Multimodal speech recognition and enhancement

Dorothea Kolossa & Steffen Zeiler
Cognitive Signal Processing Group /
Kavli Institute for Theoretical Physics

Ruhr-Universität Bochum
Fakultät für Elektrotechnik und Informationstechnik

RUB

Cognitive Signal Processing Group

Research Associate:
Steffen Zeiler

PhD students:
Benedikt Bönninghoff
Hendrik Meutzner
Christopher Schymura
Mahdie Karbasi
Dennis Orth
Lea Schönherr
Former Scientists
Ahmed Hussen Abdelaziz
Sebastian Gergen

Student Research Assistants:
Juan Rios Grajales
Jan Hünnemeyer
Tobias Isenberg
Diana Castano Marin

with many thanks to Robert Nickel, Ning Ma, Guy Brown, Ramon Fernandez Astudillo

Audiovisual speech perception

Human speech perception utilizes video information

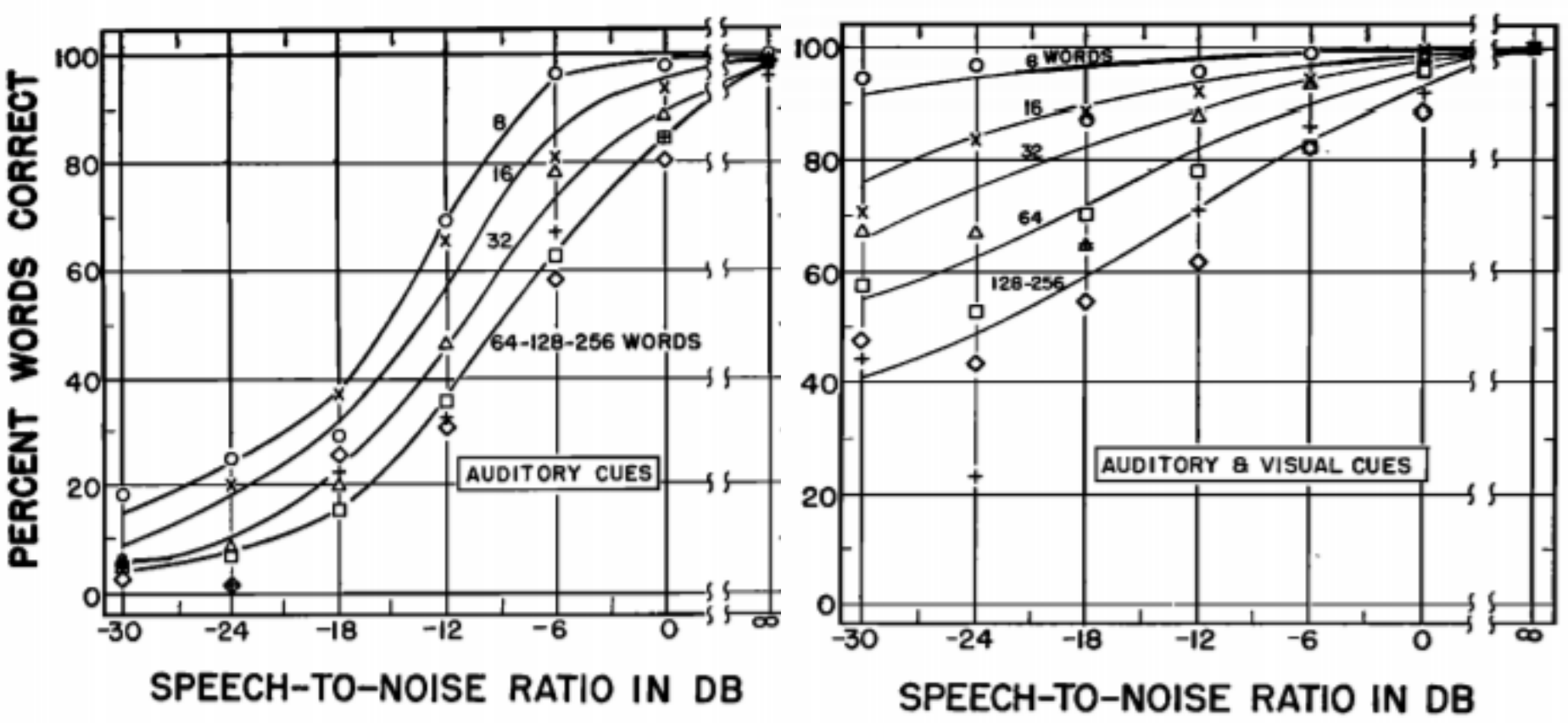
One piece of evidence:

© Myles and Alex Dainis, Bite Sci-zed

is the “McGurk Effect” [McGurk1976]

Audiovisual speech perception

Early, extensive intelligibility tests:



[Sumby1954] Sumby, Pollack: Visual Contribution to Speech Intelligibility in Noise, JASA, 1954.

Introduction & Overview

Idea:

Integrate video information in machine listening

Useful for two purposes:

- Multimodal speech recognition
- Audiovisual Speech Enhancement (to improve intelligibility)

Introduction & Overview

Outline:

- *Audiovisual* speech recognition
    - Methods and models for audiovisual integration
    - Stream weighting
- Audiovisual Speech Enhancement
- Conclusions and perspectives

Audiovisual Speech Recognition

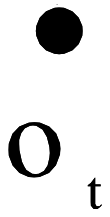
Levels of integration

Levels of integration

Graphical models [Whittaker1990, Jordan1999]

Describe statistical dependencies of multiple variables

"Visible"/"Measureable" variables are often denoted by shaded circles



Levels of integration

Graphical models [Whittaker1990, Jordan1999]

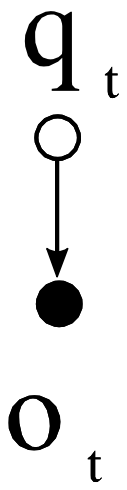
Describe statistical dependencies of multiple variables

"Hidden" variables are often denoted by empty circles

Levels of integration

Graphical models [Whittaker1990, Jordan1999]

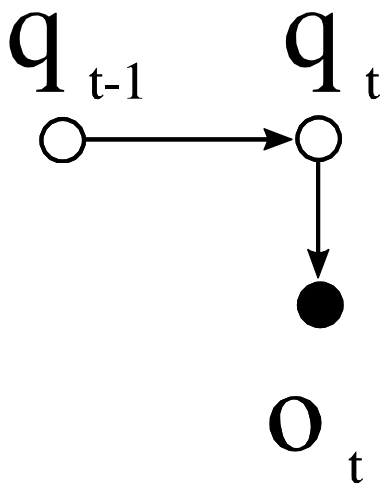
Specifically in “Bayesian Networks”, direct statistical dependencies are denoted by arrows:



Levels of integration

Graphical models [Whittaker1990, Jordan1999]

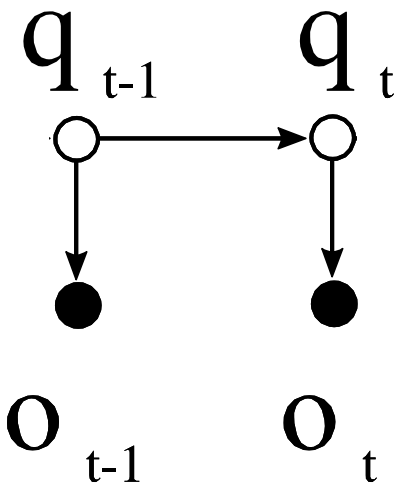
Specifically in “Bayesian Networks”, direct statistical dependencies are denoted by arrows:



Levels of integration

Graphical models [Whittaker1990, Jordan1999]

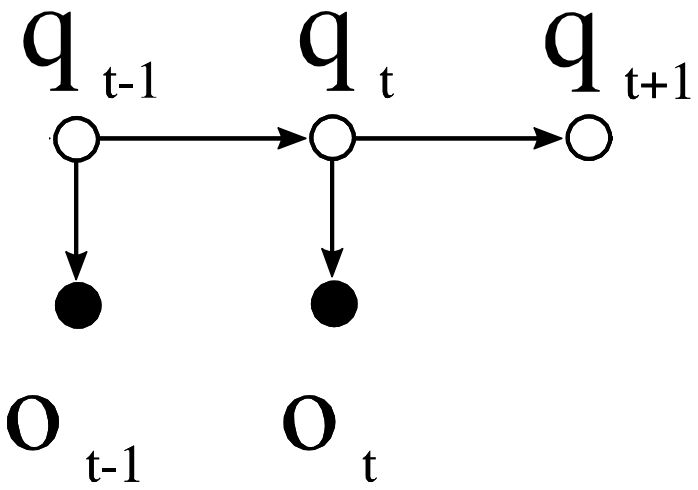
Specifically in “Bayesian Networks”, direct statistical dependencies are denoted by arrows:



Levels of integration

Graphical models [Whittaker1990, Jordan1999]

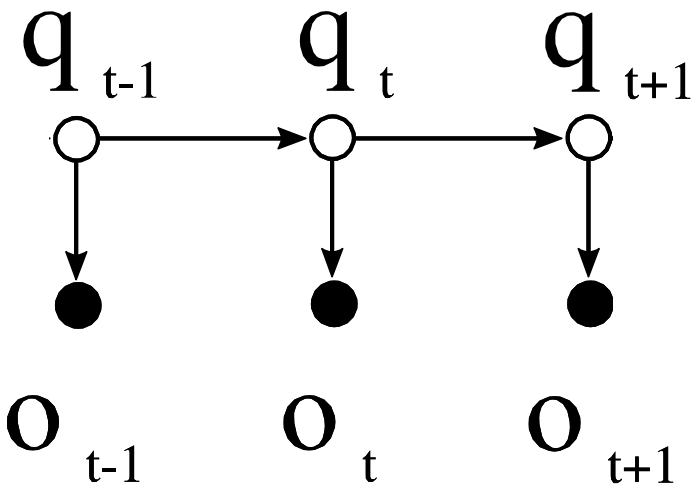
Specifically in “Bayesian Networks”, direct statistical dependencies are denoted by arrows:



Levels of integration

Graphical models [Whittaker1990, Jordan1999]

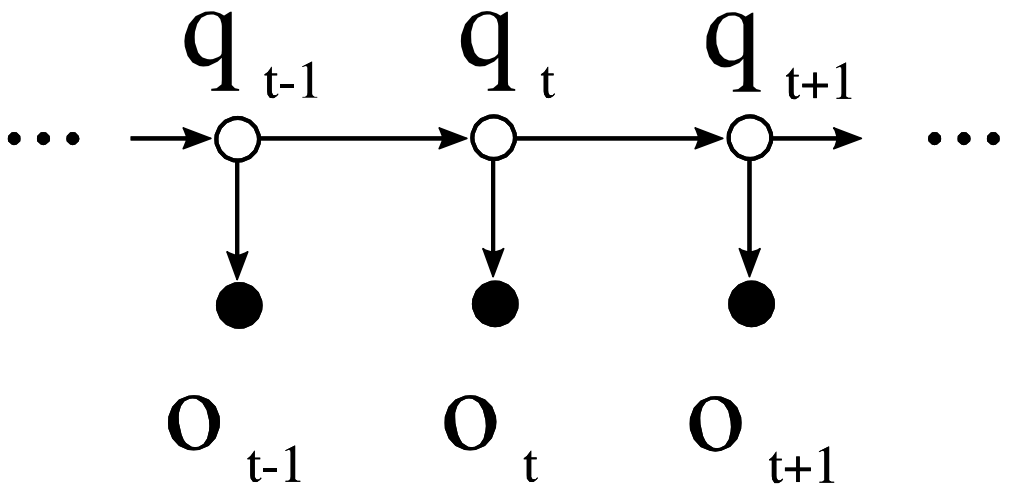
Specifically in “Bayesian Networks”, direct statistical dependencies are denoted by arrows:



