Open access • Proceedings Article • DOI:10.1109/ICASSP40776.2020.9054631

Towards Multilingual Sign Language Recognition — Source link

Sandrine Tornay, Marzieh Razavi, Mathew Magimai.-Doss

Institutions: Idiap Research Institute

Published on: 04 May 2020 - International Conference on Acoustics, Speech, and Signal Processing

Topics: Sign language, German Sign Language and Sign (mathematics)

IDIAP RESEARCH REPORT

TOWARDS MULTILINGUAL SIGN LANGUAGE RECOGNITION

Sandrine Tornay Marzieh Razavi
Mathew Magimai.-Doss

Idiap-RR-16-2019

NOVEMBER 2019

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

TOWARDS MULTILINGUAL SIGN LANGUAGE RECOGNITION

Sandrine Tornay[†‡] *Marzieh Razavi*[⋆] *Mathew Magimai.-Doss*[†]

[†] Idiap Research Institute, Martigny, Switzerland
[‡] Ecole polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland
[⋆] Telepathy Labs GmbH, Zürich, Switzerland

ABSTRACT

Sign language recognition involves modeling of multichannel information such as, hand shapes, hand movements. This requires also sufficient sign language specific data. This is a challenge as sign languages are inherently under-resourced. In the literature, it has been shown that hand shape information can be estimated by pooling resources from multiple sign languages. Such a capability does not exist yet for modeling hand movement information. In this paper, we develop a multilingual sign language approach, where hand movement modeling is also done with target sign language independent data by derivation of hand movement subunits. We validate the proposed approach through an investigation on Swiss German Sign Language, German Sign Language and Turkish Sign Language, and demonstrate that sign language recognition systems can be effectively developed by using multilingual sign language resources.

Index Terms— Sign language processing, hidden Markov models, hand movement modeling, hand shape modeling, multilingual sign language recognition

1. INTRODUCTION

Sign language (SL) is a visual mode of communication for the Deaf community, where the information is conveyed through multiple visual channels such as, hand gestures (hand shape, location and movement), facial expression, body posture, lip movement [1]. In the sign language recognition literature, the focus has largely been on extraction of the multichannel information related to hand gestures (hand shape and hand movement) from the visual signal and modeling those information to recognize signs [2]. In that regard, over the years, different approaches have evolved for sign language recognition using different machine learning techniques such as, hidden Markov models (HMM) [3–5], parallel HMM [6], relevance vector machines [7], boosting [8], sequential pattern trees [9], deep learning methods [10, 11].

Despite these advances, sign language recognition technology is still an emerging technology. Besides the challenge of extraction and modeling of multiple channel information, one of the main reasons is resource scarcity. As, unlike spoken languages, sign languages users are limited. Also, sign languages have their own vocabulary and grammar, different than the corresponding spoken language [1]. For instance, British sign language is not a signed form of British English. Furthermore, even though the spoken language can be the same, the sign languages can be different. For example, American Sign Language and British Sign Language are different sign languages. Similarly, Swiss German Sign Language (DSGS) and German Sign Language (DGS) are different sign languages. One way to address the resource scarcity challenge is to develop methods that can exploit multiple sign language resources by overcoming the limitations imposed by the differences between the sign languages. In the literature, there is limited work in that direction, more precisely with hand shape modeling only. It has been found that, given the HamNoSys annotation [12] of produced signs, a global hand shape classifier can be trained by pooling resources from multiple sign languages and hand shape information based sign language recognition systems can be developed [13]. However, hand shape is only one channel of information. There is need to model other channels such as, hand movement, which unlike hand shape is a continuous aspect or in other words are not inherently a discrete unit.

In a recent work, by drawing an analogy between speech production-perception and sign language production-perception and taking inspiration from articulatory feature based speech processing [14], an HMM-based sign language processing framework that models hand movement information and hand shape information in an integrated fashion was proposed [5]. In that framework, while modeling hand shape information with DeepHand [13], it was found that hand movement information can be modeled as discrete units using HMMs. In a more recent work, it was found that such discrete unit representation of hand movements obtained using HMMs, also referred to as hand movement subunits, tend to exhibit language independence [15]. The present paper builds upon these two recent works to investigate a multilingual sign language recognition approach, where resources from multiple sign languages are shared to model both hand movement and hand shape information for sign language recognition. We validate the proposed approach through investigations on DSGS corpus SMILE, DGS corpus and Turkish Sign Language corpus HospiSign.

