Using Viseme Recognition to Improve a Sign Language Translation System

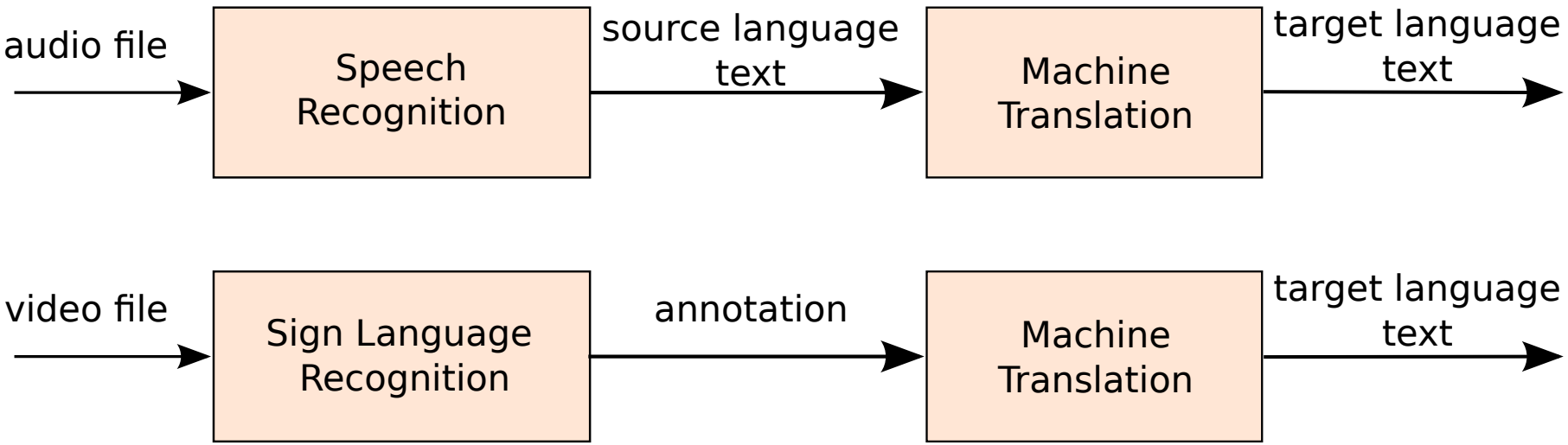
Christoph Schmidt, Oscar Koller, Hermann Ney [1]
Thomas Hoyoux, Justus Piater [2]

6.11.2013

[1] **Human Language Technology and Pattern Recognition Group**
Computer Science Department
RWTH Aachen University, Germany
{surname}@i6.informatik.rwth-aachen.de

[2] **Intelligent and Interactive Systems**
University of Innsbruck, Austria
{firstname}.{surname}@uibk.ac.at

1 Introduction



- **Main difference: Multimodality**

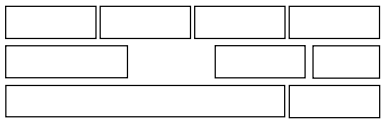
Multimodality

Express meaning simultaneously:

- **hand movements**
- **body posture**
- **facial expressions**
- **mouthing (silent pronunciation of words)**

string of symbols **string of tuples** **multiple streams**

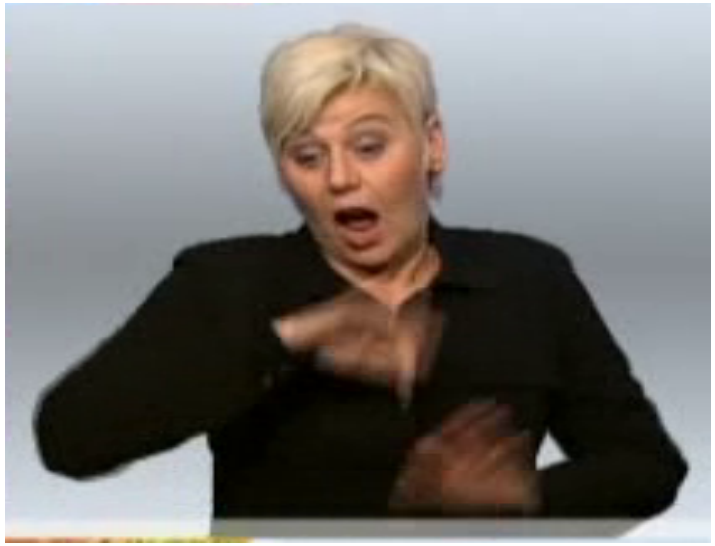
$$f_1^J \rightarrow e_1^I \qquad \begin{pmatrix} \text{hand} \\ \text{body} \\ \text{mouthing} \\ \dots \end{pmatrix}_1^J \rightarrow e_1^I \qquad \rightarrow e_1^J$$



