CONTENT-BASED OBJECTIVE EVALUATION OF ARTIFICIALLY GENERATED SIGN LANGUAGE VIDEOS

Neha Tarigopula[1,2] *Preyas Garg* [3] *Skanda Muralidhar*[1]
Sandrine Tornay[1] *Dinesh Babu Jayagopi*[3] *Mathew Magimai.-Doss*[1]

[1] Idiap Research Institute, Martigny, Switzerland
[2] Ecole polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland
[3] International Institute of Information Technology, Bangalore, India

ABSTRACT

Sign language is vital for communication within the deaf and hard-of-hearing community. Avatar-based methods and deep learning techniques like Generative Adversarial Networks have shown promise in generating sign language video content. One of the challenges in sign language generation is the evaluation of the generated video content. One possible solution is to subjectively evaluate using human raters. This is time-consuming and costly. The other possible solution is objective evaluation. In the literature, video quality metrics such as PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) and skeleton-based measures such as MSE have been proposed. A limitation of these approaches is that they do not provide information about the generated video content. In this paper, we propose a novel phonology-based approach that evaluates the generated video along different channels, namely, hand movement and handshape, which convey the linguistic information in sign language. More precisely, in this approach an objective score is obtained by extracting sequences of hand movement sub-units and handshape sub-units class conditional probabilities (posterior features) from the source and generated videos and comparing them using dynamic time warping. Our experimental studies demonstrate that the proposed objective scoring method yields a better correlation to subjective human ratings than PSNR, SSIM, and MSE-based metrics.

***Index Terms*—** Sign Language Generation, Video Evaluation, Sign Language Assessment, Generative Adversarial Networks, Subjective analysis

1. INTRODUCTION

Sign Language(SL) plays a crucial role in communication for the deaf and hard-of-hearing community. It is a visual mode of communication where information is conveyed through both manual and non-manual channels such as handshape, hand movement, body posture and facial expressions. Over recent years, demand for sign language content has surged due to increased awareness of accessibility needs in society. Since sign language is not universal, there exists variation in signs across languages, that can be attributed to various handshapes, hand movements, hand location, and orientation, and facial expressions which makes it difficult to build a unified robust system for recognition and generation of sign language data [1].

Sign language generation(SLG) encompasses generating photo-realistic sign language videos corresponding to a certain word, phrase, or sentence or simply creating more sign videos from existing data by transferring motion from a source subject to a target subject. There has been extensive use of Avatar-based methods, that synthesize sign videos using animated avatars. Rule-based approaches use the transcription of signs based either on SigML [2] representation or HamNoSys [3] representation to develop the avatar animations. [4, 5, 6, 7] are among some of the rule-based approaches. The effectiveness and acceptance of these avatar-based approaches in the deaf community are still questionable. This can be attributed to the unnaturalness of the avatar due to robotic movements and missing facial expressions. This has been addressed to a certain extent by extending JAsigning to include non-manual information [8]. For more realistic-looking avatars, MoCap data is used to animate 3D avatars [9]. With the evolution of deep learning, video generation has been revolutionized. Several image and video generation approaches that address SLG using neural architectures such as [10, 11, 12, 13, 14] have been proposed. Methods such as in [15] adapted motion transfer techniques [16] to generate sign language videos using GANs.

Having said that, like synthesis technologies such as, speech synthesis, SLG also faces the challenge of how to assess the quality and effectiveness of the generated videos. As the output of SLG is targeted to humans, one possible approach is subjective evaluations using human raters to judge correctness and acceptability. Involving humans in the loop in the development cycle of SLG poses additional challenges such as, cost, time and reproducibility. An alternate solution is use of objective assessment methods. In that direction, in the SLG literature use of of metrics such as, Peak-Signal to Noise Ratio (PSNR), Structural Similarity Measure (SSIM), temporal consistency, losses from GANs, and BLEU scores from reverse translation to evaluate their models [11, 13, 14] have been studied. Metrics such as PSNR, SSIM evaluate only the visual quality of the videos. The loss from GANs measures the pixel-wise difference between generated and real videos, it does not give any direct information on the quality of the content. Other content-based metrics include classification accuracy on the generated videos and BLEU score from reverse translation of sign language to spoken language and then comparing them to their input sequences. Such metrics evaluate the content of the videos, but the effective scores that are reported are dependent on the choice of architectures of the classification and translation models. Furthermore, such methods do not convey what aspect of SLG was correct or incorrect. Thus, beside visual quality metrics, there is a need for developing objective assessment/evaluation techniques that can assess the multiple channels

This work was funded by the SNSF through the Sinergia project SMILE-II *(Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment Phase-II)*, grant agreement CRSII5_193686. We thank all the collaborators in the project for their valuable work.

