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

Sandrine Tornay[1,2] *Marzieh Razavi*[3] *Necati Cihan Camgöz*[4]

Richard Bowden[4] *Mathew Magimai.-Doss*[1]

1 - Idiap Research Institute, Martigny, Switzerland
2 - Ecole polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland
3 - Telepathy Labs GmbH, Zürich, Switzerland
4 - University of Surrey, Guildford, UK

FNSNF
Swiss National Science Foundation
funded the project SMILE

idiap
INSTITUT DE RECHERCHE

EPFL

Sign language

Content of the video:

Human right - Freedom of language

Swiss french sign language

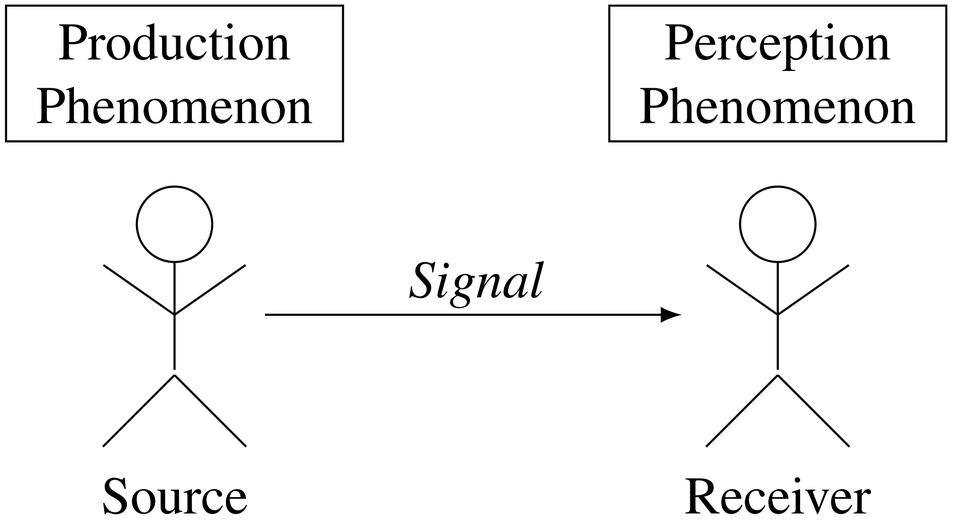
Multichannel information:

- Manual features - *hand shape, orientation, motion, position and location*
- Non-manual features - *facial expression, body posture, mouthings, mouth gestures*

Challenges:

1. How to reliably extract the multichannel information?
2. How to model the multichannel information?

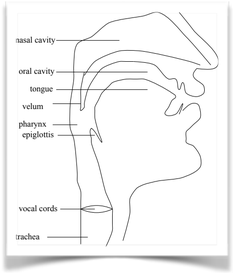
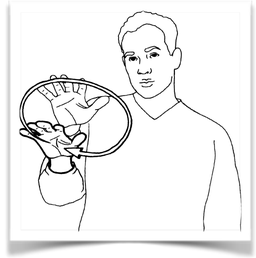
Communication process (1)



Necessitates:

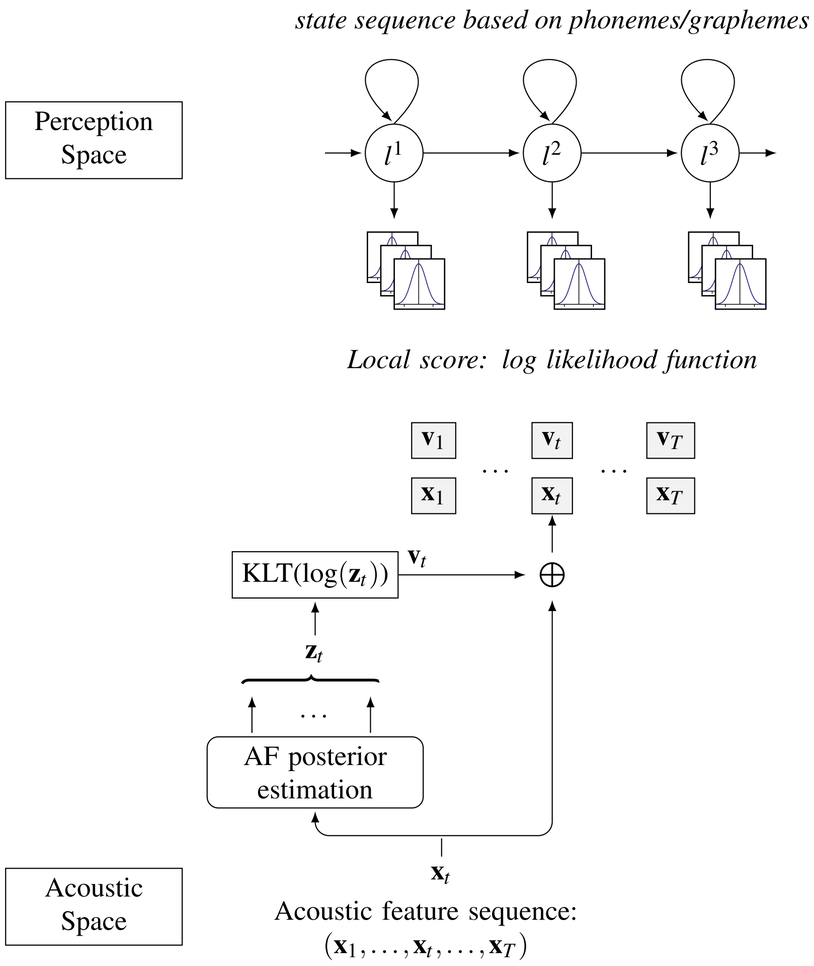
Modelling the synergy between the production phenomenon and the perception phenomenon in relation to the signal.

