Exemplar-based Recognition of Speech in Highly Variable Noise

Antti Hurmalainen[1]
Katariina Mahkonen[1]
Jort F. Gemmeke[2]
Tuomas Virtanen[1]

[1] Tampere University of Technology, Tampere, Finland

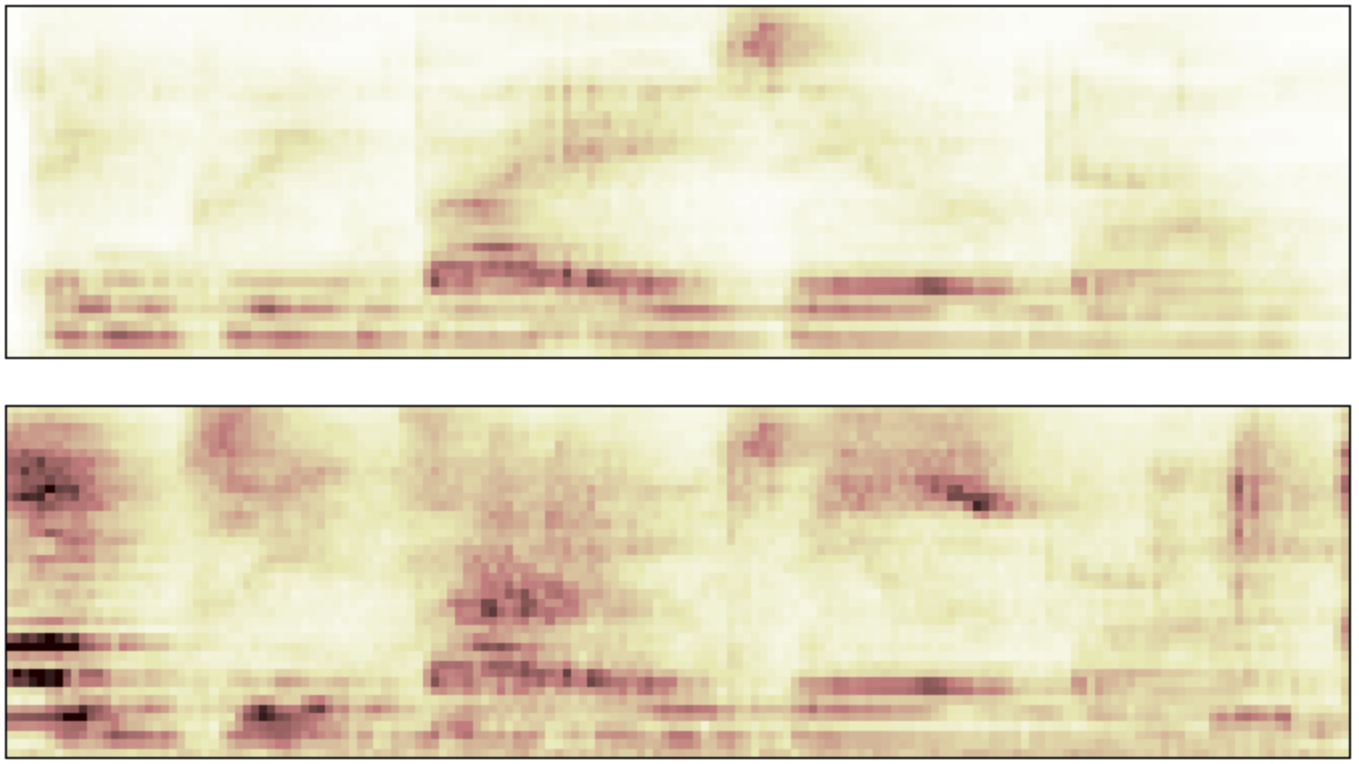
[2] Katholieke Universiteit Leuven, Belgium

Outline

- Background
- Exemplar-based framework
- Baseline results
- Variants
- Summary and conclusions

Robust ASR

Speech and noise mixture – how to handle it?