(e.g., hand movement, handshape, mouthing) that convey sign language information.

In a recent work [17], it has been shown that signs can be assessed over different channels of production like handshape, hand movement, facial expressions, etc. Two levels of assessment have been proposed 1. *Lexeme-Level*: Whether the produced sign matches the reference sign 2. *Form-Level*: Feedback on whether the form of the sign in terms of handshape, hand movement, etc. are correct or incorrect. Assessment beyond whether the produced sign is correct or incorrect is important as minor variations in production channels affect SL communication and tamper with sign interpretation. We propose to build upon this phonology-based method for sign language assessment to objectively evaluate artificially generated signs.

Our study aims to understand the feasibility of using such a method to evaluate the correctness of artificially generated sign language videos. To illustrate our approach, we employ a Generative Adversarial Network(GAN)-based video generation method to generate videos, given source videos from DSGS (DeutschSchweizerische GebärdenSprache) SL. We gather human ratings to subjectively gauge the similarity between the source and generated videos. We perform correlation studies on the scores obtained from the automatic system with the human ratings and other objective video evaluation metrics.

The paper is organized as follows: Section 2 introduces the proposed phonology-based approach for SL assessment. Section 3 presents the experimental setup and Section 4 provides the results and analysis of our studies and we conclude the paper in Section 5.

2. PROPOSED APPROACH

In this section, we provide a brief introduction to the assessment framework that was developed in [17, 18]. We then extend it to build an assessment pipeline for sign language generation.

2.1. Phonology-based Sign Language Recognition and Assessment

Tornay *et al.* [17] introduced a phonology-based sign language assessment based on Kullback-Leibler based Hidden Markov Model (KL-HMM) [19]. To assess a sign production, it is matched against an expected sign reference. The sign references are modelled using KLHMMs for sign recognition, the models are trained using a stack of posterior features corresponding to the channels of production (handshape, hand movement etc). As illustrated in Figure 1, the assessment is done by first matching the KL-HMM corresponding to the reference sign that was trained only on acceptable signs, with the stacked posterior feature sequences(handshape $z_{t,hshp}$, hand movement $z_{t,hmvt}$ etc.) estimated from the visual signal of test sign video using dynamic time warping(DTW) with local score based on Symmetric KL-divergence(SKL). Lexeme-level assessment is carried out by applying a threshold on the path length normalized global score $S(N,T)$. Form-level assessment, i.e. assessment of the different channels, is carried by factoring out the score of each channel from the global score. For further details, the reader is referred to [17, 18]. The form-level scores obtained from the DTW matching can be used to quantify the deviation from the reference sign video. This assessment framework as such can integrate all the channels in sign language. In this work, we limit ourselves to handshape and hand movement channels, as reliably estimating other channels such as, facial expression, mouthing information is still an open research. It is worth mentioning that the KL-HMM that is generating the reference posterior feature sequences can be replaced by an "instance" of acceptable sign production [20, Chapter 7].

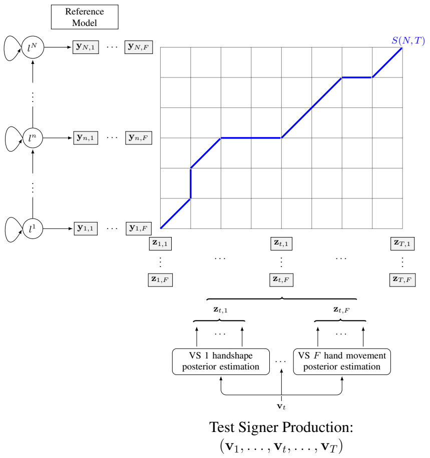

Fig. 1: Sign Language Assessment framework. The global score $S(N,T)$ is used for lexeme assessment and the local score on the grid decomposed into handshape and hand movement score is used for form-level assessment.

