Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Demonstrations Track

A Speech-driven Sign Language Avatar Animation System for Hearing Impaired Applications

Li Hu , Jiahui Li , Jiashuo Zhang , Qi Wang , Bang Zhang , Ping Tan
XR Lab, Alibaba Group
{hooks.hl, jiahui.lijiahui, jiashuo.zjs, wilson.wq, zhangbang.zb, xingye.tp}@alibaba-inc.com

Abstract

Sign language is the communication language used in hearing impaired community. Recently, the research of sign language production has made great progress but still need to cope with some critical challenges. In this paper, we propose a system-level scheme and push forward the implementation of sign language production for practical usage. We build a system capable of translating speech into sign language avatar. Different from previous approach only focusing on single technology, we systematically combine algorithms of language translation, body gesture animation and facial avatar generation. We also develop two applications: Sign Language Interpretation APP and Virtual Sign Language Anchor, to facilitate easy and clear communication for hearing impaired people.

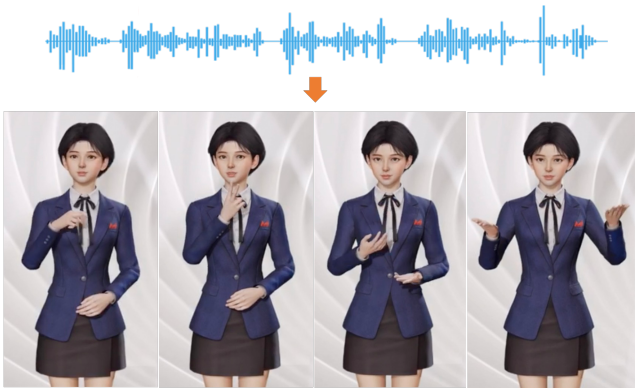

Figure 1: Our proposed system can translate speech into sign language avatar.

1 Introduction

Sign language, the language of communication for the hearing impaired community, is a visual language with complex grammatical structures that includes gestures, expressions and body movements. According to the World Health Organization (WHO) report in 2020, there are more than 466 million deaf people in the world[Kushalnagar, 2019]. Sign language production, converting spoken language to continuous sign sequences, is therefore essential in involving the deaf in the predominantly spoken language of the wider world.

In recent years, academia has proposed several schemes for sign language production. After Avatar Approaches constructs sign language data, sign language generation is realized through animation production technology, but these generated data cannot be expanded, and professional knowledge is required to check the generated data[Bangham *et al.*, 2000; Cox *et al.*, 2002; Ebling and Glauert, 2013]. Then generative models along with some graphical techniques, such as Motion Graph, are being recently employed which could better organize data-driven sign language generation, but still require a large amount of data and become more computationally complex as the amount of data increases[Lee and Shin, 1999]. To address these issues and improve computational performance, [Camgöz *et al.*, 2018; Guo *et al.*, 2018; Stoll *et al.*, 2020] adopts a deep learning model-based approach, Neural Machine Translation (NMT) approaches, which solve sign language synthesis as a translation task, and use traditional machine translation methods to handle the problem. However, this method performs poorly on long sentences, very common vocabulary and unseen sentence patterns. The method of Conditional image/video generation directly generates videos/pictures[Kataoka *et al.*, 2016; Karras *et al.*, 2019; van den Oord *et al.*, 2016; Vasani *et al.*, 2020], which makes sign language production more intuitive and the process simpler, but its model is more complex, and it is difficult to find a suitable objective function and corresponding optimization method.

Another problem is that previous approaches mainly focus on single topic of gesture generation. However, there are two significant issues that prevent sign language production from practical application. First, the grammar of sign language is entirely different from natural language. Translating natural language words one by one is not sign language, but gesture language. Besides, sign language expression is not only limited to body pose or hand gesture. Facial expression is equally important in sign language. Thus, we aim to tackle the problem systematically, combining technology of natural language processing, computer graph and machine learning. We hope to make our efforts to benefit society through practical applications.

