Multi-Task Audio-Driven Facial Animation

Youngsoo Kim*
Game Dev. AI Team
NARC, Netmarble
kimys@netmarble.com

Shounan An*
Game Dev. AI Team
NARC, Netmarble
ethan.an@netmarble.com

Youngbak Jo
Game Dev. AI Team
NARC, Netmarble
howisee@netmarble.com

Seungje Park
Game Dev. AI Team
NARC, Netmarble
psj2354@netmarble.com

Shindong Kang
Game Dev. AI Team
NARC, Netmarble
shindong1992@netmarble.com

Insoo Oh
Magellan Division
NARC, Netmarble
ioh@netmarble.com

Duke Donghyun Kim
Netmarble AI Revolution
Center (NARC), Netmarble
duke@netmarble.com

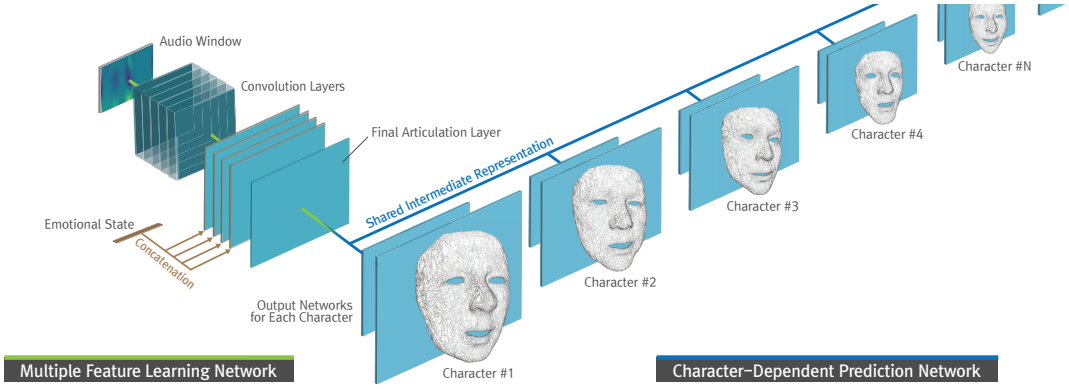

Figure 1: The network architecture of MulTaNet for multi-task audio-driven facial animation.

ABSTRACT

We propose an effective method to solve multiple characters audio-driven facial animation (ADFA) problem in an end-to-end fashion via deep neural network. In this paper each character's ADFA considered as a single task, and our goal is to solve ADFA problem in multi-task setting. To this end, we present MulTaNet for multi-task audio-driven facial animation (MTADFA), which learns a cross-task unified feature mapping from audio-to-vertex that capture shared information across multiple related tasks, while try to find within-task prediction network encoding character-dependent topological information. Extensive experiments indicate that MulTaNet generates more natural-looking and stable facial animation, meanwhile shows better generalization capacity to unseen languages compare to previous approaches.

CCS CONCEPTS

• **Computing methodologies** → **Deep neural networks**; **Animation**; **Audio processing**;

*Both authors contributed equally to this research.

SIGGRAPH '19 Posters, July 28 - August 01, 2019, Los Angeles, CA, USA

ACM ISBN 978-1-4503-6314-3/19/07.
https://doi.org/10.1145/3306214.3338541

KEYWORDS

Facial animation, multi-task learning, multiple feature learning

ACM Reference Format:
Youngsoo Kim, Shounan An, Youngbak Jo, Seungje Park, Shindong Kang, Insoo Oh, and Duke Donghyun Kim. 2019. Multi-Task Audio-Driven Facial Animation. In *Proceedings of SIGGRAPH '19 Posters.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3306214.3338541

1 INTRODUCTION

In game industry it is a labour-intensive job to produce high quality facial animation in various interactive scenarios such as dialogue and cinematic scenes. For usually there are lots of game characters and different characters has its own unique characteristics. Consequently it is vital to produce natural-looking facial animation for each game character. Recently deep learning based audio-driven facial animation (ADFA) generation is getting popular for its useful behaviour. The typical ADFA method [Karras et al. 2017] is to learn a mapping from input audio signal to output 3D vertex positions of a single character's fixed topology mesh via deep neural networks, and we consider ADFA as a single task in this paper.