2.2. Extension to Sign Language Generation Assessment

Our proposed approach for assessment of artificially generated sign language videos builds upon the approach briefed in Section 2.1, where the KL-HMM is replaced by an "instance" that represents the source video and comparison with generated video is carried out. The block diagram of the method we propose for evaluation of generated videos against source videos is shown in Figure 2. The source sign video is used to extract the pose skeleton, which serves as an input to a GAN-based [15] method for generation of images. We then extract posterior features for different channels such as handshape and hand movement for the source video and the generated video. We replace the reference KL-HMM models with the posterior sequences of the source video. The DTW matching based on SKL cost function described in Section 2.1 is used to match the sequence of posteriors from the source and generated video. The form-level scores for handshape and hand movement are obtained by factoring out the score for each channel from the global DTW score. We hypothesize that low SKL scores imply higher similarity between the source and generated video.

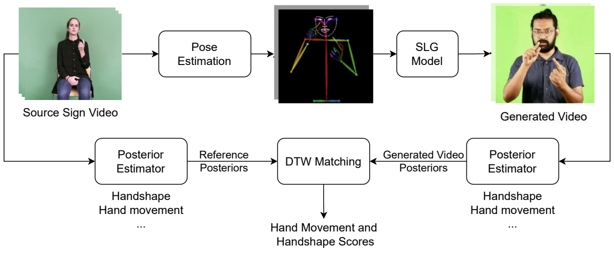

Fig. 2: Assessment of Sign Language Generation

It is worth pointing out that the proposed approach is similar to the phone class conditional probability sequences based objective speech intelligibility assessment approach proposed in [21] for assessment of speech codecs and text-to-speech systems.

3. EXPERIMENTAL SETUP

In this section, we present the experimental setup for SLG and SL assessment.

3.1. Dataset

We use the SMILE DSGS dataset [22] for the experimental studies. We use the protocol defined in [17] for sign language assessment. For sign language generation and its evaluation, we use acceptable sign productions from the test set. In the proposed approach, the hand movement posterior feature ($\mathbf{z_{t,hmvt}}$ and $\mathbf{y_{t,hmvt}}$) and the handshape posterior feature ($\mathbf{z_{t,hshp}}$ and $\mathbf{y_{t,hshp}}$) estimators need to be trained. For that, we use the train set that contains acceptable sign productions from signers different then the test set.

3.2. Sign Language Generation

In [15], Krishna *et al.* , extend the GAN-based “do as I do” motion transfer model EBDN [16], for Indian Sign Language generation. A combination of two generators, one for the body and one specifically for the hand are trained, to focus on accurate handshape generation. A smoothing network is additionally used to seamlessly combine the outputs generated for the body and hands. Briefly, given a video of a source person and another video of a target person, the method involves generating a video of the target person performing the action as the source. Figure 4 gives an overview of the training pipeline. The body GAN G_B, generates the image of the whole body of the target individual, given the pose skeleton of the target, similarly, the hand GAN G_H generates the image of the hand, given the skeleton crop of the hand. P represents the pose estimator. For adversarial training, three discriminators at different scales are used. Temporal consistency is ensured by generating consecutive frames alongside with a discriminator that distinguishes between the real correspondences between pairs of poses (x_{t-1}, x_t) and images (y_{t-1}, y_t) and fake pairs of (x_{t-1}, x_t) and $(G(x_{t-1}), G(x_t))$ Both G_B and G_H are trained using the same objective.

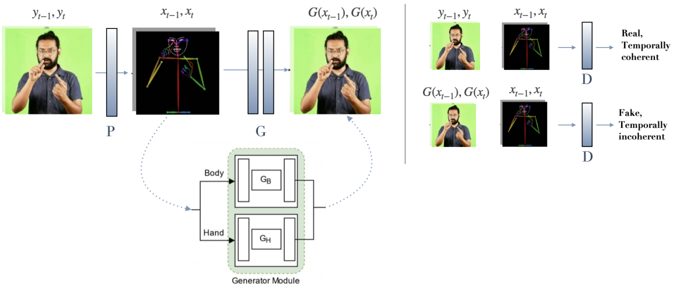

Fig. 3: Training pipeline of Sign Language video generation from pose skeleton to images

$$\min_G((\max_D \sum_{k=1}^{3} \mathcal{L}_{temporal}(G, D_k)) + \lambda_{FM} \sum_{k=1}^{3} \mathcal{L}_{FM}(G, D_k) + \lambda_p \mathcal{L}_p(G(x), y)$$

Where $\mathcal{L}_{temporal}$ is given by Equation 1, $\mathcal{L}_{FM}$ is the discriminator feature-matching loss [10] and $\mathcal{L}_p$ is the perceptual reconstruction loss that compares pre-trained VGG [23] features at different layers of the model.

