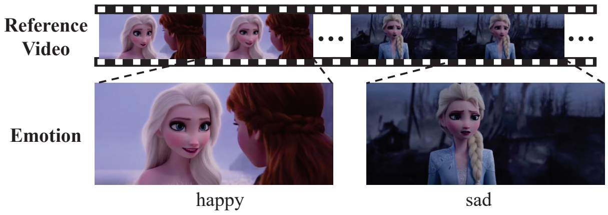


Figure 6. Samples of the character's emotion (*e.g.*, happy and sad) involved in the reference video. Here, we take *Elsa* (a character in movie *Frozen*) as an example.

A.6. List of Animated Movies and Characters

As shown in Figure 7, we report all the names of our collected animated movies with their corresponding characters/speakers on the V2C-Animation dataset.

B. V2C-Animation vs. Related Datasets

To compare the differences between the collected V2C-Animation dataset and several related datasets (*i.e.*, LJ Speech, LibriSpeech and LibriTTS), we visualise the pitch tricks of the samples from our dataset and others. Due to the varying lengths of audios, for a fair comparison, we cut two seconds of audio from each sample. As shown in Figure 8, the audio pitches from the existing datasets are more smooth and their ranges of frequency (Hz) are narrower than ours.

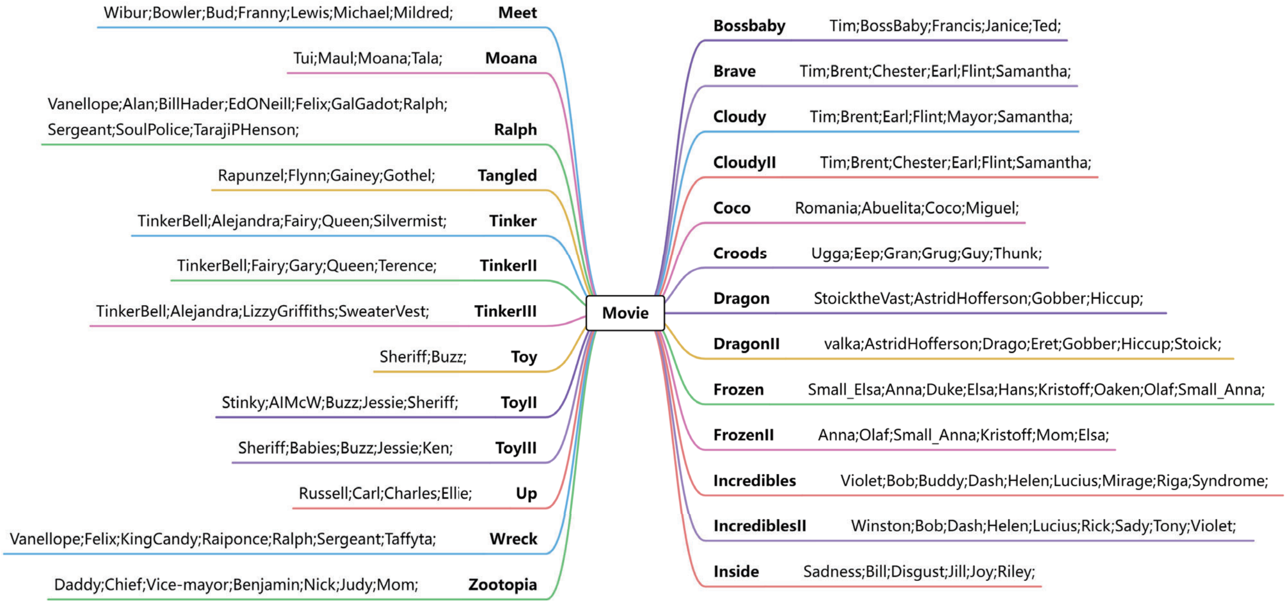

Movie
Meet
Wibur;Bowler;Bud;Franny;Lewis;Michael;Mildred;
Moana
Tui;Maul;Moana;Tala;
Ralph
Vanellope;Alan;BillHader;EdONeill;Felix;GalGadot;Ralph;
Sergeant;SoulPolice;TarajiPHenson;
Tangled
Rapunzel;Flynn;Gainey;Gothel;
Tinker
TinkerBell;Alejandra;Fairy;Queen;Silvermist;
TinkerII
TinkerBell;Fairy;Gary;Queen;Terence;
TinkerIII
TinkerBell;Alejandra;LizzyGriffiths;SweaterVest;
Toy
Sheriff;Buzz;
ToyII
Stinky;AlMcW;Buzz;Jessie;Sheriff;
ToyIII
Sheriff;Babies;Buzz;Jessie;Ken;
Up
Russell;Carl;Charles;Ellie;
Wreck
Vanellope;Felix;KingCandy;Raiponce;Ralph;Sergeant;Taffyta;
Zootopia
Daddy;Chief;Vice-mayor;Benjamin;Nick;Judy;Mom;
Bossbaby
Tim;BossBaby;Francis;Janice;Ted;
Brave
Tim;Brent;Chester;Earl;Flint;Samantha;
Cloudy
Tim;Brent;Earl;Flint;Mayor;Samantha;
CloudyII
Tim;Brent;Chester;Earl;Flint;Samantha;
Coco
Romania;Abuelita;Coco;Miguel;
Croods
Ugga;Eep;Gran;Grug;Guy;Thunk;
Dragon
StoicktheVast;AstridHofferson;Gobber;Hiccup;
DragonII
valka;AstridHofferson;Drago;Eret;Gobber;Hiccup;Stoick;
Frozen
Small_Elsa;Anna;Duke;Elsa;Hans;Kristoff;Oaken;Olaf;Small_Anna;
FrozenII
Anna;Olaf;Small_Anna;Kristoff;Mom;Elsa;
Incredibles
Violet;Bob;Buddy;Dash;Helen;Lucius;Mirage;Riga;Syndrome;
IncrediblesII
Winston;Bob;Dash;Helen;Lucius;Rick;Sady;Tony;Violet;
Inside
Sadness;Bill;Disgust;Jill;Joy;Riley;

Figure 7. Movies with the corresponding speakers/characters on the V2C-Animation dataset.

C. More Visual Results of Mel-spectrogram

