IMPROVING VOCABULARY INDEPENDENT HMM DECODING RESULTS BY USING THE DEC

Mikko Kurimo
Helsinki University of Technology
Neural Networks Research Centre
P.O.Box 2200, FIN-02015 HUT
Espoo, Finland

email: Mikko.Kurimo@hut.fi

Our current speech recognition system

- 80-dimensional "context vectors" as input: five averaged mel-cepstra and power

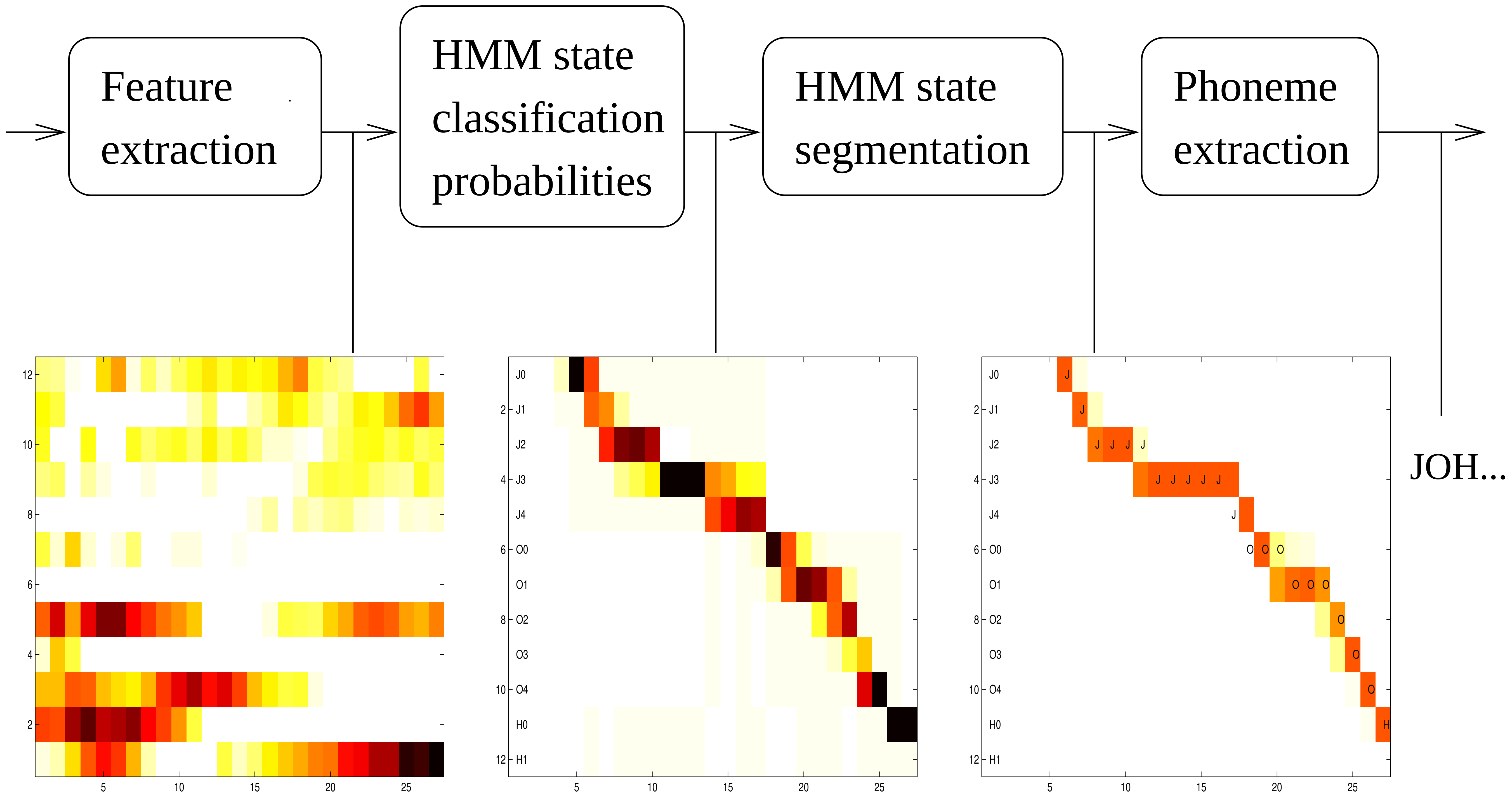
- 5-state **mixture density HMMs** as context-independent phoneme models