Robust ASR (2)

Alternative routes:

- Training speech models on noisy data
- Signal level enhancement
- Spectral enhancement or separation, followed by synthesis
-

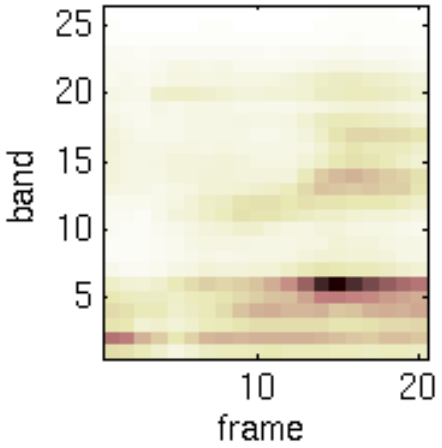
Robust ASR (2)

Alternative routes:

- Training speech models on noisy data
- Signal level enhancement
- Spectral enhancement or separation, followed by synthesis
- *Spectral separation with direct classification*

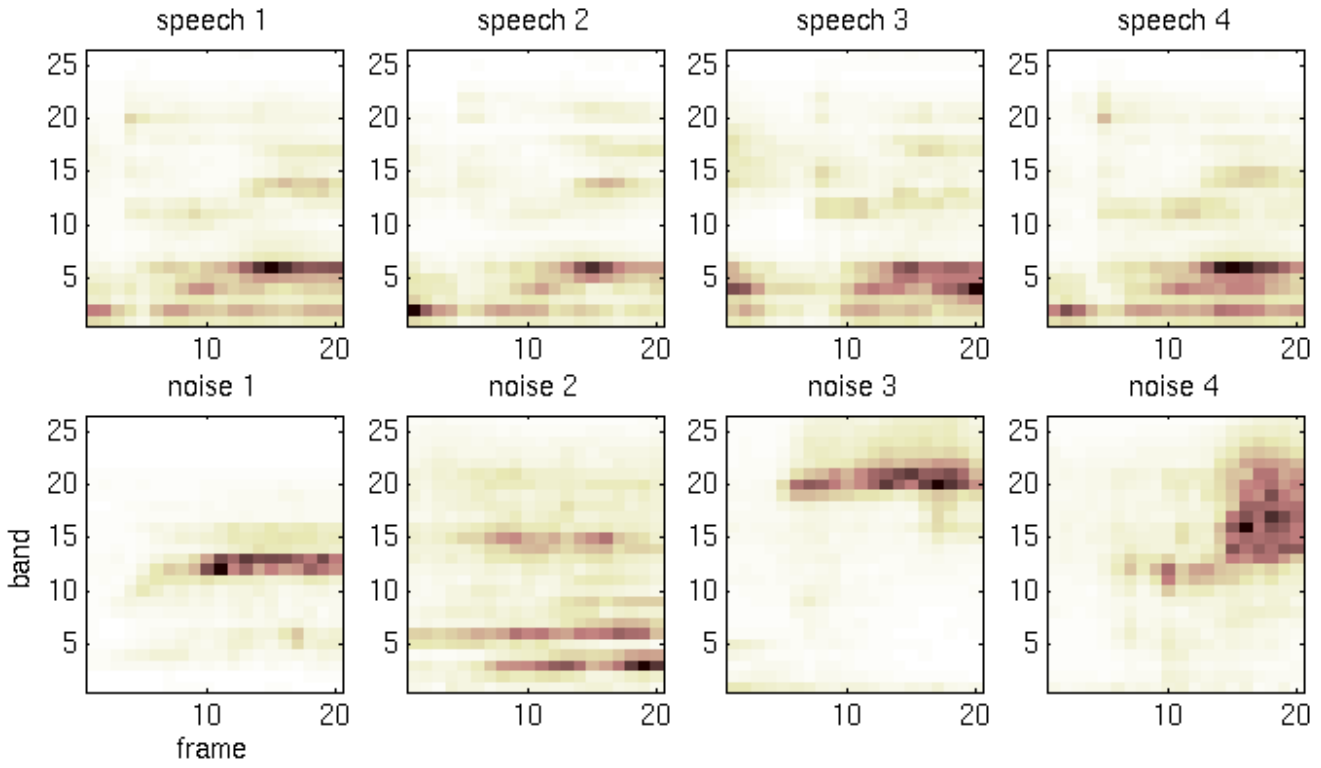
Exemplar-based framework

- Mel-scale spectral magnitude features
- Speech and noise are modelled with *exemplars* – spectrogram windows spanning multiple frames (B x T)



