HMM-BASED APPROACHES TO MODEL MULTICHANNEL INFORMATION IN SIGN LANGUAGE INSPIRED FROM ARTICULATORY FEATURES-BASED SPEECH PROCESSING

Sandrine Tornay[†‡] *Marzieh Razavi*[⋆⋆] *Necati Cihan Camgoz*[⋆]
Richard Bowden[⋆] *Mathew Magimai.-Doss*[†]

[†] Idiap Research Institute, Martigny, Switzerland
[‡] Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[⋆⋆] Telepathy Labs GmbH, Zürich, Switzerland
[⋆] University of Surrey, Guildford, UK

ABSTRACT

Sign language conveys information through multiple channels, such as hand shape, hand movement, and mouthing. Modeling this multichannel information is a highly challenging problem. In this paper, we elucidate the link between spoken language and sign language in terms of production phenomenon and perception phenomenon. Through this link we show that hidden Markov model-based approaches developed to model "articulatory" features for spoken language processing can be exploited to model the multichannel information inherent in sign language for sign language processing.

Index Terms— Sign language, Subunits, Articulatory Features, Hidden Markov Model

1. INTRODUCTION

Sign language (SL) is a visual mode of communication for the Deaf community akin to speech being a mode of communication for the Hearing community. In SL, the information is conveyed through multiple visual channels such as hand gestures (hand shape, location, position and movement), facial expressions, body postures, and lip movements. SL processing presents two main challenges: (1) robust extraction of the multichannel information and (2) modeling of the multichannel information.

Different machine learning techniques have been investigated for modeling signs for sign language recognition (SLR) such as, hidden Markov models (HMM) [1], parallel HMM (PaHMM) [2], relevance vector machines [3] and deep learning methods [4, 5]. The early work of Vogler and Metaxas [6] borrowed heavily from the studies of SL by Liddell and Johnson [7], splitting signs into motion and pause sections. While their later work [2], used PaHMM on both hand shape and motion subunits, as proposed by the linguist Stokoe [8]. This paper focuses on the latter challenge, i.e. modeling of the multichannel information.

SL processing faces data scarcity issues. Thus, the studies have also concentrated on learning sign models in an effective manner from low number of examples. Lichtenauer et al. [9] presented a method to automatically construct a SL classifier for a previously unseen sign. Their method works by collating features for signs from many people then by comparing the features of the new sign to that set. They then construct a new classification model for the target sign. This relies on a large training set for the base features (120 signs by 75 people) yet subsequently allows a new sign classifier to be trained using one shot learning. Bowden et al. [10] also presented a SLR system capable of correctly classifying new signs given a single training example. Their approach used a two-stage classifier bank, the first of which used hard coded classifiers to detect hand shape, arrangement, motion and position "subunits". The second stage removed noise from the 34 bit feature vector (from stage 1) using Independent Component Analysis (ICA), before applying temporal dynamics to classify the sign. Kadir et al. [11] extended this work with head and hand detection based on boosting (cascaded weak classifiers), a body-centered description (normalized movements into a 2D space) and then a two-stage classifier where stage 1 classifier generates linguistic feature vector and stage 2 classifier uses Viterbi on a Markov chain for highest recognition probability. Cooper and Bowden [12] continued this work still further with an approach to SLR that does not require tracking. Instead, a bank of classifiers are used to detect "phonemic" parts of sign activity by training and classifying (AdaBoost cascade) on certain sign subunits. These were then combined into a second stage word-level classifier by applying a first order Markov assumption. The results showed that the detection rates achieved with a large lexicon and few training examples were almost equivalent to a tracking based approach. With the advances in deep learning methods, there has been effort in modeling signs in the framework of hybrid HMM/ANN (Artificial Neural Network) [4] and in the framework of connectionist temporal classification [5]. However these efforts have mainly focused on modeling hand shape information.

This paper develops approaches to model multichannel information in the visual signal for SL processing taking inspirations from spoken language processing. Specifically, we elucidate that when modeling linguistically motivated speech production knowledge, i.e. "articulatory" features (AFs), it is a multichannel information modeling problem akin to SL processing. Through that understanding, we show that the methods developed to model articulatory features (AFs) can be scaled to model the multichannel information for SL processing.

The remainder of the paper is organized as follow: Section 2 presents the proposed approaches of modeling multichannel information in the SL framework. Section 3 presents the experimental setup and Section 4 presents the results and analysis. Finally, Section 5 presents the conclusion and directions for future research.