Levels of integration

Graphical models [Whittaker1990, Jordan1999]

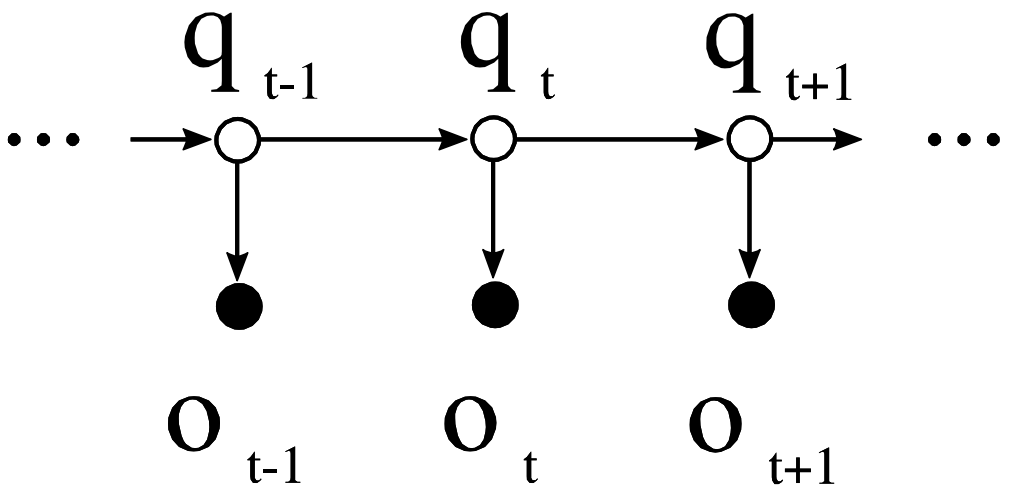
Specifically in “Bayesian Networks”, direct statistical dependencies are denoted by arrows:



Levels of integration

Graphical models [Whittaker1990, Jordan1999]

Indirect statistical dependencies are not:

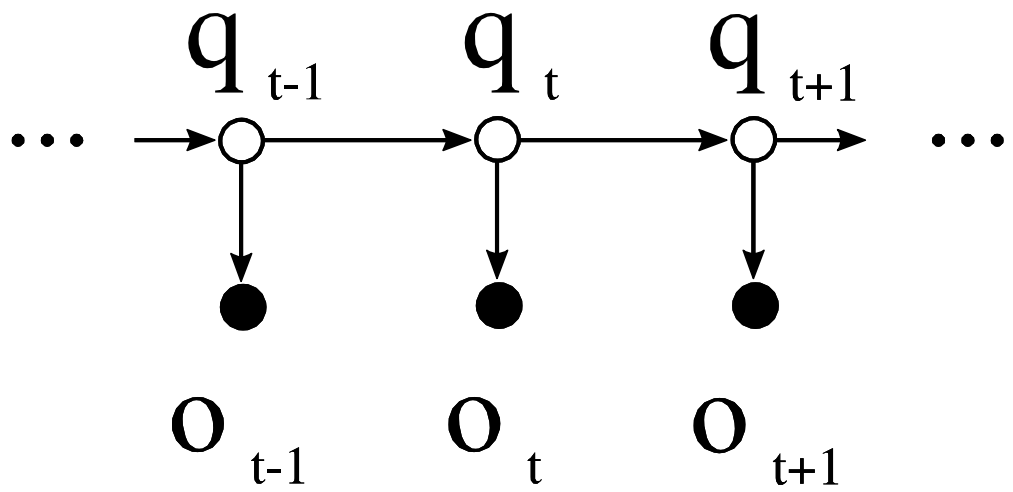


This model encodes the dependency assumptions of (1st order) Hidden Markov Models in speech recognition.

Levels of integration

Multimodal speech recognition can take place at three levels

a) Early integration = Feature fusion

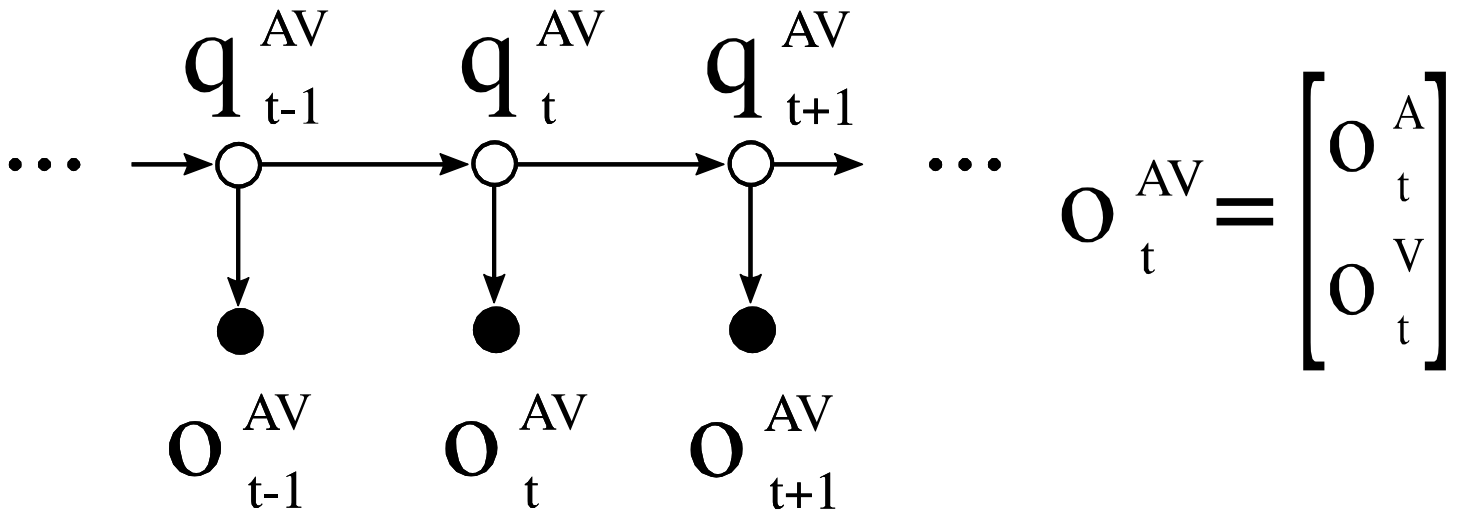


Levels of integration

Multimodal speech recognition can take place at three levels

a) Early integration = Feature fusion

Graphical Model of Audiovisual Speech Recognition with Feature Fusion



System

"Standard" speech recognition setup

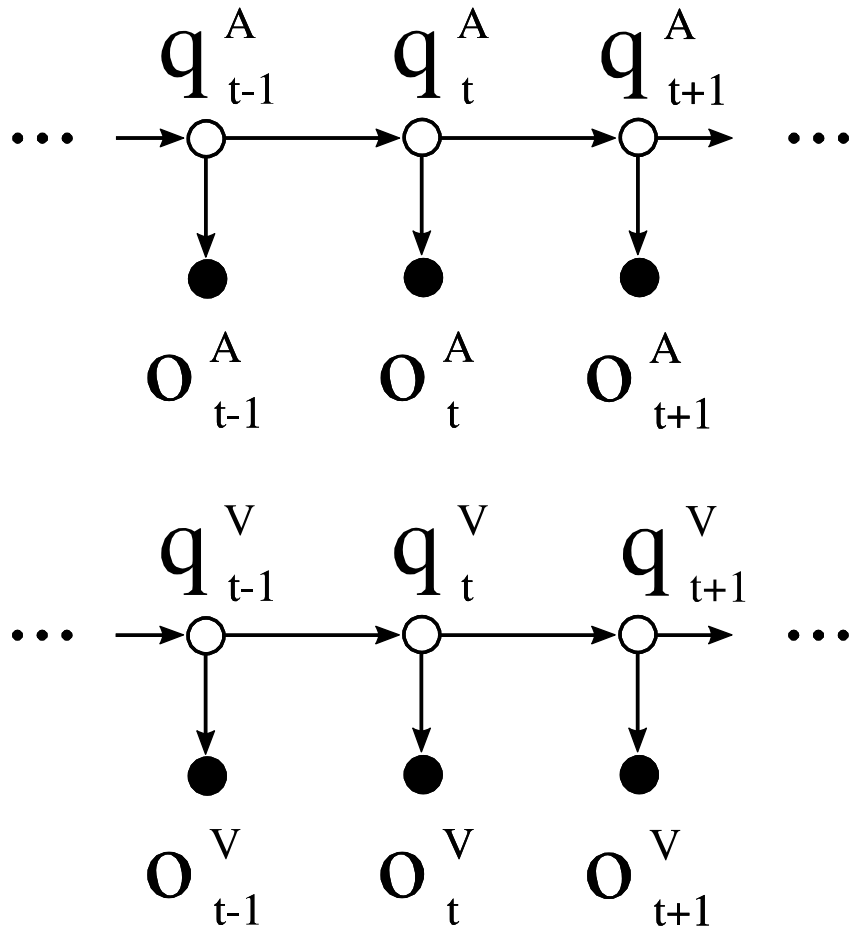
Levels of integration

Multimodal speech recognition can take place at three levels

a) Early integration = Feature fusion

b) Late integration = combine multiple recognition results (ROVER) [Fiscus1997]

Graphical model for audiovisual speech recognition with late integration



Levels of integration

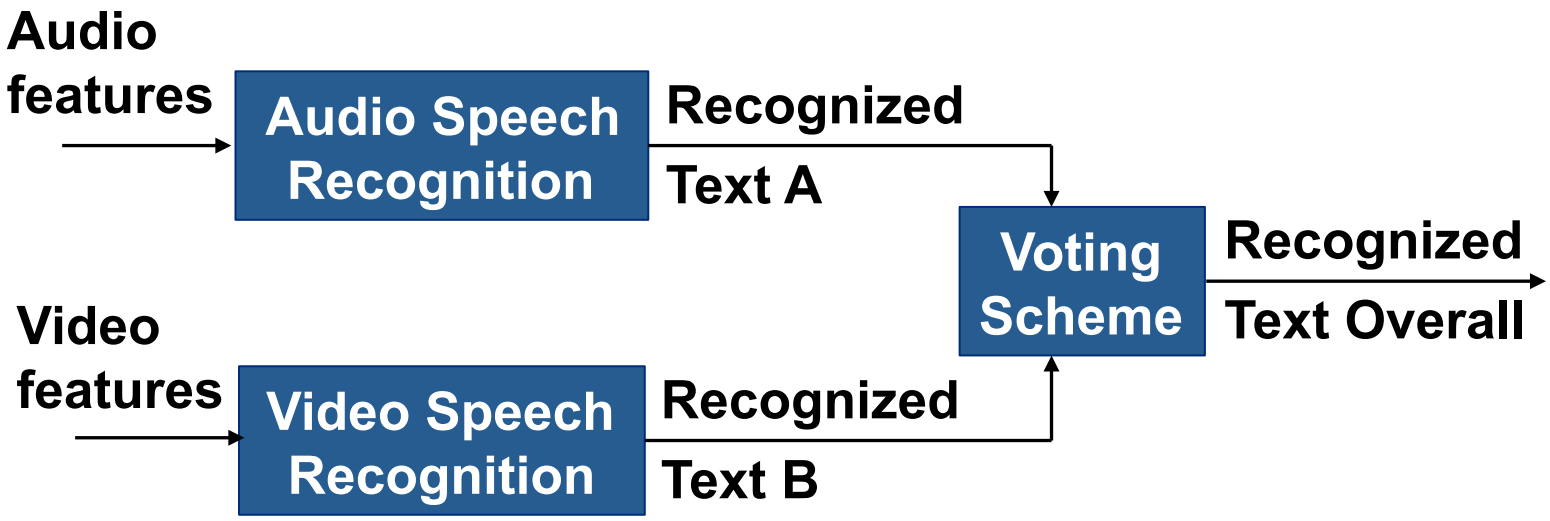
Multimodal speech recognition can take place at three levels

a) Early integration = Feature fusion

b) Late integration = combine multiple recognition results (ROVER) [Fiscus1997]

