



# Speaker adaptation, segmentation and tracking

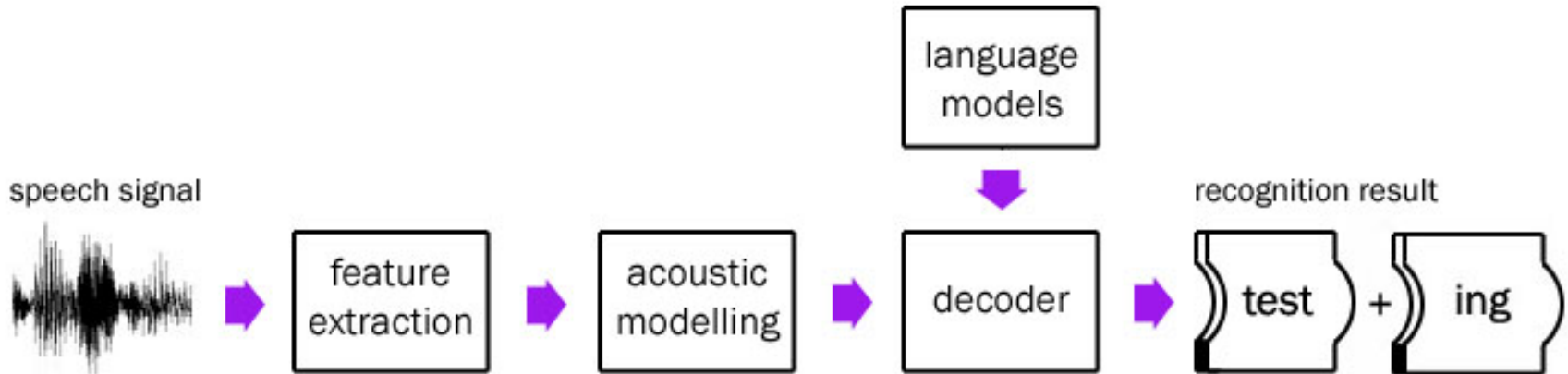
Ulpu Remes

TKK

## contents

- Speaker adaptation
  - Vocal tract length normalisation, VTLN
  - Constrained maximum likelihood linear regression, CMLLR
- Speaker segregation
  - Speaker segmentation / speaker change detection
  - Speaker tracking
- Speech recognition results on read sentences and Finnish and English broadcast news

## speech recognition



- Our large vocabulary continuous speech recognition system models the language as sequences of statistical morphemes trained on book and newspaper data, and the pronunciation of each language unit is modelled as a sequence of context dependent phonemes

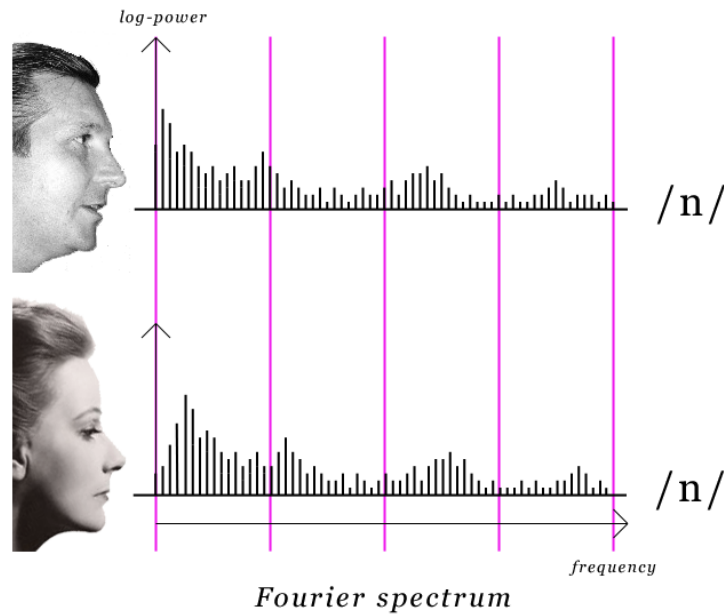
## speaker adaptation

- Speech recognition systems work well when trained for a specific speaker, but in most applications there are multiple speakers and they are unknown to the system
- Speaker adaptation modifies the acoustic models to better match a new speaker



