A Phonology-based Approach for Isolated Sign Production Assessment in Sign Language

Sandrine Tornay
Idiap Research Institute
Martigny, Switzerland
Ecole polytechnique fédérale de Lausanne
Lausanne, Switzerland

Necati Cihan Camgoz
University of Surrey
Guildford, UK

Richard Bowden
University of Surrey
Guildford, UK

Mathew Magimai.-Doss
Idiap Research Institute
Martigny, Switzerland

ABSTRACT

Interactive learning platforms are in the top choices to acquire new languages. Such applications or platforms are more easily available for spoken languages, but rarely for sign languages. Assessment of the production of signs is a challenging problem because of the multichannel aspect (e.g., hand shape, hand movement, mouthing, facial expression) inherent in sign languages. In this paper, we propose an automatic sign language production assessment approach which allows assessment of two linguistic aspects: (i) the produced lexeme and (ii) the produced forms. On a linguistically annotated Swiss German Sign Language dataset, SMILE DSGS corpus, we demonstrate that the proposed approach can effectively assess the two linguistic aspects in an integrated manner.

CCS CONCEPTS

• **Human-centered computing**;

KEYWORDS

sign language assessment, sign language verification

ACM Reference Format:
Sandrine Tornay, Necati Cihan Camgoz, Richard Bowden, and Mathew Magimai.-Doss. 2020. 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), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3395035.3425251

1 INTRODUCTION

In recent years, there is growing interest in developing assistive systems that can help in bridging the gap or breaking the barrier between Hearing and Deaf communities through multimodal systems. In that direction, as sign languages (SLs) are under-studied and under-resourced languages, there is interest in developing interactive applications that could aid in SL acquisition. Currently, existing platforms test comprehension and vocabulary through pre-recorded videos, while SL production tests are realized by online recording for later analysis, which is both expensive and time consuming. Existing interactive e-learning platforms that contain production testing use either self-correctness, such as the web-based e-learning resource SignAssess [4] which allows to compare the recorded user's video to a pre-recorded reference one, or real-time SL verification which assesses if the produced sign is correct or incorrect, such as SignAll [16] technology, ISARA [6] application. Assessing whether a produced sign is correct or incorrect would not be sufficient by itself to aid SL learners. The reason being that SL consists of different channels of information corresponding to manual components (hand position, hand movement and hand shape) and non-manual components (mouthing, facial gesture, posture). So, for realistic adoption of SL learning applications, there is need for a framework that enables assessment of those multiple channels of information in a linguistically valid manner.

In a recent work [15], a Hidden Markov Model (HMM) based SL processing framework was proposed that enables modeling of the multi-channel information present in the SL, akin to modeling of multi-channel articulatory information in speech production [13]. The present paper builds upon that work to propose a SL assessment approach that, in an integrated manner, can assess sign production at: (a) lexeme level, i.e. verify whether a produced sign is targeting the right reference sign or not and (b) form level, i.e. assessing separately the different form channels of a sign, such as hand movement and hand shape. We demonstrate the potential of the proposed approach through a validation study on Swiss German Sign Language.

The remainder of the paper is organized as follow: Section 2 provides a background on the phonology-based SL processing frameworkSection 3 presents the proposed SL assessment approach. Section 4 presents the experiment setup and Section 5 the results and analysis. Finally, we conclude in Section 6.

2 BACKGROUND

In [15], a phonological approach for SL processing was proposed, based on the understanding that, in both SL and spoken language, there is a production phenomenon that generates a signal and there is a perception phenomenon, which interprets the generated signals in terms of elements of “language", e.g. words, phrases. Given this

ICMI '20 Companion, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-8002-7/20/10…$15.00
https://doi.org/10.1145/3395035.3425251

relationship, the articulatory feature based speech processing study developed in [13] in the framework of Kullback-Leibler divergence based HMM (KL-HMM) [1, 2] was adapted to SL processing.