Section 2 presents the proposed approach. Section 3 presents the experiment setup. Section 4 presents results and analysis. Finally, in Section 5, we conclude.

2. PROPOSED APPROACH

In [5], taking inspiration from articulatory feature-based speech recognition [14], a sign language processing approach using Kullback-Leibler divergence HMM (KL-HMM) was developed. Briefly, in KL-HMM [16, 17] the feature observations are probabilistic (posterior distributions). Each HMM state is parameterized by a categorical distribution of the same dimension as the feature observations and these parameters are estimated through embedded Viterbi expectation maximization algorithm with a cost function based on

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. We thank our collaborators from the University of Surrey for giving access to the DGS corpus.

Kullback-Leibler (KL) divergence [18] between the feature observations and the state categorical distribution. The decoding step is the same as standard HMM-based approach where the log likelihood of state is replaced by the KL-divergence between the feature observations and the state categorical distribution.

As illustrated in Figure 1, in this approach the feature observation is a stack of posterior features $\mathbf{z}_t := \left[\mathbf{z}_t^{\text{hshp}}\ \mathbf{z}_t^{\text{hmvt}}\right]^T$. $\mathbf{z}_t^{\text{hshp}}$ and $\mathbf{z}_t^{\text{hmvt}}$ denote the probabilistic features corresponding to hand shape and hand movement, respectively. The HMM state s^i is parameterized by a stack of categorical distribution $\mathbf{y}_{s^i} := \left[\mathbf{y}_{s^i}^{\text{hshp}}\ \mathbf{y}_{s^i}^{\text{hmvt}}\right]^T$. The local score $S(\mathbf{y}_{s^i}, \mathbf{z}_t)$ is based on KL-divergence. It is worth mentioning that, in principle, the stack of posterior features can be expanded to model other channel of information such as, mouthing, facial expression, as and when needed or available. Each of the posterior feature vector corresponds to a set of subunits or discrete units corresponding to the channel of information.

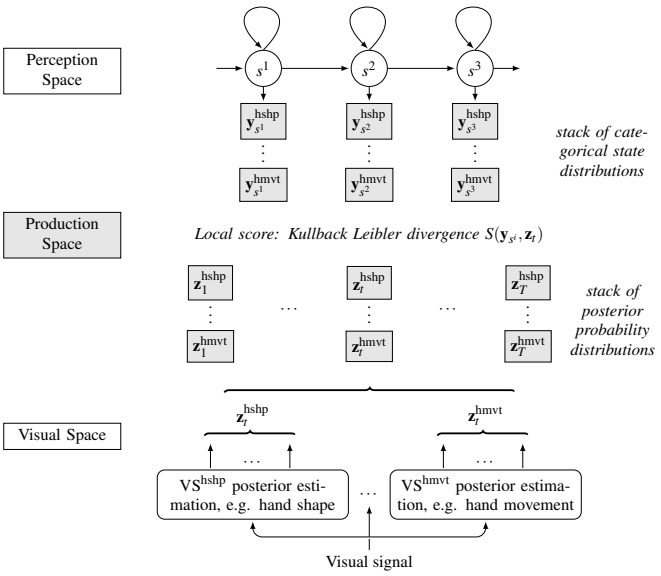

Fig. 1. Schematization of the Kullback Leibler divergence-based Hidden Markov Model (KL-HMM) applied for sign language processing. VS denotes Visual Subunits.

At a high level, similar to speech recognition using KL-HMM [19, 20], this framework can be visualized as matching of a sequence of multichannel information obtained through bottom-up modeling (visual signal-to-hand gestures) with a sequence of multichannel information obtained through top-down modeling (lexeme-to-hand gestures). In the case of speech recognition, it has been found that resource constraints can be effectively addressed by using auxiliary or non-target language resources for bottom-up modeling and using the target language resources only for top-down modeling [19].

Given that understanding, a question that arises is: can we achieve the same for sign language recognition? In other words, $\mathbf{z}_t$ estimators are trained with target language independent data and $\mathbf{y}_{s^i}$ is estimated on target language data. In [5], this was achieved for hand shape component through the use of DeepHand net to estimate $\mathbf{z}_t^{\text{hshp}}$, but not for hand movement modeling. One possible way to overcome that challenge is to use hand movement subunits that can be shared across languages.