Multi-task learning (MTL) [Caruana 1997] is a prominent methodology for learning a relationship of multiple related tasks to improve the overall generalization performance compared to learning each of them independently. Given audio and visual data recorded from different characters, our motivation is that related tasks share many common audio-to-vertex mapping patterns, while each task also has its own unique 3D topology characteristics. However, in literature

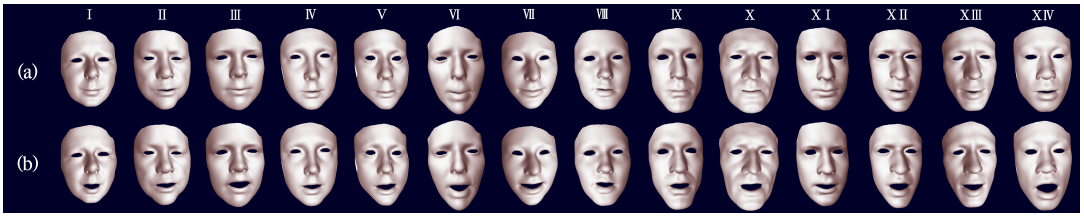

Figure 2: MulTaNet results of all 14 characters from BIWI dataset. (a) closed mouth (b) open mouth

there is no concrete work to resolve multi-task audio-driven facial animation (MTADFA) problem. If we exploit the task relationship, it will be beneficial to learn multiple tasks jointly since the knowledge contained in a task can be leveraged by other tasks. To this end, we propose MulTaNet to solve MTADFA. The contributions of this paper is summarized as follows.

- To the best of our knowledge, this is the first work try to solve ADFA problem in multi-task setting.
- MulTaNet jointly learns a unified latent representation of cross-task, which captures shared information across multiple related tasks and within-task prediction network try to encode task-dependent characteristics.
- MulTaNet produces more stable facial animation and shows better generalization capacities to unseen languages such as Korean, Chinese, Japanese.

2 MULTANET

MulTaNet (Figure 1) mainly consists of two parts, the first part is multiple feature learning (MFL) network that tries to learn a unified speaking style encoding feature map from multiple related tasks. The second one is a character-dependent prediction (CDP) network, including one-hot vector representation to encode character identity and final 3D vertex position prediction layers for each character respectively.

The architecture of MFL network is similar to formant analysis + articulation network of [Karras et al. 2017], and we call [Karras et al. 2017] method as KarrasNet in this paper. Following KarrasNet we choose linear predictive coding (LPC) [Deng and O'Shaughnessy 2003] as audio features to feed into MFL network, and convert 256ms audio signal with 16k sampling rate into 64×32 LPC feature. To distinguish the ambiguities of audio track from various facial expressions, emotional state vector was trained for each frame as well. Different with KarrasNet, in MFL network we don't concatenate emotional vector to the last convolution layer.

In CDP network we have one-hot vector to assign character identity to each pair of audio and facial animation training samples. During training each CDP network will be trained with corresponding character's data only. Consequently CDP network encodes its own unique 3D topology characteristics.

3 EXPERIMENTS

We evaluate the effectiveness of MulTaNet with publicly available audiovisual dataset: BIWI [Fanelli et al. 2010]. For comparison we implemented KarrasNet, and found it is quite difficult to train the full network end-to-end. Because the output of KarrasNet has very high dimensionality (e.g. 15066 outputs) and the network is quite deep, hence emotional state vectors trained faster than formant analysis and articulation network. Consequently emotional state vector encodes most of the information and dominate the training process to produce 3D facial vertex positions directly even without input LPC features. To solve these technical issues, final output layers of each CDP network was initialized with principal component analysis (PCA) [Turk and Pentland 1991] projection vectors of training data from each character respectively. Then we freeze the weights of last layer in CDP network for 8 epochs to train MFL network first, in order to make MFL network learns a unified mapping from audio feature to 3D vertex even in a multi-task environments. About training procedure, we set learning rate as 10^{-3} for the first 3 epochs and 10^{-4} until 4700 epochs, from epoch 8 all layers start to be trained and we fine tuned MulTaNet for last 300 epochs with 10^{-5}.

Exact same loss terms were used as KarrasNet: position, motion and regularization. With the training details as mentioned earlier, all three loss terms converge smoothly throughout training procedure. The results of final facial animation from all 14 characters are illustrated (Figure 2). For fair comparison, we also implemented ensembling method [Karras et al. 2017] for KarrasNet. However KarrasNet still shows unstable facial animation e.g. lip and mouth tremor was observed when we fed voice data from other characters and/or languages which is different from training data.

4 CONCLUSIONS AND FUTURE WORK

We have presented MulTaNet for learning cross-task and within-task knowledge in audio-driven facial animation application to improve the overall stability and generalization capacity to unseen languages. Experiments on benchmark BIWI dataset confirm the useful behaviour of our proposed method. An interesting future research direction of this work might be to design a neural emotion transfer architecture to further improve expressiveness of facial animation with controllable emotions.

REFERENCES

Rich Caruana. 1997. Multitask Learning. *Machine learning* 28, 1 (July 1997), 41–75.

Li Deng and Douglas O'Shaughnessy. 2003. *Speech Processing: A Dynamic and Optimization-Oriented Approach.* CRC Press.

Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. 2010. A 3D Audio-Visual Corpus of Affective Communication. *IEEE Transactions on Multimedia* 12, 6 (Oct. 2010), 591–598.

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion. *ACM Transactions on Graphics* 36, 4 (July 2017). https://doi.org/10.1145/3072959.3073658

Matthew Turk and Alex Pentland. 1991. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience* 3, 1 (1991), 71–86.