Adding Crowd Noise to Sports Commentary using Generative Models

NEIL SHAH, DHARMESHKUMAR M AGRAWAL, and NIRANJAN PEDANEKAR,

TCS Research, Tata Consultancy Services Pvt. Ltd., India

Crowd noise forms an integral part of a live sports experience. In the post-COVID era, when live audiences are absent, crowd noise needs to be added to the live commentary. This paper exploits the correlation between commentary and crowd noise of a live sports event and presents an audio stylizing sports commentary method by generating live stadium-like sound using neural generative models. We use the Generative Adversarial Network (GAN)-based architectures such as Cycle-consistent GANs (Cycle-GANs) and Mel-GANs to generate live stadium-like sound samples given the live commentary. Due to the unavailability of raw commentary sound samples, we use end-to-end time-domain source separation models (SEGAN and Wave-U-Net) to extract commentary sound from combined recordings of the live sound acquired from YouTube highlights of soccer videos. We present a qualitative and a subjective user evaluation of the similarity of the generated live sound with the reference live sound.

CCS Concepts: • **Information systems** → **Multimedia content creation**; • **Applied computing** → *Sound and music computing*.

Additional Key Words and Phrases: sports commentary, audio stylization, ambient noise generation, neural source separation, generative networks

ACM Reference Format:

Neil Shah, Dharmeshkumar M Agrawal, and Niranjan Pedanekar. 2021. Adding Crowd Noise to Sports Commentary using Generative Models. In *LIQUE 2021: Life Improvement in Quality by Ubiquitous Experiences Workshop, together with IMX 2021: ACM International Conference on Interactive Media Experiences, June 21–23, 2021, NY*. ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Live events include sports matches, music concerts, award ceremonies, stage plays, and talks. They are typically attended by a few hundred to several thousand spectators. The most significant and ever-present portion of the reactions occurs as sounds. Crowd noise is believed to be influencing game outcomes [9], player performance [7], predicting attendance [8], and viewer engagement [6]. The recent crisis of the COVID-19 pandemic has impacted live events in a significant way [1]. A sports telecast with the only commentary and without crowd noises is not as effective as both of them combined. To re-create the live experience, many solutions exist in this regard [2–4]. However, these solutions rely on an audio engineer mixing the sound in real-time or audiences cheering in real-time and can be computationally expensive.

Exploring the correlation between the audio dynamics of a commentary and the crowd noise is a central part of our study. For example, when a commentator explicitly describes in an emotionally charged speech that a goal has occurred in a football match, the crowd also bursts into an uproar. Since we do not have separate channel recordings for audio commentary and crowd noise, as broadcasters often have, we use a neural source separation method to separate the commentary from a live recording and then employ generative networks for producing a stadium-like sound experience. We believe that the automatic generation of crowd noise leads to an interesting path of investigation which

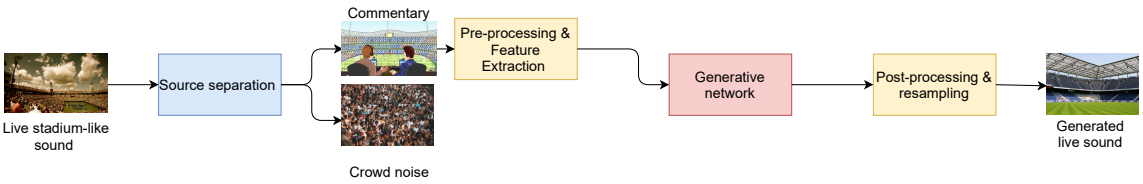

Fig. 1. A block diagram of the proposed system.

has the potential to benefit several applications in the entertainment field (like live concerts, karaoke performances, and festivals). Our contribution is:

- We outline a method (a first to our knowledge) to recreate live stadium-like sound including crowd noise and commentary using a recorded sports commentary as an input. we also present qualitative analysis of the generated live sound.
- We present an initial survey with 31 users to evaluate the ability of the generated live sound to match the ground truth.

2 PROPOSED WORK

Figure 1 shows a schematic representation of the proposed method. It contains two parts: a neural source separation model for separating the commentary from the combined sound and a generative network for generating the live stadium-like sound from the given commentary. We use Speech Enhancement Generative Adversarial Network (SEGAN) [10] and Wave-U-Net [13] to extract a commentary sound from a combined sound. We also study the impact of employing Cycle-GAN [14], and Mel-GAN [11] in generating a live stadium-like sound experience.