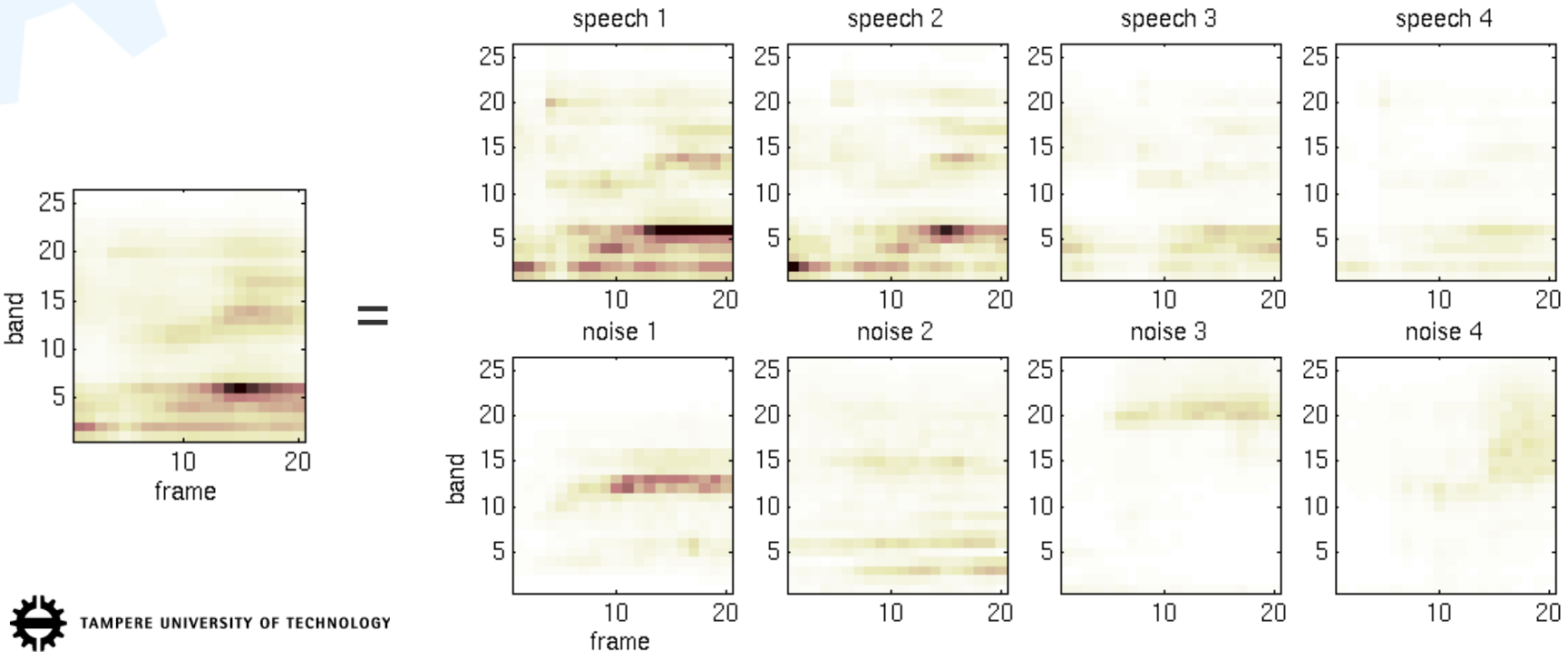
Exemplar-based framework (2)

- Exemplars are sampled from speech and noise, and combined to form the *basis* (or '*dictionary*').