$$\mathcal{L}(G, D) = \mathbb{E}_{x,y}[\log D(x_{t-1}, x_t, y_{t-1}, y_t)] + \mathbb{E}_x[\log(1 - D(x_{t-1}, x_t, G(x_{t-1}), G(x_t))] \quad (1)$$

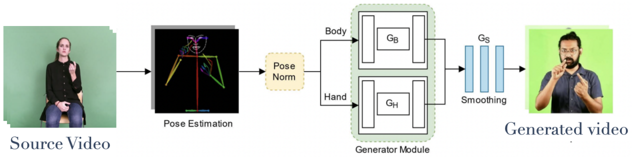

Fig. 4: Sign Language video generation from source to target person

We extend this work to generate signs for DSGS sign language. Pose skeletons obtained from OpenPose [24] are used as input poses to the GAN, to generate the target person images. Figure 4 shows the pipeline for video generation, the *Pose Norm* module normalizes the pose skeleton of the source person with respect to the target person. Figure 5 shows an example of a video generated for the DSGS sign ”ABER”.

3.3. Assessment of Sign Videos

For the subjective evaluation of the generated videos, 22 student raters were given 20 real/generated video pairs and rated 3 questions for each video pair on a five-point Likert scale.

1. How different is the hand motion in the generated video compared to the original video? (1:different - 5: similar)
2. How different is the handshape in the generated video compared to the original video? (1:different - 5: similar)
3. Rate the overall quality of the generated video (Higher the better)

Table 1: Intraclass Correlation Coefficient and statistics for rater agreement for each of the questions

Question	ICC3,k	mean	std	skew
Q1	0.94	4.36	0.89	-1.47
Q2	0.96	4.0	0.93	-0.84
Q3	0.95	3.42	1.06	-0.32

Intraclass Correlation Coefficient (ICC) [25] was used to get the agreement between the raters. A high ICC value (close to 1) indicates a high similarity between values. In our data, we observed high ICC values(ICC$\{3,k\} > 0.90$) indicating that all the raters had a high agreement and indicate excellent reliability of the raters. Table 1 summarizes the annotated questions, the ICC(3,k), and their respective descriptive statistics.

Baseline objective metrics: We use the following baseline metrics to gauge the similarity between the source video and the generated video:

1. PSNR: The ratio of the maximum possible value of the image and the power of noise that affects its quality [26].
2. SSIM: Computes the similarity between two images in terms of their structural information, it considers the strong interdependency between a group of pixels. It computes a similarity score based on how well the images match in terms of structural information [26].
3. MSE Skeleton: Mean-squared error between the sequence of 3D skeletons from source and generated videos obtained from Mediapipe [27]

Proposed Metrics: We need to extract posterior estimates of handshape and hand movement channels, in order to evaluate the generated videos. To estimate hand movement posterior features $\mathbf{z}_{t,hmvt}$, we used the same procedure as in [28] and [17]. More precisely, this involves two steps, (i) **hand movement subunit inference**: a

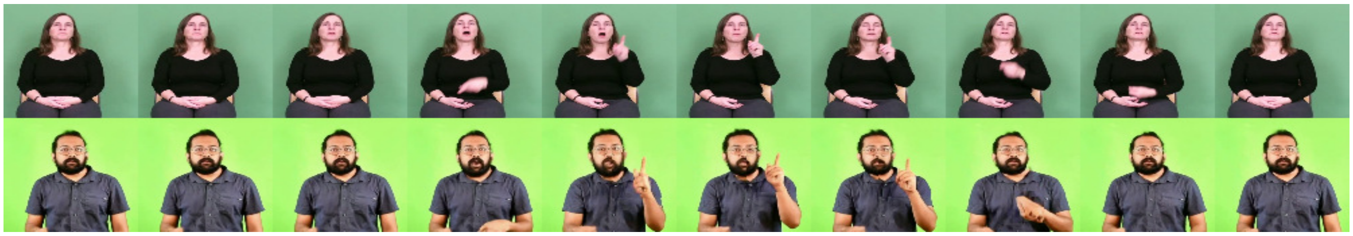

Fig. 5: First row shows the sequence of frames in the source video for the sign ”ABER”, the second row shows the corresponding frames generated by the GAN

