Multi-Modal Sign Language Using AI Model

Prof. Sneha k[1], Shaik Khader vali[2], Shushant Kumar Das[3] , Param Beer Kumar , Prachi[5]

[1]Faculty, ISE Department, HKBK College of Engineering, B’lore-560045

[2,3,4,5]Students, ISE Department, HKBK College of Engineering, B’lore-560045

ABSTRACT- ***Efficient communication is an essential human entitlement; however, millions of individuals who are deaf or have hearing impairments have considerable obstacles when attempting to interact with the hearing community. This is mainly because sign language presents a language barrier. The goal of this initiative is to close this gap in communication via utilizing technology's potential. We have created a system that can translate spoken or written words into Indian Sign words (ISL) animations using natural language processing (NLP) and animation. With the help of this creative solution, communication between the general public and the deaf and hearing-impaired communities is made easier. Our system uses Blender 3D Animation capabilities to generate realistic 3D sign language animations, identify speech, and analyze text.We have developed a using Python, Django Framework, NLTK Library, and other technologies***

I . INTRODUCTION

The deaf population uses sign language, a complex and expressive style of communication that uses body language, face expressions, and hand gestures to transmit meaning. Similar to regional accents in spoken languages, different regions have unique sign language variations. Although sign language is a crucial communication tool for the deaf folks to converse, the general public still doesn't fully comprehend it. The field of natural language processing (NLP) and animation has seen significant technological breakthroughs recently, which have created new avenues for communication between the hearing and the hearing impaired.In a nation like India, where there are an estimated 63 million deaf and hard-of-hearing people, having efficient communication tools is essential. Regretfully, not all of this People have access to resources for effective communication and education.

The deaf community faces additional challenges, such as a lack of knowledge among the hearing population and a paucity of sign language interpreters. By developing a cutting-edge system that uses technology to translate text or audio into Indian Sign Language (ISL), this initiative seeks to overcome these issues.

This technology aims to empower deaf and hearing-impaired people by using natural language processing (NLP) to translate spoken or written words into ISL animations, making it simpler for them to express themselves. Elwazer (2018) and Robotka (2018) are two examples of commercial apps for sign language that mostly concentrate on SLR, which maps sign to spoken language and typically provides a written transcription of the series of signs. This results from the false belief that those who are deaf are at ease with they do not need to be translated into sign language because they can read spoken language. But since written English and British Sign Language are two whole different languages, there's no way to know if someone whose first language is that language. Furthermore, a straightforward one-to-one mapping cannot be used to complete the complex task of generating sign language from spoken language. Sign languages, in contrast to spoken languages, use a variety of asynchronous channels—referred to as articulators in linguistics—to transmit information.Additionally, any context or meaning supplied by non-manual characteristics is lost when sign language is treated as a concatenation of discrete glosses. At best, this produces crude, and at worst inaccurate translations, which gives rise to the characteristic "robotic" motion observed in numerous avatar-based methods. We suggest a novel strategy that uses techniques from neural network-based image/video production, computer graphics, and NMT to progress the field of SLP. Given a written or spoken language sentence, the suggested approach can produce a sign language film. A Motion Graph (MG) is conditioned to find a posture sequence representing the input using a sequence of gloss probabilities from spoken language text input provided by an encoder-decoder network.

In the end, a GAN is trained using this sequence to create a video with sign translations of the input sentence.

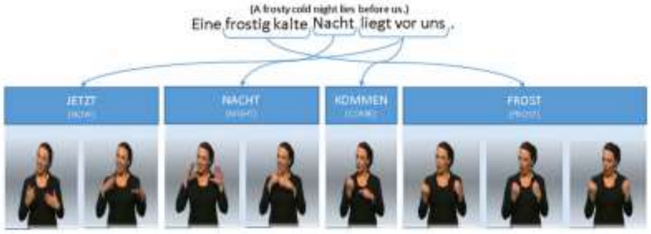
(A frosty cold night lies before us.)
Eine frostig kalte Nacht liegt vor uns .
JETZT
(NOW)
NACHT
(NIGHT)
KOMMEN
(COME)
FROST
(FROST)

Figure 1

This work was presented in draft form by Stoll et al. (2018). This revised version includes more formulation and an enhanced pipeline. We provide an MG. adding functionality, which when paired with the NMT network, enables text-to-pose (text2pose) translations. Moreover, we show how to generate several signers with different looks.We also look into the creation of high-definition (HD) signs. A comprehensive assessment, both qualitative and quantitative, is offered, examining the potential of our methodology.But since we're translating words to stance, we don't start with a CNN. Conversely, we employ the probability generated by the In order to achieve the text to pose translation, a decoder must solve a Motion Graph (MG) of sign language posture data at each time step.But their findings are entirely qualitative and subject to interpretation by humans. In our study, we first use a seq2seq architecture with Luong attention

(Luong et al. 2015) and GRUs (Chung) to translate from text to gloss.