Exemplar-based framework (3)

- The observation window is factorised to a sum of exemplars, producing *activation weights*.
- NMF with sparsity constraints



Exemplar-based framework (4)

Currently sampled features are used “as is”:

- No statistical modelling
- No exemplar learning
- No time warping
- Basis kept fixed during NMF iterations

Basis generation

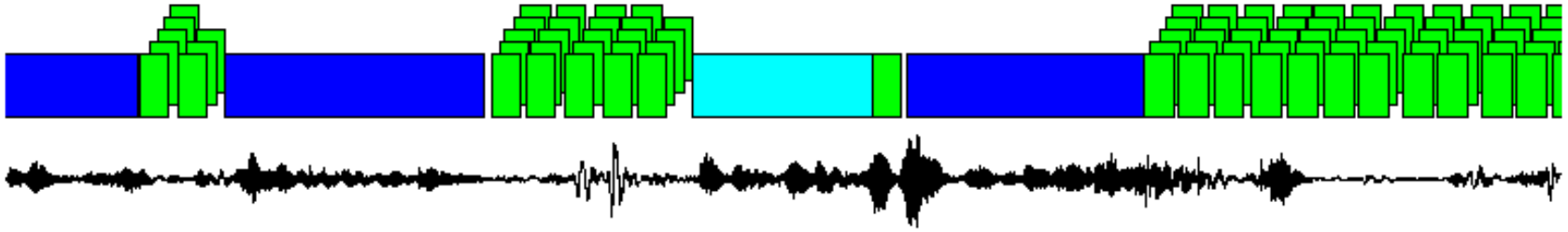
Two different speech basis types:

- Speaker-dependent: Collected for each speaker separately by sampling their training utterances
- Speaker-independent: A mixture of speaker-dependent bases

Bases are built with partially overlapping sampling and then reduced to a manageable size (5000 exemplars) with word probability flattening.

Basis generation (2)

- Noise basis is adaptive, generated for each utterance from its nearby noise context (using the 'embedded' files)
- Also 5000 exemplars



Cyan = current utterance

Blue = other utterances

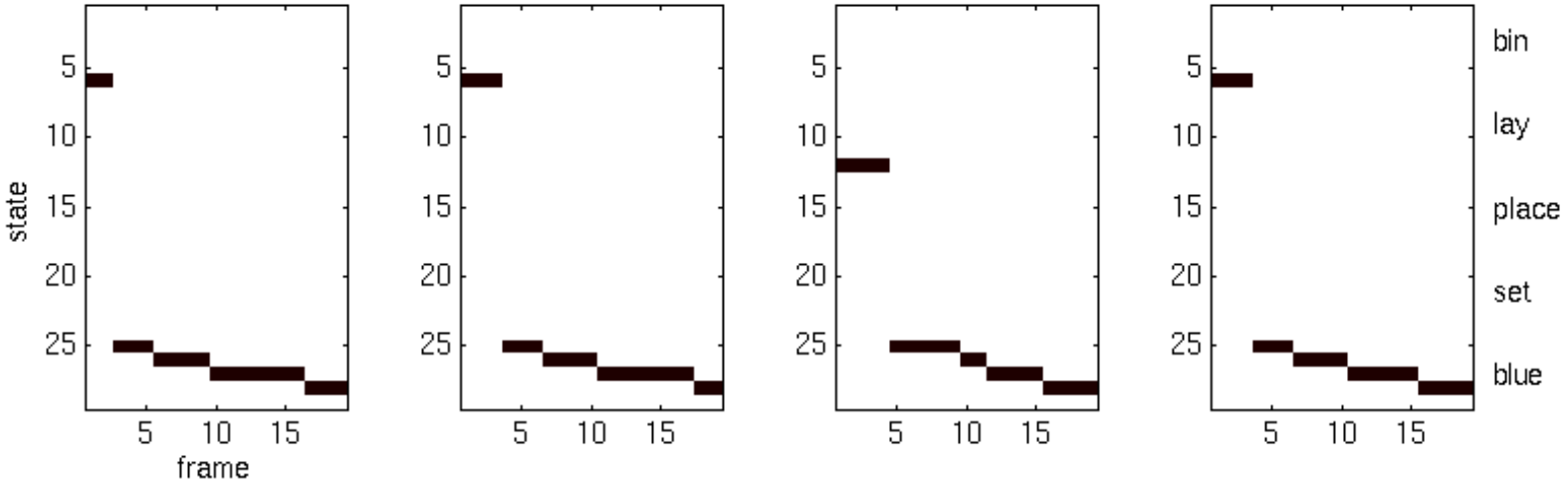
Green = exemplars

Factorisation

- Multiplicative NMF update
- Fixed basis
- Fixed number of iterations (300)
- Minimisation of KL-divergence
- Weighted L_1 penalty for non-zero activations (induces sparsity)