System for late integration

Two “standard” ASR systems, whose outputs are later combined



Levels of integration

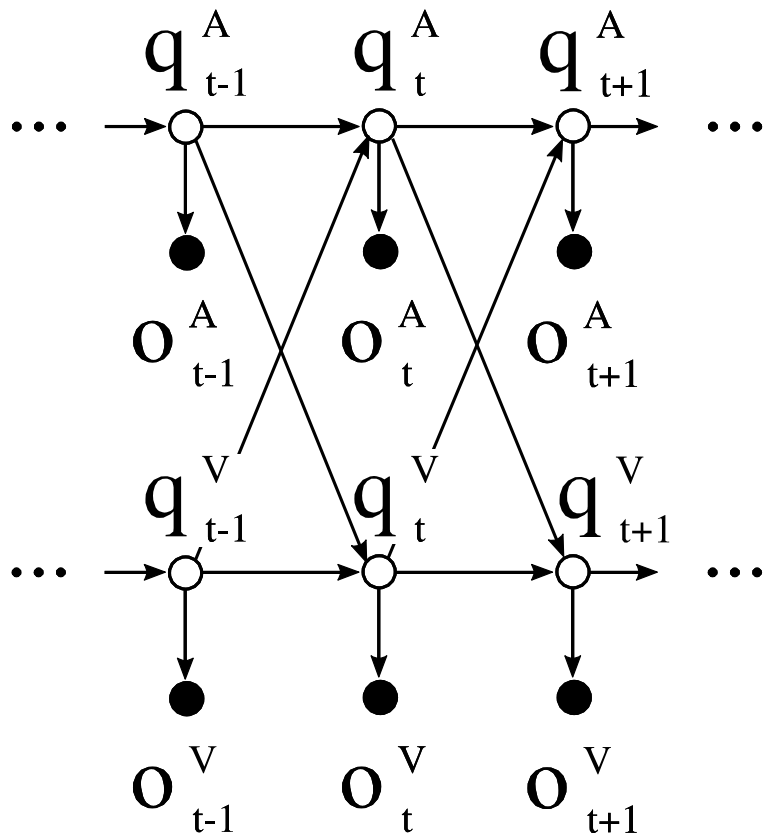
Multimodal speech recognition can take place at three levels

a) Early integration = Feature fusion

b) Late integration = combine multiple recognition results (ROVER) [Fiscus1997]

c) Intermediate integration = within the classifier/DNN

Graphical model, intermediate integration



Levels of integration

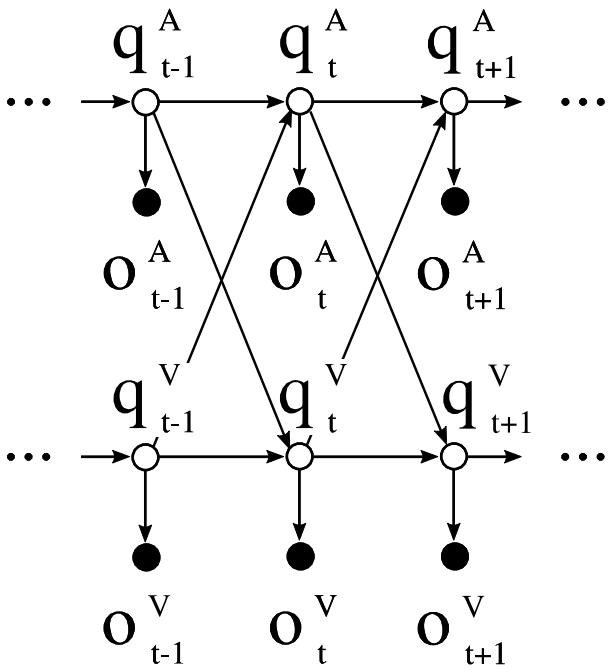
Multimodal speech recognition can take place at three levels

a) Early integration = Feature fusion

b) Late integration = combine multiple recognition results (ROVER) [Fiscus1997]

c) Intermediate integration = within the classifier/DNN

Graphical Model



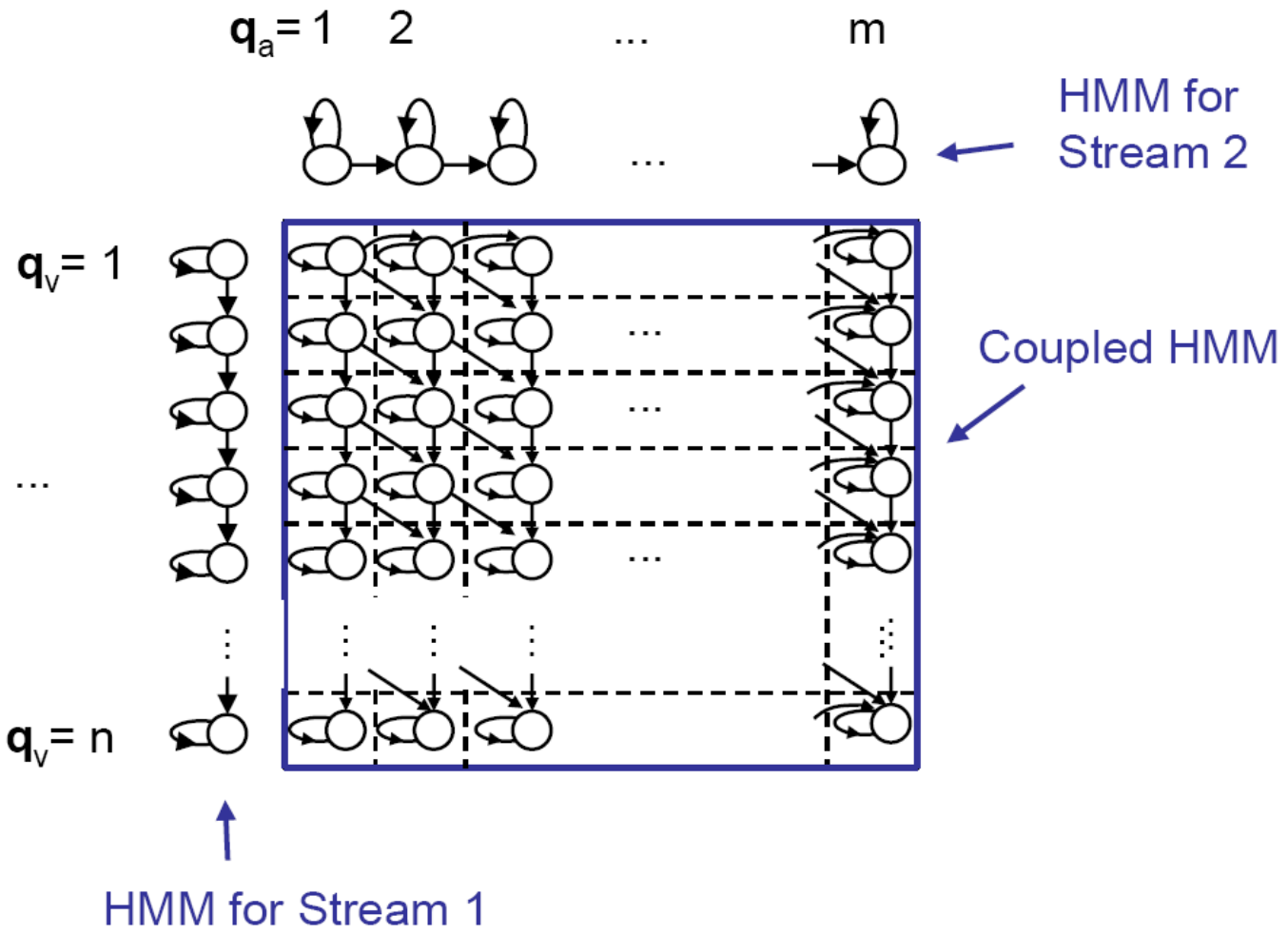
Most successful model in wide range of experiments [Nefian2002a, Zeiler 2016, Receveur2016]

Coupled Hidden Markov Models

An example of intermediate integration

Coupled HMMs for asynchronous Audio- & Video Streams

Cartesian product of audio and video HMM can cope with time-varying delay of audio and video.



[Luettin2001] J. Luettin, G. Potamianos and C. Neti: "Asynchronous Stream Modelling for Large Vocabulary Audio-Visual Speech Recognition", Proc. ICASSP, pp. 169-172, May 2001.

Audiovisual speech recognition

[Vorwerk2011] A. Vorwerk, S. Zeiler, D. Kolossa, R. Fernandez Astudillo and D. Lerch: “Use of Missing and Unreliable Data for Audiovisual Speech Recognition”, in: D. Kolossa, R. Haeb-Umbach (eds.): „Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications“, Springer Verlag, pp. 345-375, July 2011.

Coupled HMMs for asynchronous Audio- & Video Streams

Audiovisual Speech Recognition

- Audiovisual recognition using coupled HMMs always outperforms audio-only and video-only ASR when stream weights (more later!) are appropriately set.

	Keyword Error Rates (%) on CHiME 2 Corpus						
SNR	-6dB	-3dB	0dB	3dB	6dB	9dB	avg.
Video	27.8	27.8	27.8	27.8	27.8	27.8	27.8
Audio	27.9	23.0	18.1	15.4	12.8	10.4	17.9
Audiovisual CHMM	17.2	14.1	12.0	10.1	9.0	7.7	11.7

[Zeiler2016] S. Zeiler, R. Nickel, N. Ma, G. J. Brown, D. Kolossa: “Robust audiovisual speech recognition using noise-adaptive linear discriminant analysis,” Proc. ICASSP 2016, Shanghai, March 2016.