## VTLN

- Differences in vocal tract length cause the spectrum to be stretched or compressed
- Vocal tract length normalisation (VTLN) methods scale the frequency axis reversely



## CMLLR

- Constrained maximum likelihood linear regression
- Linear transformation on features:  $\mathbf{o}' = \mathbf{A} \cdot \mathbf{o} + \mathbf{b}$ , where  $\mathbf{o}$  are the features
- We wish to choose the transformation that maximises  $p(\mathbf{o}' \mid \text{spoken text})$
- As spoken text is not known, we use a decoder given hypothesis instead

## results / SPEECON

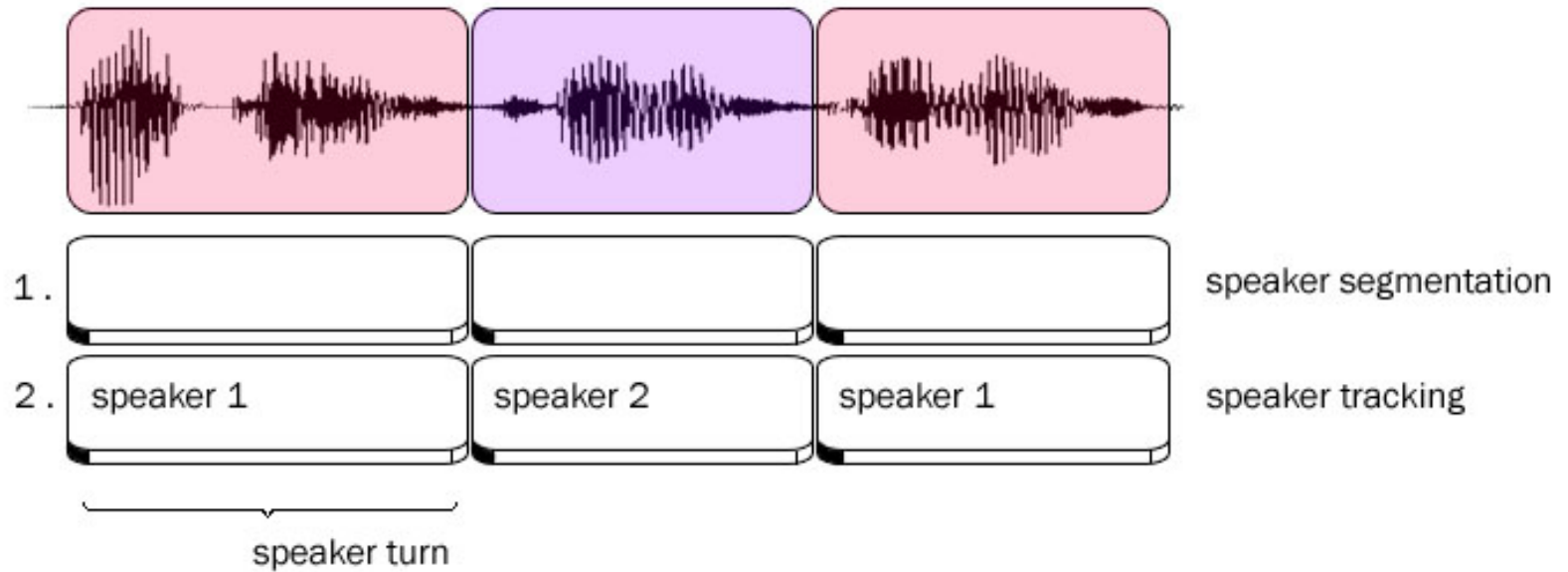
	baseline	VTLN	trained	CMLLR	trained
WER	16.4	15.9	15.5	14.4	12.6
LER	4.6	4.4	4.2	3.7	3.0

- Test data: total of 910 read sentences from 31 adult speakers, no background noise
- RER for VTLN 4.3 % with the baseline models and 8.6 % with VTLN trained models
- RER for CMLLR with baseline models 19.5 %, with speaker adapted models 34.8 %