Mouthing Variants

MOUNTAIN

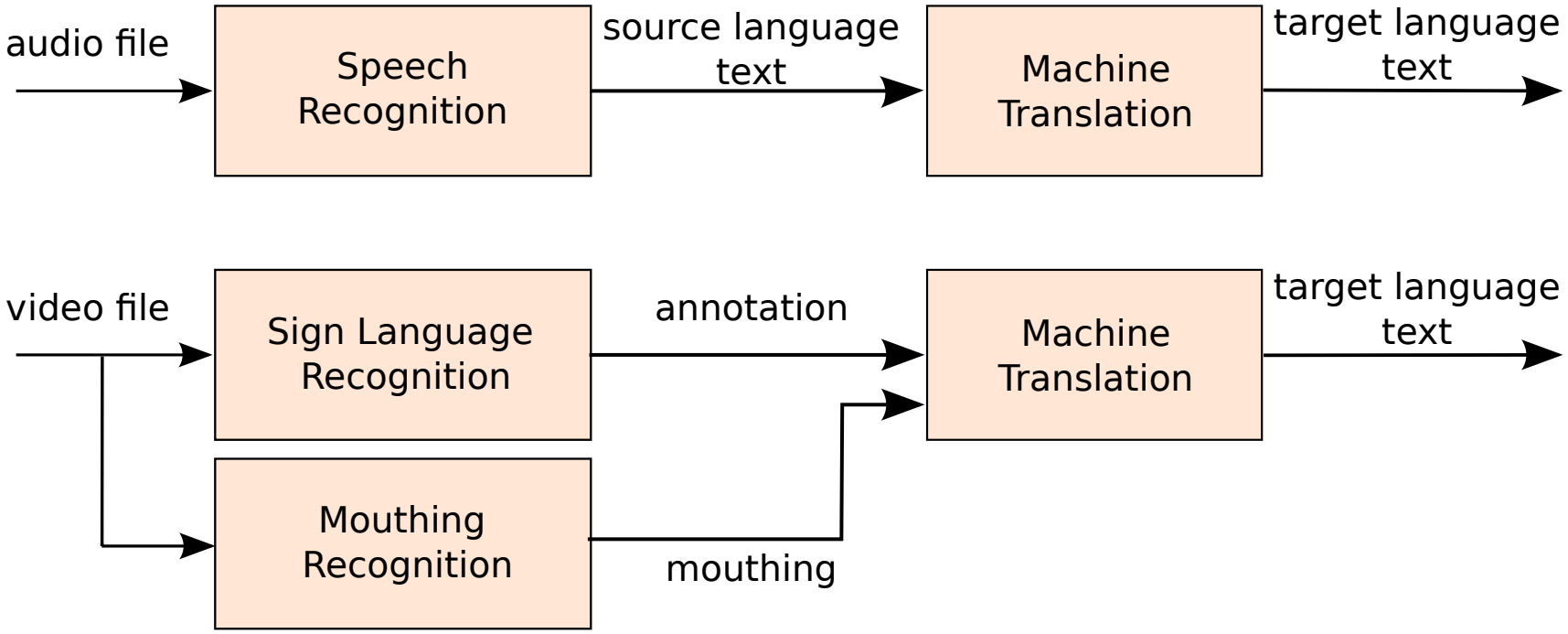
mountain (mouthing: “Berg”)



