HMM-based approaches to model multichannel information in sign language inspired from articulatory features-based speech processing

Sandrine Tornay\textsuperscript{1,2} Marzieh Razavi\textsuperscript{3} Necati Cihan Camgöz\textsuperscript{4}
Richard Bowden\textsuperscript{4} Mathew Magimai.-Doss\textsuperscript{1}

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

funded the project SMILE
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?
Necessitates:
Modelling the synergy between the production phenomenon and the perception phenomenon in relation to the signal.
**Communication process (2)**

<table>
<thead>
<tr>
<th></th>
<th><strong>Spoken Language</strong></th>
<th><strong>Sign Language</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Production</strong></td>
<td>Movement of tongue, lips and jaw, vocal fold vibration</td>
<td>Hand gestures, body posture, mouthing, facial expression, etc.</td>
</tr>
<tr>
<td><strong>Signal</strong></td>
<td>Audio Sequential <em>(Cepstral feature seq.)</em></td>
<td>Video Parallel + Sequential</td>
</tr>
<tr>
<td></td>
<td>Parallel + Sequential <em>(Articulatory features)</em></td>
<td></td>
</tr>
<tr>
<td><strong>Perception</strong></td>
<td>Auditory Seq. of phonemes, words and sentences</td>
<td>Visual Seq. of words and phrases</td>
</tr>
</tbody>
</table>
Inspiration from speech processing

**Tandem approach**

- **Perception Space**
- **Production Space**
- **Acoustic Space**

State sequence based on phonemes/graphemes

Local score: log likelihood function

KLT(log(z_t)) → AF posterior estimation, e.g. manner

... → AF posterior estimation, e.g. place

Acoustic feature sequence: (x_1, ..., x_t, ..., x_T)

**KL-HMM based approach**

- **Perception Space**
- **Production Space**
- **Acoustic Space**

State sequence based on phonemes/graphemes

Local score: Kullback-Leibler divergence S(y_t, z_t)

... → AF posterior estimation, e.g. manner

... → AF posterior estimation, e.g. place

Acoustic feature sequence: (x_1, ..., x_t, ..., x_T)


*Citation numbering comes from the paper*
**Proposed approaches**

### Tandem approach

**How to define the state sequence?**

![Diagram showing the tandem approach](image)

**Local score: log likelihood function**

$\text{KLT}(\log(z_{ht}))$

$\text{VS}_{\text{hsp}}$ posterior estimation, e.g. hand shape

$\text{VS}_{\text{hmvt}}$ skeleton information

**Visual Space**

**Visual feature sequence:** $(x_1, \ldots, x_t, \ldots, x_T)$

### KL-HMM based approach

**How to define the state sequence?**

![Diagram showing the KL-HMM based approach](image)

**Local score: Kullback Leibler divergence $S(y_s, z_t)$**

$\text{VS}_{\text{hsp}}$ posterior estimation, e.g. hand shape

$\text{VS}_{\text{hmvt}}$ posterior estimation, e.g. hand movement

**Visual Space**

**Visual feature sequence:** $(x_1, \ldots, x_t, \ldots, x_T)$

**Stack of categorical state distributions**

**Stack of posterior probability distributions**
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;
Proposed approaches

Tandem approach

How to define the state sequence?

Perception Space

Local score: log likelihood function

Production Space

KL-HMM based approach

How to define the state sequence?

Perception Space

Local score: Kullback Leibler divergence \( S(y_s, z_t) \)

Production Space

Visual feature sequence: \( (x_1, \ldots, x_t, \ldots, x_T) \)

Visual feature sequence: \( (x_1, \ldots, x_t, \ldots, x_T) \)
We used the Deep hand net developed by [23]:

\[ \text{Images} \xrightarrow{\text{Crop Hands}} \text{CNN} \xrightarrow{\text{KL-HMM based approach}} z_{hshp} \]

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

\[ z_{hshp} \xrightarrow{\text{Tandem approach}} \text{KLT} (\log(z_{hshp})) \xrightarrow{\text{Kahunen Loeve Transform (KLT)}} x_{hshp} \]