Briefly, in this approach, first posterior probabilities of subunits $\mathbf{z}_{t,f} = [P(vs_f^1|\mathbf{v}_t) \cdots P(vs_f^d|\mathbf{v}_t) \cdots P(vs_f^{D_f}|\mathbf{v}_t)]^{\mathrm{T}}$ corresponding to different channels $f \in \{1, \cdots F\}$ are estimated given the visual signal $(\mathbf{v}_1, \cdots \mathbf{v}_t, \cdots \mathbf{v}_T)$, where vs_f^d denotes visual subunit, d corresponding to channel f. The stacked posterior probability vectors from different channels, $\mathbf{z}_t = [\mathbf{z}_{t,1} \cdots \mathbf{z}_{t,F}]^{\mathrm{T}}$, are then used as feature observations for HMM, whose state emission distributions are parameterized by categorical distributions $\mathbf{y}_i = [\mathbf{y}_{i,1} \cdots \mathbf{y}_{i,F}]^{\mathrm{T}}$, for $i \in \{1, \ldots, I\}$ where I is the number of HMM states. Following the investigations in speech processing, the transition probabilities are assumed to be 0.5 to stay on the same state and 0.5 to transit from the state [1, 2]. The state emission distributions i.e. the categorical distributions are estimated by minimizing a cost function based on Kullback-Leibler (KL) divergence [1, 2, 15]. When decoding, such as in the case of sign language recognition (SLR), Viterbi search is performed with local scores based on KL-divergence.

3 PROPOSED SIGN LANGUAGE ASSESSMENT APPROACH

The present paper develops an automatic SL assessment approach by building upon the phonological approach for SL processing using KL-HMM approach presented in the previous section. More precisely, as illustrated in Figure 1, to assess sign production, the proposed approach compares the different channels of information by matching sequence of stacked probability distributions $Z = (\mathbf{z}_1 \cdots \mathbf{z}_t \cdots \mathbf{z}_T)$ corresponding to the test sign production and the sequence of stacked categorical distributions $Y = (\mathbf{y}_1 \cdots \mathbf{y}_n \cdots \mathbf{y}_N)$ corresponding to the KL-HMM representing the target reference lexeme (i.e. the sign expected to be produced), through dynamic programming and thresholding the resulting score or cost.

Formally, the match is obtained by dynamic programming with the following recursion,

$$S(n,t) = l(\mathbf{y}_n, \mathbf{z}_t) + \min\left[S(n,t-1) + c_{\mathrm{trans}}, S(n-1,t-1) + c_{\mathrm{trans}}\right], \quad (1)$$

where $c_{\mathrm{trans}} = -\log(0.5)$ is the transition cost and $l(\mathbf{y}_n, \mathbf{z}_t)$ is the local score defined by symmetric KL-divergence (SKL) between the probability distributions, i.e.,

$$l(\mathbf{y}_n, \mathbf{z}_t) = \sum_{f=1}^{F} SKL(\mathbf{y}_{n,f}, \mathbf{z}_{t,f}), \quad (2)$$

$$SKL(\mathbf{y}_{n,f}, \mathbf{z}_{t,f}) = \frac{1}{2} \cdot \sum_{d=1}^{D_f} \left(\mathbf{y}_{n,f}^d \log(\frac{\mathbf{y}_{n,f}^d}{\mathbf{z}_{t,f}^d}) + \mathbf{z}_{t,f}^d \log(\frac{\mathbf{z}_{t,f}^d}{\mathbf{y}_{n,f}^d})\right). \quad (3)$$

$\mathbf{y}_{n,f}^d$ and $\mathbf{z}_{t,f}^d$ denote d^{th} element in the vectors $\mathbf{y}_{n,f}$ and $\mathbf{z}_{t,f}$, respectively. The best matching path with the begin/end time frames t_n^b and t_n^e, respectively of each state n in the reference lexeme, can be obtained as part of the dynamic programming recursion.

Given the best matching path, the state duration normalized lexeme-level score $\mathcal{S}_{lex}$ can be estimated as,

$$\mathcal{S}_{lex} = \frac{1}{N} \cdot \sum_{n=1}^{N} \frac{\sum_{t=t_n^b}^{t_n^e} l(\mathbf{y}_n, \mathbf{z}_t)}{t_n^e - t_n^b + 1}; \quad (4)$$

while the state duration normalized form-level score $\mathcal{S}_{form}^f$ for each channel f can be estimated as,

$$\mathcal{S}_{form}^f = \frac{1}{N} \cdot \sum_{n=1}^{N} \frac{\sum_{t=t_n^b}^{t_n^e} SKL(\mathbf{y}_{n,f}, \mathbf{z}_{t,f})}{t_n^e - t_n^b + 1}. \quad (5)$$