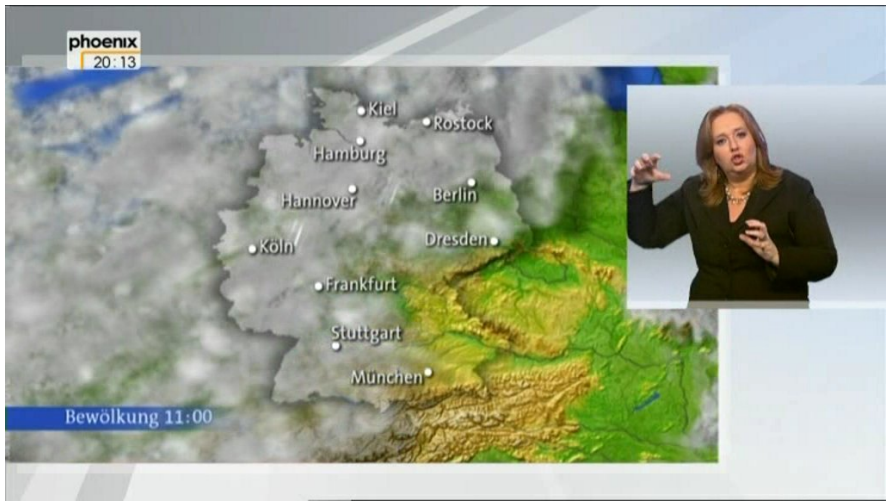
Alps (mouthing: “Alpen”)

- **Some signs only differ in mouthing / mouth gestures**
- **Prevalent in countries with oralistic education**
- **Annotation of corpus focused on hand-based features**
- **Manual refinement of annotation time consuming**

Sign language translation



2 RWTH-PHOENIX-Weather Corpus

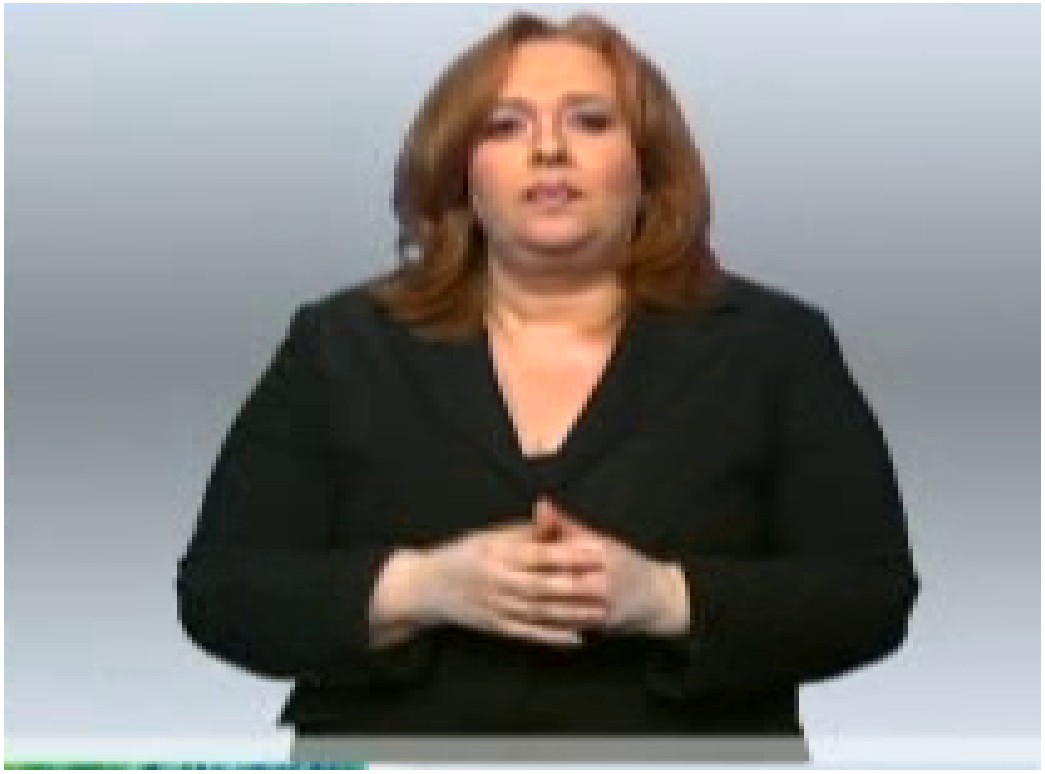


	DGS	German
signers	7	
editions	190	
duration[h]	3.25	
frames	293,077	
sentences	2,552	
glosses, words	17,771	30,860
vocabulary size	911	1,452
singletons	120	337

- video-based, large vocabulary corpus
- weather forecasts from public TV news, interpreted into DGS
- annotation: glosses, German text time boundaries on gloss level
- focus on hand-based features

- Teaser:
  new version with 645 editions submitted to LREC 2014 !