Vector dimension = 61

[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:
  \[ p_t^C = \frac{\text{hand} - \text{C}}{|\text{head} - \text{neck}|/4}, \text{where } C \in \{\text{head}, \text{shoulder}, \text{hip}\}; \]

- **Velocity features** given by delta features:
  \[ v_t^C = p_t^C - p_{t-2}^C. \]

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

Tandem approach

\[ x^{hmvt} \]

Vector dimension = 36

KL-HMM based approach

\[ 1 \rightarrow 2 \rightarrow \cdots \rightarrow n \]

sign-based models

\[ x^{hmvt} \rightarrow \text{GMM} \rightarrow z^{hmvt} \]

Priors

Vector dimension = 849
Proposed approaches

Sandrine Tornay

Modelling Multichannel information in Sign Language

Tandem approach

KL-HMM based approach

Perception Space

Production Space

Visual Space

Visual feature sequence: \((x_1, \ldots, x_t, \ldots, x_T)\)

Visual feature sequence: \((y_1, y_2, y_3)\)

Local score: \(\text{log likelihood function}\)

Local score: \(\text{Kullback-Leibler divergence } S(y_s, z_t)\)

How to define the state sequence?

How to define the state sequence?

VS\(^{\text{hshp}}\) posterior estimation, e.g. hand shape

VS\(^{\text{hmvt}}\) posterior estimation, e.g. hand movement

VS\(^{\text{hshp}}\) posterior estimation, e.g. hand shape

VS\(^{\text{hmvt}}\) posterior estimation, e.g. hand movement

Stack of posterior probability distributions

Stack of categorical state distributions

KLT(\(\log(z_{hshp}^t)\))

x\(^{hshp}^s\) \(\oplus\) x\(^{hmvt}^s\)

x\(^{hshp}^t\)

x\(^{hmvt}^t\)

\(\vdots\)

\(\vdots\)

\(\vdots\)

\(\vdots\)

x\(_1^{hshp}\)

x\(_1^{hmvt}\)

x\(_T^{hshp}\)

x\(_T^{hmvt}\)

VS\(^{\text{hshp}}\) skeleton information

VS\(^{\text{hmvt}}\) skeleton information

\(\vdots\)

\(\vdots\)

\(\vdots\)

\(\vdots\)

11/16
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

<table>
<thead>
<tr>
<th>Model</th>
<th>Movement</th>
<th>Shape</th>
<th>Movement + Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM/GMM</td>
<td>51.6</td>
<td>50.3</td>
<td>44.3</td>
</tr>
<tr>
<td>TANDEM</td>
<td>66.8</td>
<td>50.3</td>
<td>63.1</td>
</tr>
<tr>
<td>KL-HMM</td>
<td>51.6</td>
<td>50.3</td>
<td>32.8</td>
</tr>
</tbody>
</table>

This table compares the sign recognition accuracy for different models: HMM/GMM, TANDEM, and KL-HMM. The accuracy figures indicate the percentage of correctly recognized signs.
## Analysis - Hand Movement Feature

### Modelling Multichannel Information in Sign Language

Sandrine Tornay

### Sign Language

Speech Inspiration

Proposed Approaches

Experimental Setup

Results and Analysis

Summary

### Proposed Approaches

- **HMM/GMM**
- **Tandem KL-HMM with GMM**
- **KL-HMM with MLP**

### Experimental Setup

<table>
<thead>
<tr>
<th>Model</th>
<th>Movement</th>
<th>Shape</th>
<th>Movement + Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM/GMM</td>
<td>66.8</td>
<td>32.8</td>
<td>47.3</td>
</tr>
<tr>
<td>TANDEM</td>
<td>51.6</td>
<td>32.8</td>
<td>44.3</td>
</tr>
<tr>
<td>KL-HMM with GMM</td>
<td>63.1</td>
<td>32.8</td>
<td>47.3</td>
</tr>
<tr>
<td>KL-HMM with MLP</td>
<td>71.9</td>
<td>32.8</td>
<td>47.3</td>
</tr>
</tbody>
</table>
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
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?