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.

2. PROPOSED APPROACHES

In both SL and spoken language

(a) there is a production phenomenon that generates a signal. In the case of spoken language, it is movement of articulators like vibration of vocal folds, movement of tongue, lips and jaw that produce time varying 1D acoustic signal. In the case of SL, it is hand gestures, mouthing, body postures and facial expressions that produce time varying 2D visual signal; and

(b) there is a perception phenomenon, which interprets that generated signal in terms of elements of "language", e.g. word, phrases.

Linguistically, the perception phenomenon is better understood in spoken language than SL. More precisely, in spoken language, it is well understood that the time structure of word units can be defined as a sequence of subword units, e.g. phonemes, syllables, which are "perceptual" in nature (i.e. can be heard and distinguished); can be related to the movement of articulators; and can be modelled by parameterizing the spectral characteristics of the speech signal. Such an understanding, however, does not exist yet in the case of SL. More precisely, how hand gestures, facial expressions, body postures, mouthing together create a subword unit like a time structure is not clear yet. It is still an open research problem in sign linguistics.

In spoken language processing, despite the success of spectral feature based approach, there is interest in modeling the production phenomenon related information through AFs [13–15]. More precisely, defining each phoneme in terms of AFs like manner of articulation or degree of constriction, place of articulation, voicing, nasality, rounding, height of tongue, frontness of tongue; estimating these AFs from the speech signal; and then modeling the multichannel AFs through sequential models such as HMM. The AFs in speech processing are synonymous to the "subunits" in SL. This close similarity can be exploited to scale methods developed for AF based processing to SL processing. We study two such methods, namely, standard HMM based approach and Kullback-Leibler divergence HMM (KL-HMM) based approach [16, 17].

2.1. Standard HMM based Approach

One of the common approach to model AFs is to estimate these features using ANNs; transform them using tandem feature extraction technique; concatenate them with the acoustic feature; and model them with HMMs [14, 15, 18, 19]. We can adopt a similar approach for SL processing where the features representing different channels of information are extracted, concatenated $\mathbf{x}_t := \left[\mathbf{x}_t^{\text{hshp}} \ \mathbf{x}_t^{\text{hmvt}} \ \cdots \ \mathbf{x}_t^{\text{facial}}\right]^T$ and then modeled by an HMM. $\mathbf{x}_t^{\text{hshp}}$, $\mathbf{x}_t^{\text{hmvt}}$, $\mathbf{x}_t^{\text{facial}}$ denote the features corresponding to hand shape, hand movement and facial expression, respectively. We will see later in Section 3.2 that the hand movement features can be extracted in the measurement space while the hand shape features can be extracted from the probabilistic representation of the subunits using tandem technique [20].

2.2. KL-HMM based Approach

Another approach is to model AFs as probabilistic features using KL-HMM [15]. Briefly, KL-HMM [16, 17] is an approach where the feature observations are probabilistic (posterior distributions). Each HMM state is parameterized by a categorical distribution of the same dimension as the feature observations. These parameters are estimated through embedded Viterbi expectation maximization algorithm with a cost function based on Kullback-Leibler (KL) divergence [21] between the feature observations and the state categorical distribution. The decoding step remains 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, we can adopt the KL-HMM based AF modeling framework for SL processing, where for each channel we extract probabilistic features and stack them to get the feature observation $\mathbf{z}_t := \left[\mathbf{z}_t^{\text{hshp}} \ \mathbf{z}_t^{\text{hmvt}} \ \cdots \ \mathbf{z}_t^{\text{facial}}\right]^T$. $\mathbf{z}_t^{\text{hshp}}$, $\mathbf{z}_t^{\text{hmvt}}$ and $\mathbf{z}_t^{\text{facial}}$ denote the probabilistic features corresponding to hand shape, hand movement and facial expression, respectively. The HMM state s^i is parameterized by a stack of categorical distribution $\mathbf{y}_{s^i} := \left[\mathbf{y}_{s^i}^{\text{hshp}} \ \mathbf{y}_{s^i}^{\text{hmvt}} \ \cdots \ \mathbf{y}_{s^i}^{\text{facial}}\right]^T$. The local score $S(\mathbf{y}_{s^i}, \mathbf{z}_t)$ is based on KL-divergence [21]. We will see later that the probabilistic features can be based on a form of subunits representation.