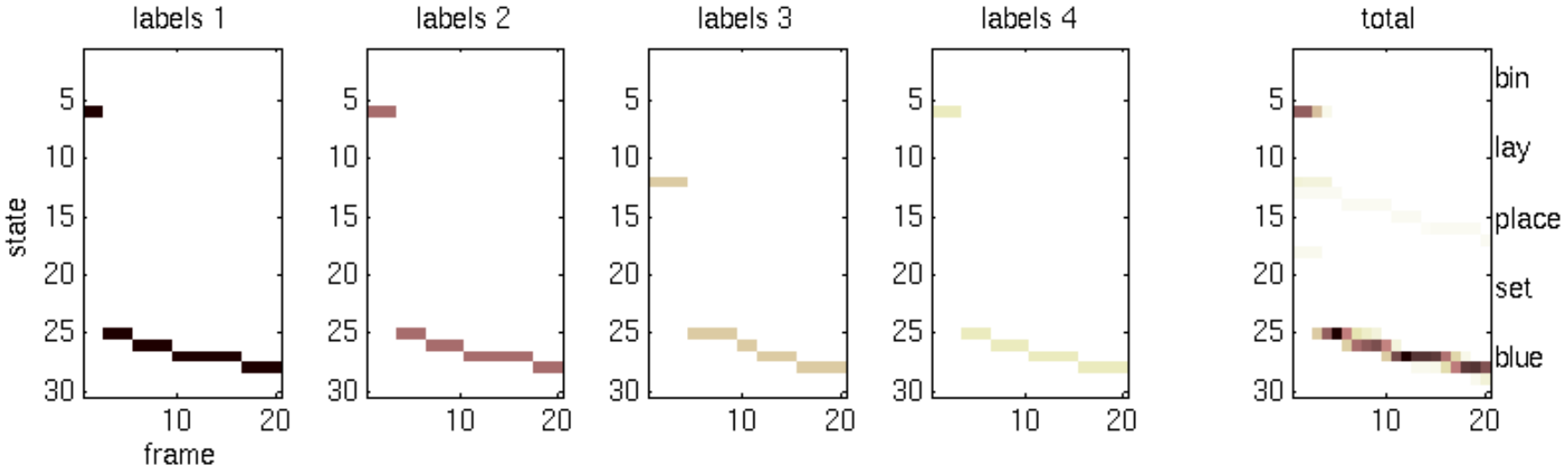
Decoding

Each speech exemplar is given a *label sequence*, acquired from forced alignment transcription of the corresponding training utterance.



Decoding (2)