Communication process (2)

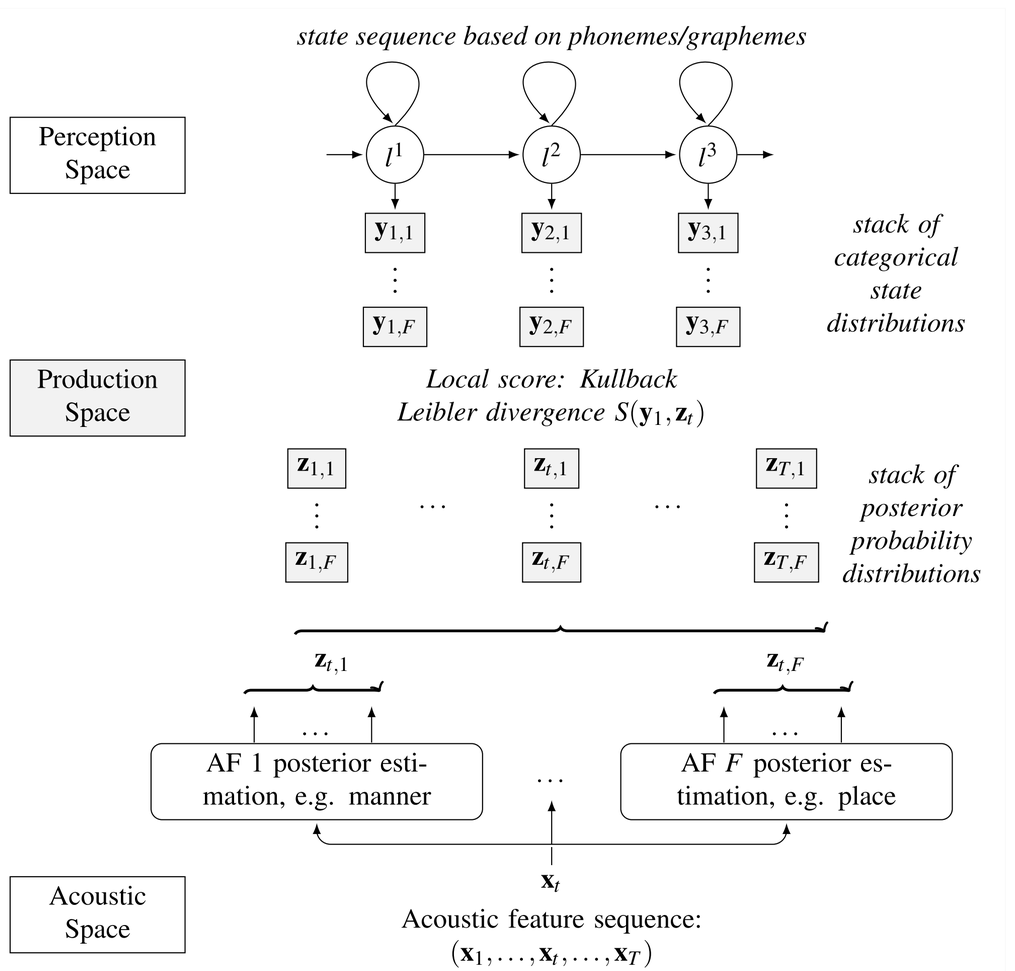
	Spoken Language			Sign Language		
	Production	→ Signal →	Perception	Production	→ Signal →	Perception
Communication process		→ Speech Signal →	Sequence of phonemes, words and sentences		→ Visual Signal →	Sequence of words and phrases
Production	Movement of tongue, lips and jaw, vocal fold vibration			Hand gestures, body posture, mouthing, facial expression, etc.		
Signal	**Audio** Sequential *(Cepstral feature seq.)* Parallel + Sequential *(Articulatory features)*			**Video** Parallel + Sequential		
Perception	Auditory *Seq. of phonemes, words and sentences*			Visual *Seq. of words and phrases*		

Inspiration from speech processing

Tandem approach



KL-HMM based approach



[14]* K. Livescu et al., « Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU Summer Workshop », in Proc. of the 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, 2016

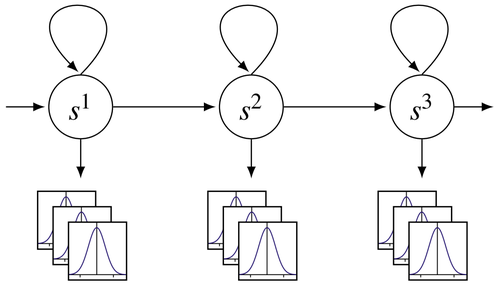
*Citation numbering comes from the paper

Proposed approaches

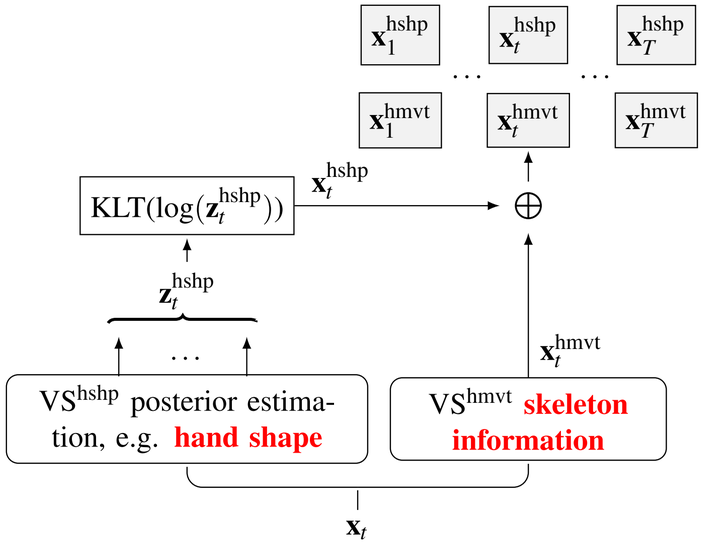
Tandem approach

How to define the state sequence?

Perception Space



Local score: log likelihood function



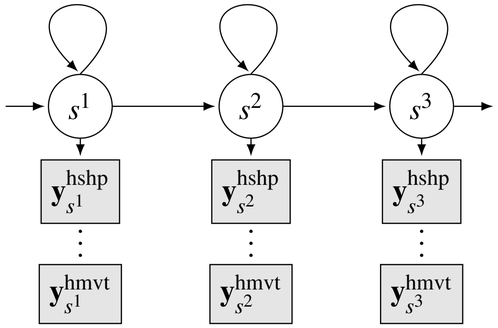
Visual Space

Visual feature sequence:
$(\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T)$

KL-HMM based approach

How to define the state sequence?