Coupled HMMs for asynchronous Audio- & Video Streams

Audiovisual Speech Recognition

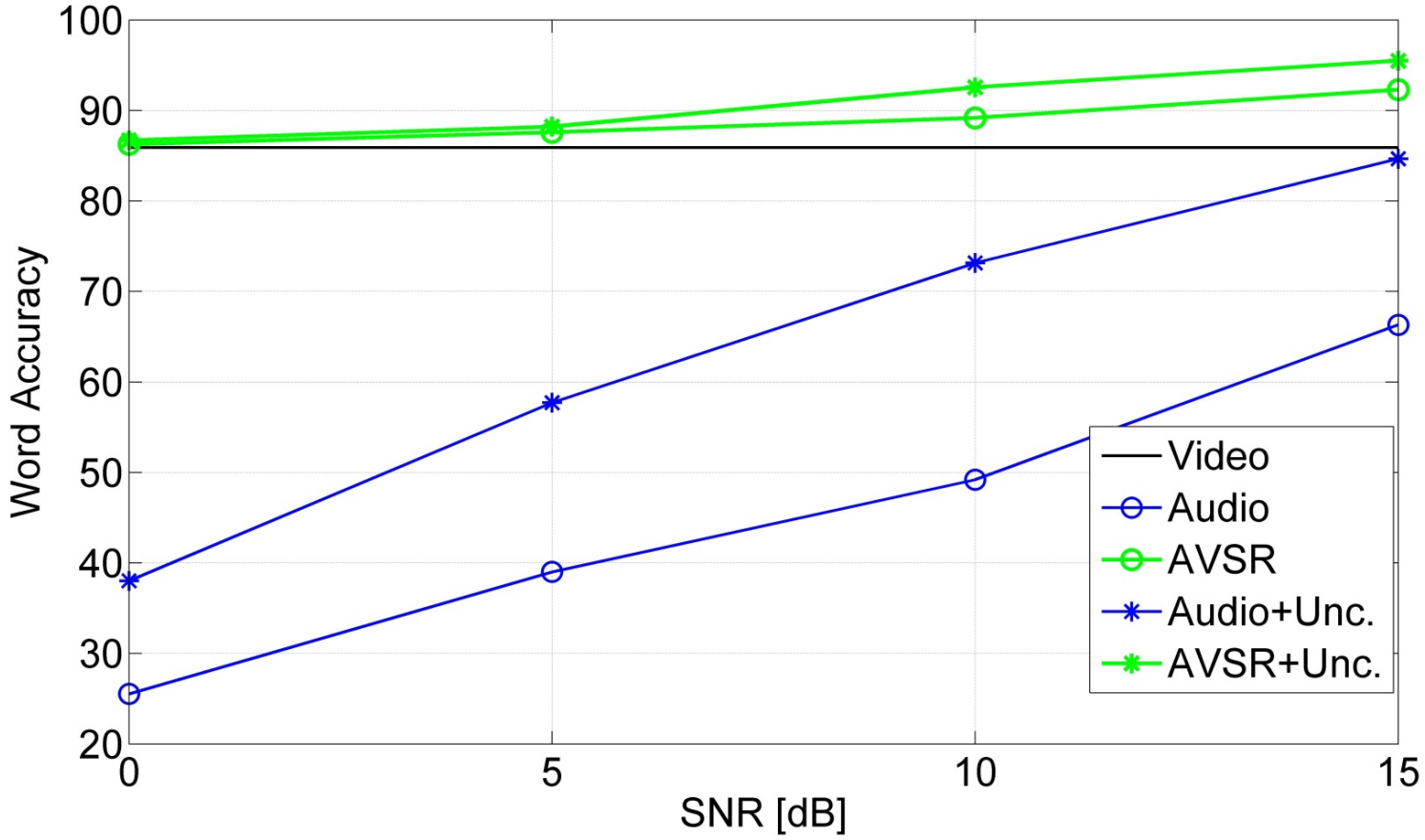
- Audiovisual recognition using coupled HMMs always outperforms audio-only and video-only ASR when stream weights (more later!) are appropriately set.
- Best results are achieved with noise-adaptive LDA + ground truth uncertainties

	Keyword Error Rates (%) on CHiME 2 Corpus						
SNR	-6dB	-3dB	0dB	3dB	6dB	9dB	avg.
Video	27.8	27.8	27.8	27.8	27.8	27.8	27.8
Audio	27.9	23.0	18.1	15.4	12.8	10.4	17.9
Audio + NALDA	17.3	13.2	11.5	9.3	7.8	7.6	11.1
Audiovisual CHMM	17.2	14.1	12.0	10.1	9.0	7.7	11.7
Audiovisual CHMM + NALDA	12.7	10.8	8.7	7.4	6.3	6.1	8.7

[Zeiler2016] S. Zeiler, R. Nickel, N. Ma, G. J. Brown, D. Kolossa: “Robust audiovisual speech recognition using noise-adaptive linear discriminant analysis,” Proc. ICASSP 2016, Shanghai, March 2016.

Coupled HMMs for asynchronous Audio- & Video Streams

Audiovisual speech recognition always outperforms audio-only and video-only ASR when stream weights are appropriately set, even under complete mismatch, here, training on clean & testing on noisy data.



Stream weighting

Can't live with it, can't seem to live without it...

Stream Weighting for Audiovisual Speech Recognition

Emission probabilities of coupled HMM:

$$p(\mathbf{o} \mid q) = b_a(o_a \mid q_a)^{\lambda} \cdot b_v(o_v \mid q_v)^{1-\lambda}$$

The $b_{a/v}(o_{a/v} \mid q_{a/v})^{\lambda}$ are observation likelihoods, λ the ***stream weight***.

Stream weighting not only applicable in coupled model but in all early and intermediate integration schemes including deep neural network-based ones.

Question: Is this really necessary?

Most recently, e.g. [Ninomiya2015, Ngiam2011,Tamura2015, Noda2015, Meutzner2017]

Question 2: If yes, how?

Stream Weighting for Audiovisual Speech Recognition

Idea of dynamic stream weighting system

Train neural network or logistic regression function to map some reliability features onto stream weights, using optimal dynamic stream weights as training targets.

During test time, this trained regression model or DNN will then map reliability measures (frame by frame) onto frame-wise stream weights

Reliability measure features

- Estimated observation uncertainties
- Estimated SNR
- Soft and hard VAD cues based on IMCRA noise estimation
- Dispersion and entropy of audio and video HMM

[Abdelaziz2015] A. Hussen Abdelaziz, S. Zeiler and D. Kolossa: "Learning Dynamic Stream Weights For Coupled-HMM-based Audiovisual Speech Recognition", IEEE Trans. Audio Speech and Language Processing, 2015.

Coupled HMMs for asynchronous Audio- & Videostreams

Results of dynamic stream weighting, comparing three strategies

- Equal weights, $\lambda = 0.5$ ("Bayes Fusion")
- Exponential Function [Estellers 2012]
- MLP: Dynamic stream weight estimation using multiple reliability features

Noise Type	SNR [dB]	Audio only	Video only	Audio-visual
				Bayes Fusion
Babble	15	0.8516	0.8476	0.9401
	10	0.6853		0.8840
	5	0.4675		0.7523
	0	0.3065		0.6040
White	15	0.8399		0.9385
	10	0.6819		0.8854
	5	0.5133		0.8130
	0	0.3701		0.7296
Clean	-	0.9886		0.9856
Avg.	-	0.6339		0.8369

[Abdelaziz2015] A. Hussen Abdelaziz, S. Zeiler and D. Kolossa: "Learning Dynamic Stream Weights For Coupled-HMM-based Audiovisual Speech Recognition", IEEE Trans. Audio Speech and Language Processing, 2015

Stream Weighting in Deep Neural Networks

Stream Weighting in Deep Neural Networks

Fundamental question:

Shouldn't we just train one large neural network?

Two considered alternatives

***1) Concatenation* of uncertainties**

Train one large network with uncertainties as an additional input.

2) Explicit *stream weighting*

Train two networks and fuse their posterior probabilities according to

$$\log \mathrm{p}(\mathbf{o}^{AV} \mid q) = \gamma \log(\mathrm{b}^{A}(o^{A} \mid \mathrm{q})) + (1-\gamma) \log(b^{V}(o^{V} \mid \mathrm{q}))$$

Stream Weighting in Deep Neural Networks

Evaluation:

Again, on the CHiME 2 data, as above.

Kaldi recipe based on Wall-Street-Journal training scripts*, using

1) Concatenation of uncertainties
2) Explicit stream weighting

https://github.com/hmeutzner/kaldi-avsr

**Hybrid system, so the DNN estimates state posteriors. Trained starting by GMM/HMM training, including LDA, fMLLR & speaker-adaptive training and continuing onto DNN/HMM. For this purpose, we use a topology with 11 frames of context, for 440d input, 6 hidden layers with 2048 neurons each, 1453 neurons in softmax output layer. RBM layer-wise pre-training is followed by minimum-cross-entropy training, followed by minimum Bayes risk fine-tuning.*

Stream Weighting in Deep Neural Networks

Concatenation of uncertainties

Features	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
Filter-bank|DCT	13.01	11.90	11.14	9.95	8.33	9.10	10.57
Filter-bank|DCT|Uncertainty	13.86	12.59	10.80	8.93	**7.40**	7.23	10.14
MFCC|DCT	17.94	16.67	15.73	14.63	13.27	12.33	15.09
MFCC|DCT|Uncertainty	14.71	13.18	11.39	10.03	9.18	8.33	11.14
Rate-map|DCT	**11.99**	**11.48**	**10.29**	8.08	7.91	8.08	**9.64**
Rate-map|DCT|Uncertainty	14.29	12.16	10.63	**7.65**	7.65	**6.97**	9.89

may or may not help

Explicit stream weighting

Features	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
Rate-map	28.36	23.45	15.17	11.55	7.93	6.72	15.53
DCT	27.50	27.50	27.50	27.50	27.50	27.50	27.50
Rate-map|DCT	**13.79**	**12.76**	**10.00**	**8.88**	8.88	8.79	**10.52**
Rate-map|DCT|Uncertainty	14.91	13.88	11.21	**8.88**	**8.79**	**7.93**	10.93
λ set per sentence, oracle-SNR-based	27.67	18.10	11.12	6.64	5.86	6.72	12.69
λ set per frame, uncertainty-based	**13.28**	**11.12**	**8.10**	**6.81**	**5.60**	**4.22**	**8.19**

does help

Stream Weighting in Deep Neural Networks

Intermediate Conclusion

With appropriate stream weighting, audiovisual recognition can reliably give accuracies that are equal to or better than the single best modality.

Stream weighting can be guided by reliability measures composed of recognition confidence measures and observation uncertainties. The composition is better than the single best measure.

Such stream weighting also appears to be helpful in the fusion of audiovisual multi-stream DNNs.

Next question

How can such audiovisual recognition systems benefit speech enhancement (e.g. for extremely noisy environments)?

...and moving on to the second part:

Audiovisual speech enhancement

...and many thanks for your attention!

References

[Abdelaziz2015] A. Hussen Abdelaziz, S. Zeiler and D. Kolossa: “Learning Dynamic Stream Weights for Coupled-HMM-based Audio-visual Speech Recognition”, IEEE Trans. Audio Speech and Language Processing, 2015.

[Estellers 2012] Estellers, M. Gurban, and J.-P. Thiran, “On dynamic stream weighting for audio-visual speech recognition,” IEEE Trans. Audio, Speech, and Language Processing, vol. 20, no. 4, pp. 1145–1157, 2012.

[Fiscus1997] J. Fiscus: “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER),” IEEE Workshop on Automatic Speech Recognition and Understanding, 1997, pp. 347 –354.

[Jordan1999] Michael Jordan (ed.): “Learning in Graphical Models,” MIT Press, Cambridge, MA.

