Brno University of Technology
Faculty of Information Technology
Speech@FIT

Igor Szöke, Michal Fapšo, Martin Karafiát,
Lukáš Burget, František Grézl, Petr Schwarz, Ondřej Glembek,
Pavel Matějka, Stanislav Kontár, Honza Černocký

szoke@fit.vutbr.cz

http://www.fit.vutbr.cz/speech

BRNO UNIVERSITY OF TECHNOLOGY

FIT FACULTY OF INFORMATION TECHNOLOGY

NIST STD 2006 workshop, December 14.-15. 2006, Gaithersburg

Outlines

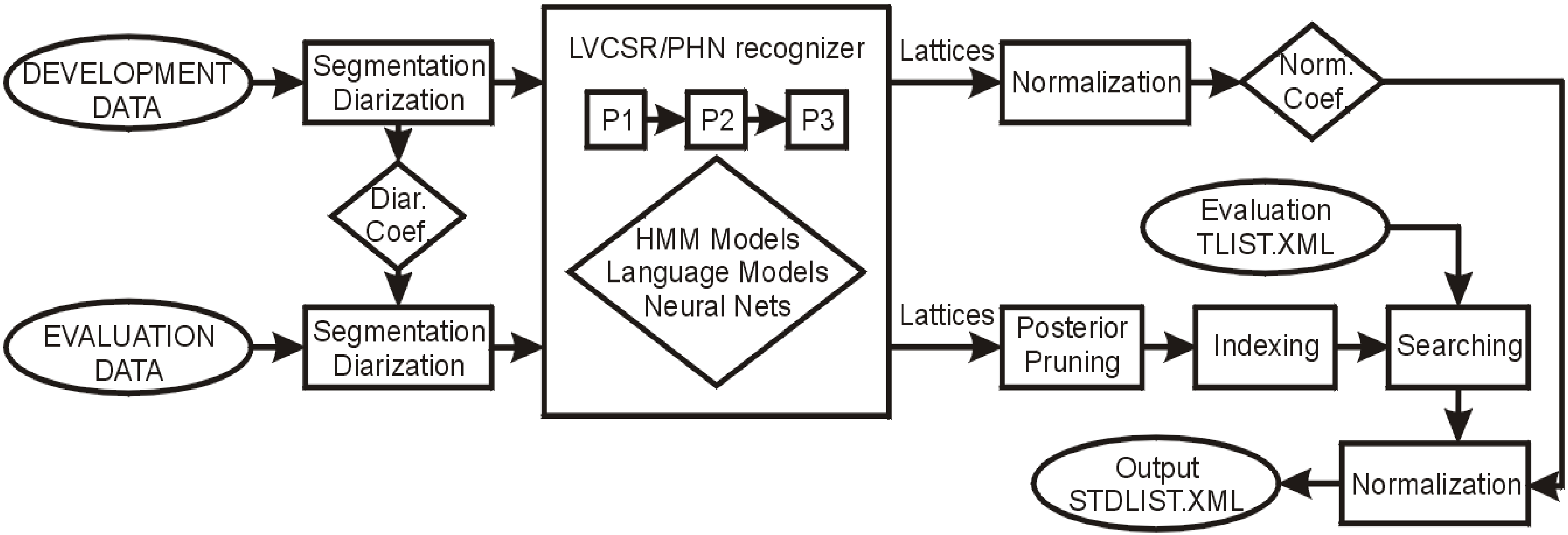
- **System overview**
- **LVCSR/Phoneme recognizer**
- **Indexing and searching**
- **Results and discussion**

English: Broadcast News, Conversational Telephone Speech, Conference Meetings

Arabic: Broadcat News, Conversational Telephone Speech

See the system description for details.