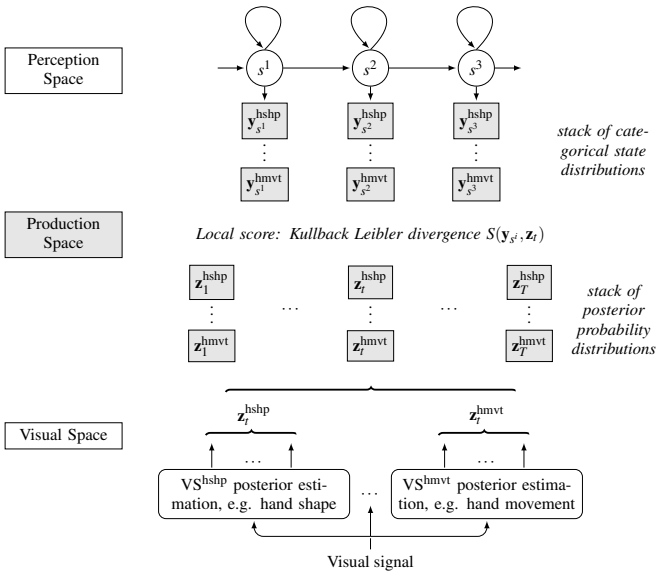

Fig. 1. Schematization of the Kullback Leibler divergence-based Hidden Markov Model (KL-HMM) applied on Sign Language; VS for Visual Subunits.

2.3. HMM topology

In both speech processing and SL processing, the left-to-right HMM serves as the perception space. In speech processing, this space can be defined based on a lexicon that transcribes each word as a sequence of subword units and minimum duration constraints. In SL, however, there is no such luxury. We will show later that this space can be dynamic, i.e. left-to-right HMMs with different number of states $n \in \{N_{min}, \ldots, N_{max}\}$ can be trained and dynamically selected during the recognition phase.

3. EXPERIMENTAL SETUP

We validated both the proposed approaches on a signer independent isolated sign language recognition (SLR) task. For the sake of simplicity, we demonstrate the approach on hand shape (hshp) and hand movement (hmvt) information. The remainder of the section presents the experimental setup.

3.1. SMILE Swiss German Sign Language Dataset

We validate the proposed systems on the large-scale SMILE Swiss German Sign Language Dataset (referred as SMILE dataset in the following) presented in [22]. The SMILE dataset 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. 11 adult L1 signers and 19 adult L2 learners performed each item three times and only the second pass was manually annotated. The SMILE dataset was collected with the Microsoft Kinect v2 sensor and the high speed and high resolution GoPro video cameras. The color videos, depth maps, user masks and 3D pose information obtained from the Kinect, the body pose, facial landmarks, and hand pose information extracted using the deep-learning-based key point detection library OpenPose are provided.

In our experimental setup, we only used the second pass annotated as Category 1 or 2 according to the 'Category of sign produced' annotation of the SMILE transcription/annotation scheme (presented in [22]). This annotation evaluates, through six categories, the acceptability of a sign according to linguistic criteria; Category 1 and 2 being acceptable signs with the same or slightly the same form. We did not make any difference between the L1 and L2 signers in our experiment. 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. Feature Estimation

3.2.1. Hand Shape Features

The estimator of the hand shape component used in this paper was the Deep Hand approach developed by Koller and al. in [23]; i.e. the convolutional neural network (CNN) associated with the EM algorithm. The CNN has been trained on the one-million hands dataset [23]. The input features are sequences of images of a cropped hand and the output of the estimator is the hand shape class-conditional posterior probabilities $\mathbf{z}_t^{\text{hshp}}$, where the hand shape 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/`.

The hand shape feature $\mathbf{x}_t^{\text{hshp}}$ for the standard HMM approach was extracted by transforming $\mathbf{z}_t^{\text{hshp}}$ using tandem feature extraction technique. Briefly, $\mathbf{x}_t^{\text{hshp}} := \text{KLT}(\log(\mathbf{z}_t^{\text{hshp}}))$, where KLT denotes Kahunen Loeve Transform [20].

3.2.2. Hand Movement Features

Inspired from [24], two types of feature observations were used as input of the hand movement posterior feature estimator: (i) the 3D skeleton position of both hands obtained in three different coordinate systems (based on the head center or the hand corresponding shoulder or hip center) normalized by the head width and (ii) the corresponding velocities of the three coordinate systems computed by subtracting the position features at time t to them at time $t-2$. The resulting vector, $\mathbf{x}_t^{\text{hmvt}}$, is of size 36 ($= (3 \text{ dimensions} \times 2 \text{ hands}) \times 3$ coordinate systems $+18$ velocity features).