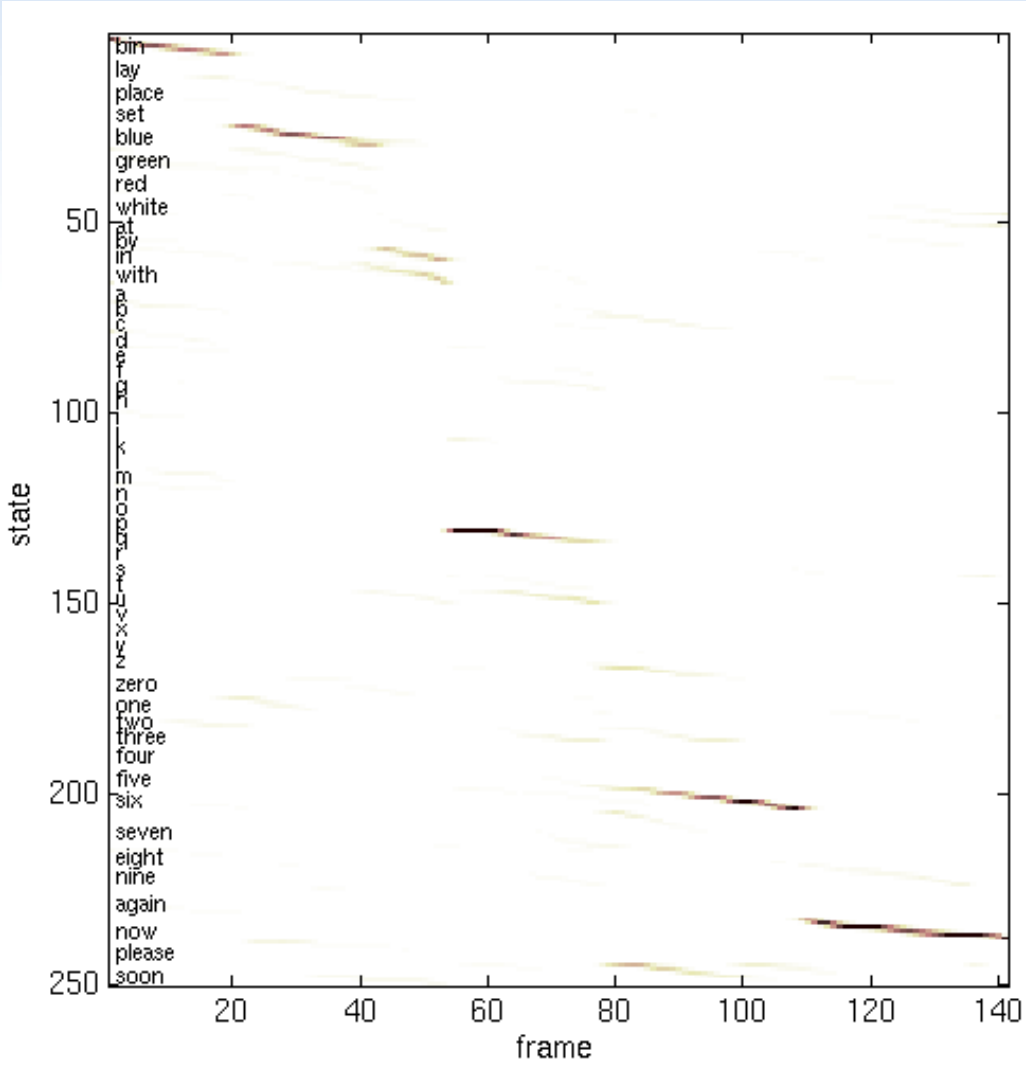
Label sequence matrices are summed according to the activation weights of the corresponding speech exemplars in the factorised window.



Decoding (3)

After repeating the process for each observation window of the utterance, the total *likelihood matrix* can be generated and decoded.

('bin blue in Q 6 now')



Decoding (4)

- There is no synthesis step or GMM evaluation – weighted labels produce the likelihoods directly.
- The matrices are decoded using a modified HVite binary, which reads the likelihood files.

Note: even the likelihood matrix building can be omitted, if a decoder is trained on the activation weights themselves!

Configurable parameters and options

- Spectral range and resolution (26 bands, 16 kHz)
- Temporal resolution (25 ms frames, 10 ms shift)
- Number of frames in a window ('T' = 10, 20 or 30)
- Stereo or downmixing to mono

- NMF iterations and sparsity penalty weight
- Basis normalisation / band weighting
- Activation and likelihood scaling
- ...

Baseline results

Speaker-independent recognition (%):

	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB
GMM	82.1	70.8	61.3	52.0	39.8	34.7
T=10	69.9	66.0	58.7	52.4	42.9	37.8
T=20	77.3	72.8	68.2	62.7	51.1	44.0
T=30	76.0	73.5	68.2	61.8	52.7	44.7

Baseline results (2)

Speaker-dependent recognition (%):

	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB
GMM	82.4	75.0	62.9	49.5	35.4	30.3
T=10	91.3	88.3	85.8	80.8	71.4	62.3
T=20	91.6	89.2	87.6	84.2	74.7	68.0
T=30	88.8	88.1	86.3	82.9	75.1	68.3

Variants: Temporal models

- In the baseline model, we use overlapping windows to cover the whole utterance.
- Each window is factorised as an independent entity, whereafter the results are averaged.

Pros: Multiple independent estimates for each frame. Robust against occasional errors in single windows.

Cons: Requires exemplars with correct temporal alignment to match each window.