Spoken Term Detection System



Segmentation

- **Speech/nonspeech detection was done using LC/RC long temporal context phoneme recognizer [Schwarz06,Matejka06]**
- **Segments were separated by using silences longer than 0.5s.**
- **Segmentation for CTS was done using comparison of short time energy in both channels. Segment is labeled as silence if:**
  - **the average energy in 'speech' segment is 30 dB less than the maximum energy of the utterance**
  - **the energy in the other channel is higher than maximum energy minus 3dB in the processed channel**
- **Diarization for BCN and MTG done by David van Leeuwen and Matěj Konečný at TNO.**

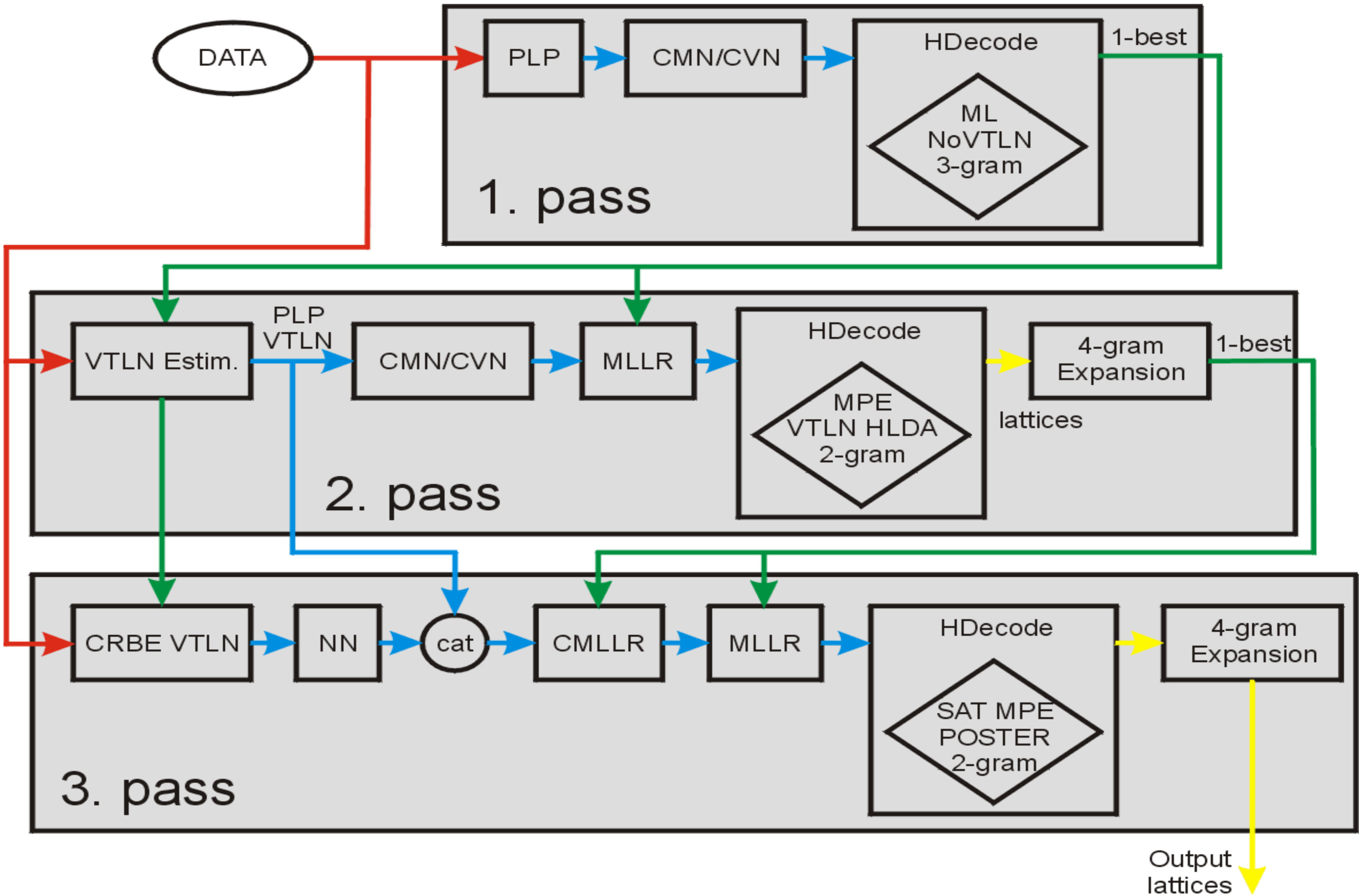
Diarization

Bayesian Information Criterion (Chen & Gopalakishnan, 1998)