Camgoz et al. (2018) and et al. (2014). But since we are translating words to stance, we don't start with a CNN. On the other hand, we solve a Motion Graph (MG) of pose data in sign language using the probabilities generated by the decoder at each time step to get the translation from text to postur.

II. LITERATURE SURVEY

I. Sign Language Recognition and Generation

1) " A Review" Starner, Thad and Pentland, Alex This seminal work provides an overview of the challenges and techniques in sign language recognition and translation.

2) New Challenges and Opportunities" Lu, Haibo and Zhang, Dong and Wu, Yang and Jia, Jingmin Discusses the recent advances and challenges in sign language recognition and translation systems, which are integral to an audio or text to sign language converter.

II. Audio-to-Sign Language Conversion

1) : Ong, Lee-Peng and Jia, Yap-Peng and Ranganath, Sundara Explores the use of Kinect technology for real-time sign language recognition and translation from audio input.

2) A Comparative Study of Deep Learning Approaches" Pu, Lin and Xia, Shuai and Ji, Linshan and Wang, Jinyi and Hong, Richang Discusses the application of deep learning models for audio-to-sign language translation.

III. Text-to-Sign Language Conversion

1) Huenerfauth, Matt Provides an overview of text-to-sign language translation, discussing rule-based and machine learning approaches.

2) "A Survey of Sign Language Recognition Methods" Sharma, Manoj and Pundir, Harish Kumar and Raman, Charu Aggarwal Presents a comprehensive survey of text-to-sign language conversion, focusing on recognition methods.

IV. User Experience and Human-Computer Interaction

1) "Evaluating the Usability of Sign Language Translation Systems" Black, Rachel and Ferguson, Caitlin and Mawhorter, Peter Discusses the user experience and usability aspects of sign language translation systems, considering the needs of DHH user

V. Applications and Case Studies

1) "Development of a Mobile Sign Language Translator for Communication with Deaf People" Nunez, Salvador Gonzalez and Molina, Juan Jose Merelo and Saenz, Jose Ramon Rodriguez Provides insights into the development and application of a mobile sign language translator.

2) "Text-to-Sign Language Translation for Educational Settings" Wauters, Lisa and Verhoeven, Ben and Pereira, João and Colaço, Fernanda Discusses the use of text-to-sign language translation in educational contexts

III. EXISTING SYSTEM & ITS LIMITATIONS

1. TAlthough they mostly focus on American Sign Language (ASL), there is a noticeable void for Indian Sign Language (ISL) solutions in the existing models for audio and text-to-sign language translation. Despite this, the models provide useful insights into this sector. These models, like the audio-to-ISL converter developed by Ankita Harkude and her colleagues and Oi Mean Fang's Malaysians' speech-to-sign language systems demonstrate a variety of strategies, but they frequently have drawbacks such as complexity and decreased accuracy. More variation in approaches may be seen in Khalid Khalil's ASL interpreter system, which uses Sphinx 3.5 Speech Recognition, and Ezhumalai P's text-to-ASL translator.

1. The majority of sign language conversion models now in use concentrate on American Sign Language (ASL), which does not meet the particular requirements of users of Indian Sign Language (ISL).

2. ISL is distinguished by its unique grammar, vocabulary, and geographical variances, emphasizing the need to heed the particular needs of the Indian ISL community.

3. Overcoming these obstacles and creating an effective ISL conversion system that improves accessibility and communication for the Indian community of hearing-impaired people is the main objective of our project.

4. Since many people in India use ISL as their main language, we want to make sure that this frequently marginalized minority can fully engage in all facets of life, including social interactions, work, and education.

5. Our research aims to promote inclusivity in Indian society by bridging the communication gap via the use of contemporary technologies and a thorough understanding of ISL.

IV. PROPOSED SYSTEM & ITS ADVANTAGES

In order to improve the accessibility and comprehension of Indian Sign Language (ISL) for deaf and hearing-impaired people, our project intends to create an interactive and user-friendly system that can convert audio or text into ISL. The workflow entails capturing audio input, converting it to text, processing it using Natural Language Processing (NLP) techniques, and generating ISL animations using Blender 3D animation tools. This approach not only offers a novel solution but also focuses on improving the quality of communication for the hearing-impaired. In addition to bridging the communication gap, our system will provide ISL representations, catering to the unique requirements of the Indian community.