[Kolossa2009] D. Kolossa, S. Zeiler, A, Vorwerk, R. Orglmeister: ”Audiovisual Speech Recognition with Missing or Unreliable Data”, Proc. AVSP, 2009

[Luettin2001] J. Luettin, G. Potamianos and C. Neti: "Asynchronous Stream Modelling for Large Vocabulary Audio-Visual Speech Recognition", Proc. ICASSP, pp. 169-172, May 2001.

[McGurk1976] H. Mc Gurk and J. MacDonald: “Hearing lips and seeing voices,” Nature 264, pp. 746-748, 1976.

[Meutzner2017] H. Meutzner, N. Ma, R. Nickel, C. Schymura, and D. Kolossa, “Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates,” in Proc. ICASSP, 2017.

[Nefian2002a] A.V. Nefian, L. Liang, X. Pi, X. Liu, K. Murphy: “Dynamic Bayesian Networks for Audio-Visual Speech Recognition,” EURASIP Journal on Applied Signal Processing, 2002.

[Nefian2002b] A.V. Nefian, L. Liang, X. Pi, X. Liu, K. Murphy: “A coupled HMM for Audio-Visual Speech Recognition,” Proc. ICASSP, 2002, pp. 2013--2016.

References

[Ngiam2011] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning.” in ICML 2011, pp. 689–696.

[Ninomiya2015] H. Ninomiya, N. Kitaoka, S. Tamura, Y. Iribe, and K. Takeda, “Integration of deep bottleneck features for audio-visual speech recognition,” in Proc. Interspeech, 2015.

[Noda2015] K. Noda, Y. Yamaguchi, K. Nakadai, H. Okuno, and T. Ogata, “Audio-visual speech recognition using deep learning,” Appl. Intell., vol. 42, pp. 722–737, 2015.

[Receveur2016] S. Receveur, R. Weib, T. Fingscheidt: “Turbo Automatic Speech Recognition, IEEE/ACM Trans. Audio, Speech & Language Processing, vol. 24, no. 5 2016, pp. 846-862.

[Sumby1954] Sumby, Pollack: Visual Contribution to Speech Intelligibility in Noise, JASA, 1954.

[Tamura2015] S. Tamura, H. Ninomiya, N. Kitaoka, S. Osuga, Y. Iribe, K. Takeda, and S. Hayamizu, “Audio-visual speech recognition using deep learning,” in Proc. APSIPA Annual Summit and Conference, 2015.

[Vorwerk2011] A. Vorwerk, S. Zeiler, D. Kolossa, R. Fernandez Astudillo and D. Lerch: “Use of Missing and Unreliable Data for Audiovisual Speech Recognition”, in: D. Kolossa, R. Haeb-Umbach (eds.): „Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications“, Springer Verlag, pp. 345-375, July 2011.

[Whittaker1990] Joe Whittaker: «Graphical models in applied multivariate statistics”, Wiley, 1990.

[Zeiler2016] S. Zeiler, R. Nickel, N. Ma, G. J. Brown, D. Kolossa: “Robust audiovisual speech recognition using noise-adaptive linear discriminant analysis,” Proc. ICASSP 2016, Shanghai, March 2016.

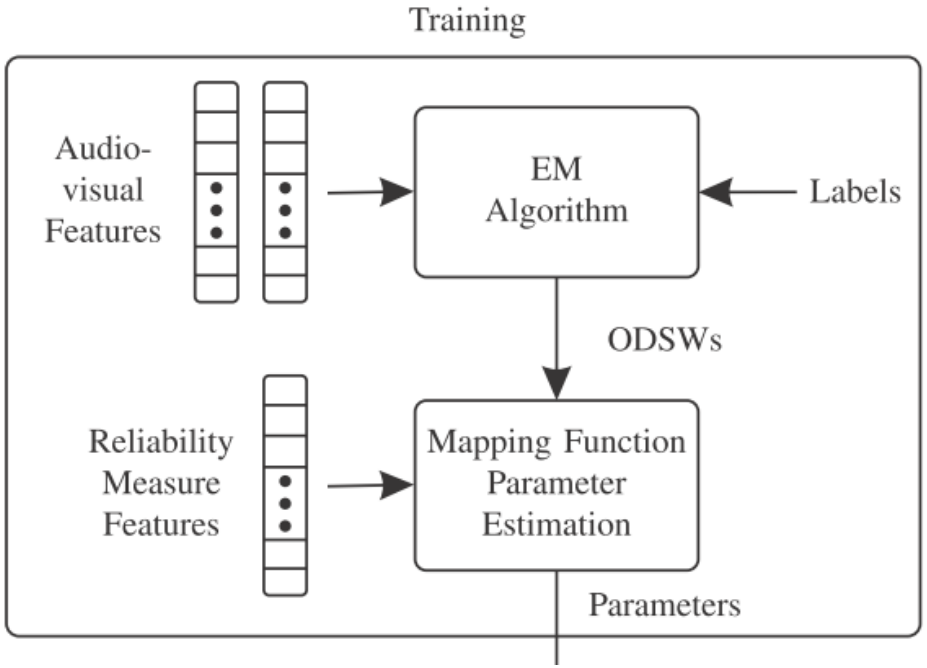
Coupled HMMs for asynchronous Audio- & Video Streams

Block diagram of dynamic stream weighting system

Train neural network or logistic regression function using oracle dynamic stream weights (ODSWs) as training targets.

Reliability measure features

- Estimated observation uncertainties
- Estimated SNR
- Soft and hard VAD cues based on IMCRA noise estimation
- Dispersion and entropy of audio and video HMM



[Abdelaziz2015] A. Hussen Abdelaziz, S. Zeiler and D. Kolossa: "Learning Dynamic Stream Weights For Coupled-HMM-based Audio-visual Speech Recognition", IEEE Trans. Audio Speech and Language Processing, 2015.

Introducing the Turbo-Twin-HMM for Audio-Visual Speech Enhancement

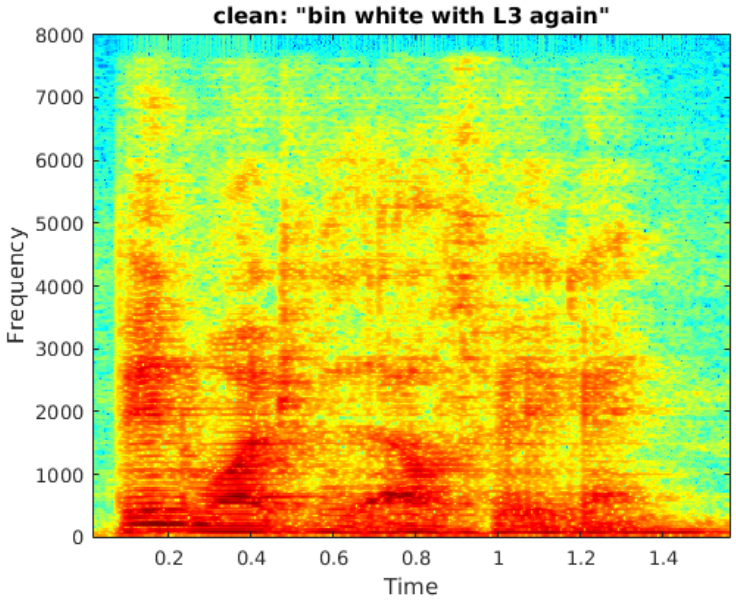
Steffen Zeiler, Hendrik Meutzner, Ahmed Hussen Abdelaziz, Dorothea Kolossa

June 28th 2017

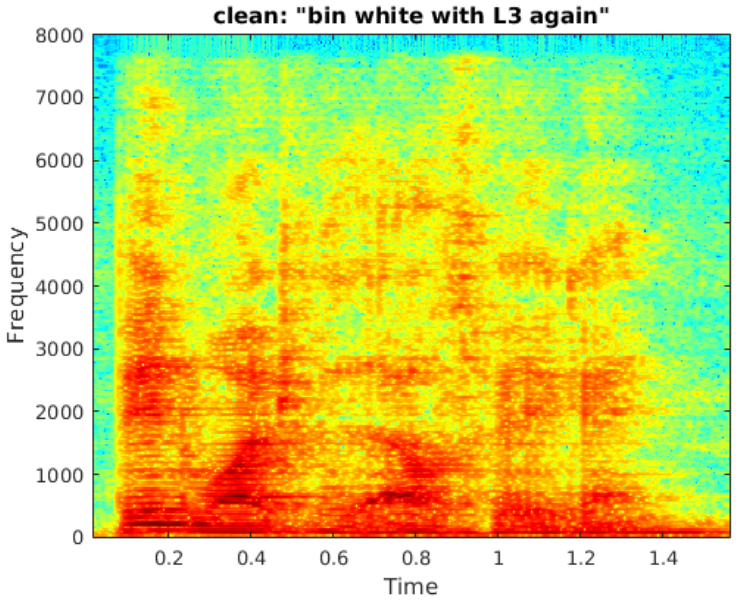
Cognitive Signal Processing Group
Institute of Communication Acoustics

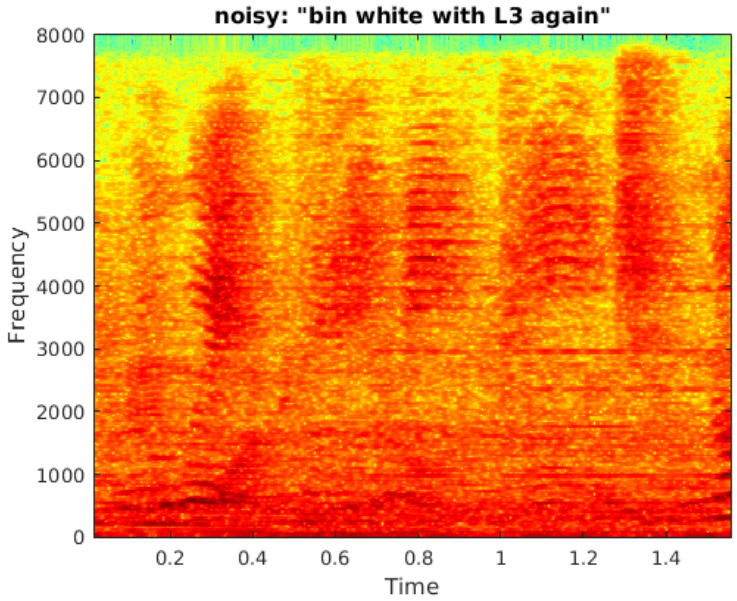
How can we recover speech from noise?

How can we recover speech from noise?