- **1 full covariance Gaussian model per segment/cluster, 13 PLP features, 16 ms frames**
- **compare self-likelihood data on model, between separate and merged segments/clusters, compensate for model complexity**

- **Segmentation**
  - **speech activity detection (only for meetings)**
  - **segment break considered every 0.1 s (6 frames)**
- **Clustering**
  - **Initialize clusters with segments found above**
  - **Agglomerative merging of clusters with smallest Gish distance**
  - **BIC stopping Criterion**
- **Viterbi re-segmentation**
  - **Build 16-Gaussian GMMs using clusters found above**
  - **include model for non-speech (silence)**

Description of The LVCSR



Description of The LVCSR

- **We cooperate on development of LVCSR with AMI partners**
- **System (derived from AMI) uses 3-pass decoding:**

  1. pass: PLP, CMN/CVN, ML models, 3-gram decoding, 1-best output

  2. pass: PLP, VTLN, CMN/CVN, HLDA, MPE models, MLLR speaker adaptation, 2-gram decoding, expansion to 4-gram, 1-best output

  3. pass: NN features + PLP, VTLN, CMN/CVN, SAT MPE models, CMLLR/MLLR speaker adaptation, 2-gram decoding, expansion to 4-gram, lattices output

- **Posterior pruning was applied on final lattices.**

For details see:System description and AMI LVCSR paper [Hain06]

LVCSR Training Data

Acoustic:

- **CTS: 277h of SWB1, part of SWB2, CHE.**
- **MTG: 63h of MDM meeting corpora (NIST, ISL, ICSI, AMI). The crosstalk parts were removed and beamforming to one superchannel was done (superchannel generated by IDIAP used for NIST RT05).**
- **BCN: 112h of IHM meeting corpora (NIST, ISL, ICSI, AMI). No BCN data were used!**

LM: SWB, Fisher, Web, BBC, HUB4, SDR99, Enron email, ICSI/ISL/NIST/AMI. Total - 1.49GW

Perplexity was maximized for each task independently.