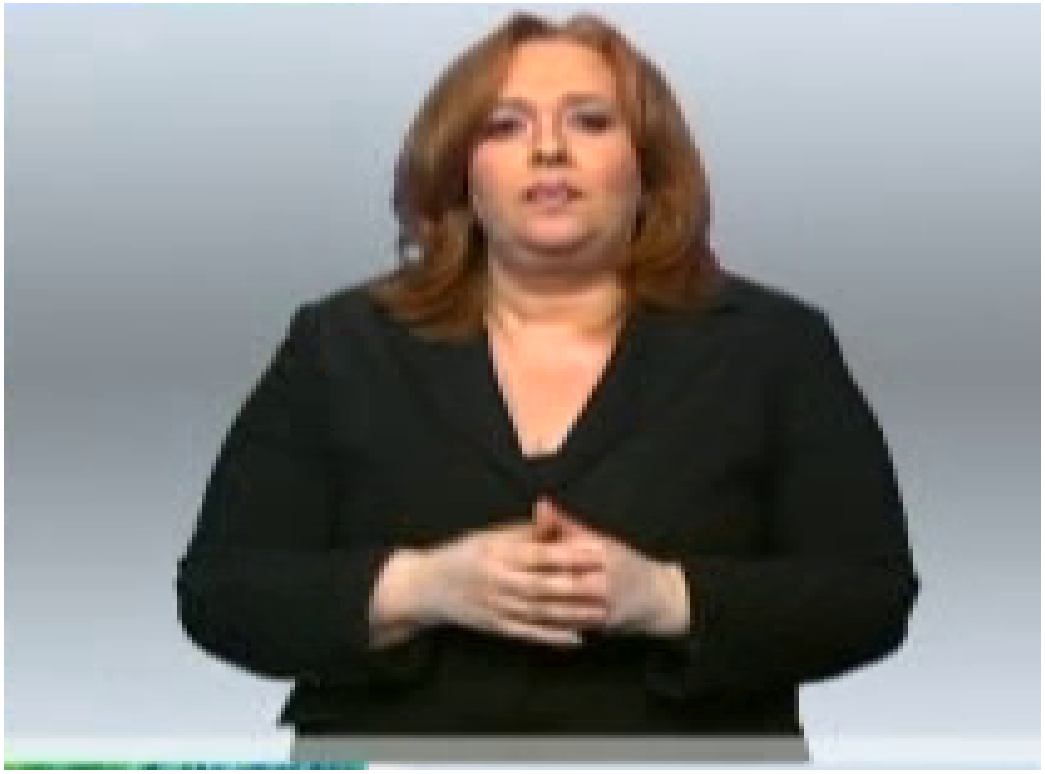
Guessing the mouthing from bilingual data



Gloss: HEUTE ABEND FLUSS DREI MINUS SECHS BERG
Mouthing: HEUTE NACHT ODER DREI MINUS SECHS ALPEN

German: Heute Nacht drei Grad an der Oder , minus sechs Grad an den Alpen .