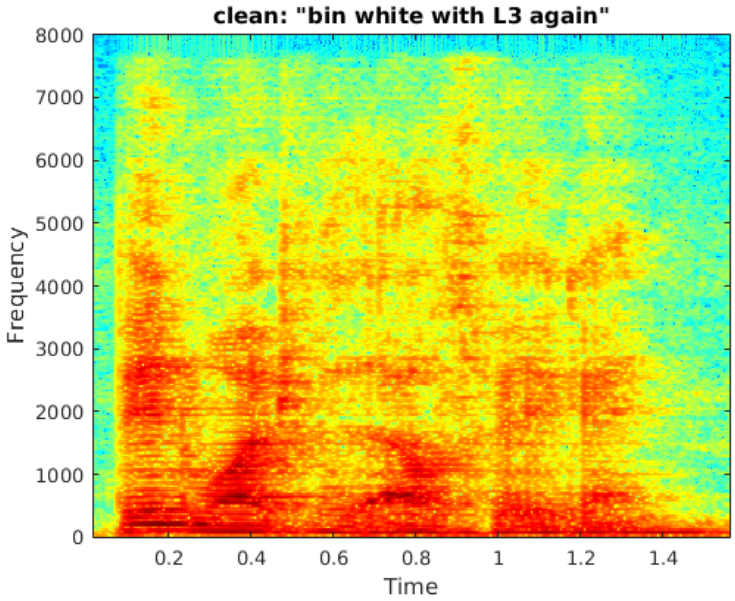


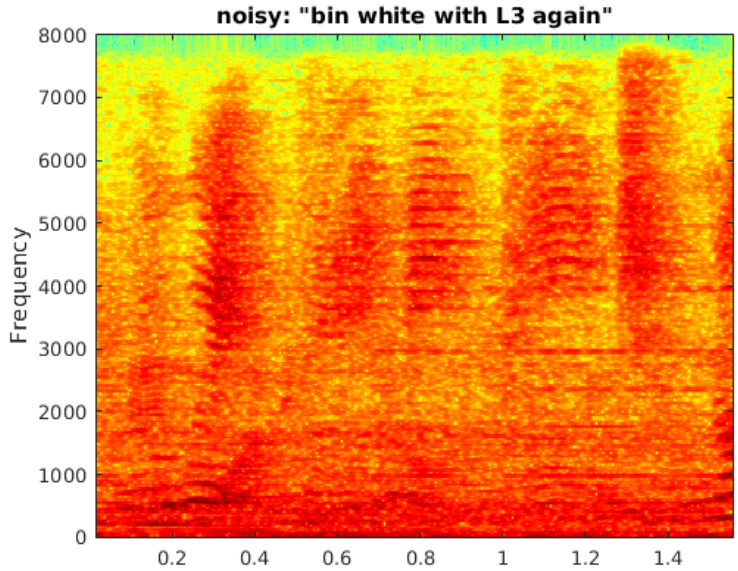
How can we recover speech from noise?

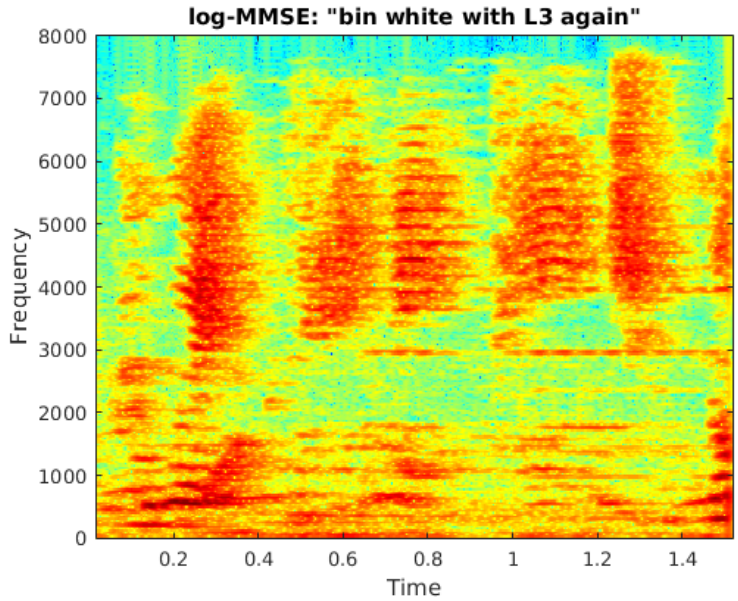




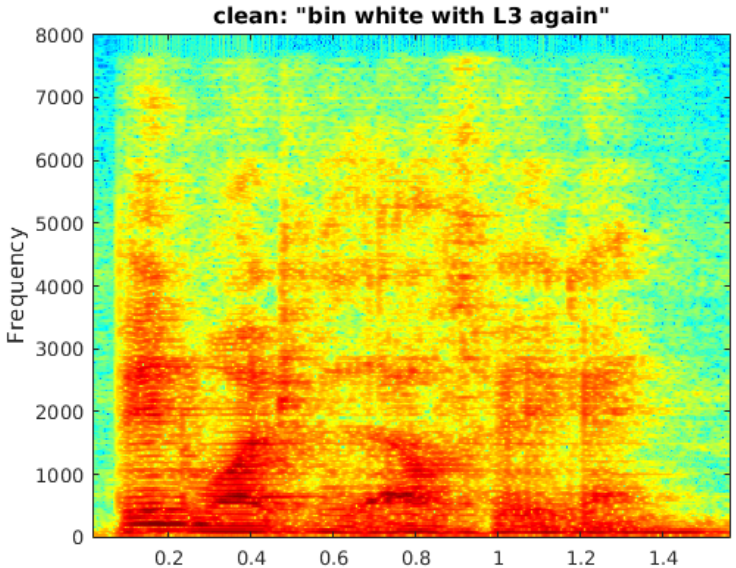
How can we recover speech from noise?

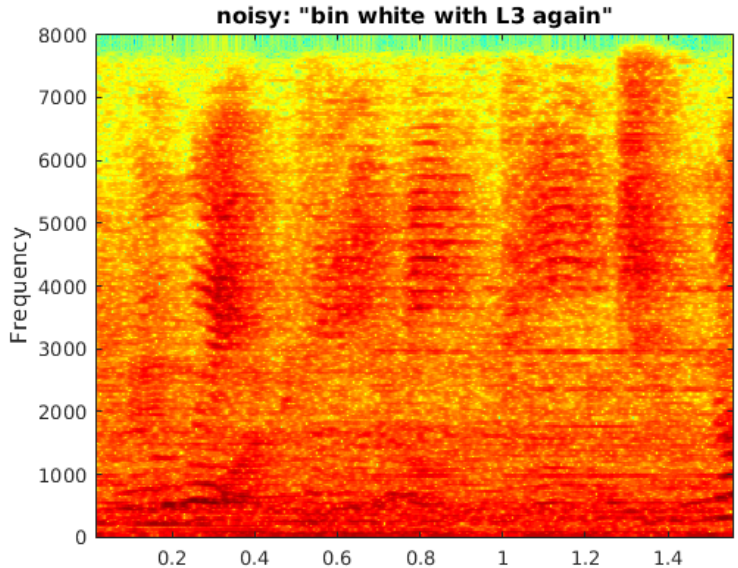


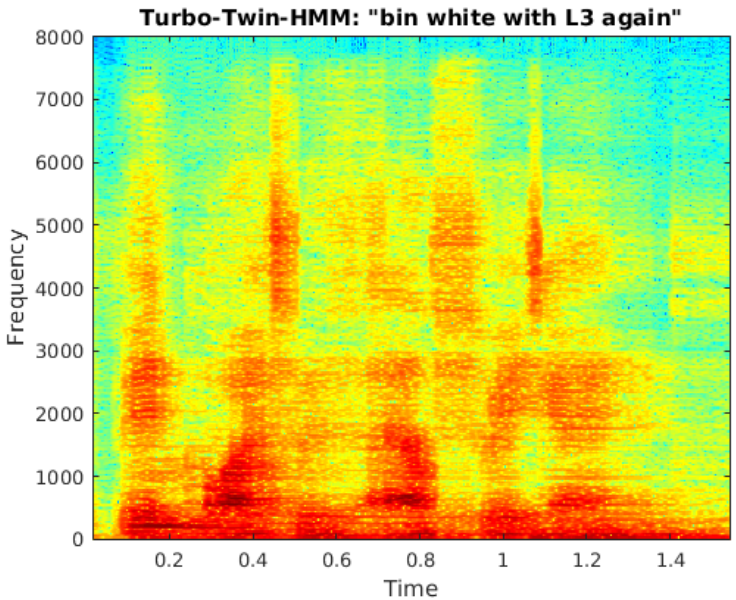


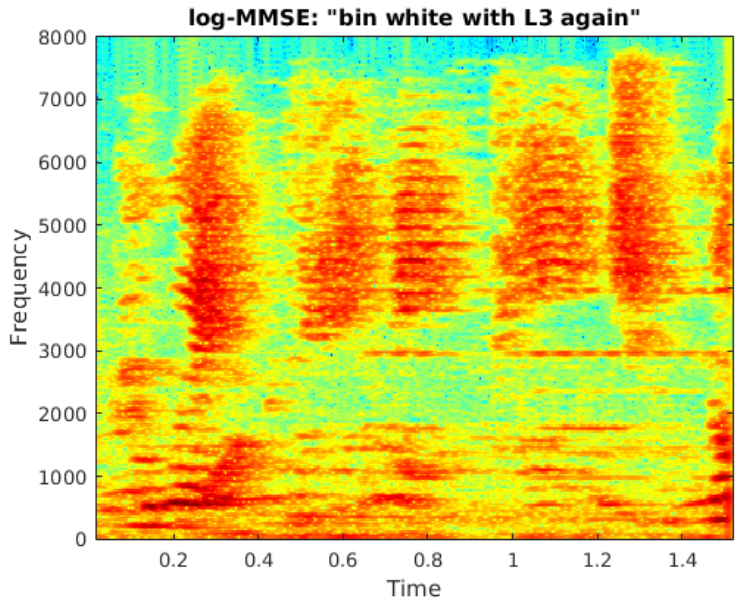


How can we recover speech from noise?