sequence of hand movement feature vectors based on skeletal information are extracted for each sign. The feature vector consists of data corresponding to the coordinates of the left and right hands relative to the head, hip, and shoulder and their velocities. To normalize the variation between signers, the neck joints of all the signers are aligned with respect to a randomly chosen signer and scaled by the shoulder width. Given the sequence of features, the hand movement subunits are inferred by training left-to-right HMMs with different numbers of states for each of the signs and selecting the number of states that yields the best performance on a development set. (ii) **hand movement subunit posterior probability ($\mathbf{z}_{t,hmvt}$) estimator training**: this is done by obtaining an alignment of the features with the HMM states and training of a multilayer perceptron (MLP) to classify the HMM states with a cost function based on cross-entropy. The architecture of the MLP is determined in a cross-validation manner. A similar method using a hand skeleton is applied to handshape classification in order to extract *handshape posterior features* $\mathbf{z}_{t,hshp}$. The dimension of handshape posterior feature vector and hand movement posterior feature vector are 540 and 2351, respectively. The extracted posterior features from the generated videos are matched against the reference posteriors to obtain the SKL scores. The SKL scores obtained from DTW matching reflect the objective scores corresponding to the subjective evaluation of sign videos.

4. RESULTS AND ANALYSIS

Figure 6 shows the frame-wise comparison of the source video and generated video over different evaluation metrics for two DSGS signs - ABER and ANGESTELLT. We use PSNR, SSIM, SKL scores from the assessment system mentioned in 2.1 and also the mean squared errors(MSE) between the estimated skeletons as evaluation metrics. From the figure, we can see that there is low variability of PSNR and SSIM across time for both the videos. Whereas the SKL scores corresponding to handshape and hand movement and MSE of skeletons show significant variability. The degree of variability of these metrics, reflect the similarity between the source and generated videos. Higher the variation in the SKL scores, lower the similarity between videos. To validate this, we do a correlation study of the objective measures mentioned and the mean opinion scores of the raters.

Table 2 shows the correlation values obtained for all the combinations of user ratings and objective similarity measures. The values depicted in bold represent the correlation coefficients that are statistically significant ($p-value < 0.05$). *The SKL scores are negatively correlated as low SKL corresponds to higher similarity, i.e., higher mean opinion scores*. We observe that PSNR and MSE shows not significant or no correlation with the user ratings. The SSIM and SKL scores show moderate to high correlation with the user ratings.

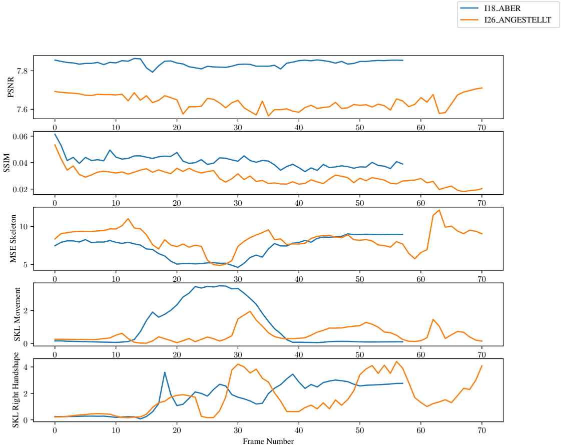

Fig. 6: Frame-wise comparison of source video and generated video across different metrics for two signs - ABER and ANGESTELLT

Table 2: Spearman's Correlation Coefficient and corresponding p-values between different video evaluation metrics and user ratings

		Movement Rating	Handshape Rating	Overall Rating
PSNR	corr	0.18	0.067	-0.085
	pvalue	0.626	0.853	0.815
SSIM	corr	**-0.51**	**-0.74**	-0.55
	pvalue	**0.041**	**0.084**	0.19
MSE	corr	-0.54	-0.42	**-0.59**
Skeleton	pvalue	0.106	0.228	**0.069**
SKL	corr	**-0.58**	**-0.58**	**-0.45**
Movement	pvalue	**0.047**	**0.032**	**0.019**
SKL	corr	**-0.89**	**-0.84**	**-0.79**
Handshape	pvalue	**0.049**	**0.002**	**0.006**

5. CONCLUSION

In this paper, we investigated the use of the phonology-based assessment system to objectively evaluate the quality of artificially generated sign language videos. We validated the proposition by conducting SLG and evaluation study with subjective ratings from 22 raters and comparing the proposed approach against PSNR, SSIM and MSE skeleton. We find that the scores obtained from the assessment system show a moderate to high correlation with the mean opinion scores of the raters, and yield better assessment than PSNR, SSIM and MSE skeleton. With respect to SLG, having evaluation scores specific to hand movement and handshape is highly desirable as variations in these aspects can affect sign language perception.

6. REFERENCES

[1] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou, "Sign language production: A review," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 3446–3456.