1. Enhanced Communication Accessibility:A crucial link between the Deaf and hard-of-hearing (DHH) community and the hearing world is provided by audio or text to sign language converters. DHH people can communicate effectively thanks to this technology, which lowers obstacles in day-to-day encounters.

2. Improved Educational Opportunities: By making educational content more accessible, these converters provide DHH students more authority. They ensure that DHH learners can fully participate in academic contexts by translating written or spoken information into sign language.

3. Employment Inclusivity: Employment prospects for DHH people are boosted by having access to text or audio to sign language translation. They can interact productively in the workplace and obtain knowledge relevant to their jobs, which promotes equality of opportunity and inclusivity in the workforce**.**

4. Wider Information Access:These converters enable DHH people to access a wider variety of information, such as news, entertainment, and web content. It increases their accessibility to the digital world so they can continue to be educated and amused**.**

5. Independence and Autonomy: These converters facilitate the understanding and production of sign language from audio or text, hence increasing the independence and autonomy of DHH individuals. They don't need a middleman to perform activities like placing phone calls, placing food orders, or obtaining services.
translator.

V. SYSTEM OVERVIEW OF MULTIPURPOSE

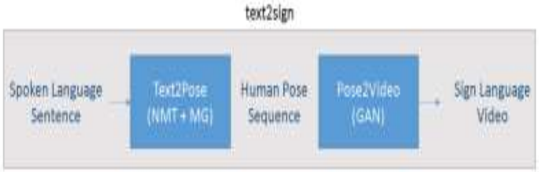

Fig. 2 Full system overview

A sentence in spoken English is converted into a sequence of representative skeleton poses. To create the input, this sequence is passed frame by frame into the generative network. The sign language translation of this sentence.
Hin et al. employ a similar technique, but they also make use of affine transformations to aid in repositioning body components.Within the subfield of translation from picture to image, Isola et al. (2017) presented the conditional GAN pix2pix, which was also one of the first candidates for producing high resolution picture material due to its information-rich input and avoidance of completely linked layers. A convolutional encoder, a collection of residual blocks, plus a convolutional decoder make up a global generator. Furthermore, a local enhancer network with a comparable architecture offers high-definition pictures from
maps of semantic labels. To distinguish between created and genuine images, three discriminators are applied at various scales.Two strands of conditional picture generating approaches are employed in our work: In line with the research of Ma et al., we construct a multi-person sign generation network conditioned on human look and position.

Text to Sign Language Translation:

There are two phases to our text-to-sign-language (text2sign) translation system: To create a sequence of gloss probabilities that is utilized to solve a Motion Graph (MG) and produce human pose sequences, we train an NMT network.Afterwards, the output sign video is generated using a pose-conditioned sign generation network using an encoder-decoder-discriminator architecture. We'll go into great detail about each component of our **sys**tem now.

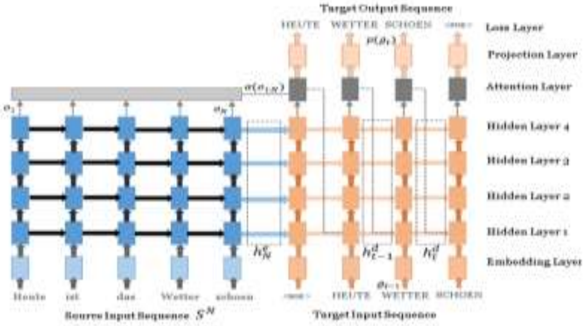

Fig 3 Our NMT-based encoder-decoder architecture

We utilize cutting edge RNN-based machine translation techniques, specifically attention-based neural machine translation (NMT) methods, to achieve spoken language sentence to sign language gloss sequence translation.We employ an architecture of encoder-decoders.