In the literature, automatic derivation of hand movement subunits has followed two strands of research. First, using HamNoSys annotations of signs [21–24]. Second, through unsupervised segmentation and clustering [8, 25–30]. The difficulty in using these methods is that it is not clear if the derived subunits are signer-independent and are language independent akin to phonemes in spoken language (which can be considered as speaker and language independent). More recently, a HMM-based approach was developed [15], where signer-independent hand movement subunits are derived based on light supervision. Through a preliminary cross-lingual study it was demonstrated that the subunits derived could be shared across languages. So, the present paper builds on that approach to model hand movement information in a language independent manner for multilingual sign language recognition. In other words, like $\mathbf{z}_t^{\text{hshp}}$ estimator, $\mathbf{z}_t^{\text{hmvt}}$ estimator is also trained with auxiliary sign language resources.

3. EXPERIMENTAL SETUP

To validate the proposed approach, we derived the language independent hand movement subunits from three different languages, namely Swiss German sign language (DSGS) in SMILE database, Turkish sign language (TSL) in HospiSign database and German sign language (DGS) in DGS database. We tested them through cross- and multi-lingual systems in the framework presented in Section 2 by modeling only the hand movement subunits and by modeling both hand movement and hand shape subunits.

3.1. SMILE Swiss German Sign Language Database

The large-scale SMILE Swiss German sign language database [31] (referred as SMILE database) was created in the context of developing an assessment system for lexical signs of Swiss German Sign Language (DSGS[1]). It has 100 isolated signs of a DSGS vocabulary production test. 30 adult signers performed each item three times and the second pass was manually annotated.

In our experimental setup, we only used the second pass annotated as "acceptable signs" (Category 1 or 2 according to the 'Category of sign produced' annotation of the SMILE transcription/annotation scheme presented in [31]). The SMILE database was collected with the Microsoft Kinect v2 sensor and the high speed and high resolution GoPro video cameras. We used the body pose information that are provided in the database which was extracted using the deep-learning-based key point detection library OpenPose as the basis of our feature extraction. To ensure enough samples for each sign (minimum 5 samples/sign), 94 signs were selected out of the 100. The resulting 94 sign data was partitioned in a signer-independent manner into 1263 training set samples from 17 signers, 249 development set samples from 3 signers and 704 test set samples from 10 signers.

3.2. Turkish Sign Language HospiSign Database

HospiSign database is a subset of 33 phrase classes of the continuous BosphorusSign database [32]. The content is Turkish Sign Language (TSL) related to the health domain. The HospiSign subset includes 6 adult signers, with each sign being repeated approximately 6 times by each signer. The database is available upon request from the authors (`https://www.cmpe.boun.edu.tr/`

[1] Deutschschweizerische Gebärdensprache

pilab/BosphorusSign/home_en.html). The database has been recorded with a Kinect camera. We have used the skeletal joint coordinates that are provided in the database as the basis for our feature extraction. In order to conduct a signer-independent experiment, we followed leave-one-signer out protocol. For each of the experiment, the average numbers of samples are: 1084 for training and 210 for testing data.

3.3. DGS Database

The DGS database contains 40 signs from German Sign Language (DGS) produced by 14 non-native right-handed signers. Each sign is repeated utmost 5 times by each person. Because it is non-native signers, the challenge of the DGS database leads in his large variety. The database has been recorded with a Kinect camera and the 3D coordinates of a human skeleton has been tracked using the OpenNI framework. The resulting skeletal joint coordinates has been shared with us by the authors of [9], which we used as the basis for feature extraction. More information about the DGS database can be found in [9].

3.4. Hand Movement Subunit Extraction

The hand movement subunits extraction was done according to the method presented in [15]. Briefly, the 3D skeleton position and velocity of both hands according to three different coordinate centers (head, shoulder and hip center) were used as feature observation (resulting a vector of size 36). In order to compensate the differences in the coordinate system in-between the three databases due to recording settings, before feature extraction, we aligned the skeletons of signers in the DSGS, HospiSign and DGS corpus w.r.t a signer from HospiSign database at the neck joint and scaled by the shoulder width. Then left-to-right HMMs with one mixture Gaussian and diagonal covariance was trained for each sign (sign-based HMM/GMM) and the HMM states are clustered by pairwise comparison of respective Gaussian distributions using the Bhattacharyya distance leading to a clustered subunits states.