## speaker segregation

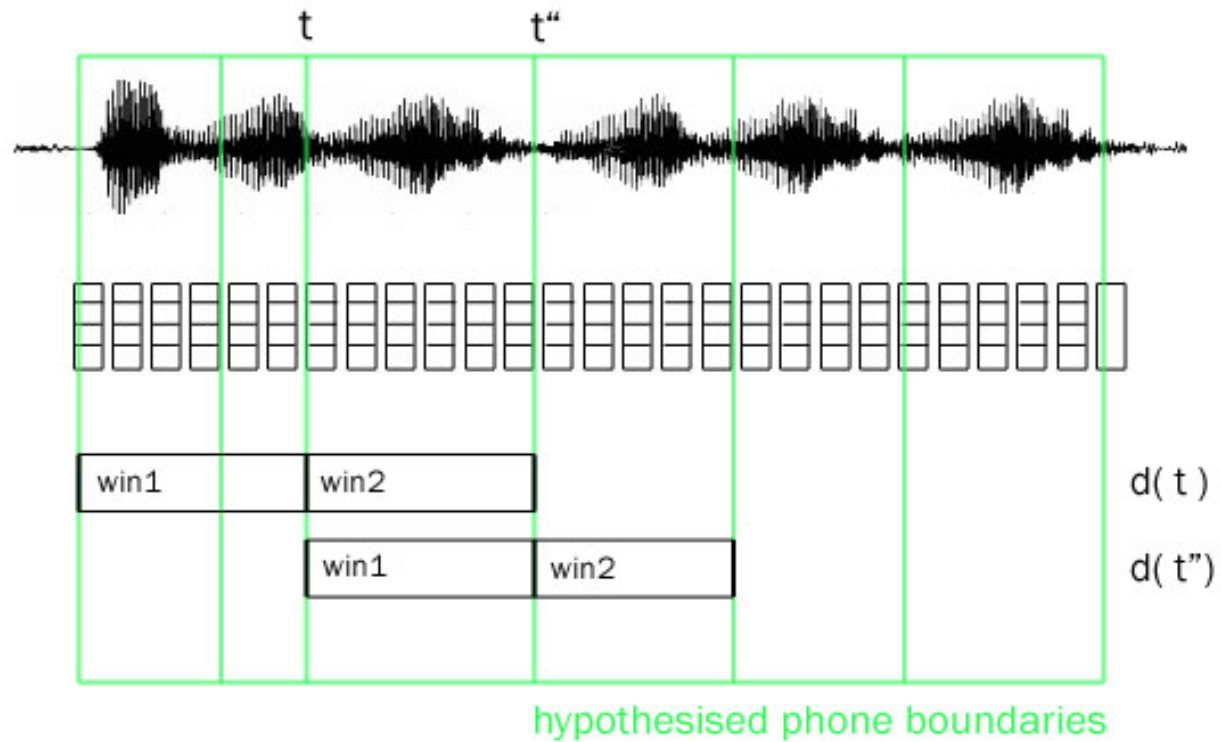
- mail-in ballots from military personnel stationed overseas were not included in the florida vote we will have details from the pentagon *the united nations says it's discussing plans with indonesia to send international-aid workers to west timor to help thousands of east timorese refugees who are stranded there* the US government wants to know whether somalis in the united states are sending money to warring clans to buy weapons hello everyone i'm paul westfeling *and i'm les carpenter* we will also bring you the latest in sports business news and ...