Guessing the mouthing from bilingual data

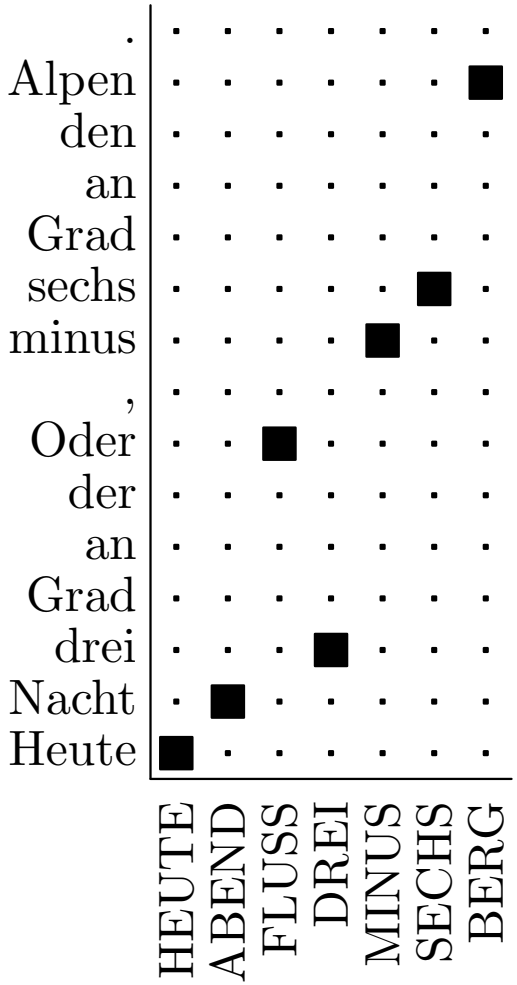


Gloss:	**TODAY**	**EVENING**	**RIVER**	**THREE**	**MINUS**	**SIX**	**MOUNTAIN**
Mouthing:	**TODAY**	**NIGHT**	**ODER**	**THREE**	**MINUS**	**SIX**	**ALPS**

English: **Tonight three degrees at the Oder, minus six degrees at the Alps.**

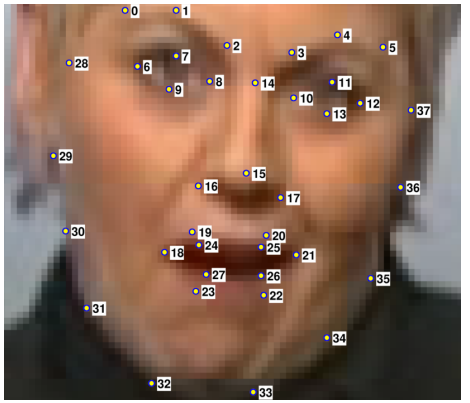
Guessing the mouthing from bilingual data

- annotation: focus on hand features
  $\rightarrow$ no mouthing information
- hypothesize mouthing from spoken translation
- align glosses and translation using GIZA++
- Issues:
  - not all signs have a mouthing
  - inflection / compounding differs
- Approach:
  - Create variants for glosses with different mouthings
  - Filtering: do not create variants for glosses with high mouthing recognition error rate

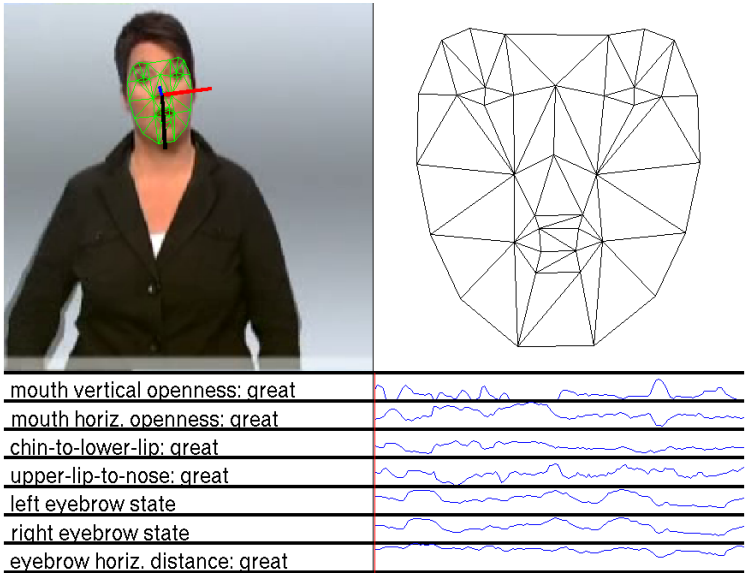


3 Active Appearance Models

- Track low-level features:
  - characteristic points on the face



- extract high-level features:
  - mouth vertical openness
  - mouth horizontal openness
  - lower lip to chin distance
  - upper lip to nose distance
  - left eyebrow state
  - right eyebrow state
  - gap between eyebrows



4 Mouthing Recognition

Use:

- **automatic speech recognition software**
- **active appearance model high level features**
- **viseme lexicon (mapped from phoneme lexicon)**
- **for each gloss, allow recognition of all variants aligned during training + garbage model**