For building the sign-based and SU-based MLPs, we first obtained the alignments in terms of the HMM states using either the sign level or the clustered subunits-based HMM/GMM systems. We then trained MLPs classifying HMM states with output non-linearity of softmax and minimum cross-entropy error criterion. We used the 36-dimensional feature observation with four frames preceding context and four frames following context as the MLP input. In our experiments, we trained MLPs with different number of hidden units (600, 800, 1000) and hidden layers (0, 1, 2, 3). The number of hidden units and hidden layers as well as other hyper-parameters such as learning rate and the batch size were chosen according to the frame-level accuracy on the development set. For HospiSign and DGS databases, the data of one signer were used as development set. The MLPs were trained using the Quicknet software [33].

This hand movement subunits extraction step was done according to each sign language separately leading to a stack of posterior probabilities $\mathbf{z}_t^{\text{hmvt}} := \left[\mathbf{z}_t^{\text{hmvt-SL}_1} \ \mathbf{z}_t^{\text{hmvt-SL}_2} \ \cdots \ \mathbf{z}_t^{\text{hmvt-SL}_N}\right]^T$. Where $\mathbf{z}_t^{\text{hmvt-SL}_n}$, denote the probabilistic features corresponding to hand movement subunits derived from sign language SL_n, $n \in \{1,2,\ldots,N\}$. The reason for that is that when we tried to extract a common set of subunits from different corpora, we noticed that during the clustering step the subunits remained separate by languages. This can be explained by the differences in the recording conditions in the different data sets.

3.5. Hand Shape Subunit Extraction

Similar to the earlier work [15], we used the DeepHand net which is trained on one-million hands dataset [13] for hand shape posterior estimation. The one-million hands is a composition of three different sign languages, namely Danish sign language, New Zealand sign language and German sign language. 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 hand shapes (linguistically inspired) presented in https://www-i6.informatik.rwth-aachen.de/~koller/1miohands-data/.

3.6. Recognition Model

Two studies were conducted: one based on the hand movement subunits solely and a second based on both the hand movement and shape subunits. All was developed by left-to-right KL-HMM system. In both cases, we extracted hand movement subunits from either one language (cross-lingual setup) or two languages (multilingual setup).

We also developed a KL-HMM monolingual reference as baseline. All the models was trained with 3 to 30 states. Due to space limitations, we reported the best system. We adopted a leave-one-signer out protocol on the DGS and HospiSign corpus.

4. RESULTS AND ANALYSIS

This section presents the language independent KL-HMM systems evaluation based on first the hand movement subunits and then based on the hand movement and shape subunits.

4.1. Hand Movement Study

Table 1 presents the results of the monolingual reference systems and the KL-HMM based cross- and multi-lingual systems in terms of recognition accuracy *RA* ($\pm$ standard deviation).

It can be observed in Tables 1(a) and 1(b) that the performance of cross- and multi-lingual systems are well above random classification but below monolingual system performance. The low performance can be due to combination of two factors: (a) Differences in recording settings. More precisely in the SMILE database the signs are performed sitting while in the DGS and HospiSign databases standing. Skeleton alignment may not fully compensate for these differences. (b) Vocabulary in each database is limited. As a consequence not all possible movements can be expected to be covered by the derived subunits. Moreover the HospiSign database is composed by phrases while the two other databases are composed by isolated signs; these can influence the nature of the subunits. This fact can explain why adding TSL subunits does not help significantly to recognize DGS or DSGS languages.

Together these results indicate that the derived subunits exhibit sign language independence characteristics. When comparing subunit-based MLP and sign-based MLP KL-HMM systems, it can be observed that the performances are comparable, despite the fact that subunit extraction leads to state reduction.

Table 1. Average *RA* (± standard deviation), over the leave-one-signer out protocol, for reference monolingual systems and cross-/multi-lingual KL-HMM systems using hand movement subunits

(a) Targeted language: DSGS (SMILE database)

hmvt MLP trained on	KL-HMM			
	sign-based MLP		SU-based MLP	
	dim.	*RA*	*dim.*	*RA*
DGS	*281*	46.6	*160*	47.3
TSL	*496*	41.6	*324*	41.5
DGS and TSL	*777*	48.2	*484*	48.4
DSGS	*2257*	57.4	*1946*	55.8