Variants: Temporal models (2)

Alternatively, we can use non-negative matrix deconvolution ('NMD')

- Activations at all temporal positions *jointly* form the estimated full utterance spectrogram.
- Several temporal positions may remain empty.

Pros: No need for so many time-shifted exemplar variants. Potentially suited for small dictionaries.

Cons: Single errors may have a larger impact.

Variants: Temporal models (3)

Results (speaker-dependent, %):

NMF

	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB
T=10	91.3	88.3	85.8	80.8	71.4	62.3
T=20	91.6	89.2	87.6	84.2	74.7	68.0
T=30	88.8	88.1	86.3	82.9	75.1	68.3

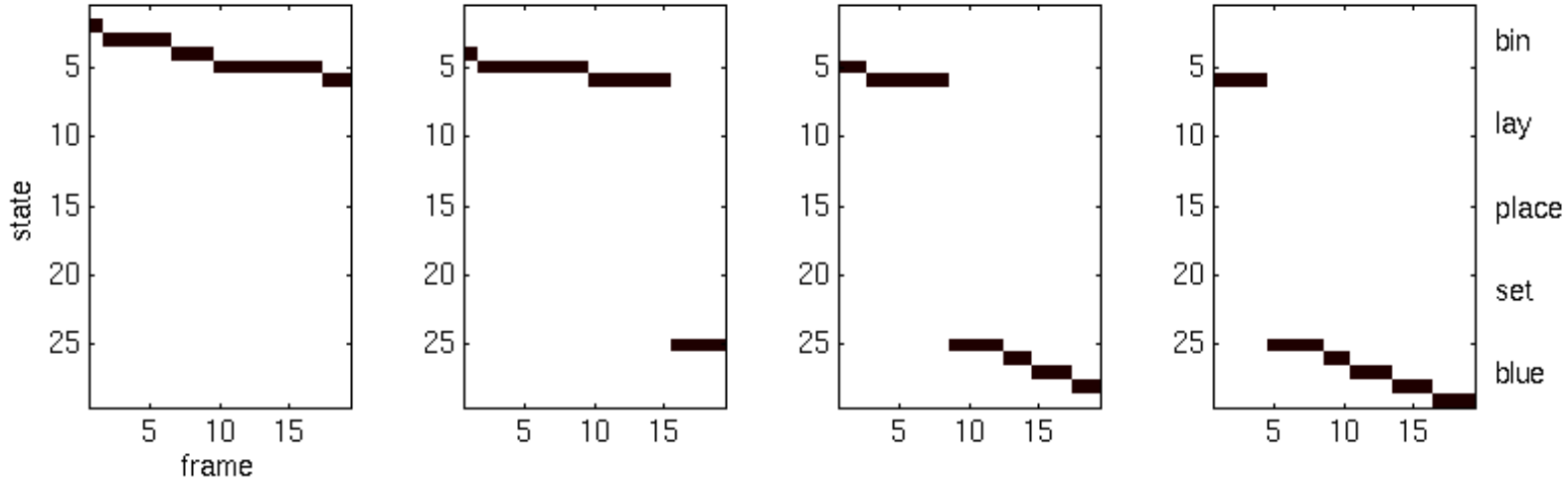
NMD

	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB
T=10	88.3	85.9	83.3	78.8	69.1	59.8
T=20	90.5	88.6	87.0	81.3	72.1	65.9
T=30	87.2	86.1	84.0	79.9	70.6	63.3

Variants: Learnt mapping

In the baseline model, the exemplar transcription (linguistic state of each exemplar frame) comes from external forced alignment.