For the KL-HMM approach, there is no well defined subunit extraction approach. So following the work presented in [25], we built a sign-specific left-to-right HMM by modeling $\mathbf{x}_t^{\text{hmvt}}$ with Gaussian mixture models (GMMs). HMM states can be expected to segment the sequence of feature observations into steady state segments or sub-movements. So we regarded those states as the movement subunits and estimated $\mathbf{z}_t^{\text{hmvt}} := [z_t^1, \ldots, z_t^I]^T$ by using the GMMs of the states and applying Bayes' rule. The total number of HMM states is $I = 849$. It can be noted that as the vocabulary or lexeme size increases I also increases. This may lead to poor model estimation due to curse of dimensionality. One way to handle this issue is by inferring hand movement subunits [25–27].

3.3. Recognition Models

We built three systems for each of the two proposed approaches. Hand movement-based system (**M**), Hand shape-based system (**S**) and Hand shape-plus-Hand movement based system (**M+S**).

The number of states in the left-to-right HMM where varied from 3 (N_{min}) to 9 (N_{max}). So each sign had 7 different HMMs. This range was found on the development set.

For the standard HMM approach, each state was modeled by 4 mixture GMMs for system **M** and by a single Gaussian for systems **S** and **M+S**. We found that increasing the number of mixture of Gaussians for **S** and **M+S** did not help in improving performance.

During **ms 3 to 9** recognition phased, the decoder selected from 94×7 sign models the sign's model that yielded the maximum likelihood in the case of standard HMM approach and the sign's model that yielded the minimum KL-divergence score.

The standard HMM based approach was implemented using HTK [28]. The KL-HMM based approach was implemented using an in-house modified version of HTK.

4. RESULTS AND ANALYSIS

This section presents the systems evaluation followed by an analysis.

4.1. Systems Evaluation

Table 1 shows the recognition accuracy for the standard HMM approach and the KL-HMM approach. We report the performance with the different left-to-right HMM topology. **ms 3 to 9** denotes the performance by dynamically selecting the model during decoding.

Table 1. Recognition accuracy of the standard HMM and the KL-HMM approaches applied on the hand movement features (**M**), the hand shape features (**S**) and combined ones (**M+S**).

	Standard HMM			**KL-HMM**		
#state	**M**	**S**	**M+S**	**M**	**S**	**M+S**
3	44.4	47.7	63.8	41.5	25.9	59.5
4	47.2	47.6	63.4	39.9	28.8	60.5
5	48.5	49.3	64.8	41.8	28.0	60.1
6	49.8	45.3	65.5	43.0	28.0	62.4
7	48.1	46.6	66.1	41.2	30.7	60.5
8	50.2	44.6	63.7	43.3	32.5	62.1
9	50.4	43.5	65.9	41.9	30.7	61.7
ms 3 to 9	51.6	50.3	**66.8**	44.3	32.8	**63.1**

It can be observed that, in both the approaches, **M+S** leads to outperform hand shape alone and hand movement alone systems. Moreover the model selection method **ms 3 to 9** yields the best system for both the approaches. When comparing across the approaches, the standard HMM approach yields better system than KL-HMM. Low performance for system **S** in KL-HMM approach can be explained from the fact the Deep Hand hand shape posterior feature estimator has not observed any SMILE dataset. However, the standard HMM approach uses SMILE training data to get the KLT matrix. Low performance for system **M** in KL-HMM approach could