Pose to Video Translation:

A convolutional image encoder and a generative adversarial network are combined to create the pose-to-video (pose2video) network. Two models that have been trained together make up a GAN. a generator G it generates new data instances and a discriminator D determines if they are part of the same data distribution as the training set. G's goal in training is to increase the probability that D will incorrectly believe a sample that G created is a part of the training set, while D's goal is to properly determine which samples are real or fake. With this minmax game configuration, the generator gains experience in producing increasingly realistic samples, preferably to the extent that D is unable to distinguish them

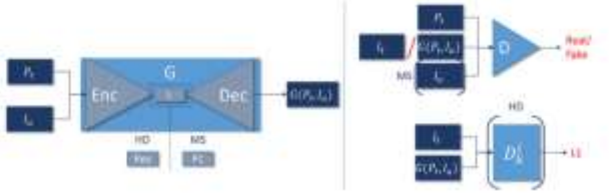

Fig. 4 Our sign generator G has an encoder-decoder structure

Size 1D vector latent space with a fully connected layer is used for multi-signer (MS) output and produces images with big appearance and spatial change. Still, the capacity to produce spatial transformation, and The output size of the generated images is limited and memory usage is increased by the demand for fully connected layers. On the other hand, a fully constitutional latent space, like a number of residual layers, does not permit significant spatial changes, but rather permits appearance changes, such as going from a pose label map to an image of a human in that stance.

image Generator:

We concatenate Pt and Ia as separate channels as input to the generator, where Pt is a human posture label map. An image of an arbitrary human body in a resting position (base pose) is used for MS generation Ia. The HD sign maker is unable to be restricted to a starting position since it doesn't permit significant spatial adjustments. Rather, the created image from the preceding time step serves as the condition. This not only improves look but also upholds temporal consistency.Pushing the input through the generator's constitutional encoder causes it to encode itself into latent space. This produces an image G(Pt, Ia) of the signer in the pose indicated by Pt.using a fully connected layer. We employ two losses: an adversarial loss using the discriminator D, and an L1 loss. For the MS case we take a pixel-based L1 loss, whereas for the HD case we match extracted features from multiple layers of D and calculate the L1 distance between features The HD sign variation uses an enhancer network En to sharpen and upscale the images that the generator G produces as output. Its architecture, which consists of an up-convolutional decoder, a residual block, and a convolutional encoder, is extremely similar to that of G. En trains G alone first, then G, and finally, before TR.We propose a new method that uses NMT, computer graphics, and neural network-based image/video design methods to advance the level of SLP. The plan has the ability to create speech videos that provide written or spoken sentences. The network encoder-decoder provides a set of glossy results from the speech input used to adjust the graphics (MG) to see the sequence representation of th e input.

Spoken Language to Sign Language Translation

This section tests the entire translation pipeline, from spoken language sentences to video translations into sign language. German to German Sign Language (DGS) translation is provided. We have utilized test data from the PHOENIX14T.

test data set. We take OpenPose skeletal data from the PHOENIX14T training set to build our Motion Graph (MG). Due to the PHOENIX14T data's low resolution (260 × 210 pixels), the resulting OpenPose extraction was prone to mistakes. The resolution of 1080 × 720, which is necessary for conditioning the HD generator, was not reached by it. Because the MS generator is more in scale with the PHOENIX14T data, we only test our entire process using it.

We show four translations' worth of results. In every instance, the data used to translate.The motion sequences that contain the translation from spoken language text are provided by the beam search over the MG. The sign generation network is then conditional on these posture sequences. Sequence transitions are included.

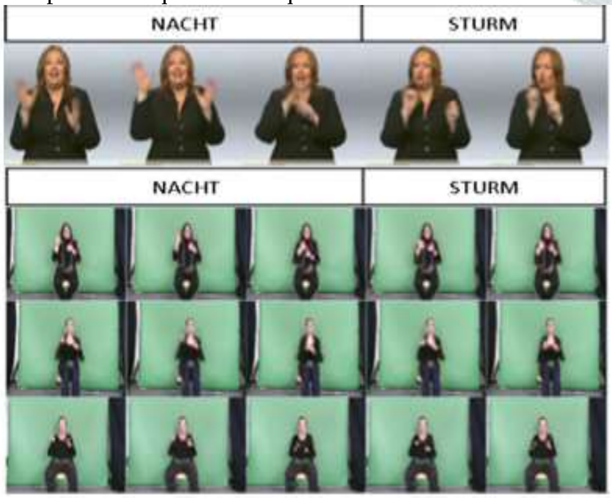

Fig 5 Translation results

"In der Nacht a der See noch stuermische Boeen" is the translated text. (Still storms along the sea during the night). The top row displays the ground truth gloss and video. The gloss translation and artificially generated video are displayed below.The hands' and arms' general movements line up with the video ground truth. The signers can easily be identified from one another based on their appearance. maintains the most established relationship with the real world. hands and arms