We provide more visualised results of our V2C-Net with comparisons against the baseline method and ground truth. As shown in Figure 9, the mel-spectrograms generated by the proposed V2C-Net are more similar to the ground-truth ones. Note that the baseline method FastSpeech2 does not take the reference videos (*i.e.*, emotions) as inputs, which may lead to some misses of the prosody involving in the videos. The results further demonstrate the effect of the reference video when generating speech with rich emotions. Besides, the ranges of pitch for the mel-spectrograms are various due to the different emotions. For example, the pitch of the mel-spectrogram would be more drastic with the emotions "*happy*" or "*sad*", while it would be more smooth if the emotion is "*neutral*".

D. Implementation Details

For the speaker encoder f_{spk}, we use the same architecture as [12], comprising three LSTM layers. The audio encoder maps a sequence of mel-spectrogram frames, derived from the reference audio, to a vector with a fixed dimension of 256. We optimise the model with a generalised end-to-end speaker verification loss, which ensure features from the same speaker are more similar than ones from different speakers. For the emotion encoder f_{emo}, we use a conventional I3D model [1], trained on our V2C-Animation dataset [1] with 64×10^3 iterations and final output a vector with 1024 dimensions. For our synthesizer, we train the text encoder f_{txt} and the synthesizer in an end-to-end manner with 16 batch size and 2×10^6 iterations on our proposed V2C-Animation dataset. We train all models on a single GPU device (GeForce RTX 3090).

E. Details of Vocoder

To synthesise the waveform of the speech from our generated mel-spectrogram, we use HiFi-GAN [7] as our vocoder. The HiFi-GAN model is based on Generative Adversarial Networks (GANs) [3], which consists of one generator and two discriminators, *i.e.*, a multi-period discriminator (MPD) and a multi-scale discriminator (MSD).

The generator of HiFi-GAN can be divided into two major modules: a transposed convolution (ConvTranspose) network and a multi-receptive field fusion (MRF) module. Specifically, we first upsample the mel-spectrogram by ConvTranspose, which seeks to take an alignment between the length of the output features and the temporal resolution of raw waveforms. Then, we feed the upsampled features into MRF module, which consists of multiple residual blocks [5], and take the sum of outputs from these blocks as our predicted waveform. Here, the residual blocks with different kernel sizes and dilation rates are used to ensure different receptive field.

For the two discriminators, the multi-period discriminator (MPD) contains several sub-discriminators, where each sub-discriminator handles a specific periodic part of the input audio. By contrast, the multi-scale discriminator (MSD) proposed in MelGAN [8], consisting of three sub-discriminators, tries to capture the consecutive patterns and long-term dependencies from input audio.