In this paper, we propose a novel system, translating speech

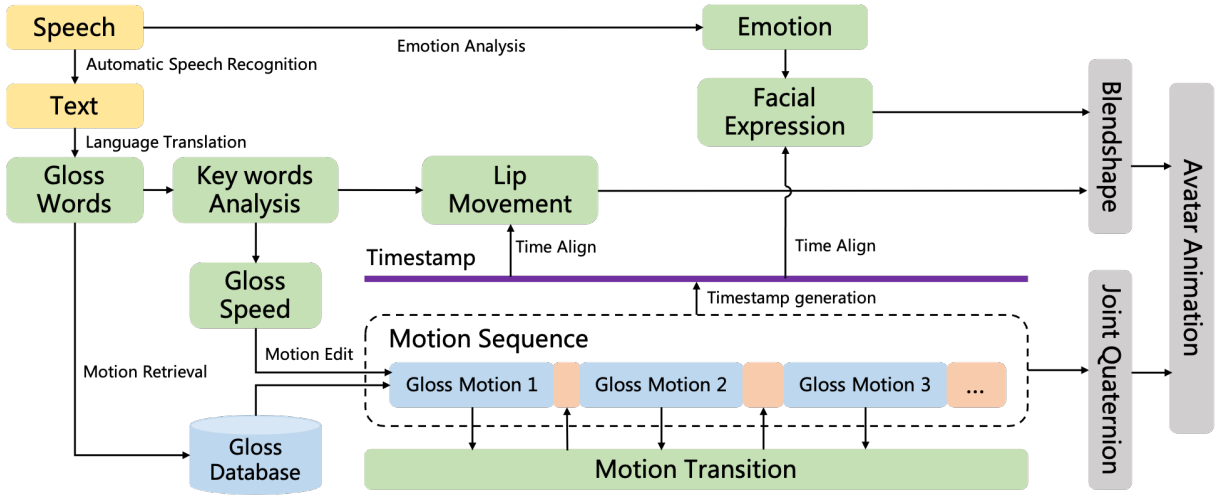

Figure 2: The overview of our proposed system design.

into sign language avatar as shown in Figure 1.

- Given speech signal, we perform automatic speech recognition and develop a language translation method to translate natural language into sign glosses.
- For sign language production, we combine avatar-based approach and learning-based approach to get an accurate and human-like avatar animation.
- We employ facial synthesis algorithm for better sign language expression.
- We introduce two applications equipped with our proposed system: Sign Language Interpretation APP and Virtual Sign Language Anchor.

2 System Design

2.1 Language Translation

Sign language expresses and conveys meanings through the combination of glosses. In this section, we aim to translate natural language into sign glosses. The natural language text is generated from speech with Automatic Speech Recognition (ASR). We apply state-of-the-art language translation method BART[Lewis *et al.*, 2020] as the network structure. The training data is manually collected and annotated from sign language practitioners and hearing impaired people. We also develop a novel data augmentation method to tackle the data deficiency problem. Specifically, we randomly delete, re-order some words in text and replace similar words or glosses based on the word embedding, generating 1 million text and pseudo gloss pairs at pre-training. Besides, different from natural language, the quantity of sign language vocabulary is much smallermany natural words need to use multiple glosses to express. Thus we establish the Expert Alignment Knowledge to tackle this problem, which can significantly reduce the difficulty of model learning. The translated glosses will be matched with an unique gloss ID which represents an unique gloss motion. We also attach an additional network branch to predict the key words probability used in the subsequent animation system. The translation network is trained in an end-to-end manner.

2.2 Sign Language Production

Previous approaches to sign language production can be divided into avatar-based approach and learning-based approach. Avatar-based approach can generate realistic sign production as the sign language action can be captured by special cameras and sensors precisely. But the combination of sign language is countless and the expensive cost is unacceptable. Learning-based approach aims to produce sign language action by training a neural networks. However, the result is uncontrollable and is far from practical application. Our approach is to combine avatar-based approach and learning-based approach. We collect 2000 gloss motions from professional sign language practitioner with motion capture devices to ensure the fine-grained accuracy. To concatenate glosses into sentences, we develop a learning-based method for motion transition.

Avatar-based Gloss Motion Retrieval and Editing

Based on the collected gloss motions, we establish an offline gloss database for Gloss Motion Retrieval. The saved gloss motion data is represented as unit quaternion of different upper body joints. The retrieval signal is the pre-defined gloss IDs calculated by proposed language translation model.

As the gloss motions are pre-recorded with a fixed speed, directly combining them will lead to an under-articulated and unnatural movements and the robotic motion of the aforementioned avatars can make viewers uncomfortable, due to the uncanny valley. Hearing impaired people often rhythmically play sign language based on the words they want to emphasize in a sentence. Thus, we tackle this problem by adaptively editing the motion speed of every gloss word. We utilize the probability predicted by language translation module. A low probability means that the gloss word is less important in this sentence then we improve its speed rate by downsampling

frames of the motion data and vice versa. The relation between key words probability and downsampling rate is manually set in our experiments.

Learning-based Gloss Motion Transition

In this subsection, we introduce how to concatenate gloss motion into sentence motion. Since there are various movement states at the start and end of different gloss motions, it is difficult for interpolation-based methods to get human-like transition speed and trajectory. Thus we develop a data-driven deep neutral network to generate the transition motion for better smoothness. We define the gloss motion transition problem as: given body joints in the last few frames of the previous gloss motion and the first few frames of the next gloss motion, predict the transitional motion.

We model the transition function with a neural network based on Transformer[Vaswani *et al.*, 2017]. When training the neural network, we randomly sample one gloss motion from the offline gloss database. To construct training samples, We randomly erase 5-20 frames in the middle of body motion. The erased body joints will be applied as the groundtruth during training. To construct the input of the network, we pad the empty body joints with zero value. The network input consists of three parts: the last 20 frames before erased frames, the padded 5-20 frames and the first 20 frames after erased frames. The loss function is mean square error(MSELoss). At inference, we first determine the length of transition frames based on the distance and instantaneous velocity of the wrist joint. Then transition result can be predicted by the trained Transformer. The learning-based transition methods can better handle the complicate and diverse sign gesture combinations.