[2] Richard Kennaway, "Avatar-independent scripting for real-time gesture animation," *CoRR*, vol. abs/1502.02961, 2015.

[3] T. Hanke, "HamNoSys - representing sign language data in language resources and language processing contexts," *Workshop proceedings : Representation and processing of sign languages*, pp. 1–6., 2004.

[4] J.A. Bangham, S.J. Cox, R. Elliott, J.R.W. Glauert, I. Marshall, S. Rankov, and M. 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)*, 2000, pp. 6/1–6/7.

[5] E Efthimiou, SE Fotinea, T Hanke, J Glauert, R Bowden, A Braffort, C Collet, P Maragos, and F Lefebvre-Albaret, "The dicta-sign wiki: Enabling web communication for the deaf," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7383 L, no. PART 2, pp. 205 – 212, 2012.

[6] Inge Zwitserlood, Margriet Verlinden, Johan Ros, and Sanny Schoot, "Synthetic signing for the deaf: Esign," 01 2005.

[7] 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 *Proc. of the ACM Conference on Assistive Technologies*, New York, NY, USA, 2002, Assets '02, p. 205–212.

[8] Sarah Ebling and John Glauert, "Exploiting the full potential of jasigning to build an avatar signing train announcements," 10 2013.

[9] Sylvie Gibet, François Lefebvre-Albaret, Ludovic Hamon, Rémi Brun, and Ahmed Turki, "Interactive editing in french sign language dedicated to virtual signers: Requirements and challenges," *Universal Access in the Information Society*, vol. 15, 09 2015.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

[11] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden, "Text2sign: Towards sign language production using neural machine translation and generative adversarial networks," *Int. J. Comput. Vision*, vol. 128, no. 4, pp. 891–908, apr 2020.

[12] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden, "Everybody sign now: Translating spoken language to photo realistic sign language video," *CoRR*, vol. abs/2011.09846, 2020.

[13] 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)*, 2020, pp. 1250–1255.

[14] Stephanie Stoll, Necati Camgoz, Simon Hadfield, and Richard Bowden, "Sign language production using neural machine translation and generative adversarial networks," 08 2018.

[15] Shyam Krishna, Janmesh Ukey, and Dinesh Babu J, "Gan based indian sign language synthesis," in *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, New York, NY, USA, 2021, ICVGIP '21, Association for Computing Machinery.

[16] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros, "Everybody dance now," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[17] S. Tornay, N. C. Camgoz, R. Bowden, and M. Magimai.-Doss, "A phonology-based approach for isolated sign production assessment in sign language," in *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion)*, Oct. 2020.

[18] S. Tornay, M. Razavi, N. C. Camgoz, R. Bowden, and M. Magimai.-Doss, "HMM-based approaches to model multichannel information in sign language inspired from articulatory features-based speech processing," in *Proc. in the IEEE ICASSP*, 2019.

[19] G. Aradilla, J. Vepa, and H. Bourlard, "An Acoustic Model Based on Kullback-Leibler Divergence for Posterior Features," in *ICASSP*, 2007, pp. 657–660.

[20] Sandrine Tornay, *Explainable Phonology-based Approach for Sign Language Recognition and Assessment*, Ph.D. thesis, EPFL Lausanne, 2021.

[21] Raphael Ullmann, Mathew Magimai-Doss, and Hervé Bourlard, "Objective speech intelligibility assessment through comparison of phoneme class conditional probability sequences," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4924–4928.

[22] S. Ebling, N. C. Camgöz, P. Boyes Braem, K. Tissi, S. Sidler-Miserez, S. Stoll, S. Hadfield, T. Haug, R. Bowden, S. Tornay, M. Razavi, and M. Magimai-Doss, "SMILE Swiss German sign language dataset," in *Proc. of the Language Resources and Evaluation Conference*, 2018.

[23] Shuying Liu and Weihong Deng, "Very deep convolutional neural network based image classification using small training sample size," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 730–734.

[24] Zhe Cao, Gines Martinez, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, pp. 1–1, 07 2019.

[25] Patrick E Shrout and Joseph L Fleiss, "Intraclass correlations: uses in assessing rater reliability.," *Psychological bulletin*, vol. 86, no. 2, pp. 420, 1979.

[26] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[27] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann, "Mediapipe: A framework for perceiving and processing reality," in *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019.

[28] S. Tornay and M. Magimai.-Doss, "Subunits inference and lexicon development based on pairwise comparison of utterances and signs," *Information*, vol. 10, 2019.