Phoneme-Viseme Mapping

Phoneme	Viseme	Examples
p, b	P	Pause, Bitte
t, d, k, g	T	Tonne, Dach, König, Gier
n, @n, l, @l	N	Nadel, raten, Liebe, Igel
m	M	Mutter
f, v	F	Finder, Vase
s, z	S	Fass, Sein
S, Z, tS, dZ	Z	Schein, Garage, Tscheche
h, r, x, N	R	Hase, Reden, Dach, Wange
j, C	C	Junge, Wicht
i:, I, e:, E:, E	E	Bier, Tisch, Weg, Räte, Menge
a:, a	A	Wagen, Watte
o:, O	O	Wolle, Wogen
u:, U	U	Buch, Runde
@, 6	Q	Bitte, Weiher
y:, Y, 2:, 9	Y	Tür, Mütter, Goethe, Götter

- Phoneme-viseme mapping (taken from [Aschenberner & Weiss])

Mouthing Recognition: Results

- evaluate recognition on set of 640 manual annotated glosses
- CER: character error rate
- Recall: fraction of glosses which are split into variants

	CER [%]	Recall [%]
initial segmentation	40.5	82.5
10× EM-realignment	35.7	47.5
after RANSAC processing	32.2	45.5
filter mouthings with high error	12.7	6.5

- false assumption: each gloss has a mouthing
- filtering removes mouthings with high error rate
- optimized on development set

5 Machine Translation

- Jane phrase-based system
- cross-validation for optimization
- train system on (gloss, mouthing) variants
- learn correspondence between mouthing and translation
  - word level: lexical smoothing
  - phrase level: indicator and count features

word level: lexical smoothing

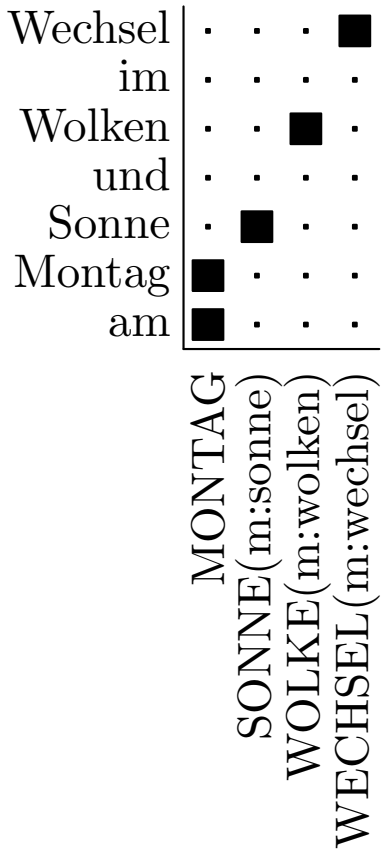
- adapt lexical smoothing [Huck & Mansour+ 11]
- increase lexical probabilities $p(\text{spoken}|\text{gloss}, \text{mouthing})$ iff spoken=mouthing by factor α
- optimize factor α on development set

$$N_s(e, f) = \sum_{e_{i_s}:e_{i_s}=e} \sum_{f_{j_s}:f_{j_s}=f, j \in \{a_i\}_s} \frac{\beta}{|\{a_i\}_s|}$$

$$N(e, f) = \sum_s N_s(e, f)$$

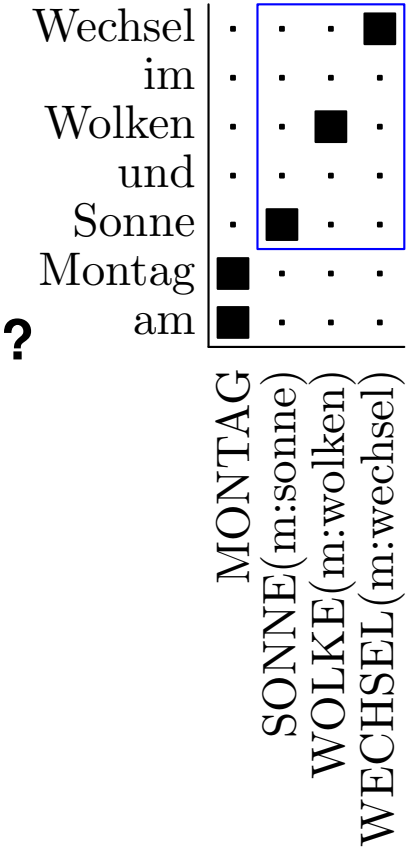
$$p(e|f) = \frac{N(e, f)}{\sum_{e'} N(e', f)}$$

$$\beta = \begin{cases} 1 + \alpha & \text{iff spoken=mouthing} \\ 1 & \text{else} \end{cases}$$