3 EXPERIMENTAL SETUP

We use live commentary audio for soccer matches downloaded from YouTube. For training, we use 9 sound files spanning 11 hours. The validation and test set comprises 1 sound files each, for 1.40 and 1.45 hours, respectively. We use available pre-trained network weights of SEGAN and Wave-U-Net. We analyze the generative network trained at 5^{th} and 10^{th} epoch for the Cycle-GAN. We conduct the analysis on different combinations of these models, viz. 5^{th} and 10^{th} epoch trained SEGAN-Cycle-GAN as SEGAN-5Cycle-GAN and SEGAN-10Cycle-GAN, and 5^{th} and 10^{th} epoch trained Wave-U-Net-Cycle-GAN as Wave-U-Net-5Cycle-GAN and Wave-U-Net-10Cycle-GAN, respectively. We also train Mel-GAN with the source separation outputs, namely, SEGAN-Mel-GAN and Wave-U-Net-Mel-GAN. For the Cycle-GAN, we extract the Short-Time Fourier Transform (STFT) features with 64 frequency bins, 2 *millisecond (ms)* hop size, and 8 *ms* window size. For the Mel-GAN, we extract 192 Mel filters with 12 *ms* hop size and 72 *ms* window size, as in [11]. We use similar training parameters for the GAN as suggested in their respective works [5, 11]. We stop the training after visually observing the validation spectrograms since further training resulted in the loss of crucial harmonics.

4 RESULTS

We show the spectrogram of the generated live sound using Wave-U-Net in figure 2, as the SEGAN fails to simulate the real-life audience experience. The Mel-GAN does not preserve any linguistic information and does not generate crowd noise across the frames between 33 – 35 seconds. However, the crowd noise generated by the Wave-U-Net-5Cycle-GAN

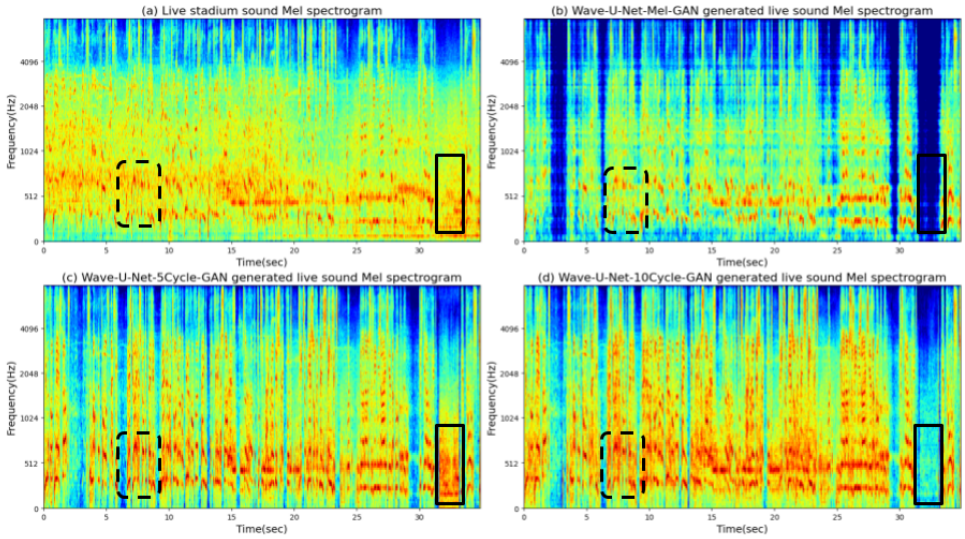

Fig. 2. Mel spectrograms of (a) ground-truth live. Generated live sound from (b) Wave-U-Net-Mel-GAN, (c) Wave-U-Net-5Cycle-GAN and (d) Wave-U-Net-10Cycle-GAN. The solid rectangle show the generated crowd noise in silence, the dash rectangle shows the linguistic presence.

in the silence region and the ability of Wave-U-Net-10Cycle-GAN in preserving the linguistic content confirms that the cycle-consistent loss enforces better forward-backward consistency.

4.1 Subjective Evaluation

The distortions introduced by the source separation network will impact the crowd noise generation. Figure 3 (a) shows the Mean Opinion Score (MOS) [12] analysis for the intelligibility of the commentary extracted from SEGAN and Wave-U-Net. We also report in 3 (b) the MOS analysis for the similarity between the ground-truth live sound and the model-generated live sound. 31 subjects (19 males and 12 females with age-group between 21 to 40 years) with no known hearing impairment participated in all the subjective tests. For both the test, we ask the subjects to rate on a five-point scale (1 = not at all intelligible/similar; 2 = hardly few seconds of audio samples are intelligible/similar; 3 = partly intelligible/similar; 4 = mostly intelligible/similar, and 5 = completely intelligible/similar).