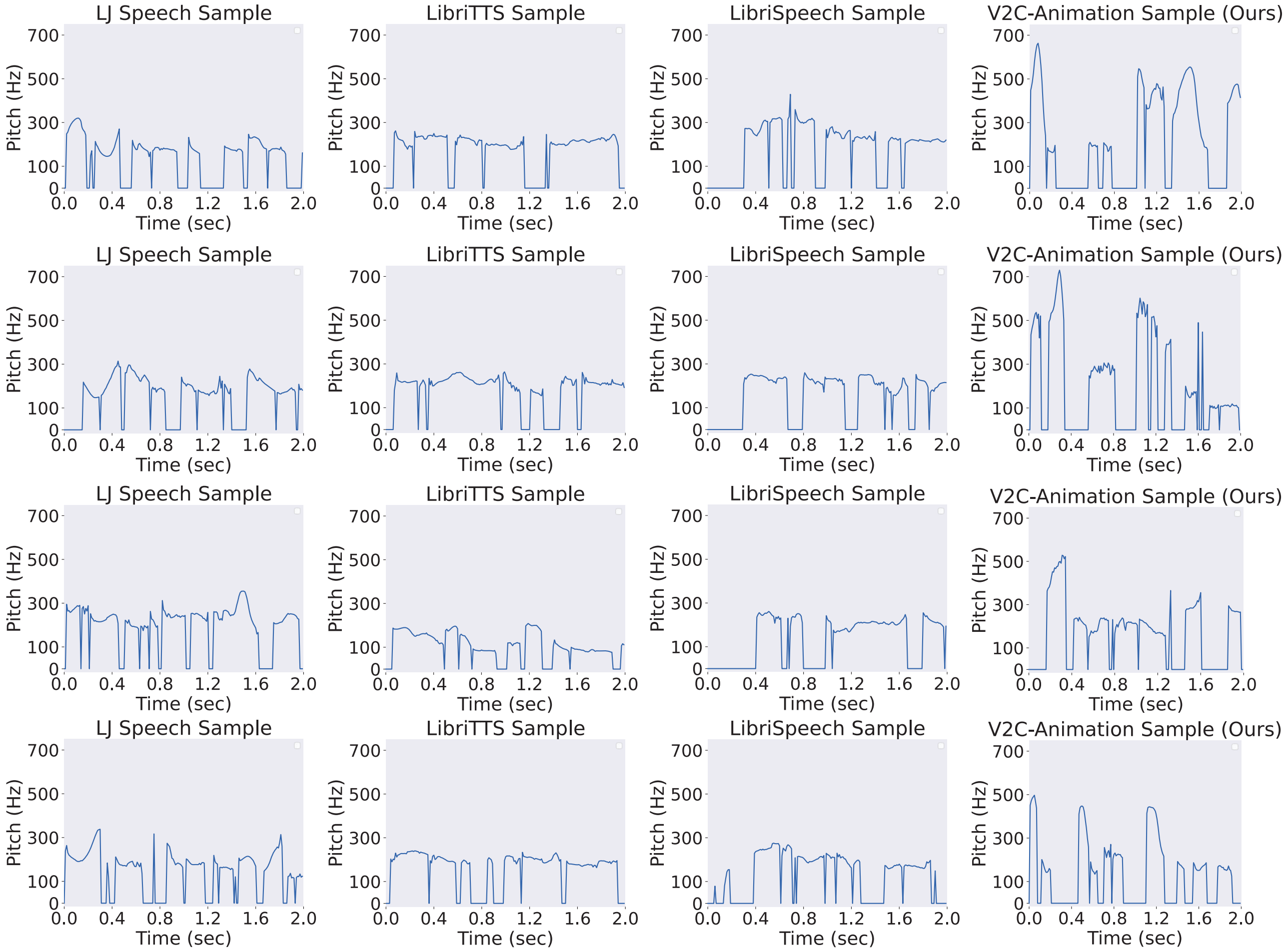


Figure 8. Visual comparison between our V2C-Animation dataset and the related datasets (*i.e*., LJ Speech, LibriSpeech and LibriTTS). A pitch of 0 Hz refers to an unvoiced segment.

The generator and discriminators are trained adversariaally, aiming to improve the training stability and the model performance. Specifically, the vocoder (*i.e*., HiFi-GAN) is optimised via the objective function that contains an LSGAN-based loss [9], a mel-spectrogram loss [6], and a feature matching loss [8]. In practice, we use the vocoder (*i.e*., HiFi-GAN) pretrained on the LibriSpeech dataset [10].

F. Limitations of MCD-DTW

As mentioned in the main paper, *"MCD-DTW would achieve a better value as long as there is a match between two speeches. This is not reasonable as a better generated speech should have a similar length with the ground truth"*. To verify that, we take an example of the generated speech with its ground truth (GT) in Figure 10. The corresponding MCD-DTW and MCD-DTW-SL (ours) are 8.69 and 11.76, respectively. Smaller value of MCD-DTW demonstrates that it focuses more on the alignment within pitch and energy while ignores whether the duration is reasonable or not. Our metric (MCD-DTW-SL) can reflect the alignment on both feature and length.