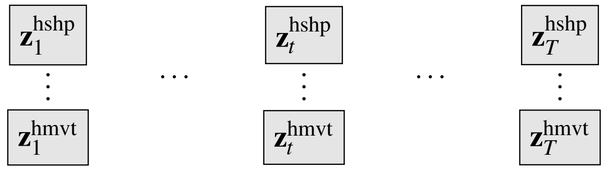
Perception Space



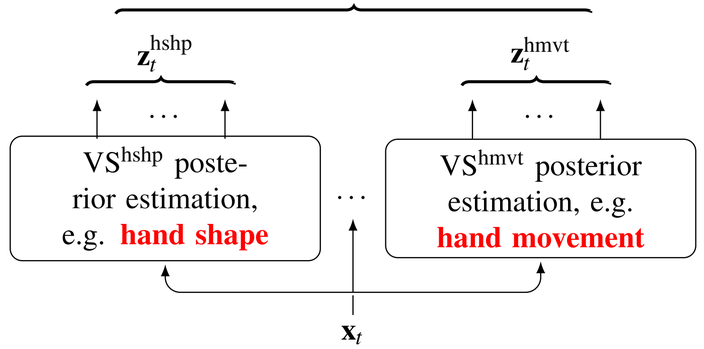
stack of categorical state distributions

Production Space

Local score: Kullback Leibler divergence $S(\mathbf{y}_{s^i}, \mathbf{z}_t)$



stack of posterior probability distributions



Visual Space

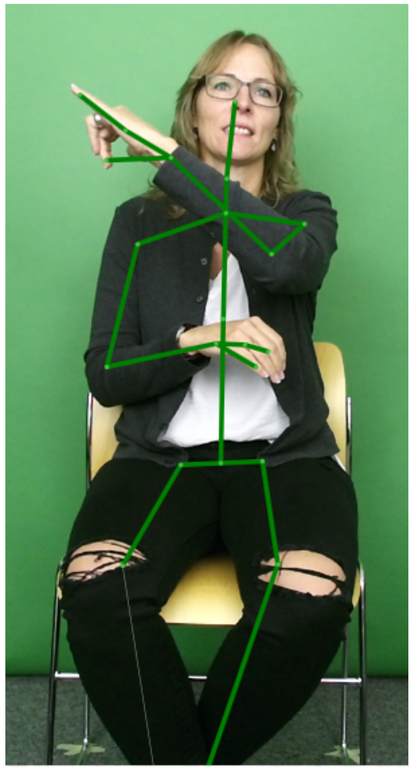
Visual feature sequence:
$(\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T)$

Dataset

SMILE Swiss German Sign Language Dataset

The SMILE dataset was created in the context of developing an assessment system for lexical signs of Swiss German sign language (DSGS)

- **100 isolated signs** of a DSGS vocabulary production test - **94 selected**;
- 11 adult L1 signers and 19 adult L2 signers = **30 signers** - (17 train, 3 dev, 10 test);
- SMILE dataset was collected with the **Microsoft Kinect v2 sensor** and the high speed and high resolution **GoPro video cameras**;
- Each sign was performed 3 times and only **the second pass was annotated -** Only the annotated «acceptable» signs of second pass was used in the following experiment;



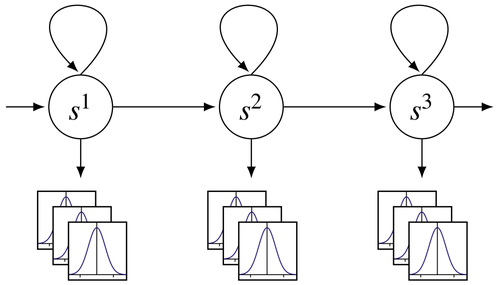
[22] S. Ebling et al., « SMILE Swiss German sign language dataset », in Proc. of the Language Resources and Evaluation Conference 2018

Proposed approaches

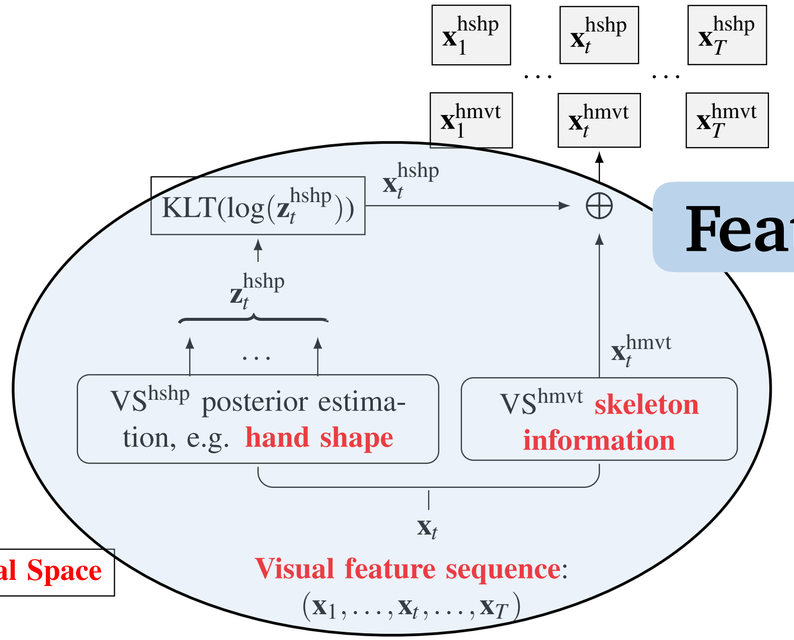
Tandem approach

How to define the state sequence?

Perception Space



Local score: log likelihood function

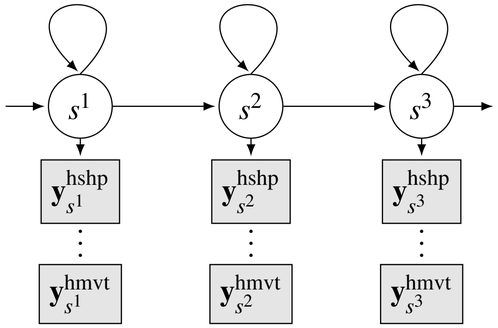


Visual Space

KL-HMM based approach

How to define the state sequence?