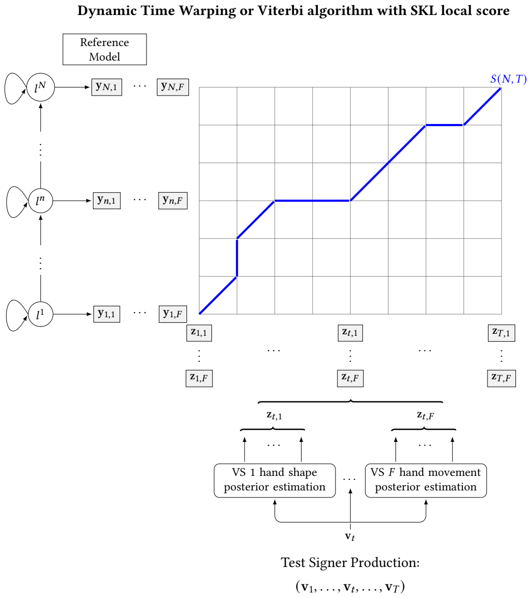
Dynamic Time Warping or Viterbi algorithm with SKL local score
Reference Model
l^N
l^n
l^1
$\mathbf{y}_{N,1}$ ··· $\mathbf{y}_{N,F}$
$\mathbf{y}_{n,1}$ ··· $\mathbf{y}_{n,F}$
$\mathbf{y}_{1,1}$ ··· $\mathbf{y}_{1,F}$
$S(N,T)$
$\mathbf{z}_{1,1}$
$\mathbf{z}_{1,F}$
$\mathbf{z}_{t,1}$
$\mathbf{z}_{t,F}$
$\mathbf{z}_{T,1}$
$\mathbf{z}_{T,F}$
$\mathbf{z}_{t,1}$
$\mathbf{z}_{t,F}$
VS 1 hand shape posterior estimation
VS F hand movement posterior estimation
$\mathbf{v}_t$
Test Signer Production:
$(\mathbf{v}_1, \ldots, \mathbf{v}_t, \ldots, \mathbf{v}_T)$

Figure 1: Illustration of the assessment framework.

As comparison of probability distributions using KL-divergence and other measures such as Bhattacharya distance is equivalent to hypothesis testing [3, 9], lexeme-level and form-level assessment can be carried out by simply applying a threshold δ_{lex} and δ_{form}^f on $\mathcal{S}_{lex}$ and $\mathcal{S}_{form}^f$ to decide correct/incorrect lexeme and forms.

4 EXPERIMENTAL SETUP

We validated the proposed approach on the large-scale SMILE Swiss German Sign Language database [5] (referred as SMILE DSGS database in the following) which was created in the context of developing an assessment system for lexical signs of Swiss German Sign Language (DSGS for Deutschschweizerische Gebärdensprache). We demonstrate the approach using two channels of information for which linguistic annotations are available, namely, hand movement (*hmvt*) and hand shape (*hshp*).

4.1 SMILE DSGS database

The SMILE DSGS database [5] is composed of 11 adult L1 signers and 19 adult L2 learners performing three times 100 isolated signs of a DSGS vocabulary production test. Only the second pass out of the three was manually linguistically annotated. The data collection was done using the Microsoft Kinect v2 sensor and the high speed and high resolution GoPro video cameras.

In our experimental setup, we only used the second pass annotated with the 'Category of sign produced' annotation of the SMILE transcription/annotation scheme (presented in [5]). Briefly, this linguistic annotation evaluates, through six categories, the acceptability of a sign according to linguistic criteria (lexeme, meaning and form), see Table 1. To ensure that enough correct samples for each sign is available (minimum 5 samples/sign annotated as cat.1 or cat.2), 94 signs were selected out of the 100. The cat.1 and cat.2, consisting of acceptable sign productions, was partitioned in a signer-independent manner into 1125 training set samples from 15 signers, 509 development set samples from 7 signers and 581 test set samples from 8 signers. We used the same test set samples for evaluating both the KL-HMM references (in terms of SLR) and the proposed assessment system. The cat.1 and cat.2 were used to build the different components of the proposed assessment system.