Phrase level: indicator, count features

- add indicator, count feature to the phrase table
- indicator feature:
  do mouthings in source phrase exist in target phrase?
- count feature:
  number of mouthings in source phrase
  which appear in target phrase
- optimize two additional weights on development set



Machine Translation: Oracle Results

System	Dev		Test	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]
Baseline	35.5	58.8	23.8	66.5
Oracle + Translation	36.8	53.4	29.8	60.1
+ word level	39.8	45.3	31.7	52.7
+ phrase level	40.8	43.6	32.6	49.9
+ word + phrase level	41.1	44.4	33.6	48.7

- Oracle machine translation results, assuming all mouthings were recognized correctly

Machine Translation: Results

System	Dev		Test	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]
Baseline System	35.5	58.8	23.8	66.5
Recognition + Translation	35.2	53.2	23.1	65.4
+ word level	36.1	54.3	24.1	65.5
+ phrase level	36.8	53.5	24.4*	64.4**
+ word + phrase level	37.5	52.6	24.8**	64.4**

- Machine translation results of systems using viseme recognition input
- Statistical significance: *: p<0.05, **:p<0.01

Machine Translation: Example (good)

Gloss:	**SUED**	**BISSCHEN**	**REGEN**	**KOENNEN**
Mouthing:	**SUEDEN**		**SCHAUER**	

Baseline: Im Sueden hier und da faellt etwas Regen .
+Mouthing rec.: Im Sueden sind ein paar Schauer moeglich .
Reference: Ganz im Sueden sind einzelne Schauer moeglich .

Gloss:	**SOUTH**	**BIT**	**RAIN**	**CAN**
Mouthing:	**SOUTH**		**SHOWER**	

Baseline: In the south here and there falls a bit rain .
+Mouthing rec.: In the south a few showers are possible .
Reference: In the very south a few showers are possible .

- **Mouthing recognition can distinguish between “rain” and “shower”**

Machine Translation: Example (bad)

Gloss:	**SONST**	**VIEL**	**SONNE**
Mouthing:	**SONST**	**VIEL**	**SONNE**

Baseline: Sonst ist es meist sonnig .
+Mouthing rec.: Sonst ist es meist viel Sonne .
Reference: Sonst ist es meist sonnig .

Gloss:	**OTHERWISE**	**MUCH**	**SUN**
Mouthing:	**OTHERWISE**	**MUCH**	**SUN**

Baseline: Otherwise it is mostly sunny .
+Mouthing rec.: Otherwise it is mostly much sun .
Reference: Otherwise it is mostly sunny .

- **System does not consider word inflections**
  $\Rightarrow$ sometimes awkward grammar

6 Conclusions / Outlook

Conclusions:

- **Integration of viseme recognition into sign language translation pipeline**
- **Mouthing as an additional knowledge source in translation**
- **Improvement over system without mouthing information**

Outlook:

- **include mouth patch as feature into viseme recognition**
- **capture different inflections (lemmatizing)**
- **incorporate other modalities besides hands and mouthing**
- **address time offset between signing and mouthing (dynamic time alignment)**

Thank you for your attention

Christoph Schmidt

schmidt@i6.informatik.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/

References

[Aschenberner & Weiss] B. Aschenberner, C. Weiss: Phoneme-Viseme Mapping for German Video-Realistic Audio-Visual-Speech-Synthesis. 11

[Huck & Mansour+ 11] M. Huck, S. Mansour, S. Wiesler, H. Ney: Lexicon Models for Hierarchical Phrase-Based Machine Translation. In *International Workshop on Spoken Language Translation*, pp. 191–198, San Francisco, California, USA, Dec. 2011. 14