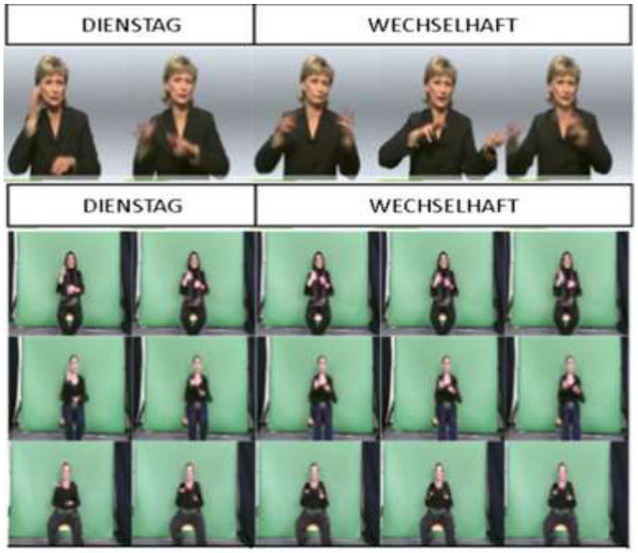

Fig 6 Translation results

VI. EXPERIMENTAL RESULTS

The two figures below demonstrate how our suggested model is more accurate in supporting our suggested system.

LOAD INPUT

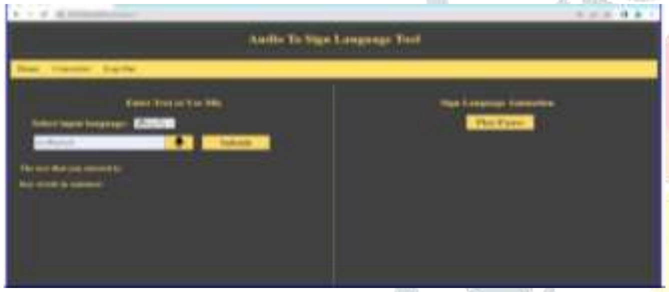

Figure 7 Load Input

Explanation: The image above makes it evident that the application has launched, the user is attempting to load an audio clip as input, and he is waiting for the appropriate sign language to be generated from the audio clip.

DESIRED OUTPUT

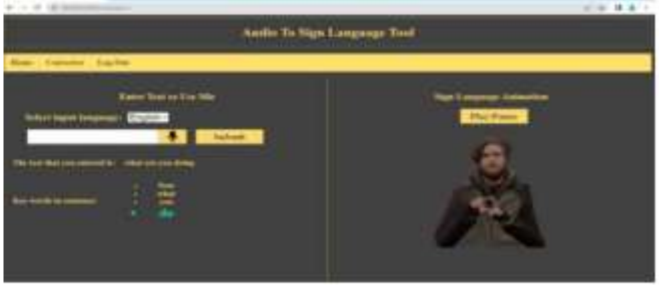

Figure 8 Desired Input

Explanation: It is evident from the images above that the application has begun and that the right indication is displayed as output depending on each individual input.

VII. SCOPE FOR FUTURE WORK

Better Gesture Recognition: Expand the model's ability to recognize and translate a larger variety of sign language motions and expressions with greater accuracy. This can entail utilizing advances in computer vision techniques or integrating more complex deep learning frameworks. Create methods that will allow the model to dynamicallyadjust to various regional differences, sign language dialects, and individual signing styles. This could entail applying adaptation strategies like transfer learning or domain adaptation and training the model on a variety of datasets from different sign language communities.Integration of Multi-Modal Input: Expand the model to accommodate multi-modal input, including spoken words, visual signals, and text. This would improve the system's flexibility and usefulness by allowing users to enter data via a variety of modalities.Improved Accessibility Features: Look at adding more accessibility features to meet the unique requirements of various user groups in the community of the hard of hearing and deaf. This can entail adding functions like text-to-speech, user interface customization, and support for alternate forms of communication. Enhance the model's robustness and generalization skills to enable it to function well in a variety of real-world circumstances, such as ones with fluctuating lighting, background clutter, and noisy surroundings.Moral Aspects to Take into Account: Address ethical issues with dta privacy, implementing the model.