[1] Deutschschweizerische Gebärdensprache

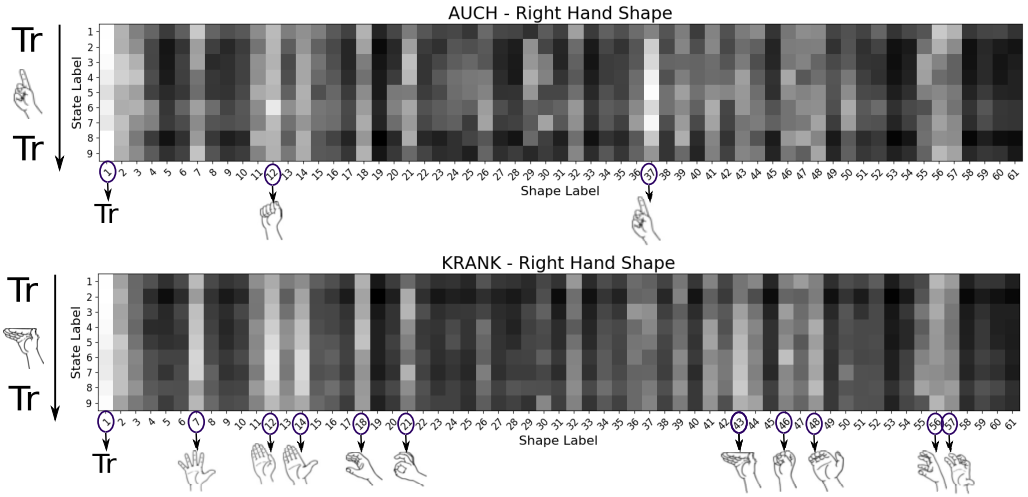

Fig. 2. Density plots of the log right hand shape categorical distribution linked to each KL-HMM states for AUCH and KRANK sign's model (the brighter meaning the more probable). **Tr** is used for the Transition shape.

be attributed to the use of GMMs to estimate the posterior probability. This can be improved by replacing the GMM based posterior estimation by ANN based posterior estimation. This line of investigation is part of our future work.

4.2. Further Analysis

The proposed approach, in particular KL-HMM approach, allows further simplifications. For instance, the hand movement can be decomposed into position and velocity and can be modeled independently with hand shape. We demonstrate that through an experiment with KL-HMM approach. We used the method used to derive the hand movement posterior feature estimator (see Section 3.2.2) to obtain the hand position $\mathbf{z}_t^{\text{hpos}}$ and the hand velocity $\mathbf{z}_t^{\text{hvel}}$ posterior feature estimators based on the hand position and velocity features separately. Table 2 presents the results. System **P** denotes modeling of $\mathbf{z}_t^{\text{hpos}}$ alone. System **V** denotes modeling of $\mathbf{z}_t^{\text{hvel}}$ alone. **P+S** and **V+S** denotes modeling of hand shape posterior feature $\mathbf{z}_t^{\text{hshp}}$ along with $\mathbf{z}_t^{\text{hpos}}$ and $\mathbf{z}_t^{\text{hvel}}$, respectively. We can observe the same trends as before that jointly modeling hand shape and hand position or hand velocity information helps. It can be observed that separating the hand movement features into position and velocity (the **P+V** system) does not affect the performance in comparison to the **M** system. Furthermore, we see that there is a slight increase in the performance of system **V+S** when compared to system **M+S**.

One of the advantages of KL-HMM approach is that the parameters i.e. the categorical distribution of HMM states can be interpreted. Figure 2 shows the hand shape categorical distributions of the 9 states **V+S** system of two signs: AUCH and KRANK. In the AUCH case, the **V** system recognized 0 samples out of 9, the **S** system 3 samples and the **V+S** one 6 samples; thus we can hypothesize that the hand shape information is the major source of information in that case. Indeed the density plot of the hand shape categorical distributions shows that the model contains relevant information since the sequence of maximum distribution by state (1, 1, 1, 37, 37, 1, 37, 1, 1) corresponds to the true label (1, 37, 1). In the KRANK case, the reverse can be observed; adding the hand shape features adds confusion in the recognition task. The **V** system recognized 5 samples out of the 7, the **S** system 1 sample and **V+S** 3 samples. The density plot confirms the fact that there is a confusion in the model itself since the resulting hand shapes are the transition shape for all the states. This can be partly attributed to high signer variations in the hand shapes used in the training data.

Table 2. Recognition accuracy of the KL-HMM approach applied on the hand position features (**P**), the hand velocity features (**V**), both features (**P+V**), and each combined with the hand shape features (·**+S**)

	KL-HMM					
#state	**P**	**V**	**P+V**	**P+S**	**V+S**	**P+V+S**
3	30.7	36.4	40.0	50.7	59.4	58.8
4	30.5	38.5	42.2	53.0	61.2	59.9
5	30.8	40.1	44.9	53.3	62.1	60.4
6	31.3	40.1	46.0	53.0	61.8	60.8
7	31.0	37.2	44.7	53.3	62.4	62.2
8	33.4	40.3	45.3	53.1	63.9	60.2
9	32.5	39.6	44.7	54.7	64.1	61.1
ms 3 to 9	32.5	40.5	43.5	54.4	**64.5**	61.4

Another relevant use, and not least, of this feature separation property is in the assessment framework, where it allows to find the type of error and also when it appears.