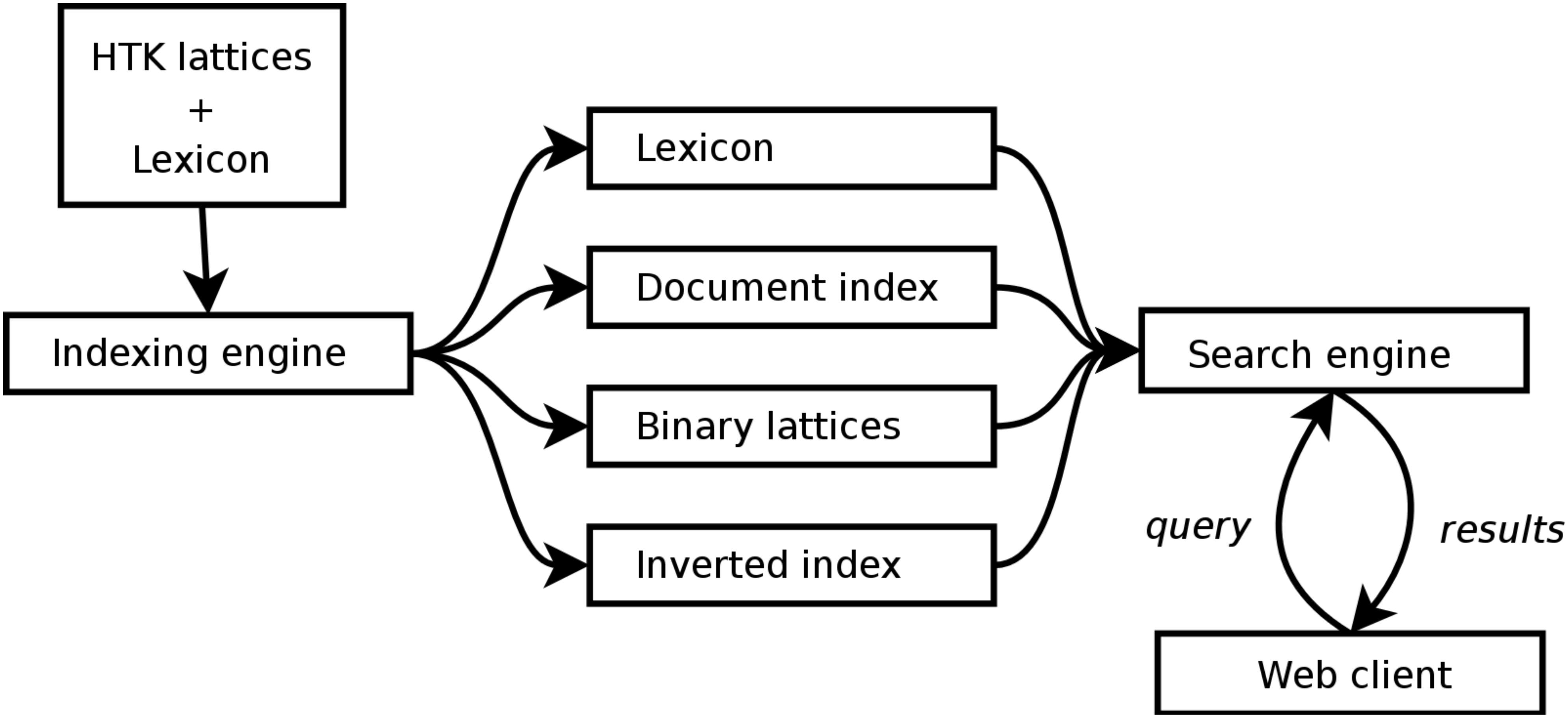
LVCSR WER and Oracle for STD Development Set

	WER	Oracle WER
BCN	21.03%	9.06%
CTS	22.83%	8.32%
MTG	46.65%	21.79%

Description of Phoneme System

- **Phoneme lattices were generated from P3 pass features and acoustic models.**
- **Word language model was replaced by a phoneme 2-gram LM.**
  - **BCN and CTS: trained on phoneme alignment of CTS corpora used for acoustic models training.**
  - **MTG: trained on phoneme alignment of meeting corpora (NIST, ISL, ICSI, AMI).**
- **Posterior pruning was applied.**

Indexing and Search



Indexing I

- **Processing lattices while computing posterior probability of links and generating a forward index. Lattices are stored in our own binary format (optimized for fast access):**
  - **nodes and links are indexed**
  - **random access has *O*(1) complexity**
  - **time index is generated for each lattice to make it possible to cut out only a small part of lattice in the verification step**
- **For word lattices, unigrams are indexed, while for phoneme lattices, indexing units are phoneme 3-grams.**

Indexing II

- **If there are overlapped words, only 1 record is stored in the forward index. It has outer time boundaries of the whole cluster and the highest confidence score (log posterior probability) of all overlapped links.**
- **Two inverted indices are generated:**
  1. **Sorted by *wordID* and *confidence score***
  2. **Sorted by *wordID*, *docID*, *time*. This index is only list of pointers to the first one (no redundant information is stored).**
- **Inverted indices store *wordID*, *docID*, *start time*, *end time* and *confidence score***

Search I

- **Set of hits is retrieved from the inverted index for each word of a term.**
- **A word with the least number of hits is selected and the corresponding set is taken.**
- **For each record in this set, hits from the other words' sets satisfying the time constraints are selected.**
  - ***$O(n^m)$***
    ***n*...number of words in the term**
    ***m*...number of word's hits**
- **This way, a list of candidates is generated.**
- **Since a set of each word's hits in the inverted index is sorted by time, binary search is used to get neighbour word's hits with a lower complexity $O(n \cdot log(m))$.**

Search II

- **The list of candidates is sorted according to an estimated confidence score.**

$$C_{est} = \min_{i=0..N} \left(\max_{j=0..M_i} (C_{ij}) \right)$$

N ... number of words in the query
M_i ... number of overlapped occurrences of the word *i* in the cluster

- **For each of the candidates, existence of valid path in lattice is verified.**
- **Precise posterior probability of each candidate is evaluated.**

OOV Search

- **If a word is not in LVCSR dictionary, G2P rules are applied for phoneme string generation.**
- **Phoneme string is converted to a sequence of overlapped phoneme trigrams, which are searched in index (phoneme trigrams).**
- **If there are 2 or more consecutive OOVs, they are processed as one word with possibility of having *sil* between them.**
- **If all trigrams satisfy time constraints (are overlapped), then the candidate is verified in phoneme lattice and posterior probability is calculated.**
- **OOVs shorter than 3 phonemes are not searched.**
- **Terms with OOVs shorter than 3 phonemes are not searched.**

Term Search

- **After OOV candidates are verified, they are handled as if they were in LVCSR index.**
- **Term is split into sequences of IV and OOV words.**
- **One word sequences are obtained directly from the index (are not verified).**
- **Two or more IV word sequences are verified in lattice.**
- **If time constraints of all sequences are satisfied, the worst confidence score of them is returned (= term nonnormalized posterior probability).**

Normalization

- **The goal is to normalize score of different keywords where we consider that the score is affected by:**
  - **length of keyword**
  - **phonemes the keyword consists of**

$$NScore(KW) = score(KW) - G - len(KW) * F - |phn1| * P_1 - |phn2| * P_2 - \dots$$

- **score(kw) is confidence score of keyword (log posterior probability)**
- **len(KW) is length of the keyword (in frames)**
- **|phnN| is count of phoneme N in the keyword**
- **G is global offset to shift optimal threshold to 0**
- **G, F, P1, P2, ..., PN are constants to be estimated on development data.**

Normalization

- **For large set of KWs, we derived scores for HITs and FAs on the development set.**
- **The scores corresponding to each keyword are used to construct pairs of (HIT,FA).**
- **For each pair, an equation in the following form is created:**

$$(score(HIT)+score(FA))/2=G+len(HIT)*F+|phn1|*P_1+|phn2|*P_2+\dots$$

  - **The left side represents an optimal threshold for given (HIT, FA) pair.**
- **We solve the over-deffined set of equations in minimum square error sence.**