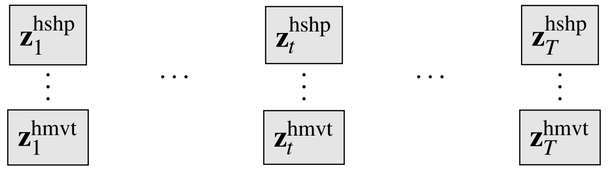
Perception Space



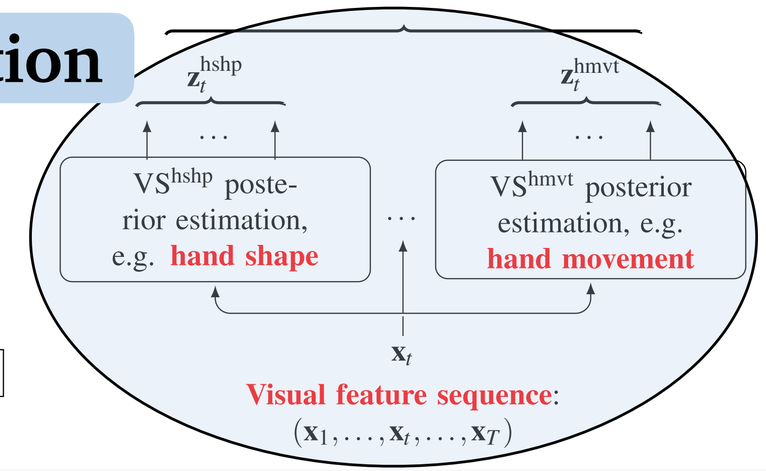
stack of categorical state distributions

Production Space

Local score: Kullback Leibler divergence $S(\mathbf{y}_{s^i}, \mathbf{z}_t)$



stack of posterior probability distributions

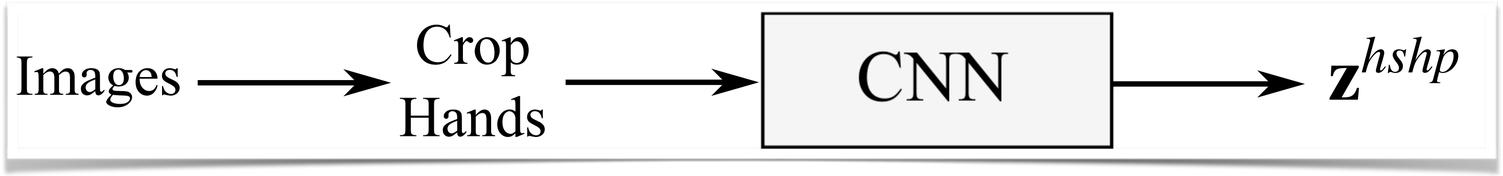


Visual Space

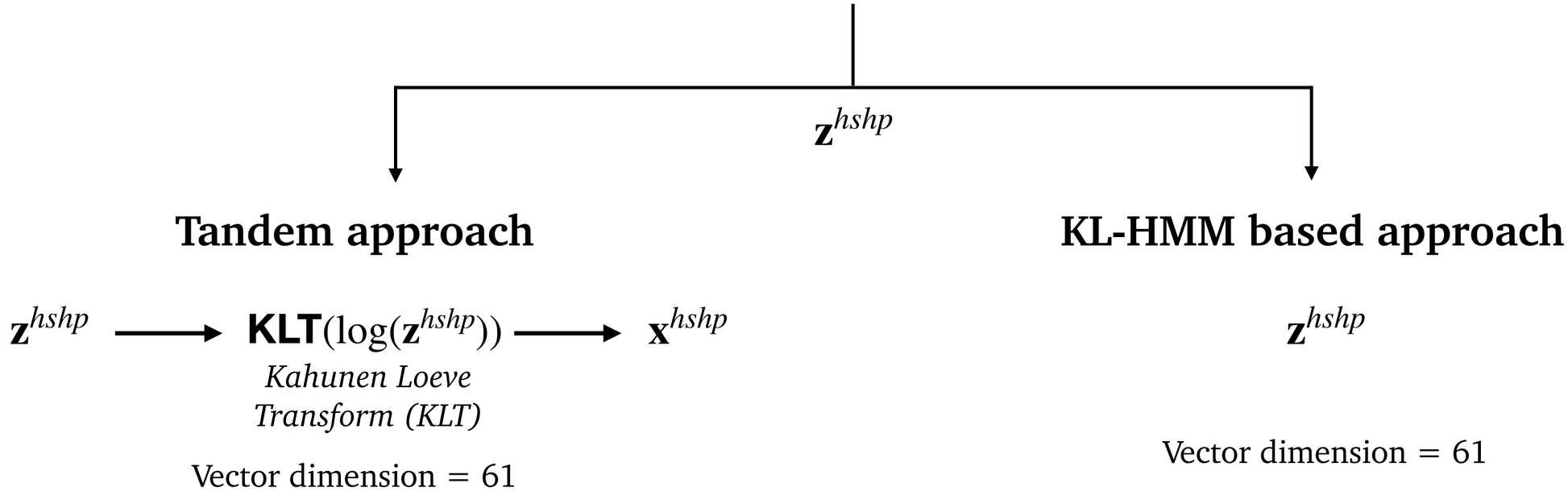
Feature Extraction

Feature extraction - Hand shape

We used the Deep hand net developed by [23] :

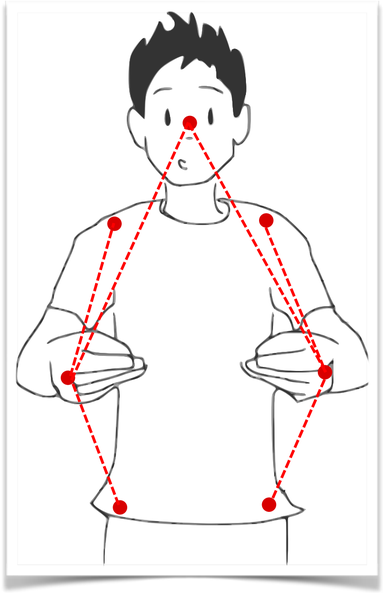


The CNN is trained on the **1-Million-Hands dataset [23]** containing 60 hand shapes + 1 transition shape.