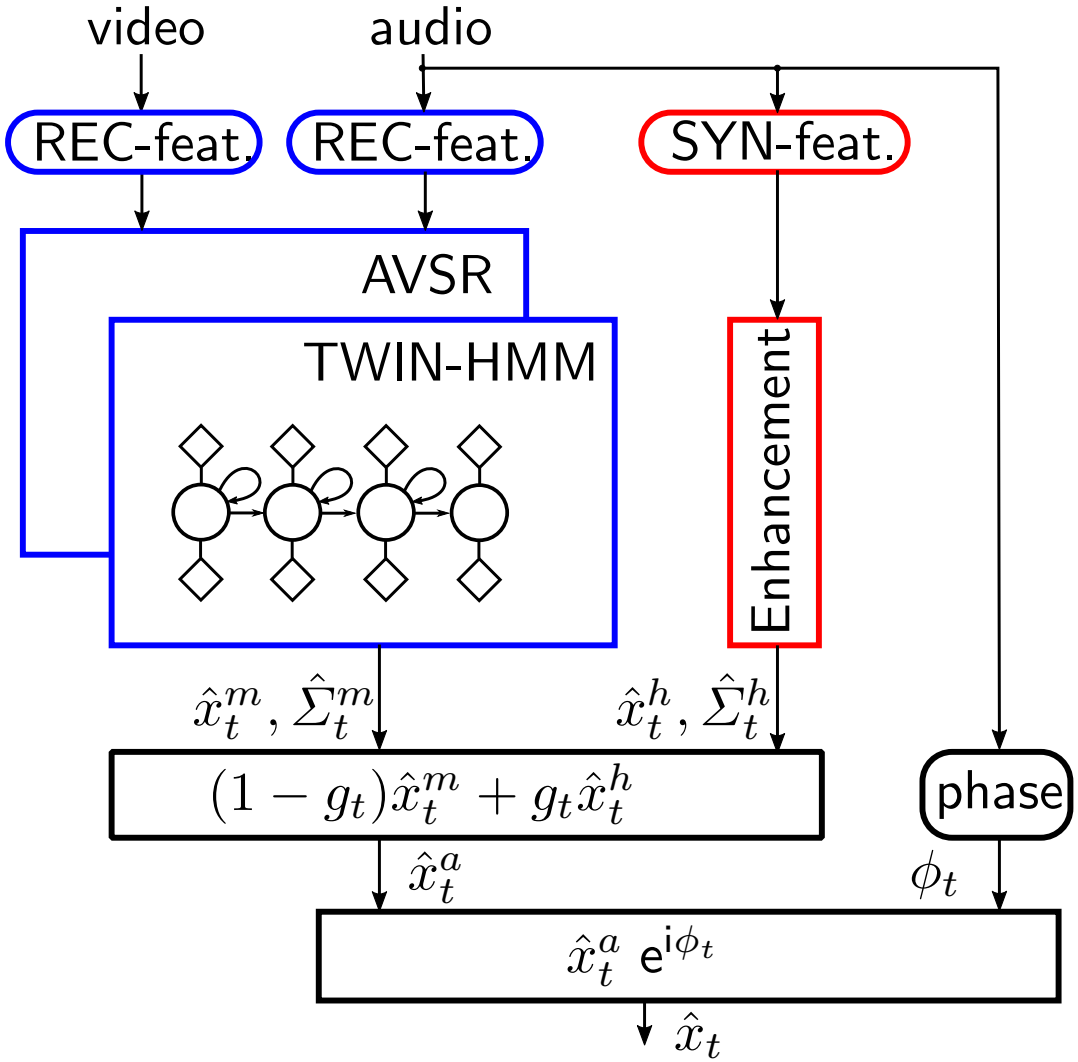




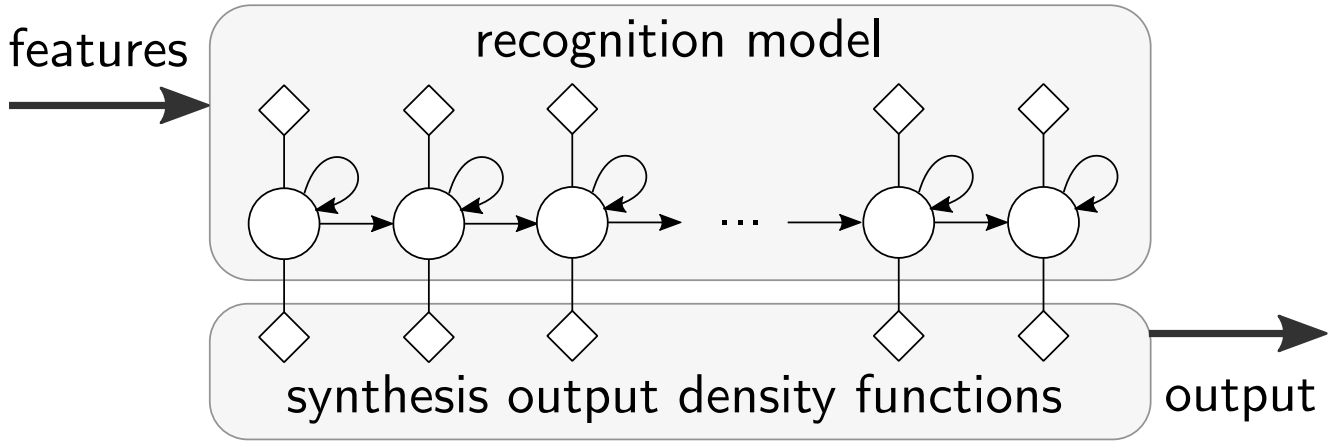




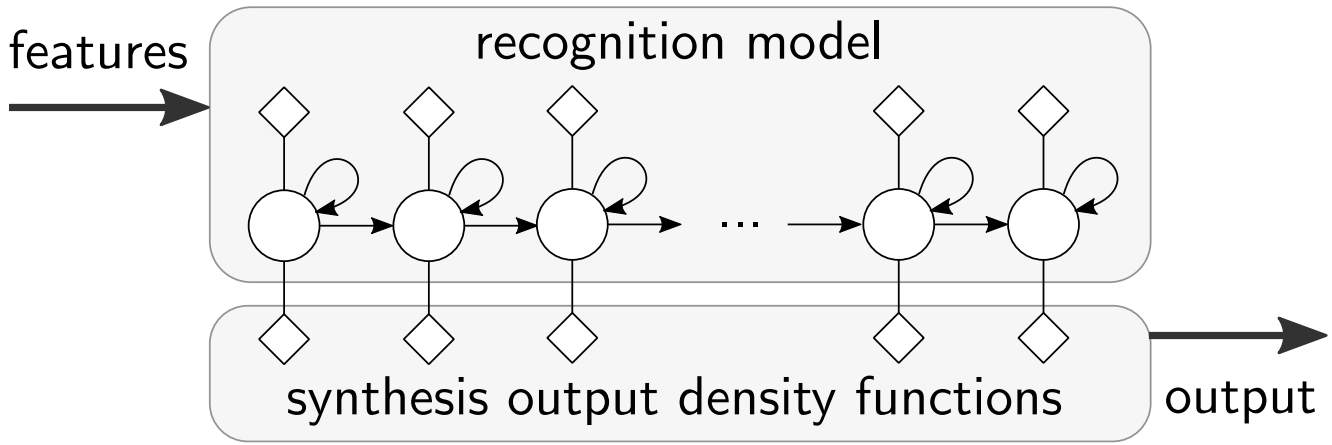
System Overview



The Twin-HMM


features
recognition model
...
synthesis output density functions
output

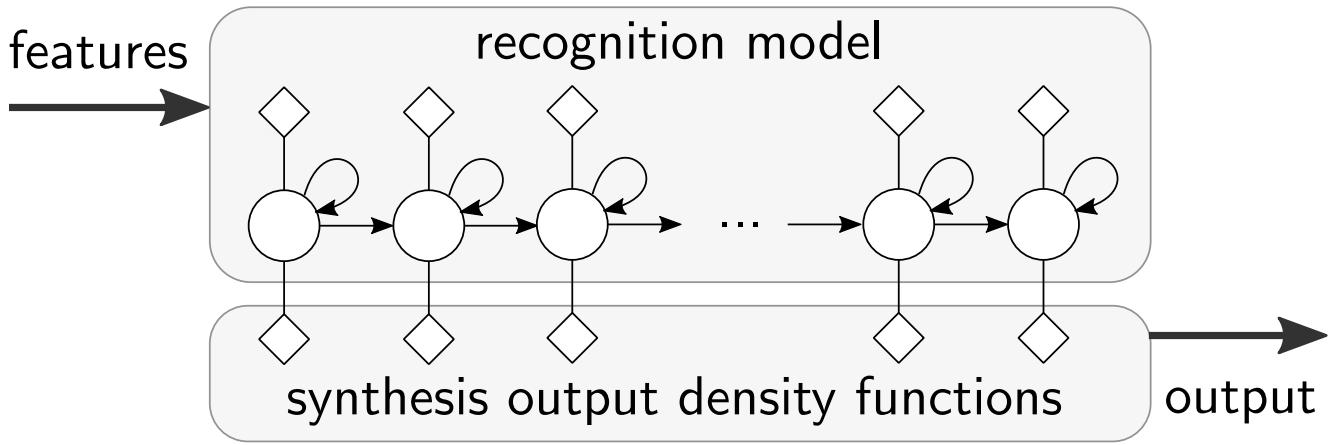
The Twin-HMM



AP: MMSE estimate of the clean speech amplitude spectrum

$$\hat{x}(t) = \mathsf{E}(x_t|o) = \sum_{i=1}^{N} p(q_t = i|o)\ \mathsf{E}(x_t|q_t = i)$$

The Twin-HMM



AP: MMSE estimate of the clean speech amplitude spectrum

$$\hat{x}(t) = \mathsf{E}(x_t|o) = \sum_{i=1}^{N} p(q_t = i|o)\ \mathsf{E}(x_t|q_t = i)$$

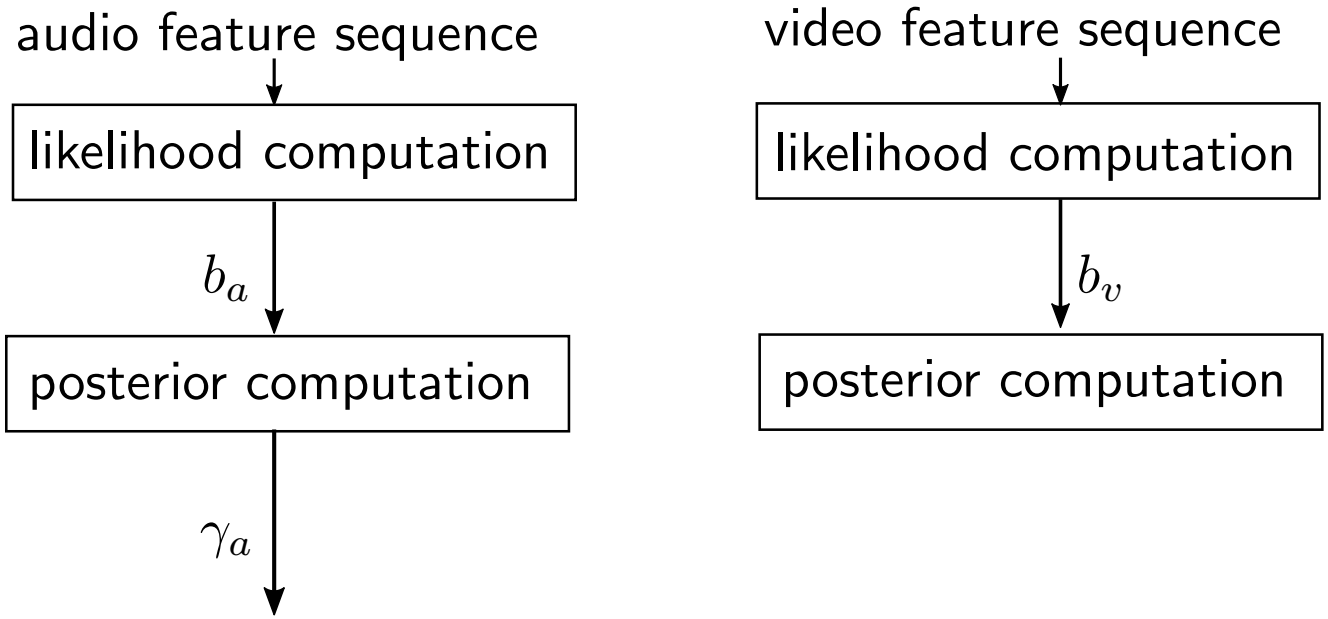
BP: use the most probable state i_t^* in each frame t

$$\hat{x}(t) = \mathsf{E}(x_t|q_t = i_t^*)$$

Turbo Decoding[1]