- The 140-component Gaussian mixtures are trained by segmental SOM and LVQ3 [M. Kurimo, Ph.D. thesis, 1997]

- The resulting phoneme strings can be postprocessed by **Dynamically Expanding Context** (DEC)

- SOM, LVQ3: [T. Kohonen, Self-Organizing Maps, Springer, 1995, 1997]

The main phases in recognition



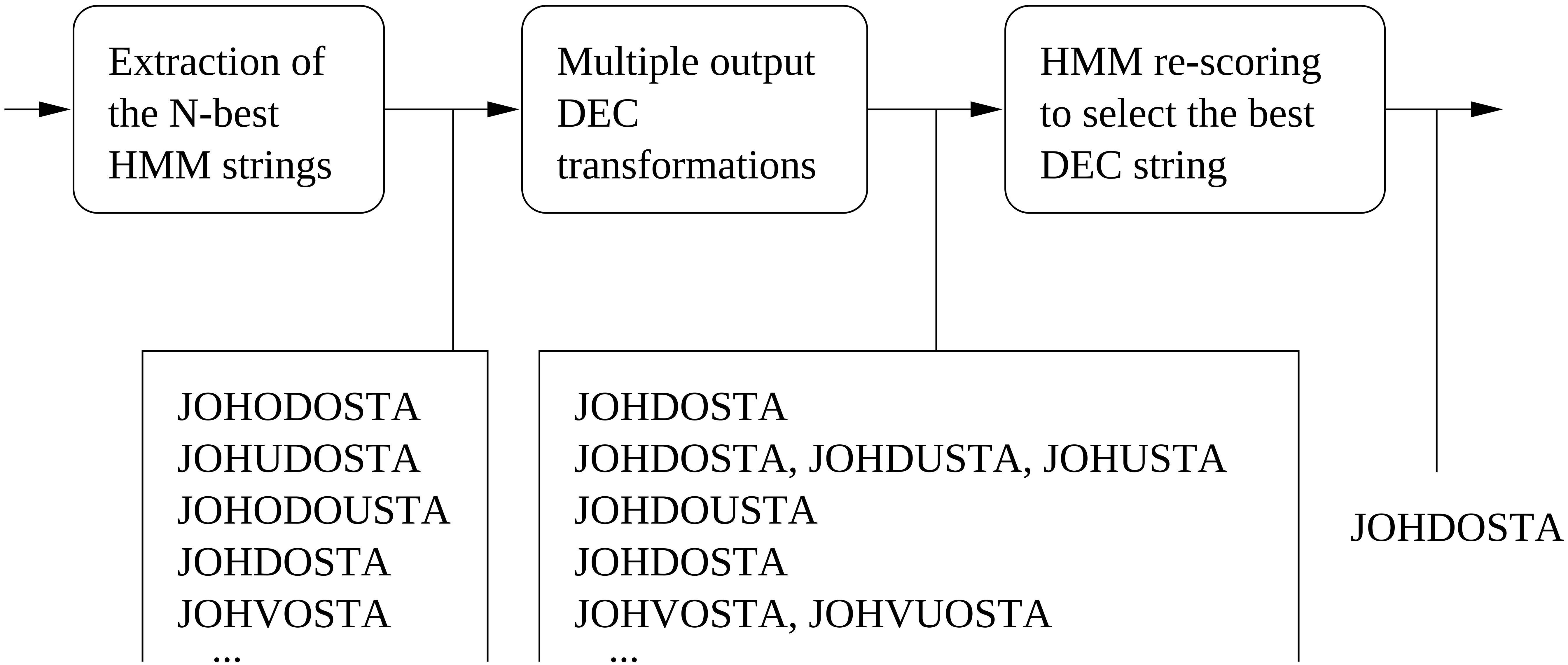
Vocabulary-independent post-processing

- The idea is to find out the frequent changes in spoken words, as /kolome/ $\rightarrow$ /kolme/, and generate **transformation rules** to correct them
- The rules are not stored for whole words, but only for the **minimum amount of local phoneme context**
- **DEC learns** the rules from a list of string pairs: *(recognition result, correct string)*
- First, only a narrow context, e.g. l(o)m $\rightarrow$ l()m, is stored
- More context is added each time the rules conflict
- DEC:[T. Kohonen, Proc. ICPR, 1986]

Multiple output DEC

- There is an upper limit of stored context length
- If the DEC transformation of a string is ambiguous, the most frequently occuring string is normally chosen
- Here, **feed all** the alternative DEC outputs forward to the next decision stage
- For the unlimited vocabulary recognition task, **choose the string that fits best** to the acoustic features
- Measure the fit using **the likelihood of the best HMM** state sequence that produces the string

N-best HMM-DEC system



N-best HMM

- Modify the Viterbi search to **get N different best-matching HMM result strings**
- Can be implemented with low additive computational cost

  e.g. [F. Soong and E-F. Huang, Proc. ICASSP, 1991]
- In postprocessing the most suitable result candidate is selected by grammatical or other task based information
- Here, feed the N result strings are to DEC **to provide a larger collection of alternative DEC results**

Experiments

- **Vocabulary-independent Finnish phoneme recognition**
- Speaker-dependent phoneme models are trained for each speaker using 3 of the 4 recorded sets of 350 words each.