[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 CVPR 2016

Feature extraction - Hand movement



+

- **Position features** given by 3D coordinate of a human skeleton:

$$\mathbf{p}_t^{\mathbf{C}} = \frac{\mathbf{hand} - \mathbf{C}}{|\mathbf{head} - \mathbf{neck}|/4}, \text{ where } \mathbf{C} \in \{\mathbf{head},\ \mathbf{shoulder},\ \mathbf{hip}\};$$

- **Velocity features** given by delta features: $\mathbf{v}_t^{\mathbf{C}} = \mathbf{p}_t^{\mathbf{C}} - \mathbf{p}_{t-2}^{\mathbf{C}}$.

=

Movement features are concatenation of position and velocity of both hands according to head, shoulder hip coordinate centers.

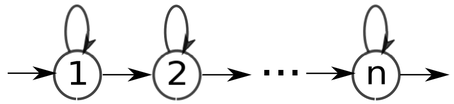
$\mathbf{x}^{hmvt}$

Tandem approach

$\mathbf{x}^{hmvt}$

Vector dimension = 36

KL-HMM based approach



$\mathbf{x}^{hmvt}$ → GMM → $\mathbf{z}^{hmvt}$

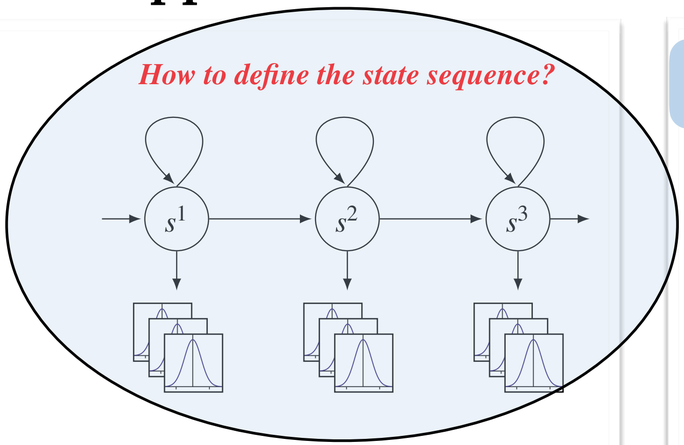
priors

Vector dimension = 849

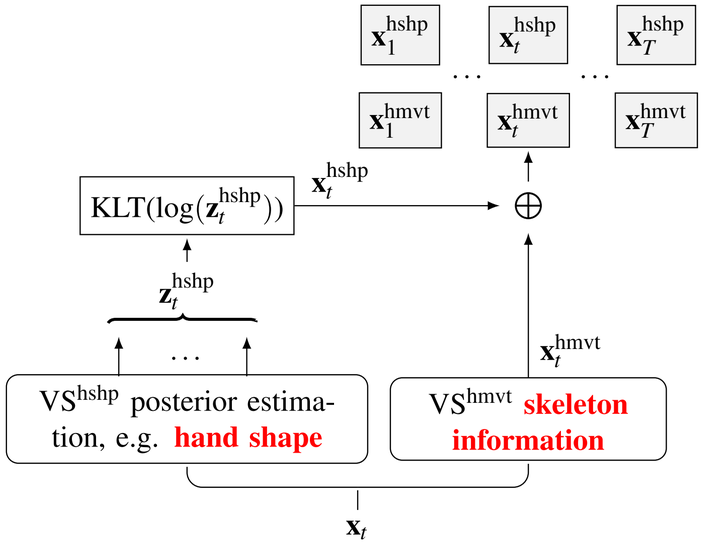
Proposed approaches

Tandem approach

Perception Space



Local score: log likelihood function



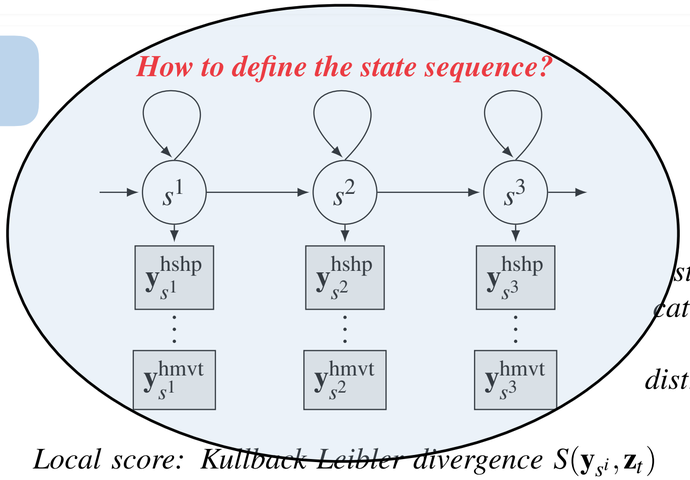
Visual Space

Visual feature sequence:
$(\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T)$