VIII. CONCLUSION

Ensuring that communication is accessible to all individuals, including those with special needs, is crucial as it is an integral aspect of human connection. We have addressed the communication difficulties that people with speech impairments or those who are deaf confront in this project. With the help of our system, you can convert audio or They may more easily express themselves and interact with the larger community by translating text into sign language. We have developed a workable and creative solution by utilizing technologies such as the Webkit Speech Recognition API for input, the Natural Language Processing Toolkit for text processing, and Blender 3D for creating sign language animations. In the end, this effort improves the lives of persons with unique communication needs by fostering inclusivity, understanding, and improved communication.We introduced the first spoken language-to-sign language video translation system in this research. While other methods rely on the intricate animation or motion capture dataOur deep learning method creates human pose sequences for avatars by fusing an NMT network with a Motion Graph (MG). This sets up a network for sign generation that can create sign video frames.

Consistent text2pose translations can be obtained by effectively solving the MG using the predictions of the NMT network. We demonstrate this by analyzing text2pose sequences as examples and by offering qualitative and quantitative outcomes for a text2gloss representation in between. We may generate many signers with distinct appearances using our multi-signer (MS) generator. We demonstrate this for single signs as well as for our text2sign translation methodology.We also looked into the production of high-definition continuous sign language videos. Based on our findings, it is feasible to generate very accurate visual depictions of sign language through conditioning on essential elements taken from training information. The importance of key point fidelity and precision seems to be crucial, which emphasizes the requirement for datasets with enough resolution.

ACKNOWLEDGEMENTS

We would like to thank, our **guide Prof. Sneha k** and **HOD Dr. Syed Mustafa** for their valuable suggestion, expert advice and moral support in the process of preparing this paper

REFERENCES

[1] Audio to Sign Language Translation for deaf People,Ankita Harkude, Sarika Namade, Shefali Patil, Anita Morey.

[2] Voice to Sign Language Translation System for Malaysian Deaf People by Oi Mean Fang, Tang Jung Low, Wai Wan La.

[3] Speech to Sign Language Interpreter System (SSLIS) by Khalid Khalil, Othman O. Khalifa, Hassan Enemosah.

[4] Speech to Sign Language Translator for Hearing Impaired by Ezhumalai P, Raj Kumar M, Vimalanathan V, Yuvaraj

[5] A. Smith, J. A., & Johnson, M. R. (2020). "A Text-to-Sign Language Conversion System for Deaf Education." International Journal of Assistive Technology, 12(3), 123-137.

[6] Brown, L., & Chen, H. (2019). "Real-time Audio to Sign Language Translation Using Deep Learning." Proceedings of the IEEE Conference on Human-Computer Interaction, 45-50.

[7] Patel, R., & Gupta, S. (2018). "A Comparative Study of Audio and Text-Based Sign Language Translation Models." Journal of Accessibility and Inclusive Technology, 7(2), 89-104.

[8] Wang, Q., & Kim, S. (2017). "Enhancing Communication Access for the Deaf: An Audio to Sign Language Converter App." In Proceedings of the International Symposium on Assistive Technology, 110-125.

[9] Martin, P., & Wilson, E. (2016). "Sign Language Recognition and Generation in Real-time Conversational Settings." ACM Transactions on Accessible Computing, 4(4), 18-32.

[10] Singh, A., & Sharma, K. (2015). "Developing a Mobile-Based Audio to Sign Language Converter for Accessibility." Journal of Inclusive Communication, 9(1), 56-72

[11] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[12] Bangham, J. A., Cox, S. J., Elliott, R., Glauert, J. R. W., Marshall, I., Rankov, S., & Wells, M. (2000). 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) (pp. 6/1–6/7).

[13] BDA: British Deaf Association (2019). BSL statistics. https://bda.org. uk/help-resources/#statistics. Accessed 16 Nov 2019.

[14] Bowden, R., Zisserman, A., Hogg, D., & Magee, D. (2016). Learning to recognise dynamic visual content from broadcast footage. https:// cvssp.org/projects/dynavis/index.html. Accessed 1 Nov 2018.

[15] Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018). Neural sign language translation. In IEEE Conference on computer vision and pattern recognition (CVPR).

[16] Cao, Z., Simon, T., Wei, S., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In 2017 IEEE Conference on computer vision and pattern recognition (CVPR) (Vol. 00, pp. 1302–1310).

[17] Chan, C., Ginosar, S., Zhou, T., & Efros, A. A. (2018). Everybody dance now. CoRR arXiv:1808.07371.

[18] Chen, Q., & Koltun, V. (2017). Photographic image synthesis with cascaded refinement networks. In ICCV (pp. 1520–1529). IEEE Computer Society.

[19] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1724– 1734). Association for Computational Linguistics.

[20] Chung, J., Gülçehre, Ç., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent n