# Advances in Phonetics-based Sub-Unit Modeling for Transcription, Alignment and Sign Language Recognition.

Vassilis Pitsikalis[1], Stavros Theodorakis[1], Christian Vogler[2] and Petros Maragos[1]

[1] School of Electrical and Computer Engineering, National Technical University of Athens

[2] Institute for Language and Speech Processing/Gallaudet University

# Overview

1. Gestures, signs, and goals
2. Sign language data and visual processing
3. Data-Driven Sub-Units without Phonetic Evidence for Recognition
4. Phonetic modeling
   - What is it?
   - Annotations vs phonetics
   - Conversion of annotations to structured phonetic description
   - Training and alignment
5. Recognition experiments
6. Conclusions

# 1. Gestures versus Signs

**Gestures**

- Isolated hand, body, and facial movements
- Can be broken down into primitives (but rarely are in gesture recognition work)
- Few constraints, other than convention

**Signs**

- Hand body and facial movements, both in isolation and as part of sentences
- Can be broken down into primitives (cheremes/phonemes/phones)
- Numerous phonetic, morphological, and syntactic constraints

# SL Recognition vs Gesture Recognition

- Continuous SL recognition is invariably more complex than gestures, but:
  - Isolated sign recognition (i.e. the forms found in a dictionary) is ***essentially the same*** as gesture recognition
  - Methods that work well on isolated sign recognition should work well on gesture recognition
  - Exploit 30+ years of research into structure of signs

# Subunit Modeling

- Two fundamentally different ways to break down signs into parts:
  - Data-driven
  - Phonetics-based (i.e. linguistics)
- Similar benefits:
  - Scalability
  - Robustness
  - Reduce required training data

# Goals of this Presentation

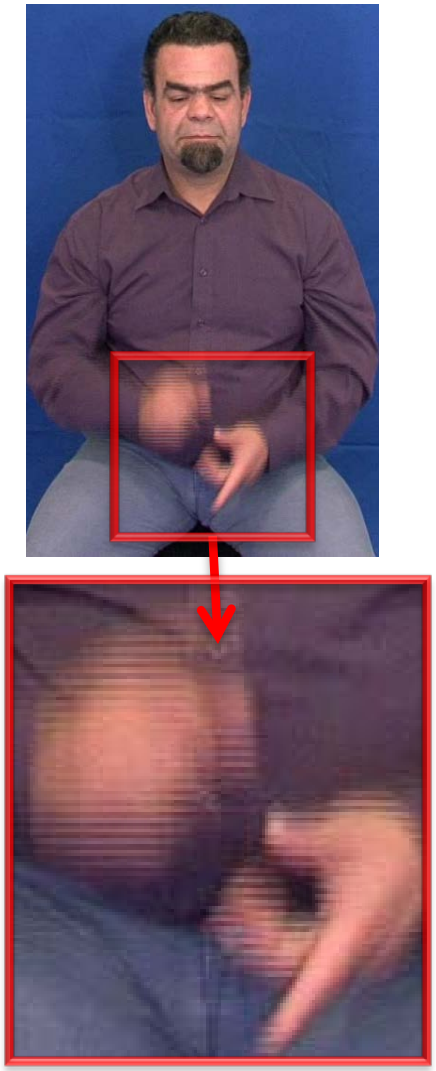
- Work with large vocabulary (1000 signs)
- Compare data-driven and phonetic breakdown of signs into subunits
- Advance state of the field in phonetic breakdown of signs

# 2. Sign Language Data

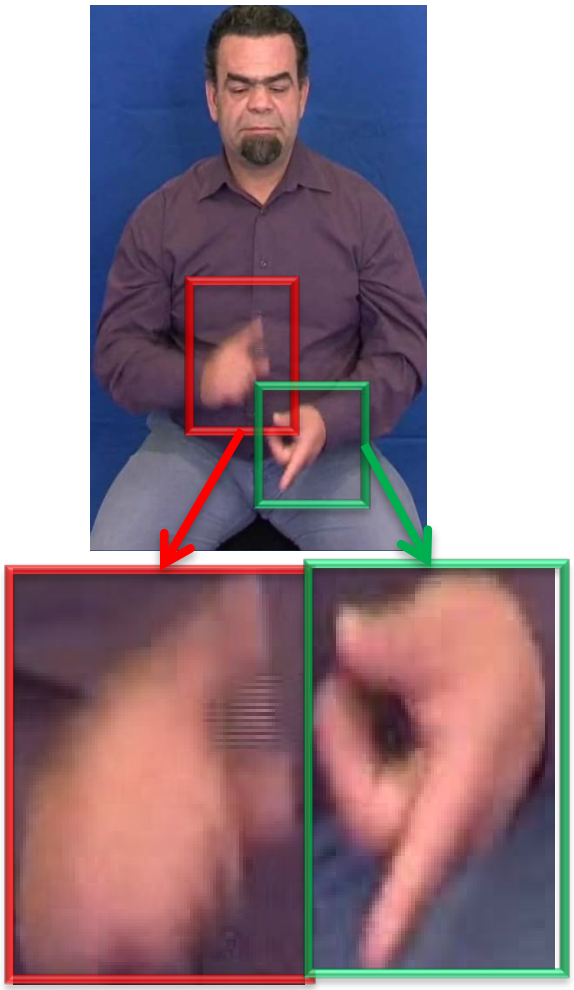
- Corpus of 1000 Greek Sign Language Lemmata
  - 5 repetitions per sign
  - Signer-dependent, 2 signers (only 1 used for this paper)
  - HD video, 25 fps interlaced
- Tracking and feature extraction
  - Pre-processing, Configuration, Statistics, Skin color training

# Interlaced data and pre-processing

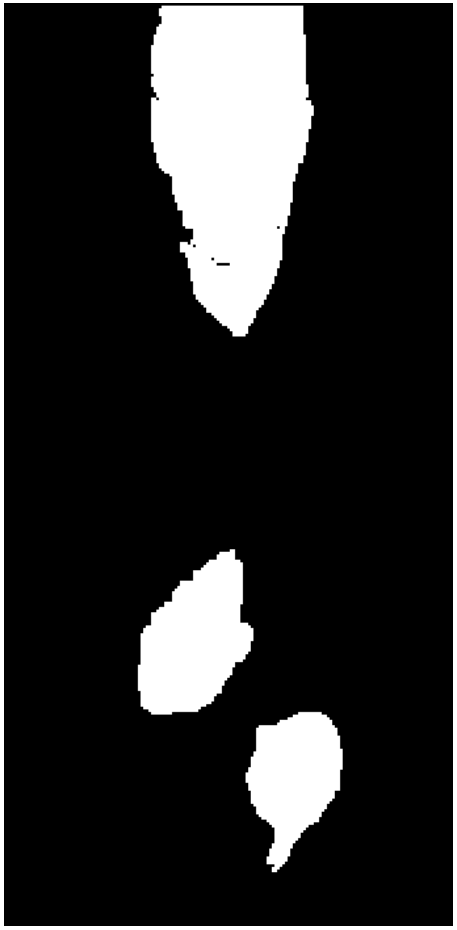
Interlaced



De-interlaced

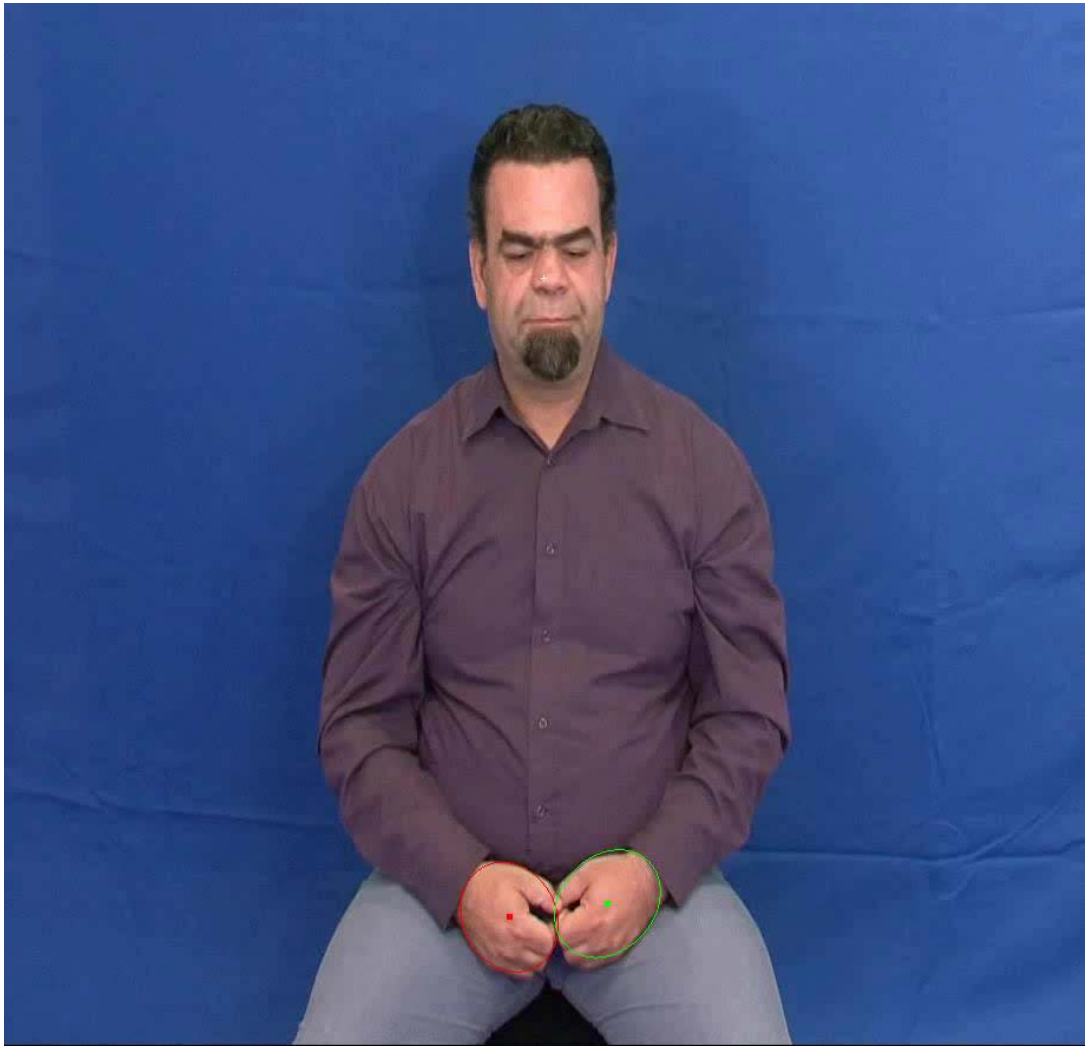


Refined
Skin color masks



2nd Version
Full Resolution, Frame rate

# Tracking Video, GSL Lemmas Corpus



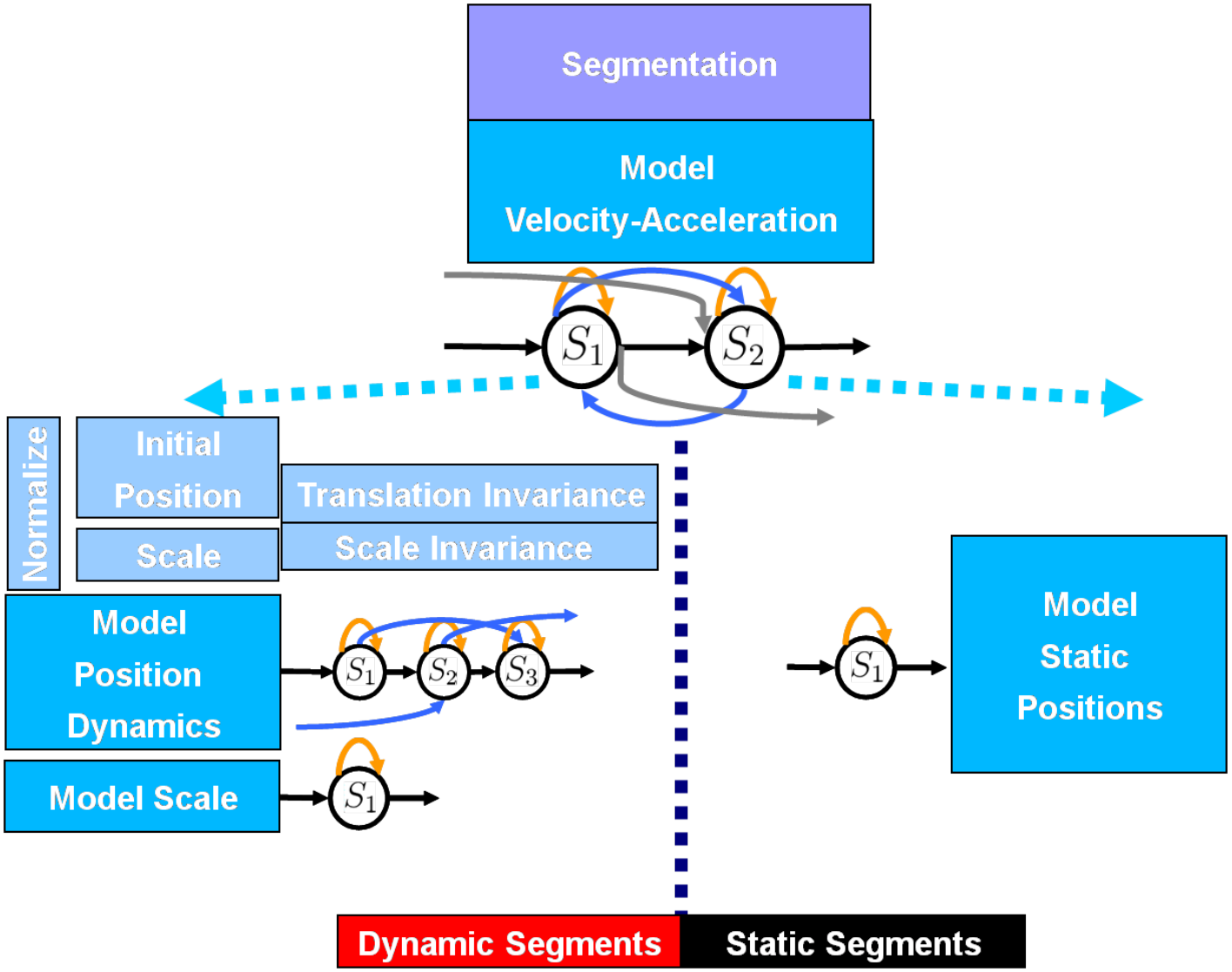
# 3. Data-Driven Subunit Modeling

- Extremely popular in SL recognition lately
- Good results
- Different approaches exist
  - ours is based on distinguishing between dynamic and static subunits

# Dynamic-Static SU Recognition

- **Dynamic (Movement)-Static (Position)**
  Segmentation: **Intuitive**, Segments + Labels
- Separate Modeling, SUs, Clustering wrt.
  Feature type (e.g. static vs. dynamic features);
  Parameters (e.g. Model Order) and Architecture (HMM, GMM); Normalize features
- Training, Data-Driven Lexicon
- Recognize SUs, Signs

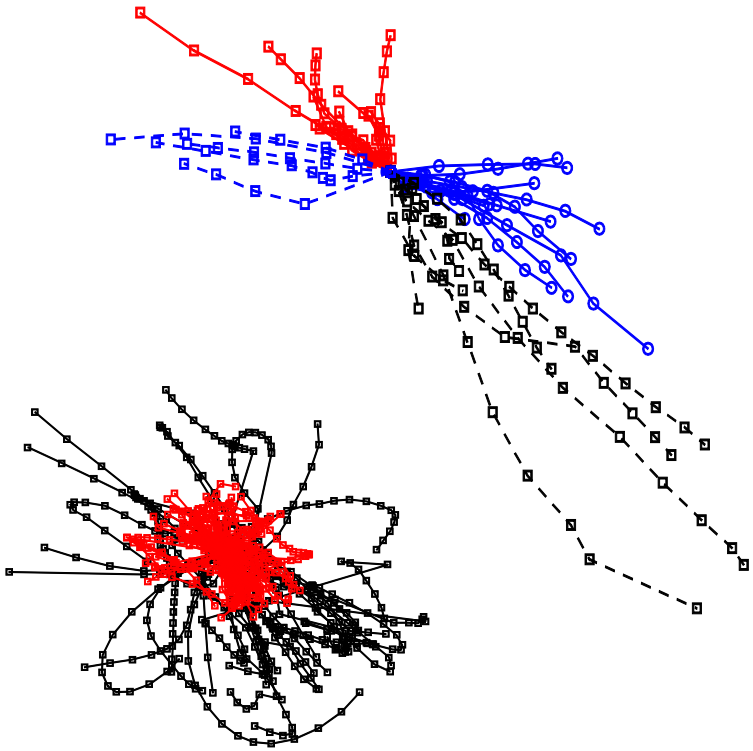
# Dynamic-Static SU Extraction



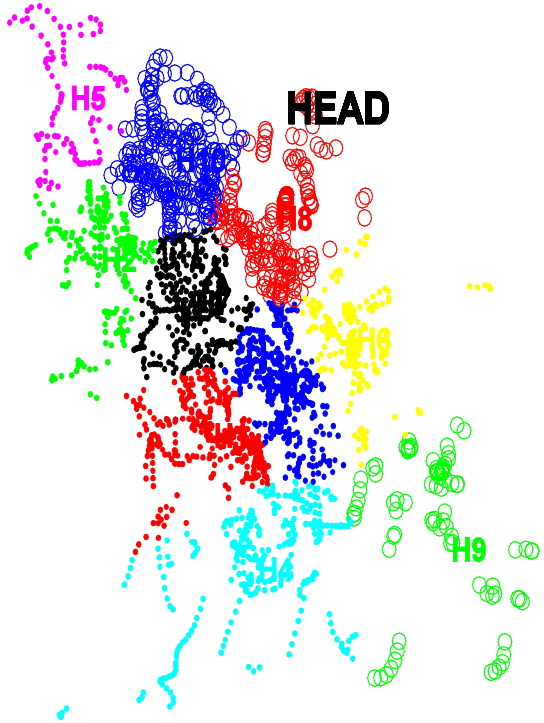

V. Pitsikalis, S. Theodorakis and P. Maragos, **Data-Driven Sub-Units and Modeling Structure for Continuous Sign Language Recognition with Multiple Cues**, LREC, 2010

# Dynamic-Static SU extraction

Dynamic clusters



Static clusters



# 4. Phonetic Modeling

- Based on modeling signs linguistically
- Little recent work

# Phonetics

- Phonetics: the study of the sounds that constitute a word
- Equivalently: the study of the elements that constitute a sign (i.e. its “pronunciation”)

# The Role of Phonetics

- Words consist of smaller parts, e.g.:
  - cat → /k/ /æ/ /t/
- So do signs, e.g.:
  - CHAIR → (HS, orientation, location, movement)
  - Parts well-known: 30+ years of research
  - Less clear: a good structured model
- Gestures can borrow from sign inventory

# The Role of Phonetics in Recognition

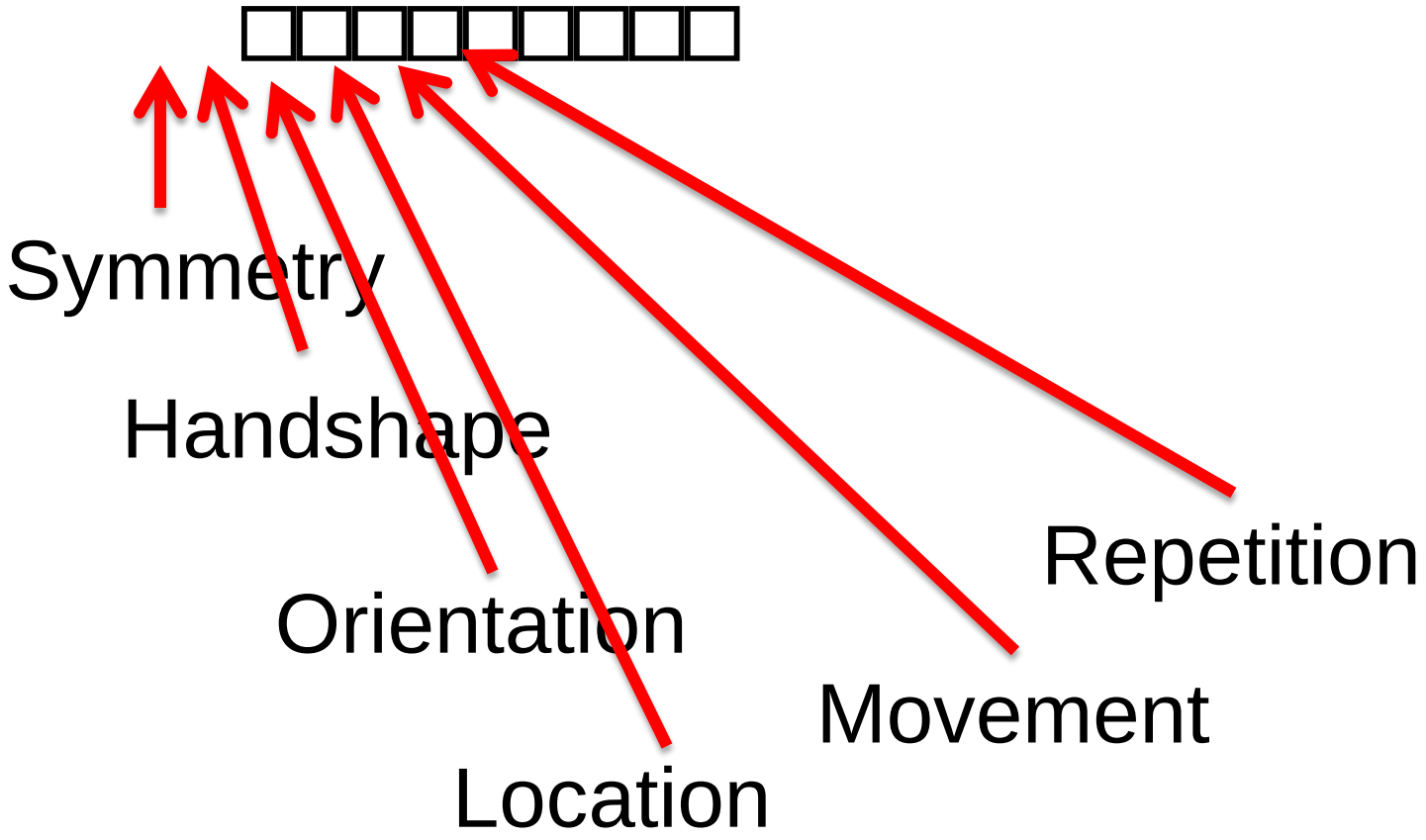
- The most successful speech recognition systems model words in terms of their constituent phones/phonemes, not in terms of data-driven subunits
  - Adding new words to dictionary
  - Linguistic knowledge & robustness
- Why don't sign language recognition systems do this?
  - Phonetics are **complex**, and phonetic annotations/lexica are **expensive** to create

# Annotation vs Phonetic Structure

- There is a difference between annotations (writing down) of a word and its phonetic structure, required for recognition
  - Annotations cannot be applied directly to recognition, although an expert can infer the full pronunciation and structure from an annotation
- Annotations for signed languages are much less time consuming than writing the full phonetic structure

# Annotation of a Sign

- Basic HamNoSys annotation of CHAIR:

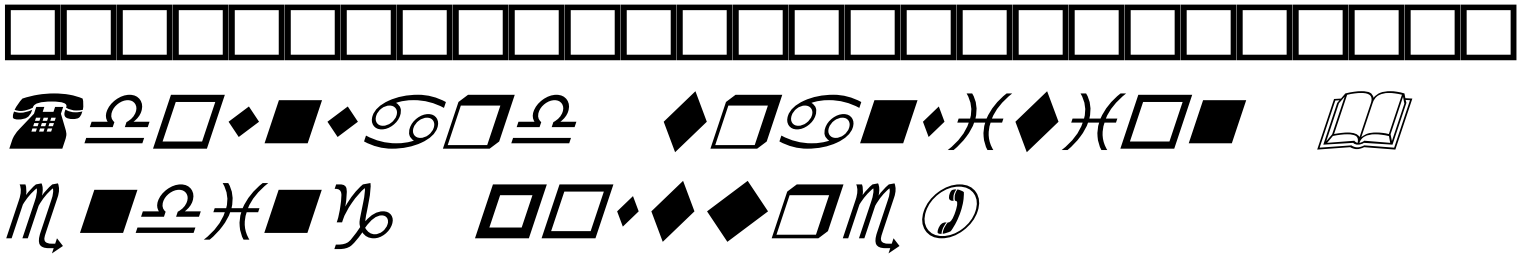
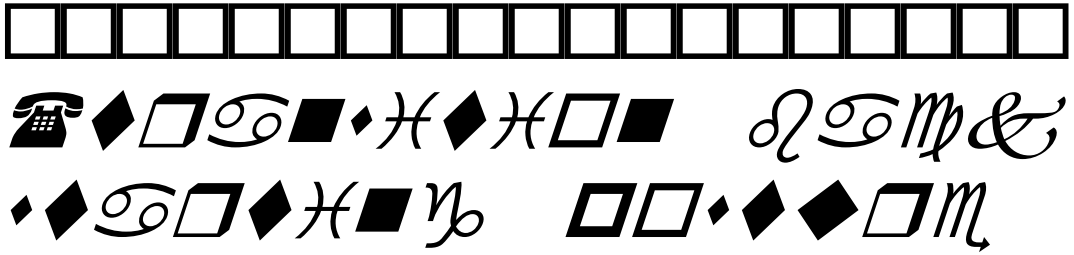


# Phonetic Structure of Signs

- Postures, Detentions, Transitions, Steady Shifts (PDTS)
  - > improved over 1989 Movement-Hold model
  - Alternating postures and transitions

| Posture | Trans | Posture | Trans | Posture | Trans | Det |
|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |
| shoulder | Straight down | chest | Back up | shoulder | Straight down | chest |

# How Expensive is Phonetic Modeling?

- Basic HamNoSys annotation of CHAIR:
  - 
- Over 70 characters compared to just 8!
  - 
  - 

# Automatic Extraction of Phonetic Structure

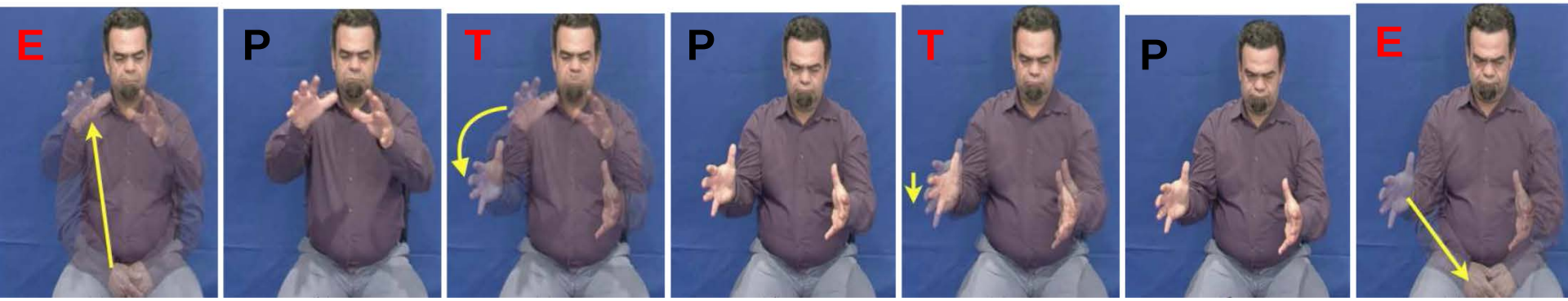
- **First contribution:** Automatically extract phonetic structure of sign from HamNoSys
- Combines convenience of annotations with required detail for recognition
- Recovers segmentation, postures, transitions, and relative timing of hands
- Based on symbolic analysis of movements, symmetries, etc.

# Training and Alignment of Phonetic SUs

- **Second contribution:** Train classifiers based on phonetic structure, and align with data to recover frame boundaries
  - Frame boundaries not needed for recognition, but can be used for further data-driven analysis
- Classifiers based on HMMs – why?
  - Proven track record for this type of task
  - No explicit segmentation required, just concatenate SUs, use Baum-Welch training
  - Trivial to scale up lexicon size to 1000s of signs

# Phonetic Models to HMM

# Phonetic Subunit Training, Alignment



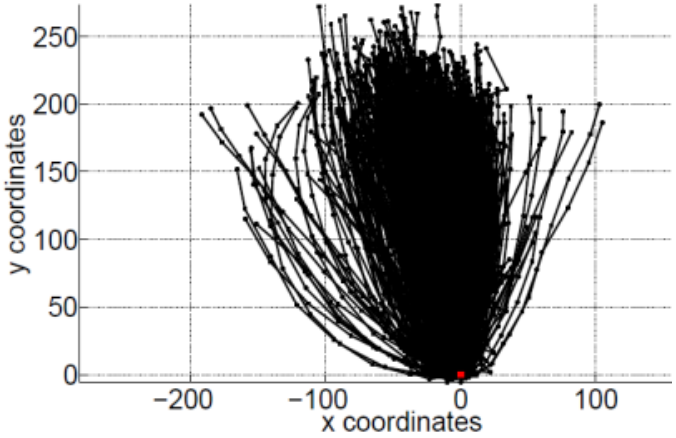

**Transition/Epenthesis Segments**
Superimposed Initial-End Frames + Arrow

**Posture/Detention Segments**
Single Frame

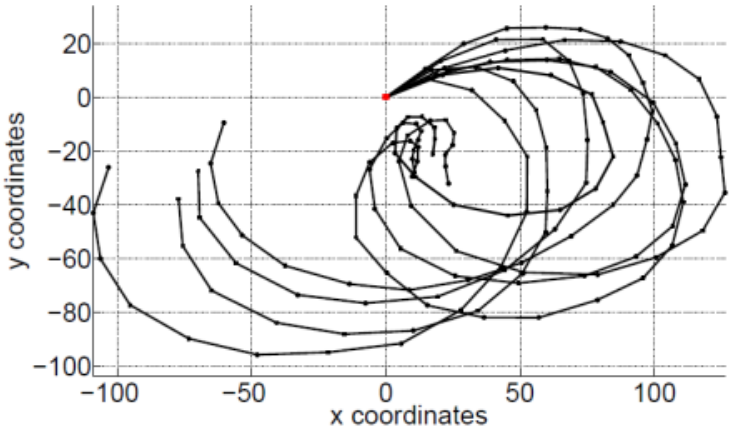
| Frames | Type | PDTS label |
|---|---|---|
| 1:12 | E | rest-position — location-head |
| 13:13 | P | location-head |
| 14:25 | T | directedmotion, curve-r, direction-o, second-direction-do, tense-true |
| 26:27 | P | location-torso, side=right_beside |
| 28:50 | T | directedmotion, direction-dr, small |
| 51:51 | P | location-torso, side=right_beside_down |
| 52:66 | E | location-torso, side=right_beside_down — rest-position |

# Phonetic Sub-Units

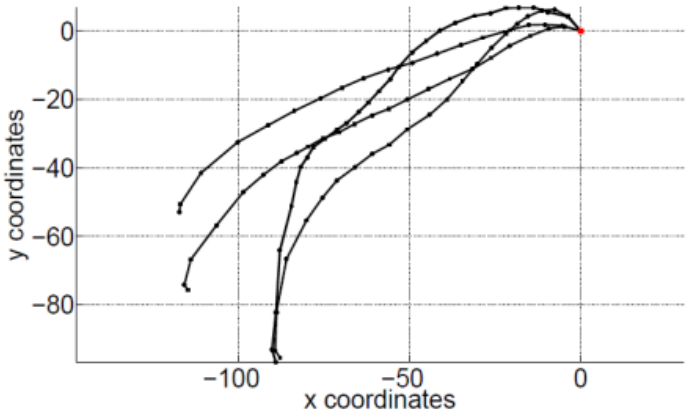
**Transition/Epenthesis**



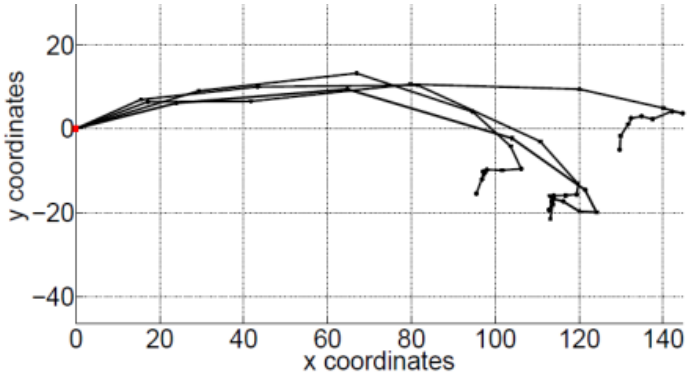

(a) E-to-head




(b) T-circular



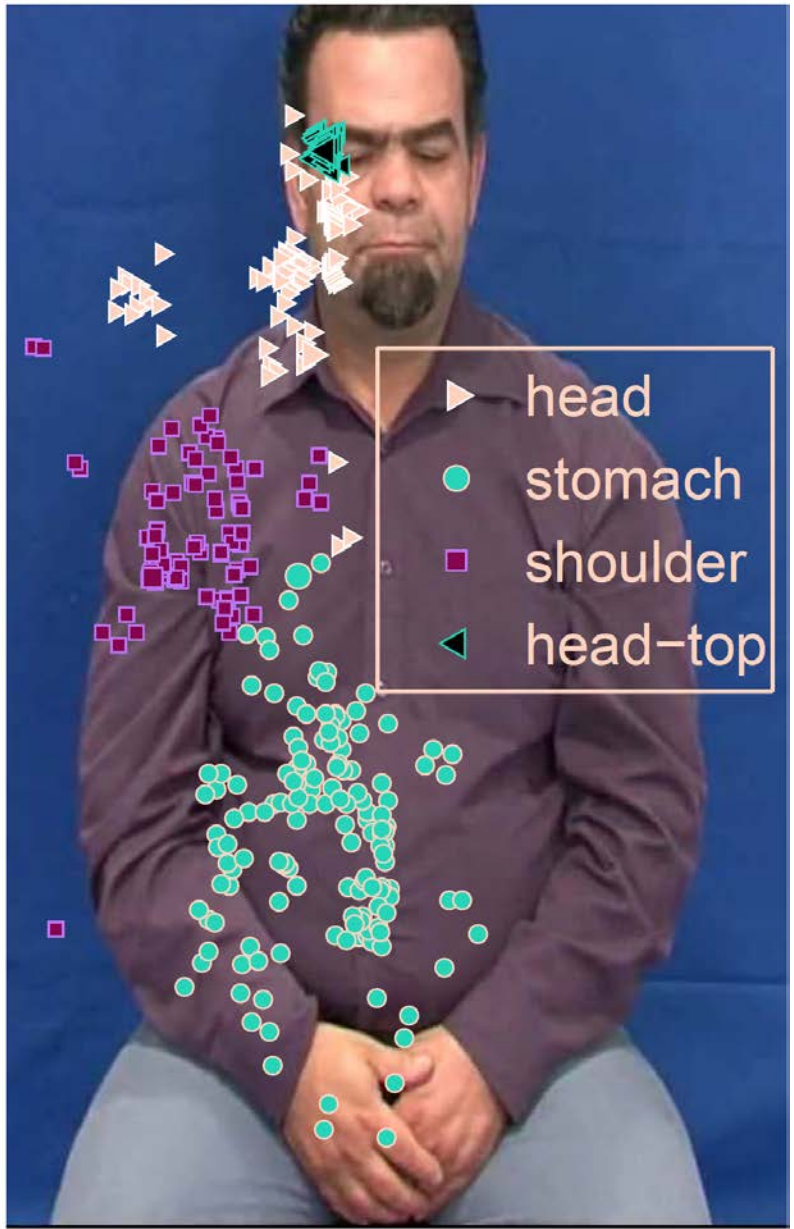

(c) T-down-right




(d) T-in-left

| PSU | Sign | Type | PDTS Label |
|---|---|---|---|
| E-to-head | * | E | rest-position — location-head |
| T-circular | EUROPE | T | circularmotion, axis=i |
| T-down-right | PILE | T | directedmotion, direction=dr, small |
| T-in-left | IMMEDIATELY | T | directedmotion, direction=il, fast=true, halt=true |

# Phonetic Sub-units

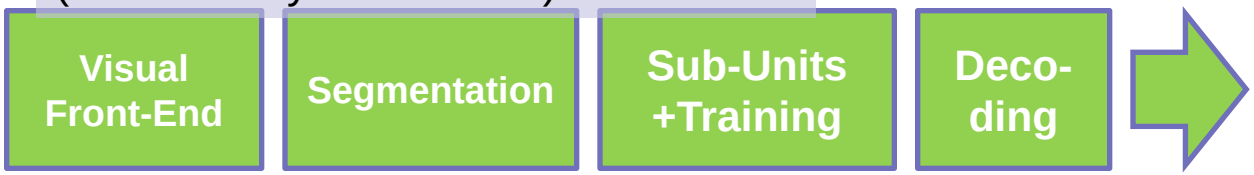



## Postures

| PSU | Type | PDTS Label |
|---|---|---|
| P-forehead | P | location=forehead |
| P-stomach | P | location=stomach |
| P-shoulder | P | location=shouldertop, side=right_beside |
| P-head-top | P | location=head-top |

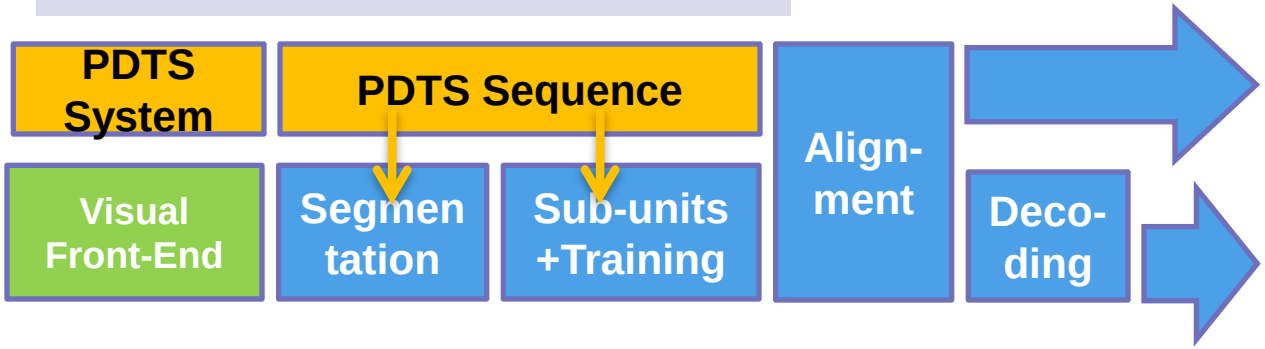
# 4. Recognition based on both Data-Driven SUs + Phonetic Transcriptions

**1. Data-Driven Subunits**
(based on Dynamic-Static)




Recognize Signs, Data-driven SUs

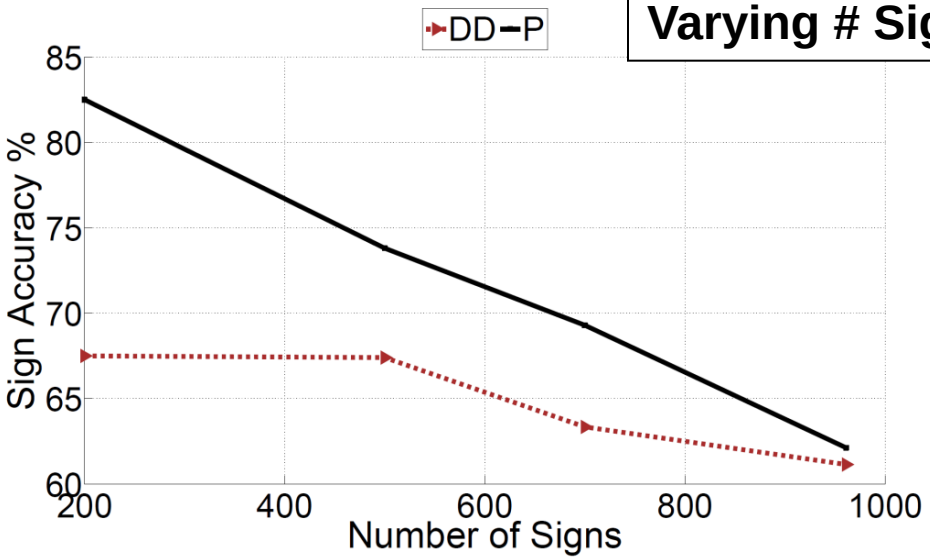
**2. Data+Phonetic Subunits**

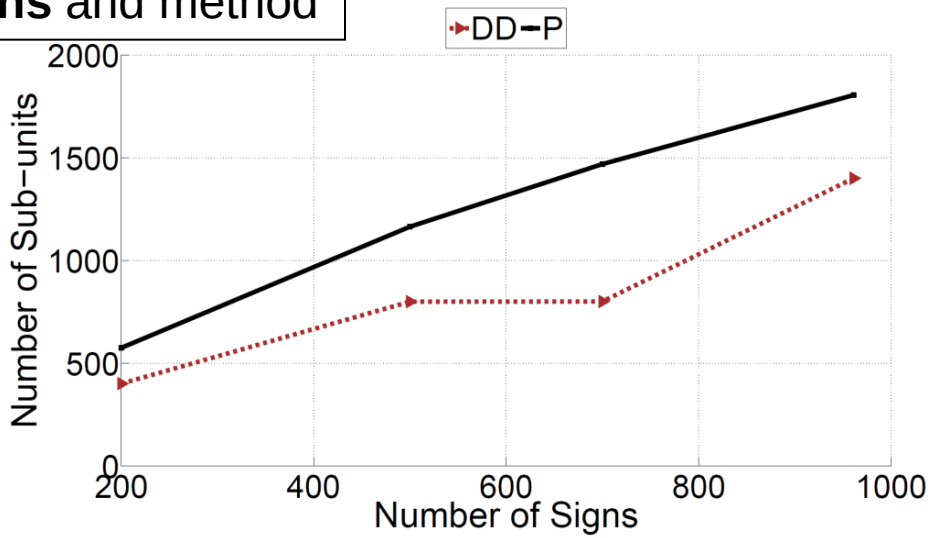


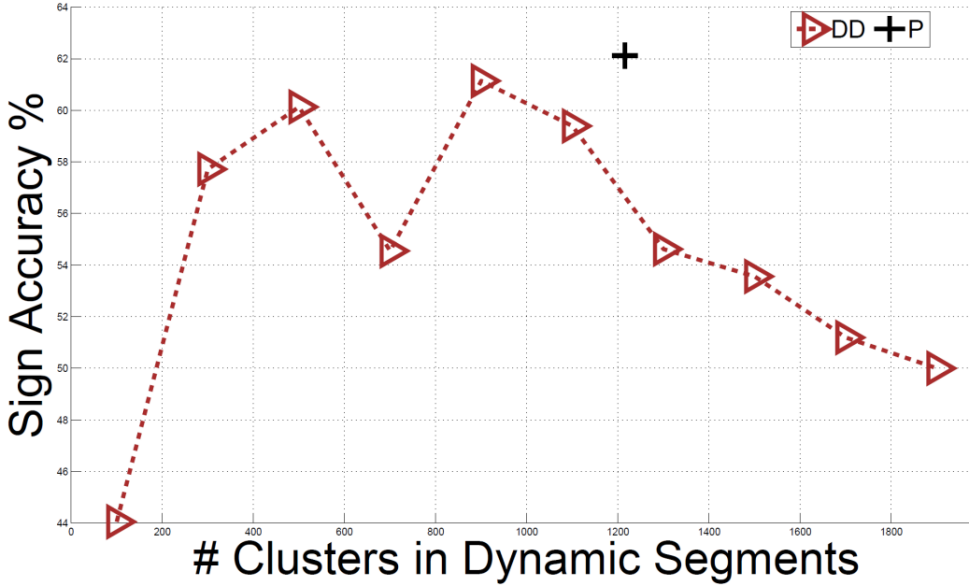

Recognize Signs, + PDTS Sequence +Alignment

# Data-Driven vs. Phonetic Subunits Recognition

**Varying # Signs** and method




**Varying Dynamic SUs** and method,
Static SUs=300, **#Signs = 961**

# 5. Conclusions and The Big Picture

- Rich SL corpora annotations are rare (in contrast to speech)
- Human annotations of sign language (HamNoSys) are expensive, subjective, contain errors, inconsistencies
- HamNoSys contain no time structure
- Data-Driven approaches Efficient but construct abstract SubUnits

- Convert HamNoSys to PDTS; Gain Time Structure and Sequentiality
- Construct meaningful phonetics-based SUs

- Further exploit the **PDTS+Phonetic-SUs**
  - Correct Human Annotations Automatically
  - Valuable for SU based SL Recognition, Continuous SLR, Adaptation, Integration of Multiple Streams

# Thank you !

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n◦ 231135. Theodore Goulas contributed the HamNoSys annotations for GSL.