## speaker segregation

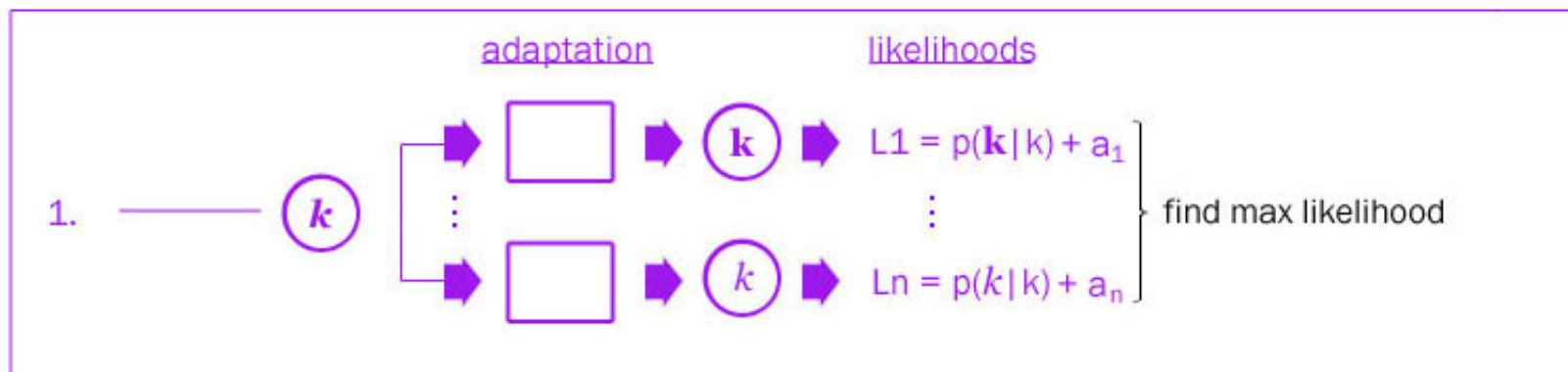




## speaker segmentation



## speaker tracking



## results / YLE

- YLE TV news audio
- Test data set : 68 min speech data, 49 speakers
- Average speaker turn length around 30 seconds
- Only planned speech from the newscasters and reporters
- Background music and other noise present in some parts

## results / YLE

	baseline	CMLLR true	CMLLR auto
WER	23.0	19.8	19.5
LER	7.9	6.0	5.9

- Speaker change boundaries: false acceptance 22.8 % and false rejection 24.1 %
- Speaker labels: average cluster purity 95.5 % and average speaker purity 84.0 %

## results / VOA

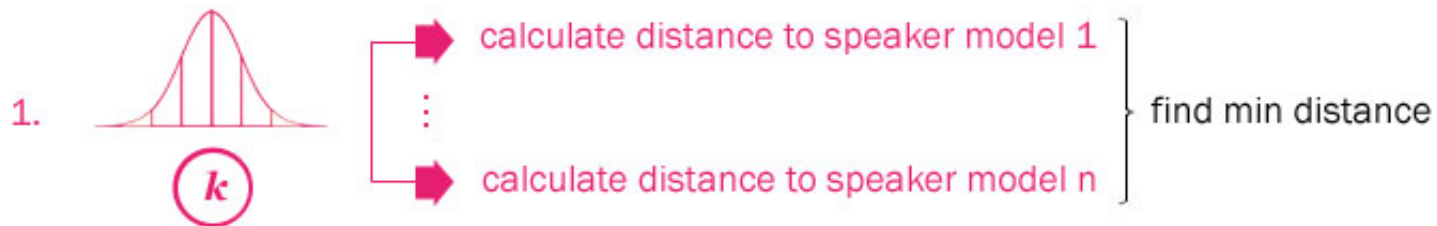
- VOICE OF AMERICA news broadcasts
- Test data sets
  - VOA 1: 19 min, 10 speakers; clean speech from reporters
  - VOA 2: 38 min, 28 speakers; also some background noise, interviews
  - VOA 3: 41 min, 27 speakers; also, speaker changes may not be clear

## results / VOA

		baseline	CMLLR true	CMLLR auto	
VOA 1	WER	25.3	22.6	22.5	
VOA 2	WER	30.3	27.3	28.6	
VOA 3	WER	29.5	25.9	26.3	

- VOA 1 : average cluster purity 97.7 % and average speaker purity 96.7 %
- VOA 2 : average cluster purity 83.0 % and average speaker purity 87.9 %
- VOA 3 : average cluster purity 83.8 % and average speaker purity 84.4 %

## speaker tracking



## results / VOA

		baseline	CMLLR true	CMLLR auto	CMLLR auto-d
VOA 1	WER	25.3	22.6	22.5	22.5
VOA 2	WER	30.3	27.3	28.6	28.2
VOA 3	WER	29.5	25.9	26.3	26.0

- VOA 1 : average cluster purity 99.9 % and average speaker purity 96.7 %
- VOA 2 : average cluster purity 91.6 % and average speaker purity 89.8 %
- VOA 3 : average cluster purity 93.4 % and average speaker purity 88.5 %

## conclusions

- CMLLR speaker adaptation introduced some significant error reductions
  - relative error reduction on SPEECON 20 % and on YLE 25 %
  - relative error reductions on our English data were about 10 %
- The difference may follow from our English models being more complex
- Solution: more complex adaptation methods like regression tree CMLLR



**thank you**