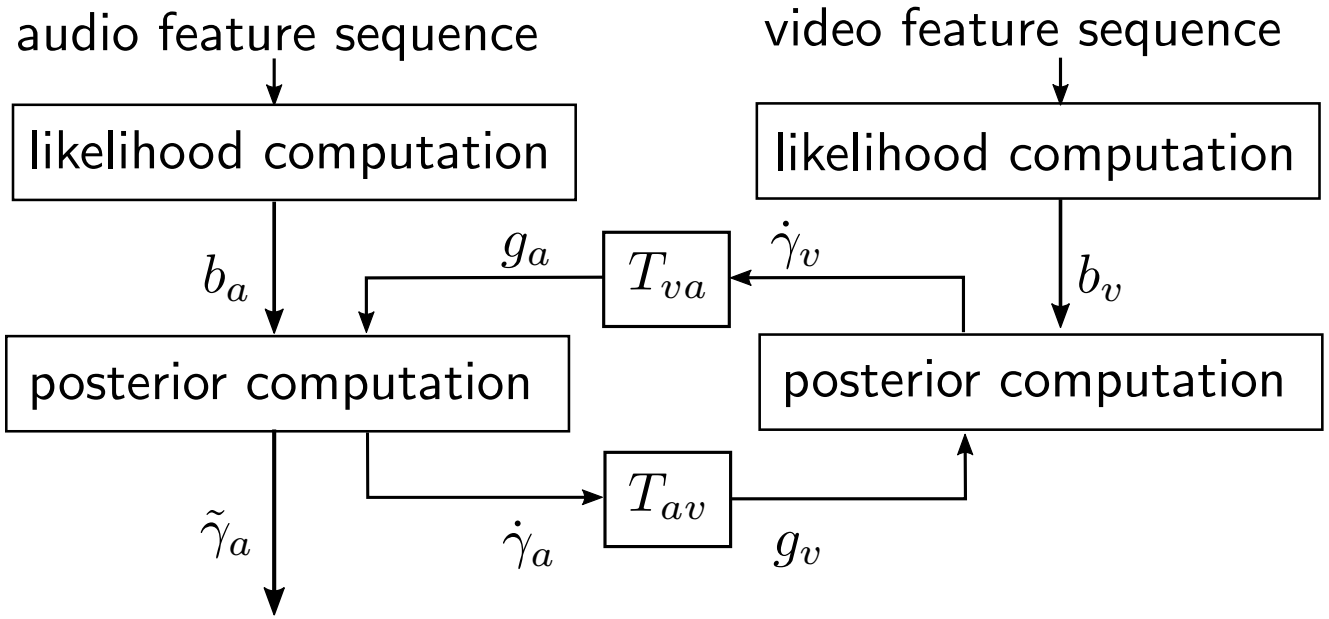
[1] Shivappa, Rao, Trivedi: *Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition*, ICASSP 2008

Turbo Decoding[1]



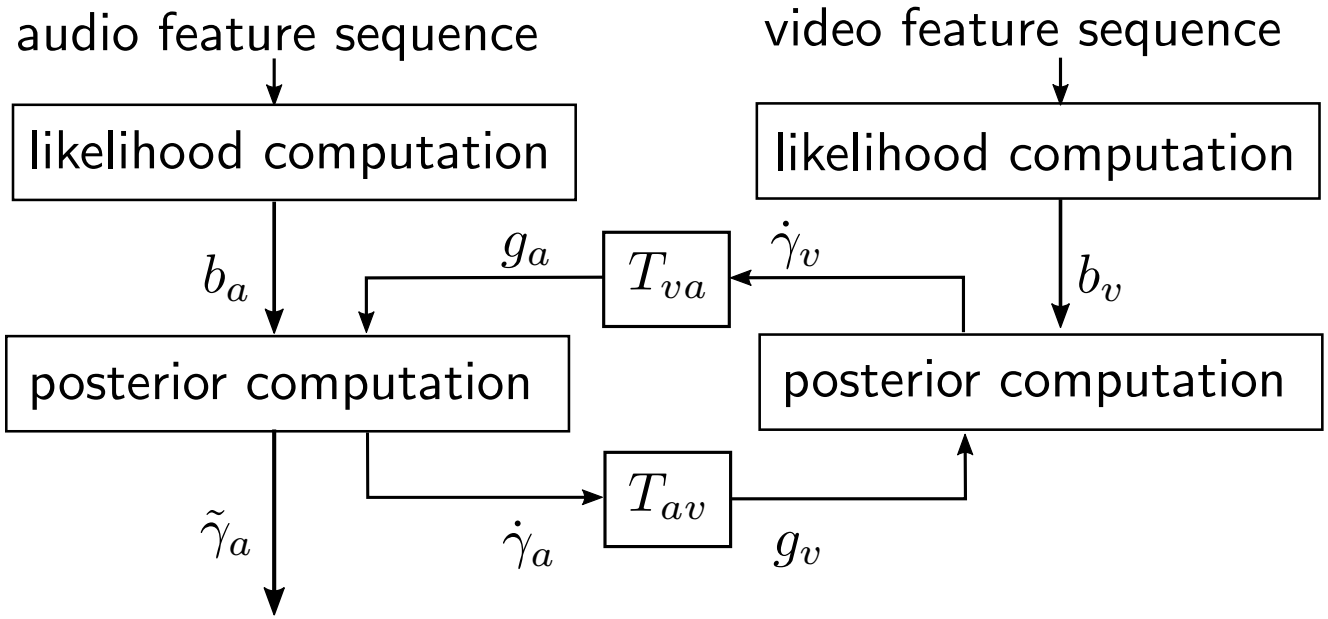
[1] Shivappa, Rao, Trivedi: *Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition*, ICASSP 2008

Turbo Decoding[1]



[1] Shivappa, Rao, Trivedi: *Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition*, ICASSP 2008

Turbo Decoding[1]



likelihood modification:

$$\tilde{b}_a(o_a|q_a) = b_a(o_a|q_a) \cdot g_a(q_a)^{\lambda_T \lambda_P},$$
$$\tilde{b}_v(o_v|q_v) = b_v(o_v|q_v) \cdot g_v(q_v)^{(1-\lambda_T)\lambda_P}$$

[1] Shivappa, Rao, Trivedi: *Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition*, ICASSP 2008

Instrumental Measures

	PESQ[2]				STOI[3]			
SNR	0 dB	-3 dB	-6 dB	-9 dB	0 dB	-3 dB	-6 dB	-9 dB
noisy	1.95	1.69	1.46	1.21	**0.70**	0.62	0.53	0.45
log-MMSE	1.90	1.58	1.36	1.06	0.66	0.57	0.49	0.41
E1AP	**2.11**	**2.02**	**1.94**	**1.83**	0.68	0.65	0.61	0.57
E1BP	2.02	1.92	1.82	1.71	0.66	0.63	0.59	0.54
E2AP	2.08	2.01	1.91	1.82	**0.70**	**0.68**	**0.64**	**0.59**
E2BP	1.99	1.91	1.80	1.68	0.67	0.65	0.60	0.56

588 files per SNR

recognizer features

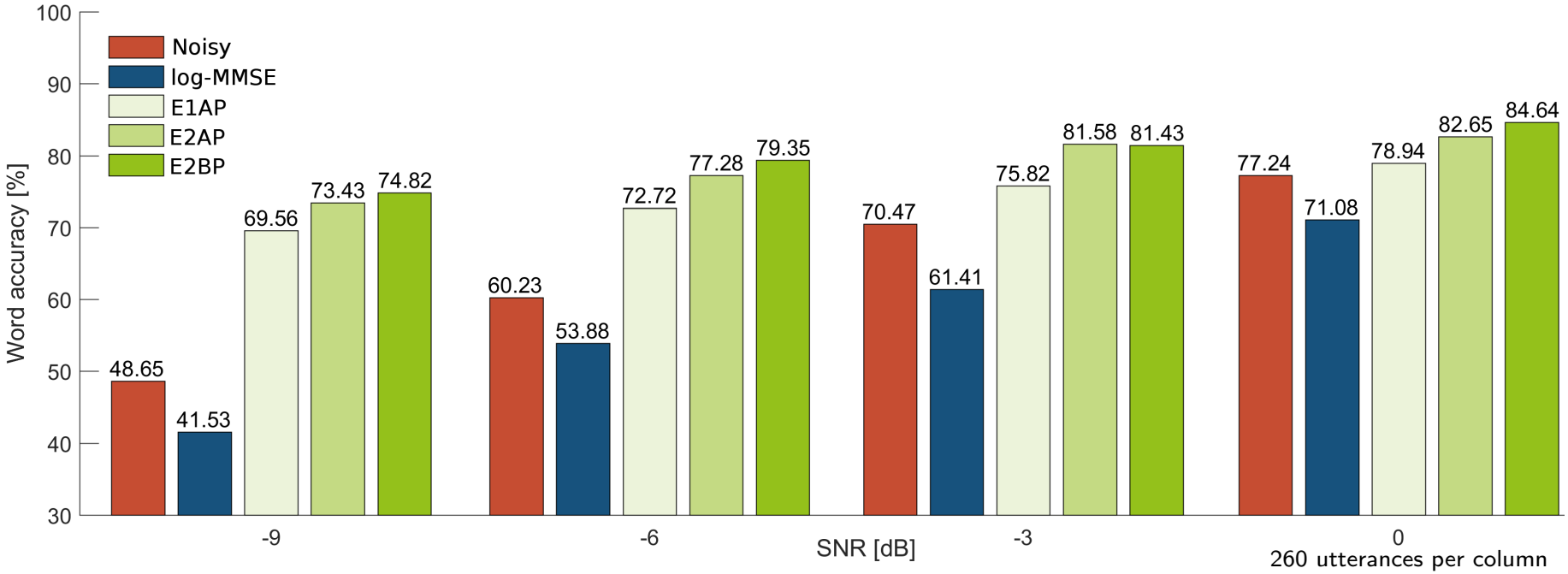
E1 minimize synthesis distortions
E2 optimize recognition results

clean speech estimation

AP all path synthesis
BP best path synthesis

[2]PESQ: Perceptual Evaluation of Speech Quality, [3]STOI: Short Time Objective Intelligibility

Listening Tests



- large-scale listening experiment (CrowdFlower)
- 690 individual participants, 27.118 transcribed utterances
- quality control to identify cheaters or language deficits

Conclusion

- video assisted single channel speech enhancement works
- our best system [E2BP] improves word accuracy of human listeners for the GRID task
  - from 48.6% to 74.8% at -9dB
  - from 77.2% to 84.6% at 0dB SNR
- predictions of reference-based objective speech intelligibility measures are *unreliable* for non-linearly processed speech

Perspectives

- better intelligibility estimators are needed - we are considering speech-recognition-based measures
- AV speech enhancement needs to be extended to open vocabularies and arbitrary recording conditions (taking video reliability information into account)
- for this purpose, and others, we are working on large-vocabulary AV speech recognition, combining our more general topologies with TensorFlow training of convolutive/recurrent nets

Cognitive Signal Processing Group
Institute of Communication Acoustics

Recognition Accuracy for variants E1 and E2

Method	-9 dB	-6 dB	-3 dB	0 dB	∞ dB
E1	87.95%	90.27%	91.13%	93.38%	97.15%
E2	89.67%	91.81%	93.98%	95.34%	98.18%

E1 : optimized for minimal distortion during synthesis

E2 : optimized for best recognition results

Listening Test Recognition Accuracy

Noisy	log-MMSE	*E1AP*	*E2AP*	*E2BP*
64.27 %	57.09 %	74.30 %	78.78 %	80.10 %

- average word accuracy over all SNRs
- Each score is based on 1037 utterances