2.3 Facial Movement Generation

When hearing impaired people playing sign language, they always come with rich facial expressions to express their mood or to show what they want to emphasize. Thus, we develop a lip movement generation module to make the virtual character simultaneously read the word in dumb when playing sign language gesture. The facial movement is represented as blendshape. If key words probability are larger than 0.5, the glosses will be selected to generate corresponding lip movements. We translate those glosses to Chinese phoneme. The pre-defined mouth shape will be mapped to the phoneme as well as be merged into lip movement sequence with cross-fade smoothness. Furthermore, we also extract emotion labels from the input speech. Emotional facial movement avatars are manually collected and animated, such as neutral, happy, confused, worried, etc. They are also merged into the whole facial movement sequence to make the virtual character more vivid.

3 Applications

In this section, we introduce our two developed applications equipped with our proposed sign language animation system. We also explain and demonstrate the system and applications via a short video at https://youtu.be/oVtZKC4H0LA.

Figure 3: Sign Language Interpretation APP on mobile device.

Figure 4: Virtual Sign Language Anchor in TV broadcast.

Sign Language Interpretation APP

The Sign Language Interpretation APP can perform the bidirectional translation work between sign language and natural speech. Our system can be embedded into mobile device for translating the natural speech into the avatar animation. Normal hearing users input a voice message into the APP, and then show the virtual character to the hearing impaired users. It is expected as the future communication tool for hearing impaired people and normal hearing people. The demonstration is illustrated in Figure 3.

Virtual Sign Language Anchor

Sign language anchor is widely applied in video broadcast such as news report, sports commentary and program host. Previous practice requires a real professional presenter to synchronously play sign language according to the speech, which is expensive on personnel costs. To tackle the problem, our Virtual Sign Language Anchor can automatically generate sign language animation based on either speech or text, which is more extensible and low-cost compared to the manual solution. Typical product is illustrated in Figure 4, where the virtual anchor commentates the Olympic Games with sign language in TV broadcast.

4 Conclusion

We propose a system capable of translating speech into sign language avatar. We systematically combine algorithms of language translation, body gesture animation and facial avatar generation and develop two applications to facilitate easy and clear communication for hearing impaired people.

References

[Bangham *et al.*, 2000] J Andrew Bangham, SJ Cox, Ralph Elliott, John RW Glauert, Ian Marshall, Sanja Rankov, and Mark Wells. Virtual signing: Capture, animation, storage and transmission-an overview of the visicast project. In *IEE Seminar on speech and language processing for disabled and elderly people (Ref. No. 2000/025)*, pages 6–1. IET, 2000.

[Camgöz *et al.*, 2018] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7784–7793. Computer Vision Foundation / IEEE Computer Society, 2018.

[Cox *et al.*, 2002] Stephen Cox, Michael Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, and Sanja Abbott. Tessa, a system to aid communication with deaf people. In Vicki L. Hanson and Julie A. Jacko, editors, *Proceedings of the ACM Conference on Assistive Technologies, ASSETS 2002, Edinburgh, Scotland, UK, July 8-10, 2002*, pages 205–212. ACM, 2002.

[Ebling and Glauert, 2013] Sarah Ebling and John Glauert. Exploiting the full potential of jasigning to build an avatar signing train announcements. In *Third International Symposium on Sign Language Translation and Avatar Technology*, 2013.

[Guo *et al.*, 2018] Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. Hierarchical lstm for sign language translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4401–4410. Computer Vision Foundation / IEEE, 2019.

[Kataoka *et al.*, 2016] Yuusuke Kataoka, Takashi Matsubara, and Kuniaki Uehara. Image generation using generative adversarial networks and attention mechanism. In *15th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2016, Okayama, Japan, June 26-29, 2016*, pages 1–6. IEEE Computer Society, 2016.

[Kushalnagar, 2019] Raja Kushalnagar. Deafness and hearing loss. In *Web Accessibility*, pages 35–47. Springer, 2019.

[Lee and Shin, 1999] Jehee Lee and Sung Yong Shin. A hierarchical approach to interactive motion editing for human-like figures. In Warren N. Waggenspack, editor, *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1999, Los Angeles, CA, USA, August 8-13, 1999*, pages 39–48. ACM, 1999.

[Lewis *et al.*, 2020] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics, 2020.

[Stoll *et al.*, 2020] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *Int. J. Comput. Vis.*, 128(4):891–908, 2020.

[van den Oord *et al.*, 2016] Aäron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Koray Kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4790–4798, 2016.

[Vasani *et al.*, 2020] Neel Vasani, Pratik Autee, Samip Kalyani, and Ruhina Karani. Generation of indian sign language by sentence processing and generative adversarial networks. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pages 1250–1255. IEEE, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.