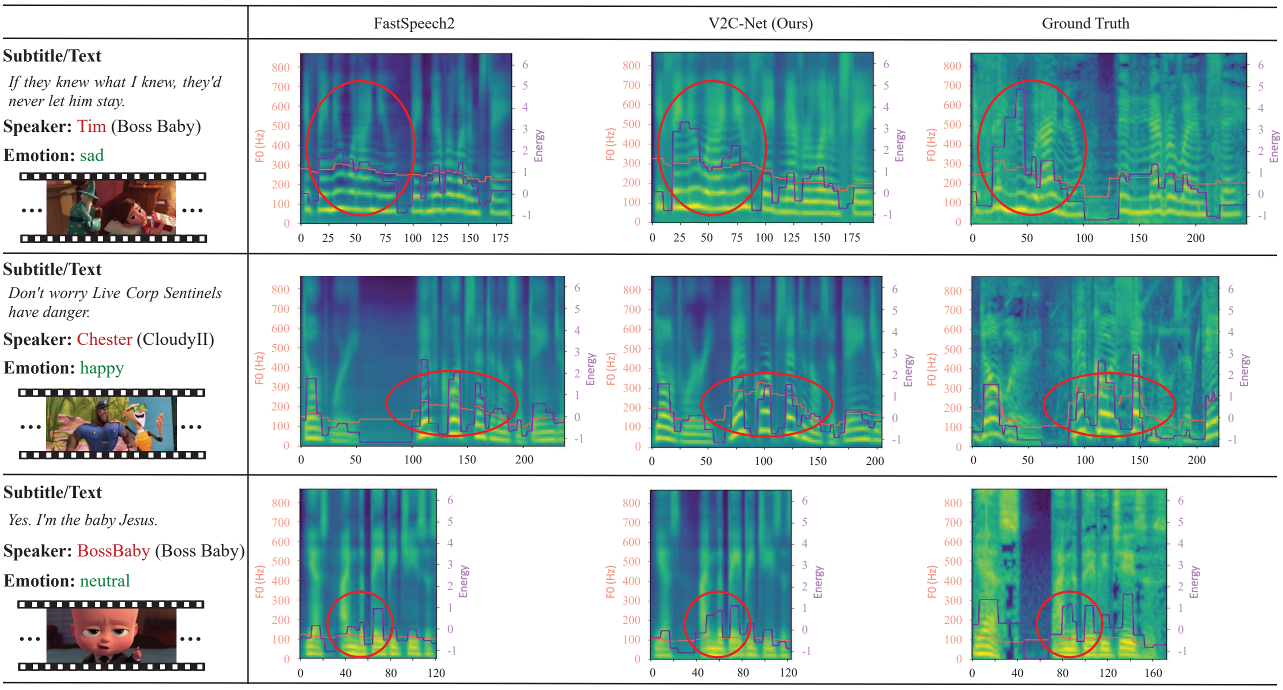
References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 3

[2] Glen Coppersmith and Erin Kelly. Dynamic wordclouds and vennclouds for exploratory data analysis. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 22–29, 2014. 1

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014. 3

[4] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Chal-



Subtitle/Text

If they knew what I knew, they'd never let him stay.

Speaker: Tim (Boss Baby)

Emotion: sad

Subtitle/Text

Don't worry Live Corp Sentinels have danger.

Speaker: Chester (CloudyII)

Emotion: happy

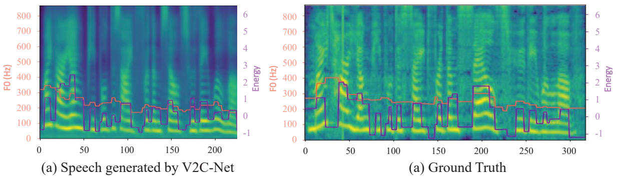
Subtitle/Text

Yes. I'm the baby Jesus.

Speaker: BossBaby (Boss Baby)

Emotion: neutral

Figure 9. More visualised mel-spectrograms of generated and ground-truth audios. The orange curves are F_0 contours, where F_0 denotes the fundamental frequency of audio. The purple curves refer to energy (volume) of audio. Horizontal axis is the duration of audio. We highlight the main difference via red circle.



(a) Speech generated by V2C-Net

(a) Ground Truth

Figure 10. Similar shape of pitch and energy curves with different length between the synthesised speech and the ground-truth one.

lenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124, 2013. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 4

[7] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *NeurIPS*, 2020. 3

[8] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *NeurIPS*, 2019. 3, 4

[9] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017. 4

[10] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *ICASSP*, 2015. 4

[11] Suramya Tomar. Converting video formats with ffmpeg. *Linux Journal*, page 10, 2006. 2

[12] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *ICASSP*, pages 4879–4883, 2018. 3