Results

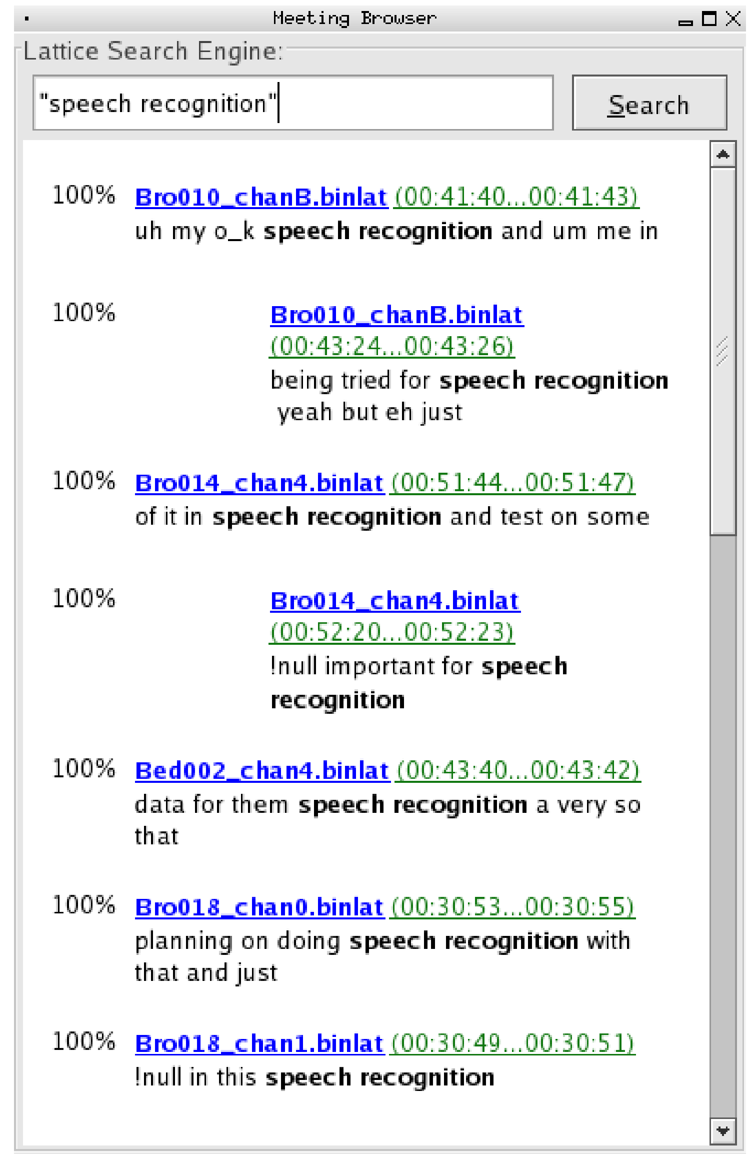
	EVAL ATWV Merged	EVAL MTWV Merged	EVAL MTWV LVCSR	EVAL MTWV PHN	DEVEL MTWV Merged	
BCN	0.6541	0.6558	0.6305	0.3625	0.7020	
CTS	0.5235	0.5344	0.5301	0.3106	0.5580	
MTG	0.0549	0.0731	0.0695	0.0540	0.2950	!
BCN	DEVEL Merged lattices + index	DEVEL Merged index	DEVEL LVCSR lattices + index	DEVEL LVCSR index	DEVEL PHN lattices + index	DEVEL PHN index
size	1716M	242,8M	395,8M	7,8M	1319M	235M
Verif *NoVerif*	0.7020	0.6880	0.6690	0.6670	0.3960	0.3770

Lessons Learned

- **Using 4-gram expansion is only slightly better than 3-gram expansion (according to TWV).**
- **Posterior pruning of LVCSR lattices shortens DET but does not decreases TWV significantly.**
- **Posterior pruning of PHN lattices shortens DET and decreases TWV only a little. TWV decreases a lot for greater pruning factors.**
- **The higher branching factor for PHN lattices, the better TWV. Using higher branching factor and then stronger posterior prunning gives better TWV.**

Search Engine Capabilities not Used in STD

- **Getting a context for each result by traversing the lattice forward and backward from the found sequence of links.**
- **Searching for unquoted queries by specifying a maximum time distance between words.**
- **Client/server architecture**
- **Graphical user interface**



Credit Outside BUT

- Thomas Hain (Sheffield) for having coordinated the AMI LVCSR.
- Vinny Wan (Sheffield) for all word language models.
- David van Leeuwen and Matěj Konečný (TNO) for diarization.
- Cambridge for providing definition of h5train03 CTS training set.
- IDIAP for beam-forming.
- Funding agencies:
  - EC
  - Czech Ministry of Defence
  - CESNET (for the HW to burn)

References

[1] Schwarz P., Matejka P. and Cernocky J.: Hierarchical Structures of Neural Networks for Phoneme Recognition, In Proceedings of ICASSP 2006, May 2006, Toulouse, France

[2] Matejka P., Burget L., Schwarz P. and Cernocky J., Brno University of Technology System for NIST 2005 Language Recognition Evaluation. Odyssey: The Speaker and Language Recognition Workshop, San Juan, Puerto Rico,Jun 2006