We select 5 random sound samples of each 15 second duration from both the system. There is a 118.3 % improvement in the MOS ratings obtained using Wave-U-Net over the SEGAN as shown in figure 3 (a). Though the spectrogram analysis between the Wave-U-Net-5Cycle-GAN and Wave-U-Net-10Cycle-GAN suggests a trade-off, the subjective tests provide a 3.52 % relative improvement in the similarity using Wave-U-Net-10Cycle-GAN. Though the best-separated commentary is rated as mostly intelligible, the generated live sound is still at best partly similar to the actual live sound.

5 CONCLUSION AND FUTURE WORK

In this paper, we present, to our knowledge, the first attempt to generate a stadium-like sound experience using neural generative models over only the commentary sound. We examine the spectrograms of generated live samples and find that the models capture to some extent the correlation between the commentary and the crowd noise. This leads to a

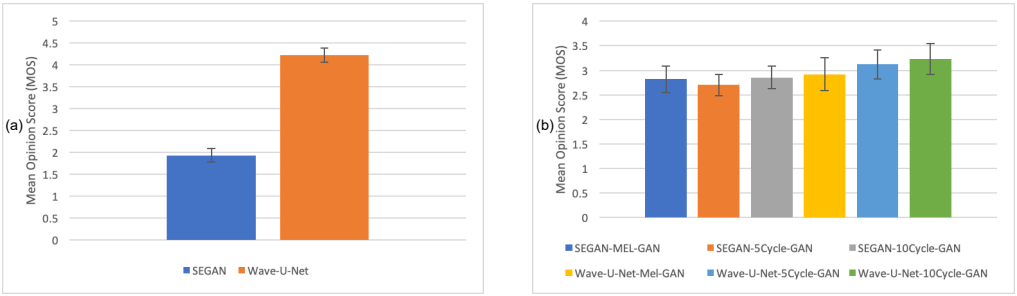

Fig. 3. (a) MOS analysis for the intelligibility of commentary extracted by various source separation systems. (b) MOS analysis for the similarity of the generated live sound by various generative networks with the ground-truth live sound.

promising research direction where we envisage realistic live sample generation with crowd noises based on the live commentary. Possible future work could include exploring the combination of visual and textual features extracted from the commentator's and audience's chatter for enhancing the stadium sound experience.

REFERENCES

[1] [n.d.]. Canceled Events Due to the Coronavirus: A Complete List. https://www.vulture.com/2020/05/events-cancelled-coronavirus.html. (Accessed on 05/11/2020).

[2] [n.d.]. FIFA 20 crowd noise to be used for real Premier League games - here's how it'll work | GamesRadar+. https://www.gamesradar.com/fifa-20-crowd-noise-to-be-used-for-real-premier-league-games-heres-how-itll-work/. (Accessed on 02/04/2021).

[3] [n.d.]. German Bundesliga broadcasts: Where the 'crowd noise' feed comes from and how they made it. https://www.espn.in/football/german-bundesliga/story/4102971/german-bundesliga-broadcasts-where-the-crowd-noise-feed-comes-from-and-how-they-made-it. (Accessed on 02/04/2021).

[4] [n.d.]. *MyApplause*: An app allowing users to cheer from their homes for enhanced sporting experience. https://myapplause.app/en/myapplause-en/. (Accessed on 03/26/2021).

[5] Gino Brunner, Yuyi Wang, Roger Wattenhofer, and Sumu Zhao. 2018. Symbolic music genre transfer with cyclegan. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 786–793.

[6] Sheranne Fairley and B David Tyler. 2012. Bringing baseball to the big screen: Building sense of community outside of the ballpark. *Journal of Sport Management* 26, 3 (2012), 258–270.

[7] Mack Hagood and Travis Vogan. 2016. The 12th man: Fan noise in the contemporary NFL. *Popular communication* 14, 1 (2016), 30–38.

[8] John Hall, Barry O'Mahony, and Julian Vieceli. 2010. An empirical model of attendance factors at major sporting events. *International Journal of Hospitality Management* 29, 2 (2010), 328–334.

[9] Alan M Nevill, Nigel J Balmer, and A Mark Williams. 2002. The influence of crowd noise and experience upon refereeing decisions in football. *Psychology of Sport and Exercise* 3, 4 (2002), 261–272.

[10] Santiago Pascual, Antonio Bonafonte, and Joan Serra. 2017. SEGAN: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452* (2017).

[11] Marco Pasini. 2019. Melgan-vc: Voice conversion and audio style transfer on arbitrarily long samples using spectrograms. *arXiv preprint arXiv:1910.03713* (2019).

[12] ITUT Rec. 1994. P. 85. a method for subjective performance assessment of the quality of speech voice output devices. *International Telecommunication Union, Geneva* (1994).

[13] Daniel Stoller, Sebastian Ewert, and Simon Dixon. 2018. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185* (2018).

[14] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.