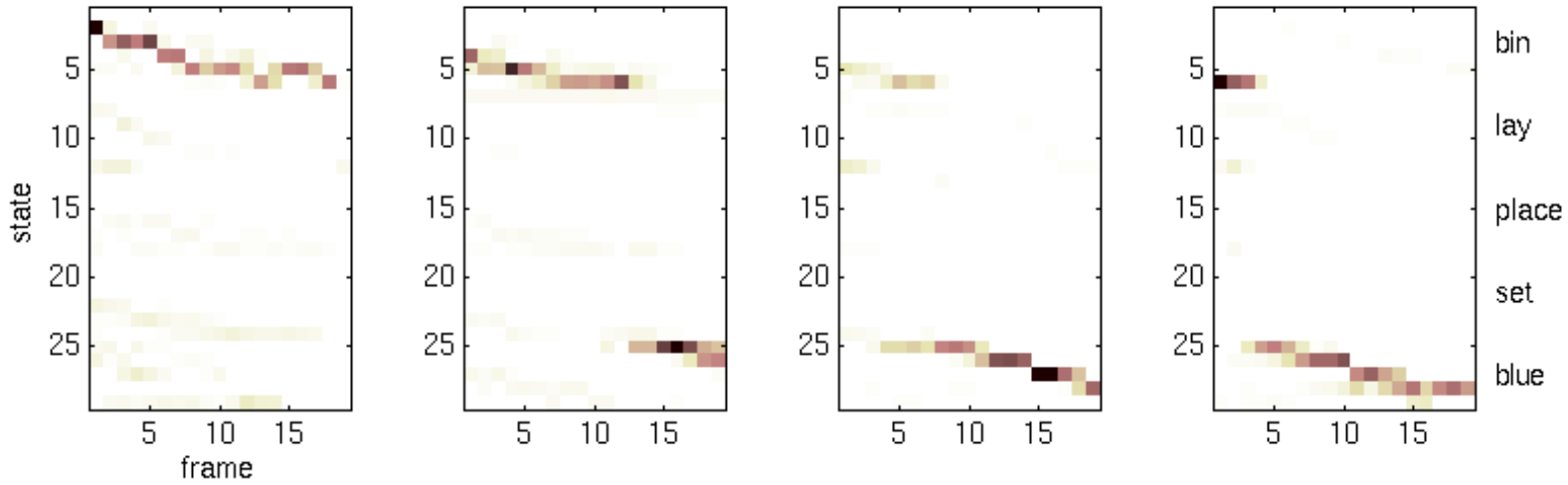
- Strict mapping – an exemplar always produces its predetermined sequence (one state per frame).



Variants: Learnt mapping (2)

Alternative: Learn the exemplar-state mapping by factorising training files and observing the relation.

- Produces fuzzy mapping matrices.
- OLS and PLS regression algorithms



Variants: Learnt mapping (3)

Learnt mapping results (speaker-independent, T=20):

	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB
GMM	82.1	70.8	61.3	52.0	39.8	34.7
labels	77.3	72.8	68.2	62.7	51.1	44.0
OLS	85.2	80.5	78.7	71.7	60.2	51.5
PLS	82.9	78.8	74.8	70.1	59.5	50.6

Variants: Learnt mapping (4)

Learnt mapping results (speaker-dependent, T=20):

	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB
GMM	82.4	75.0	62.9	49.5	35.4	30.3
labels	91.6	89.2	87.6	84.2	74.7	68.0
OLS	91.1	90.0	88.5	85.2	77.6	69.2
PLS	91.9	89.3	88.2	85.0	78.6	69.6

Summary

- An additive model of speech and noise exemplars can be used to represent noisy spectral features.
- Speech activations reveal the linguistic content directly without synthesis or GMM evaluation.
- Long temporal context (up to 300 ms) helps in discovering the underlying patterns robustly.
- Several alternative approaches exist for the factorisation and decoding steps.

Conclusions

- High separation quality can be achieved using speaker-dependent speech, adaptive noise dictionaries, and at least 200 ms context.
- The current feature space and decoding algorithms are still relatively simple.
- Some phonetically close letters cannot be distinguished reliably in this representation.
- As the main framework has become quite stable, focus can be shifted to fine-tuning and integration in pursuit for higher overall recognition rates.

Future work

- Improved feature spaces
- Advanced dictionary construction
- Phonetic models, large vocabulary
- Hybrid algorithms
- More adaptive and learning-based methods

References

JFG, TV & AH, “Exemplar-based sparse representations for noise robust automatic speech recognition”, IEEE TASLP 2011

AH, JFG & TV, “Non-negative matrix deconvolution in noise-robust speech recognition”, ICASSP 2011

KM, AH, TV & JFG, “Mapping sparse representation to state likelihoods in noise-robust automatic speech recognition”, Interspeech 2011

Thank you!