(b) Targeted language: DGS (DGS database)

hmvt MLP trained on	KL-HMM			
	sign-based MLP		SU-based MLP	
	dim.	*RA*±*std*	*dim.*	*RA*±*std*
TSL	*496*	52.5 ± 10.2	*324*	52.2 ± 9.5
DSGS	*2163*	57.3 ± 9.8	*1485*	58.1 ± 9.5
TSL and DSGS	*2659*	57.7 ± 9.8	*1809*	58 ± 10.8
DGS	*281*	65.8 ± 13.1	*217*	68.2 ± 10

(c) Targeted language: TSL (HospiSign database)

hmvt MLP trained on	KL-HMM			
	sign-based MLP		SU-based MLP	
	dim.	*RA*±*std*	*dim.*	*RA*±*std*
DGS	*281*	97.5 ± 1.4	*160*	95.4 ± 2.0
DSGS	*2163*	98.0 ± 1.1	*1485*	98.8 ± 1.0
DGS and DSGS	*2444*	98.1 ± 1.1	*1645*	98.2 ± 1.1
TSL	*300*	97.5 ± 1.7	*217*	97.3 ± 1.7

4.2. Hand Movement and Hand Shape study

Table 2 presents the results of the hand shape based KL-HMM system in terms of recognition accuracy on the three different sign languages (DSGS, DGS and TSL). As it can be observed, in the three databases, the hand shape component is not as good as the hand movement to differentiate the signs. One of the reason can be because of the hand orientation independence of the Deep Hand model. The cropped hand zone is also dependent of the quality of the joint tracking which differs for each database. Moreover the particularly low result of the DGS case can be due to the poorly wild collecting setup of the database. This poor result is the reason why we decided not to pursue the DGS study based on the hand movement and shape subunits.

Table 2. Average *RA* (± standard deviation) of the hand shape based KL-HMM systems on three sign languages (DSGS, TSL and DGS)

	hshp-based KL-HMM
DSGS	38.2
DGS	5.8 ± 2.5
TSL	83.8 ± 8.0

In the next experiment, we combined the hand movement and shape observation to train the KL-HMM system. Table 3 presents these results. As expected, the hand shape component gives complementary information to the hand movement as evidenced by the results. Moreover adding the hand shape decreases the gap in between the monolingual and the cross-/multi-lingual framework. Moreover it is relevant to notice that in [32], the best reported recognition accuracy on the HospiSign database which use hand movement and hand shape information, is 96.67% (± 1.80); and in [5] on the DSGS database is 66.8%.

Table 3. Average *RA* (± standard deviation) for reference monolingual HMM/GMM system and cross-/multi-lingual KL-HMM systems using hand movement and hand shape subunits

(a) Targeted language: DSGS (SMILE database)

hmvt MLP trained on	hshp MLP trained on	KL-HMM	
		sign-based MLP	SU-based MLP
DGS	1 million hand	72.9	72.6
TSL	1 million hand	67.3	66.1
DGS and TSL	1 million hand	72.9	73.2
DSGS	1 million hand	75.6	74.3

(b) Targeted language: TSL (HospiSign database)

hmvt MLP trained on	hshp MLP trained on	KL-HMM	
		sign-based MLP	SU-based MLP
DGS	1 million hand	98.6 ± 1.4	99.0 ± 1.1
DSGS	1 million hand	99.0 ± 1.2	99.1 ± 1.1
DGS and DSGS	1 million hand	99.4 ± 0.7	99.3 ± 0.9
TSL	1 million hand	98.9 ± 0.9	98.9 ± 1.4

5. CONCLUSION AND FUTURE WORK

This paper investigated methods to model hand movement information in a language independent manner using hand movement subunits obtained through HMMs. Our investigations showed that there is a performance gap when modeling hand movement information in a language independent manner and in a language dependent manner. However, this gap is significantly reduced when combined with hand shape information and yields competitive systems. These findings are promising and they pave the path for development of sign language processing systems by sharing multiple sign language resources. Our future work will build upon these finding to address resource-constraint issues in sign language processing such as, developing systems with reduced number of signers and examples. In addition, we will also investigate whether such a multilingual approach can be applied for sign language assessment.

6. REFERENCES

[1] Rachel Sutton-Spence and Bencie Woll, *The Linguistics of British Sign Language: An Introduction*, Cambridge University Press, 1999.

[2] H. Cooper, B. Holt, and R. Bowden, "Sign language recognition," in *Visual Analysis of Humans, 2011*.

[3] T. Starner, J. Weaver, and A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1371–1375, 1998.

[4] Oscar Koller, Necati Camgoz, Hermann Ney, and Richard Bowden, "Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 04 2019.

[5] 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*.