KL-HMM based approach

Model

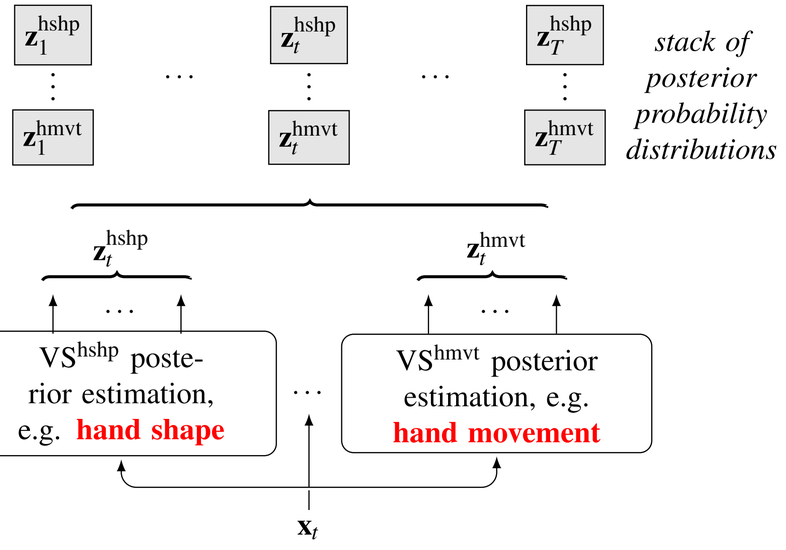
Perception Space



stack of categorical state distributions

Production Space

Local score: Kullback-Leibler divergence $S(\mathbf{y}_{s^i}, \mathbf{z}_t)$



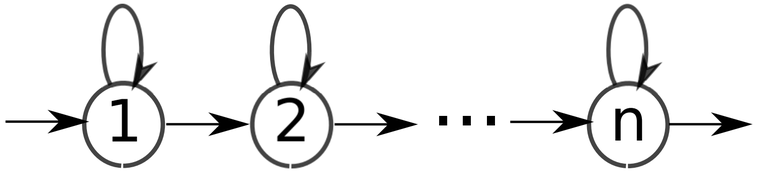
stack of posterior probability distributions

Visual Space

Visual feature sequence:
$(\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T)$

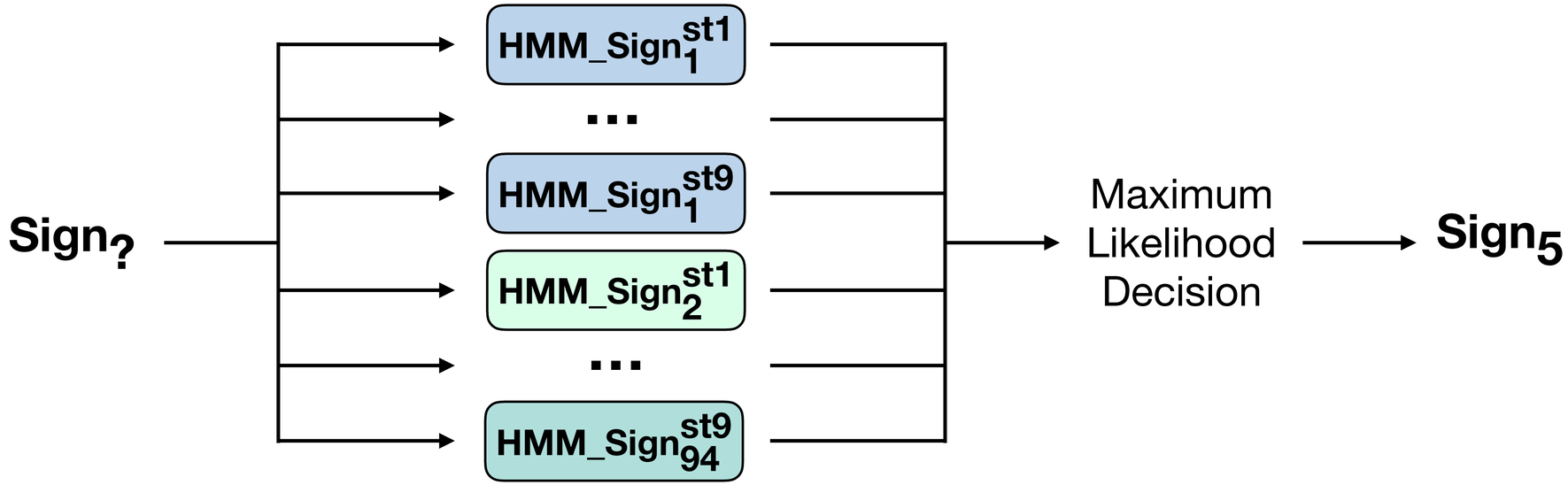
HMM-based recognition framework

- **Training:** Multiple left-to-right HMM-based systems are trained for each sign



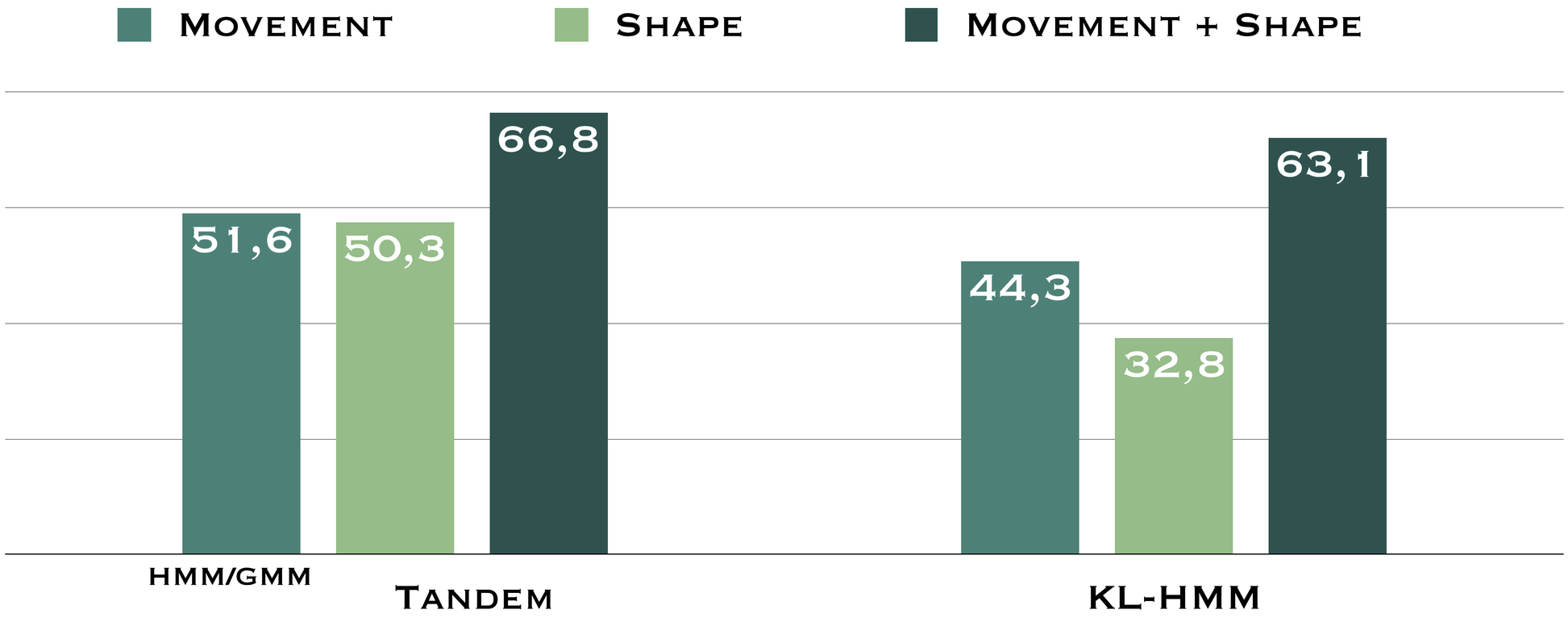
depending on the number of state n: where $3 \leq n \leq 9$