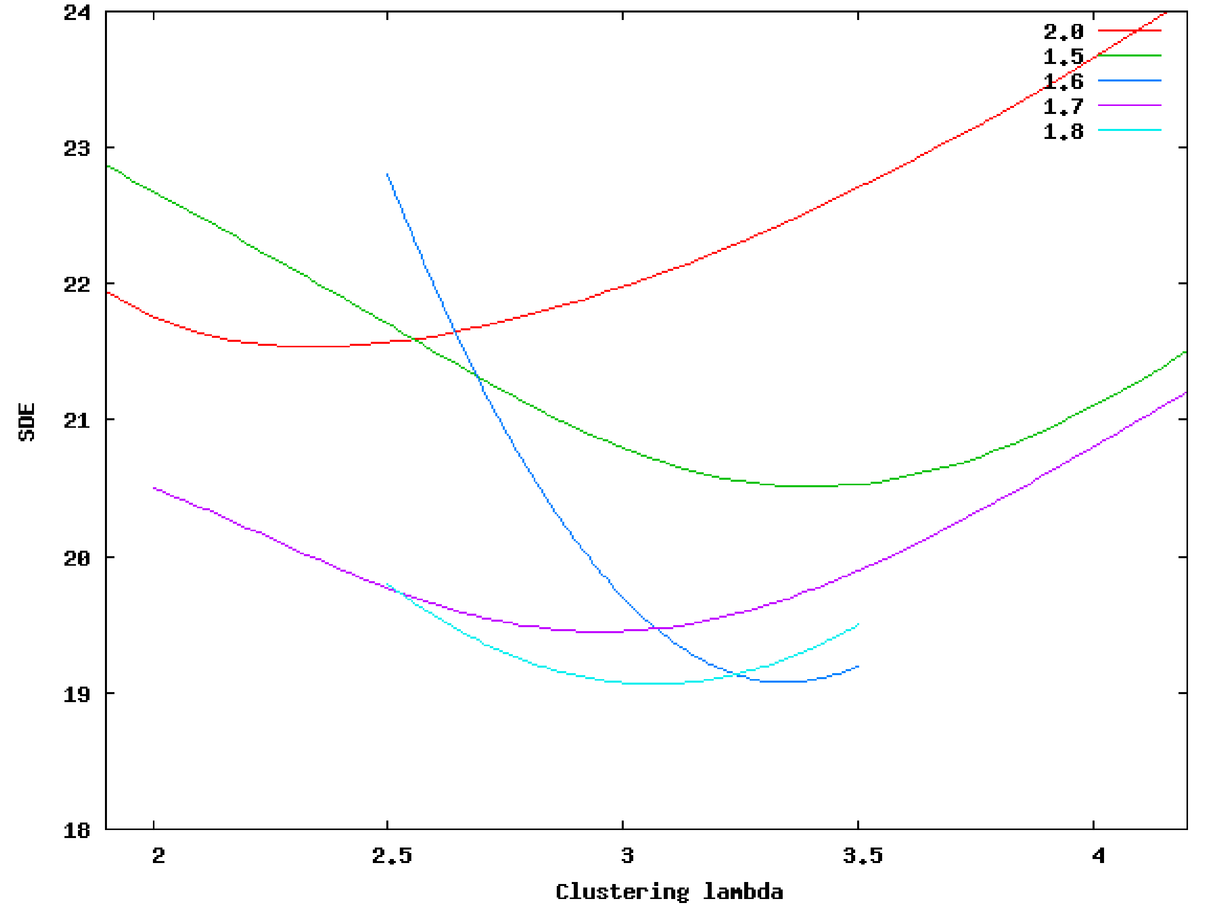
[3] Thomas Hain et al., The AMI Meeting Transcription System: Progress and Preformance, NIST RT06 evaluations, 2006

Thank You for Your attention.

Choosing λ…

(diarization bonus slide)

λ	MTG	BN
Seg	1.8	1.7
Clust	4	3



BN choice λ's

λ penalizes more parameters for separate models
higher λ: less segments, less clusters
Choice of λ optimized for minimum Speaker Diarization Error rate on devset.