5. CONCLUSION AND FUTURE WORK

This paper showed that, although spoken language and SL are different modes of communication, there is similarity when it comes to modeling the synergy between the production phenomenon and the perception phenomenon through the observed speech signal or visual signal. We showed that this similarity can be exploited to import methods developed for AF modeling in speech processing to develop methods to effectively model hand shape and hand movement information. The method as such is not restrictive to hand shape and hand movement information. Other information such as facial expression, mouthing could be modeled by feature augmentation.

Our future includes: (a) validation of the developed approaches on a continuous sign language recognition task and (b) understanding the differences between L1 and L2 signers by exploiting the KL-HMM approach.

6. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep sign: Hybrid CNN-HMM for continuous sign language recognition," in *Proc. of BMVC*, 2016.

[5] 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 ICCV*, 2017.

[6] C. Vogler and D. Metaxas, "Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods," in *Proc. of the IEEE on Systems, Man, and Cybernetics*, Orlando, FL, USA, Oct. 12 – 15 1997, vol. 1, pp. 156 – 161.

[7] S. K. Liddell and R. E. Johnson, "American sign language: The phonological base ," *Sign Language Studies*, vol. 64, pp. 195–277, 1989.

[8] W. Stokoe, "An outline of the visual communication systems of the American deaf," *Studies in linguistics: Occasional papers*, vol. 86, 1960.

[9] J. Lichtenauer, E. Hendriks, and M. Reinders, "Learning to recognize a sign from a single example," in *Proc. of the 8th IEEE International Conference on Automatic Face and Gesture Recognition*, Sep. 2008, pp. 1–6.

[10] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M Brady, "A linguistic feature vector for the visual interpretation of sign language," in *Proc. of the European Conference on Computer Vision (ECCV) 2004*. pp. 390 – 401, Springer.

[11] T. Kadir, R. Bowden, E. J. Ong, and A Zisserman, "Minimal training, large lexicon, unconstrained sign language recognition," in *Proc. of the BMVC*, 2004, vol. 2, pp. 939 – 948.

[12] H. Cooper and R. Bowden, "Large lexicon detection of sign language," in *Proc. of ICCV, Workshop: Human-Computer Interaction*, 2007, pp. 88–97.

[13] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, February 2007.

[14] K. Livescu, Ö. Çetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, S. Bezman, Dawson-Haggerty, B. Woods, J. Frankel, M. Magimai-Doss, and K. Saenko, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU Summer Workshop," in *Proc. of the IEEE ICASSP*, 2007.

[15] 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.

[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] J. Frankel, M. Magimai-Doss, S. King, K. Livescu, and Ö. Çetin, "Articulatory feature classifiers trained on 2000 hours of telephone speech," in *Proc. of Interspeech*, 2007.

[19] Ö. Çetin, M. Magimai-Doss, A. Kantor, S. King, C. Bartels, J. Frankel, and K. Livescu, "Monolingual and crosslingual comparison of tandem features derived from articulatory and phone MLPs," in *Proc. of Automatic Speech Recognition and Understanding Workshop*, 2007.

[20] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. of the IEEE ICASSP*, 2000, vol. 3, pp. 1635–1638.

[21] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 03 1951.

[22] S. Ebling et al., "SMILE Swiss German sign language dataset," in *Proc. of the Language Resources and Evaluation Conference*, 2018.

[23] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[24] O. Aran, *Vision based sign language recognition: modeling and recognizing isolated signs with manual and non-manual components*, Ph.D. thesis, Bogazici University, Istanbul, Turkey, 2008.

[25] S. Tornay, M. Razavi, and M. Magimai.-Doss, "Data-driven movement subunit extraction from skeleton information for modeling signs and gestures," Idiap Research Report Idiap-RR-02-2019, Idiap Research Institute, Switzerland, 2019, `https://publidiap.idiap.ch/downloads/reports/2018/Tornay_Idiap-RR-02-2019.pdf`.

[26] S. Sako and T. Kitamura, "Subunit modeling for Japanese sign language recognition based on phonetically depend multi-stream hidden Markov models," in *Proc. of the Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion*, Constantine Stephanidis and Margherita Antona, Eds., Berlin, Heidelberg, 2013, pp. 548–555, Springer Berlin Heidelberg.

[27] 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. of the CVPR Workshops*, June 2011, pp. 1–6.

[28] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University Engineering Department, 2002.