Table 1: SMILE annotation scheme of the 'Category of sign produced' annotation

Category	Same lexeme as target sign?	Same meaning as target sign?	Same form as target sign?	*#test samples*
cat.1	yes	yes	yes	581
cat.2	yes	yes	slightly different	
cat.3	yes	yes	no	412
cat.4	yes	slightly different	slightly different	
cat.5	no	yes	no	183
cat.6	no	no	no	

4.2 Hand movement subunits posterior probability estimation

The hand movement subunits extraction was inspired by the method presented in [14]. Briefly, a 36 dimensional position and velocity features for both hands was extracted from the 3D skeleton with three different coordinate centers (head, shoulder and hip center). A shoulder normalization was then applied to compensate the variation in-between the signers by aligning the neck joint of all the signers w.r.t a randomly chosen signer and scaling the shoulder width. Whole sign left-to-right HMMs with different number of states (3 to 30) for each sign was trained were trained using HTK [18]. The state emission distributions were modeled by single Gaussians with a diagonal covariance matrix. The development data was decoded using all the 28 whole sign-based HMM/GMMs for all the signs, and the most frequently recognized model in terms of number of states was chosen. These states served as the hand movement subunits.

We trained a multilayer perceptron (MLP) to classify the hand movement subunits based on the alignments obtained from the HMM/GMM (Gaussian Mixture Models) systems. The input to the MLP are 36-dimensional feature observation with four frames preceding and following context. The output non-linearity was softmax. The MLP was trained with cross-entropy based error criterion using the Quicknet software [8]. The trained MLP was used to extract hand movement posterior features $\mathbf{z}_t^{\text{hmvt}}$ for the KL-HMM.

4.3 Hand shape subunit posterior probability estimation

We used a residual network based Convolutional Neural Network (CNN) architectures, namely ResNeXt-101 [17], trained on the One-Million-Hands [11] dataset. The hand shape observations are the hand shape class-conditional posterior probabilities $\mathbf{z}_t^{\text{hshp}_+}$, where the classes are composed by a transition shape and the 60 linguistically inspired hand shapes presented in https://www-i6.informatik.rwth-aachen.de/~koller/1miohands-data/.

Then, as a second channel, to improve the quality and the generalization of our hand shape subunit representations, we first start by reducing the number of classes by choosing the most common hand shape classes present in the One-Million-Hands dataset. Inspired by the sample distribution of ImageNet [7], we kept the hand shape classes which have at least 1000 samples in the training set, reducing our number of classes to 27. We then collected new samples from four participants, two L2 signers and two non-signers, to help with the class imbalance. Leading to adapted hand shapes, denoted as $f = \text{hshp}\_$.

The stack of the 61 hand shapes and the 28 adapted hand shapes were used in our experiment, i.e. $\mathbf{z}_t^{\text{hshp}_+}$ and $\mathbf{z}_t^{\text{hshp}_-}$. We trained our networks using Adam optimizer [10] using a batch size of 32. We apply random rotation, zoom and colour jitter to help our networks generalize better. To further address the class imbalance issue, we re-sample the training images w.r.t. their corresponding classes and simulate a uniform distribution over all classes. Our models were implemented using the PyTorch deep learning framework [12].

To overcome the jittery wrist localization of SMILE dataset [5] we utilize a state-of-the-art 2D pose estimation method, namely OpenPose [19], to localize wrist locations. Using the wrist pixel coordinates, we crop patches around both hands and extract their posteriors over the hand shape classes for our KL-HMM framework.

4.4 Sign reference systems

We trained five KL-HMM systems to develop reference models for each sign, namely,

- the **rlS** system refers to the case where only the hand shape subunit posterior probabilities of the right and left hands estimated by the residual network based CNN are stacked and modeled.
- the **M** system refers to the case where only the posterior probabilities of hand movement subunits obtained by combining right and left hand features are modeled. In other words, distinction between dominant hand and non-dominant hand is not made.
- the **rlM** system refers to the case where hand movement subunits are obtained for the left hand and the right hand separately; two separate MLPs are trained to classify the the left hand and the right hand movement subunits; and the left and right hand movement subunits posterior probabilities estimated by the respective MLPs are stacked and modeled.
- the **rlS+M** and **rlS+rlM** systems refer to the case of using the concatenation of the hand shape and the hand movement subunit probability posteriors depending on the different setups presented above.