- **Recognition:** Model selection framework

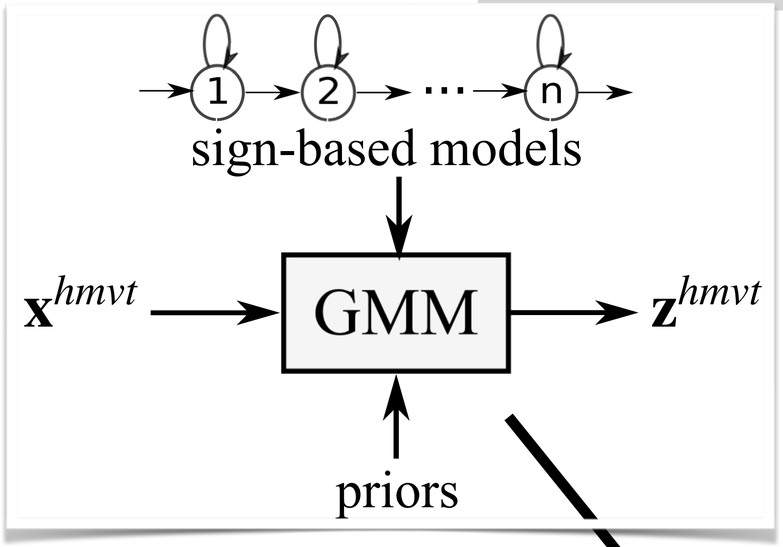


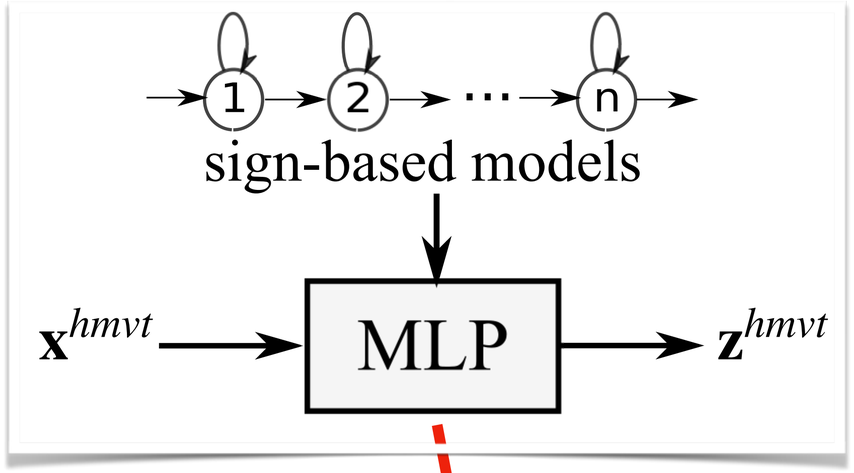
Sign language recognition results

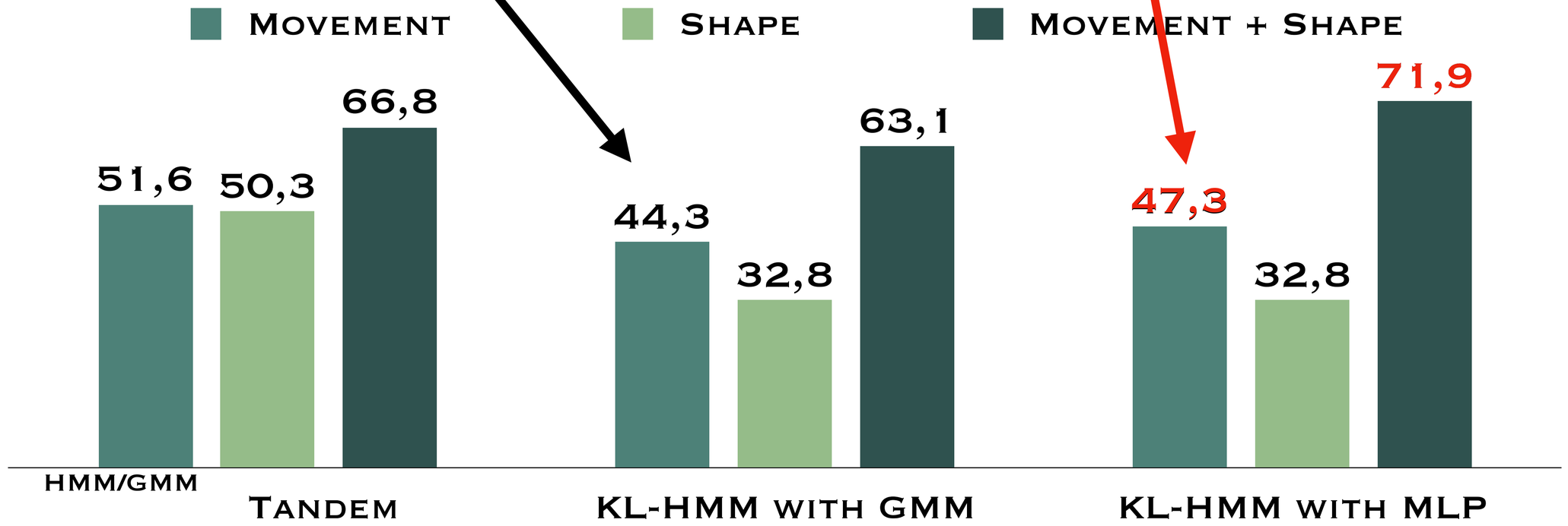
Sign recognition accuracy



Analysis - Hand movement feature





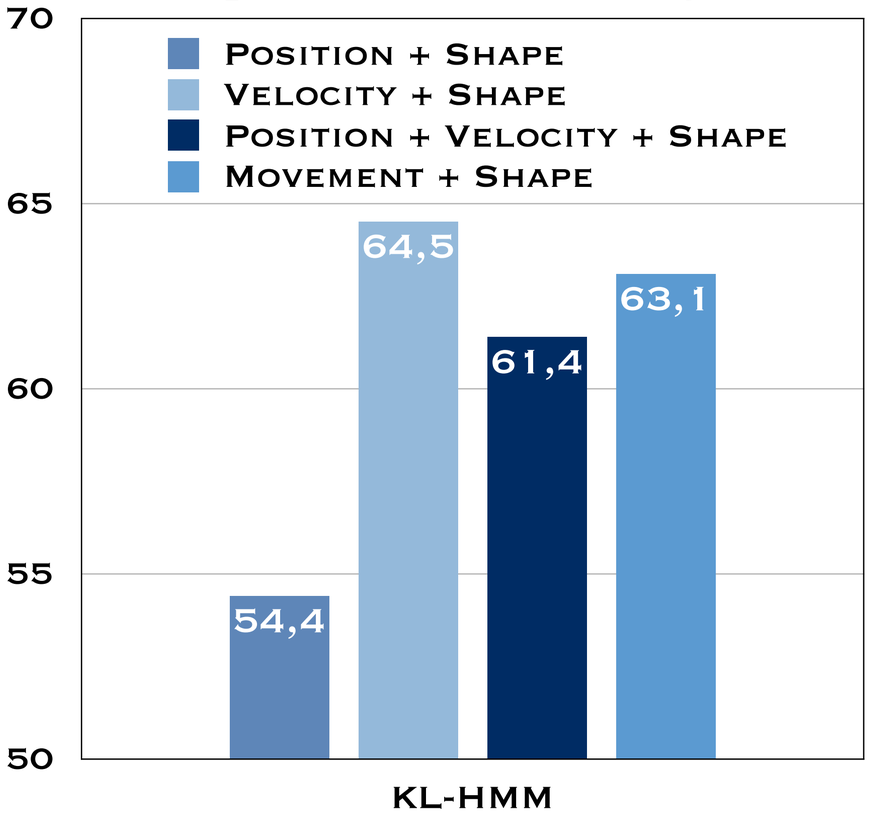


Further analysis

Advantages of KL-HMM approach

1

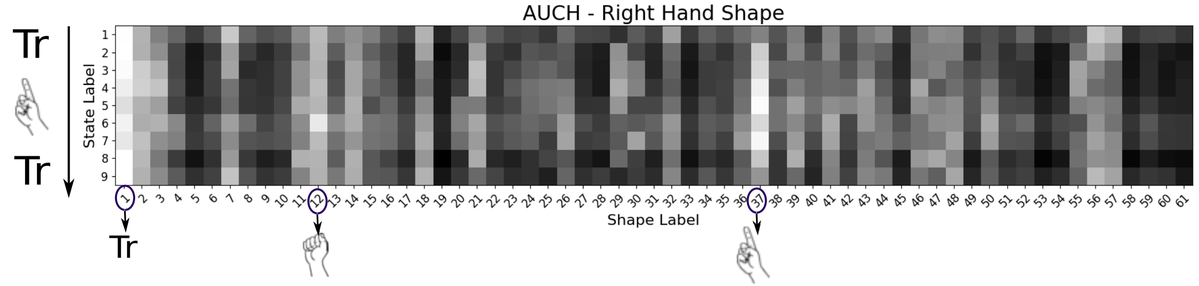
Possibility of separating the movement features into position and velocity



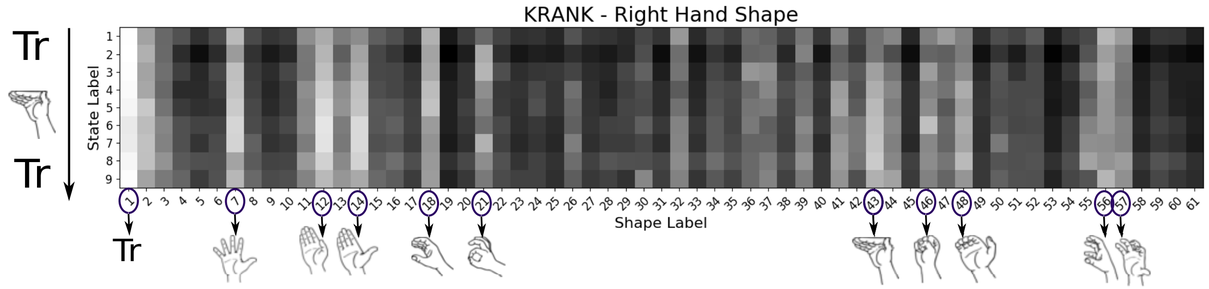
2

Possibility to analyse the trained categorical distributions ;
Here the hand shape ones

Well-recognised example



Poor-recognised example



Summary

- elucidated the link between articulatory feature-based speech processing and sign language processing
- proposed two HMM-based approaches to model multichannel information in sign language
  - Tandem approach and KL-HMM based approach
  - Both approaches yielded promising results
- demonstrated flexibility and interpretability of the KL-HMM approach

On-going Work ➡ developing an assessment system for Swiss German sign language learners (SMILE project).

A video of the assessment system demonstrator is available at:
www.idiap.ch/project/smile/news/smile-how-it-works

Questions ?

sandrine.tornay@idiap.ch