- The error rate is measured on remaining word set and averaged over 7 speakers
- Phoneme error rate $\equiv \dfrac{\text{changed+missing+extra}}{\text{correct number of phonemes}}$
- Only words with no phoneme errors are counted as correct

- HMM: The best-matching HMM string
- D-1: The corrected (DEC) best-matching HMM string
- D-1M: best re-scored string in 1-best HMM-DEC
- D-5M: The best re-scored string in 5-best HMM-DEC

Rate %	**HMM**	**D-1**	**D-1M**	**D-5M**
The original test set				
Phoneme errors	5.2	2.3	**1.9**	2.2
Correct phonemes	96.0	98.2	**98.6**	98.3
Correct words	73.8	90.3	**92.1**	90.1
Correct in top 5	79.4	-	(93.2)	97.3

Rate %	**HMM**	**D-1**	**D-1M**	**D-5M**
The new test set				
Phoneme errors	7.2	7.8	7.0	**6.6**
Correct phonemes	94.8	94.2	94.9	**95.4**
Correct words	66.8	69.1	69.8	**70.9**
Correct in top 5	72.6	-	(71.9)	80.9

- Every 4th word in the lists is separated for testing only
- Both the test samples and the test strings are now unseen
- Degraded phoneme coverage in the training material

Conclusions

- A method is presented to **find out the best result string that obeys the DEC rules**
- DEC corrects erroneous strings very nicely, if new phoneme combinations are expected to be rare
- The DEC sometimes transforms strings with new phoneme combinations in a way that increases phoneme errors, although the correct word rate increases as well
- The N-best HMM-DEC **improves both the phoneme and word rates even for** the tested extreme data with completely **new strings**

Discussion

- The DEC rules are valid only for phoneme combinations by which they were trained
- Here, the **DEC** is trained by exactly the same material as the HMM phoneme models, but it clearly **needs much more training words** to learn the contexts properly
- Compared to the simple system, the N-best HMM-DEC utilizes the training data more thoroughly, because it generates a set of likely string rivals
- In the recognition, the system cannot give the right string, even for large N, if the HMM likelihood score of another valid string is higher