[6] C. Vogler and D. Metaxas, "Parallel hidden Markov models for American sign language recognition," in *Proc. of the Seventh IEEE International Conference on Computer Vision (ICCV)*, Sep. 1999, vol. 1, pp. 116–122 vol.1.

[7] S.-F. Wong and R. Cipolla, "Real-time interpretation of hand motions using a sparse Bayesian classifier on motion gradient orientation images," in *Proc. of the British Machine Vision Conference (BMVC)*, Sept. 2005, vol. 1, pp. 379–388.

[8] G. Awad, J. Han, and A. Sutherland, "Novel boosting framework for subunit-based sign language recognition," in *Proc. of the 16th IEEE International Conference on the Image Processing (ICIP)*.

[9] E.-J. Ong, H. Cooper, N. Pugeault, and R. Bowden, "Sign language recognition using sequential pattern trees," in *Proc. of the IEEE CVPR*, 2012.

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

[11] N. C. Camgöz, S. Hadfield, O. Koller, and R. Bowden, "Subunets: End-to-end hand shape and continuous sign language recognition," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[12] 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.

[13] O. Koller, H. Ney, and R. Bowden, "Deep hand: How to train a CNN on 1 million hand images when your data is continuous and weakly labelled," in *Proc. of the IEEE CVPR*, June 2016.

[14] R. Rasipuram and M. Magimai.-Doss, "Articulatory feature based continuous speech recognition using probabilistic lexical modeling," *Computer Speech and Language*, vol. 36, pp. 233–259, 2016.

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

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

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

[18] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, 1951.

[19] R. Rasipuram and M. Magimai.-Doss, "Acoustic and lexical resource constrained asr using language-independent acoustic model and language-dependent probabilistic lexical model," *Speech Communication*, vol. 68, pp. 23–40, Apr. 2015.

[20] Marzieh Razavi, Ramya Rasipuram, and Mathew Magimai.-Doss, "On modeling context-dependent clustered states: Comparing hmm/gmm, hybrid hmm/ann and kl-hmm approaches," in *Proceedings of ICASSP*, May 2014, pp. 7659–7663.

[21] V. Pitsikalis, S. Theodorakis, C. Vogler, and P. Maragos, "Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition," in *Proc in the IEEE CVPR Workshops*, 2011.

[22] R. Elakkiya and K. Selvamani, "Extricating manual and non-manual features for subunit level medical sign modelling in automatic sign language classification and recognition," *Journal of Medical Systems*, Sep 2017.

[23] H. Cooper, E.J. Ong, N. Pugeault, and R. Bowden, "Sign language recognition using sub-units," *Journal of Machine Learning Research 13*, 2012.

[24] O. Koller, H. Ney, and R. Bowden, "May the force be with you: Force-aligned signwriting for automatic subunit annotation of corpora," in *Proc. of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (AFGR)*, 2013.

[25] B. Bauer and K.-F. Kraiss, "Towards an automatic sign language recognition system using subunits," in *Gesture and Sign Language in Human-Computer Interaction: International Gesture Workshop*, 2002.

[26] H. Junwei, A. George, and S. Alistair, "Modelling and segmenting subunits for sign language recognition based on hand motion analysis," *Pattern Recognition Letters*, vol. 30, no. 6, pp. 623 – 633, 2009.

[27] J. Han, G. Awad, and A. Sutherland, "Boosted subunits: a framework for recognising sign language from videos," *IET Image Processing*, vol. 7, no. 1, 2013.

[28] Sh. Sako and T. Kitamura, "Subunit modeling for Japanese sign language recognition based on phonetically depend multi-stream hidden Markov models," in *Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion*. 2013, pp. 548–555, Springer Berlin Heidelberg.

[29] F. Gaolin, G. Xiujuan, G. Wen, and C. Yiqiang, "A novel approach to automatically extracting basic units from chinese sign language," in *Proc. of the 17th International Conference on Pattern Recognition*, 2004.

[30] S. Theodorakis, V. Pitsikalis, and P. Maragos, "Model-level data-driven sub-units for signs in videos of continuous sign language," in *Proc. in the IEEE ICASSP*, March 2010, pp. 2262–2265.

[31] 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.

[32] N. C. Camgöz, A. A. Kındıroğlu, and L. Akarun, "Sign language recognition for assisting the deaf in hospitals," in *Proc. of the Human Behavior Understanding: 7th International Workshop*, 2016.

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