Data of the cat.1 and cat.2 (see Table 1) were used to train and test the reference models as it corresponds to "acceptable signs" annotation. All the KL-HMM systems were trained using 3 to 30 KL-HMM states per sign. The system that yielded the best recognition accuracy on the development data was chosen as the reference.

Evaluation of KL-HMM reference models: Table 2 presents the recognition accuracy (RA) of the different KL-HMM systems with the corresponding number of states as well as the corresponding number of features (D^f). It can be observed that the system modeling both hand movement and hand shape information yields the best SLR performance. These results show that the KL-HMM reference lexeme models are indeed modeling the different signs and are able to discriminate between them.

Table 2: SL recognition accuracy (RA) of the reference KL-HMM systems with the corresponding number of states (*# state*) as well as the number of features (*# feature*)

	KL-HMM References				
	rlS	**M**	**rlM**	**rlS+M**	**rlS+rlM**
RA	37.2	56.9	57.4	74.7	75.2
# state	26	18	24	29	28
# feature	*178*	*2075*	*4214*	*178+2075*	*178+4214*

4.5 Assessment systems

Lexeme assessment: to evaluate the lexeme assessment, according to the category annotation of the data summarized in Table 1, we separated the test correct/incorrect data as the following: *cat.1-2-3-4* which is correct target signs composed of cat.1 to cat.4 and *cat.5-6+* which is incorrect target signs composed of cat.5, cat.6 and since these categories contain only few data, we balanced the incorrect set by creating additional data by matching each sample of the cat.1 and cat.2 data with a randomly chosen wrong reference.
Form assessment: to evaluate the form assessment, we used the *cat.1-2* as correctly produced form data and since the targeted sign is incorrect for *cat.5-6+* we supposed that the produced form (hand movement and hand shape) was incorrect. In the present study, we did not make difference between dominant and non-dominant hand.

We determined the thresholds, δ_{lex} and δ^f_{form} for $f \in$ {hmvt, hshp} on the development set, which consists of cat.1 and cat.2 data (see Section 4.1). We created a set of correct sign scores by matching the same sign instances and a set of incorrect match scores by matching instances of different signs. δ_{lex} and δ^f_{form} for each f were set as the threshold that yielded the best F_1 score for lexeme assessment and form assessment.

5 RESULTS AND ANALYSIS

Lexeme assessment: Table 3 presents the F_1 score of the lexeme assessment study depending on the KL-HMM reference used to align the produced sign. As it can be observed, combining the hand movement and shape channels helps in the lexeme assessment since using **rlS+M** as reference gives the best assessment result. Another relevant observation is that using combined right and left hand movement (**M**,**rlS+M**) is sufficient for lexeme assessment.

Table 3: F_1 scores of the correct lexeme assessment according to the five reference KL-HMM systems

KL-HMM References				
rlS	**M**	**rlM**	**rlS+M**	**rlS+rlM**
71.6	88.3	85.0	90.0	87.4

Form assessment: Table 4 presents the F_1 score of the forms error assessment study of the hand movement channel and the hand shape channel depending on the five KL-HMM references. First, we can observe that adding the hand movement information helps in the hand shape error assessment, while the reverse is not true. Indeed using either **rlS+M** or **rlS+rlM** does not change significantly and is better than using **rlS** for hand shape form error assessment. A potential reason for that could be that the hand movement channel has more temporal variations than the hand shape channel. This can also explain why adding hand shape channel to hand movement one does not help in hand movement error assessment. In fact, hand movement form assessment using **M** or **rlS+M**, or using **rlM** or **rlS+rlM** are not significantly different. Moreover, making no distinction between dominant and non-dominant hand movement gives better form assessment results. This aspect could be further explained or understood by separating the one-handed or two-handed sign assessment results. This is part of our future work.

Table 4: F_1 scores of the forms error assessment (hand movement (hmvt) and hand shape (hshp)) according to the five reference KL-HMM systems

	KL-HMM References				
	rlS	**M**	**rlM**	**rlS+M**	**rlS+rlM**
hshp form	75.8	-	-	77.6	78.0
hmvt form	-	91.1	88.6	90.5	88.0

6 CONCLUSION

This paper presented a phonologically motivated SL assessment approach that allows to assess two different linguistic aspects of a produced sign: the lexeme and the form. In this approach, in the framework of KL-HMM, sequence of posterior probabilities of subunits/classes corresponding to different channels are stacked and compared using dynamic programming to compute lexeme level and form level scores assessment. A validation study on the SMILE DSGS dataset yielded promising lexeme level assessment and form level assessment results. Our studies also showed that the different components of the proposed assessment system can be built only using cat.1 and cat.2 data. In the present work, we limited ourselves to modeling and assessing hand movement and hand shape information. The reason being lack of linguistically annotated data sets for other channels (e.g. mouthing, facial expression) as well as lack of reliable methods to extract information related to those channels from the visual signal. Having said that, in principle the proposed approach allows integration of those channels (see Equation (2) and (5)), as and when reliable methods are available to model those channels. In our future work, we will focus on assessment of cat.3 and cat.4, where lexeme is correct but form is incorrect.

7 ACKNOWLEDGMENTS

This work was funded by the SNSF through the Sinergia project SMILE *(Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment)*, grant agreement CRSII2_160811. We thank all the collaborators in the project for their valuable work.

REFERENCES

[1] G. Aradilla, H. Bourlard, and M. Magimai.-Doss. 2008. Using KL-based acoustic models in a large vocabulary recognition task. In *Proc. of Interspeech*.

[2] G. Aradilla, J. Vepa, and H. Bourlard. 2007. An acoustic model based on Kullback-Leibler divergence for posterior features. In *Proc. of the IEEE ICASSP*.

[3] R. E. Blahut. 1974. Hypothesis Testing and Information Theory. *IEEE Trans. on Information Theory* IT-20, 4 (1974).

[4] J. Christopher. 2012. SignAssess – Online Sign Language Training Assignments via the Browser, Desktop and Mobile. In *Computers Helping People with Special Needs*, Klaus Miesenberger, Arthur Karshmer, Petr Penaz, and Wolfgang Zagler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 253–260.

[5] 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. 2018. SMILE Swiss German sign language dataset. In *Proc. of the Language Resources and Evaluation Conference*.

[6] ISARA application. [n.d.]. ISARA application. https://isara.app/features.

[7] D. Jia, D. Wei, S. Richard, L. Li-Jia, L. Kai, and L. Fei-Fei. 2009. ImageNet: A Large-scale Hierarchical Image Database. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] D. Johnson et al. 2004. ICSI Quicknet Software Package. http://www.icsi.berkeley.edu/Speech/qn.html.

[9] T. Kailath. 1967. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology* 15, 1 (February 1967), 52–60.

[10] D. P. Kingma and J. Ba. 2014. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[11] O. Koller, O. Zargaran, H. Ney, and R. Bowden. 2016. Deep sign: hybrid CNN-HMM for continuous sign language recognition. In *Proc. of the British Machine Vision Conference (BMVC)*.

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*. 8024–8035.

[13] R. Rasipuram and M. Magimai.-Doss. 2016. Articulatory feature based continuous speech recognition using probabilistic lexical modeling. *Computer Speech and Language* 36 (2016), 233–259. https://doi.org/10.1016/j.csl.2015.04.003

[14] S. Tornay and M. Magimai.-Doss. 2019. Subunits Inference and Lexicon Development Based on Pairwise Comparison of Utterances and Signs. *Information* 10 (2019). https://doi.org/10.3390/info10100298

[15] S. Tornay, M. Razavi, N. C. Camgoz, R. Bowden, and M. Magimai.-Doss. 2019. HMM-based Approaches to Model Multichannel Information in Sign Language inspired from Articulatory Features-based Speech Processing. In *Proc. in the IEEE ICASSP*.

[16] SignAll Technologies Inc. (USA). [n.d.]. SignAll. https://www.signall.us/.

[17] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2016. Aggregated Residual Transformations for Deep Neural Networks. *arXiv preprint arXiv:1611.05431* (2016).

[18] S. Young et al. 2002. *The HTK Book*. Cambridge University Engineering Department.

[19] C. Zhe, S. Tomas, W. Shih-En, and S. Yaser. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.