NTT

NTT Communication Science Laboratories

Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation

M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S. Hahm, A. Nakamura

Motivation of our system

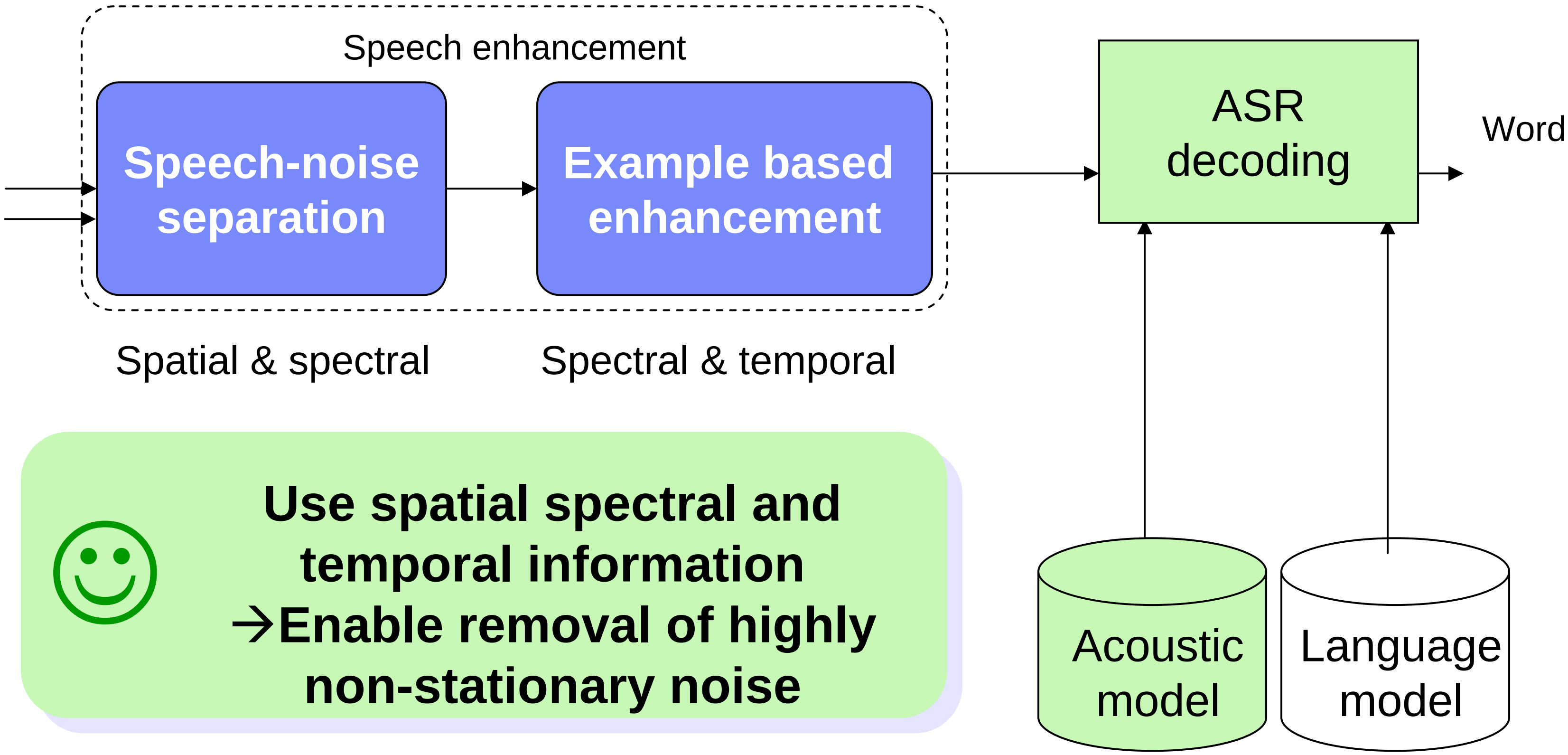
- **Speech enhancement**
  - Deal with highly non-stationary noise, using **all information** available about speech/noise
    Spatial - Spectral - Temporal
  - Realized using **two complementary** enhancement processes
- **Recognition**
  - **Interconnection** of speech enhancement and recognizer using dynamic acoustic model adaptation
  - Use of state of the art ASR technologies (**discriminative training**, **system combination**…)

Average accuracy improves 69 % → 91.7 %

Approaches for noise robust ASR

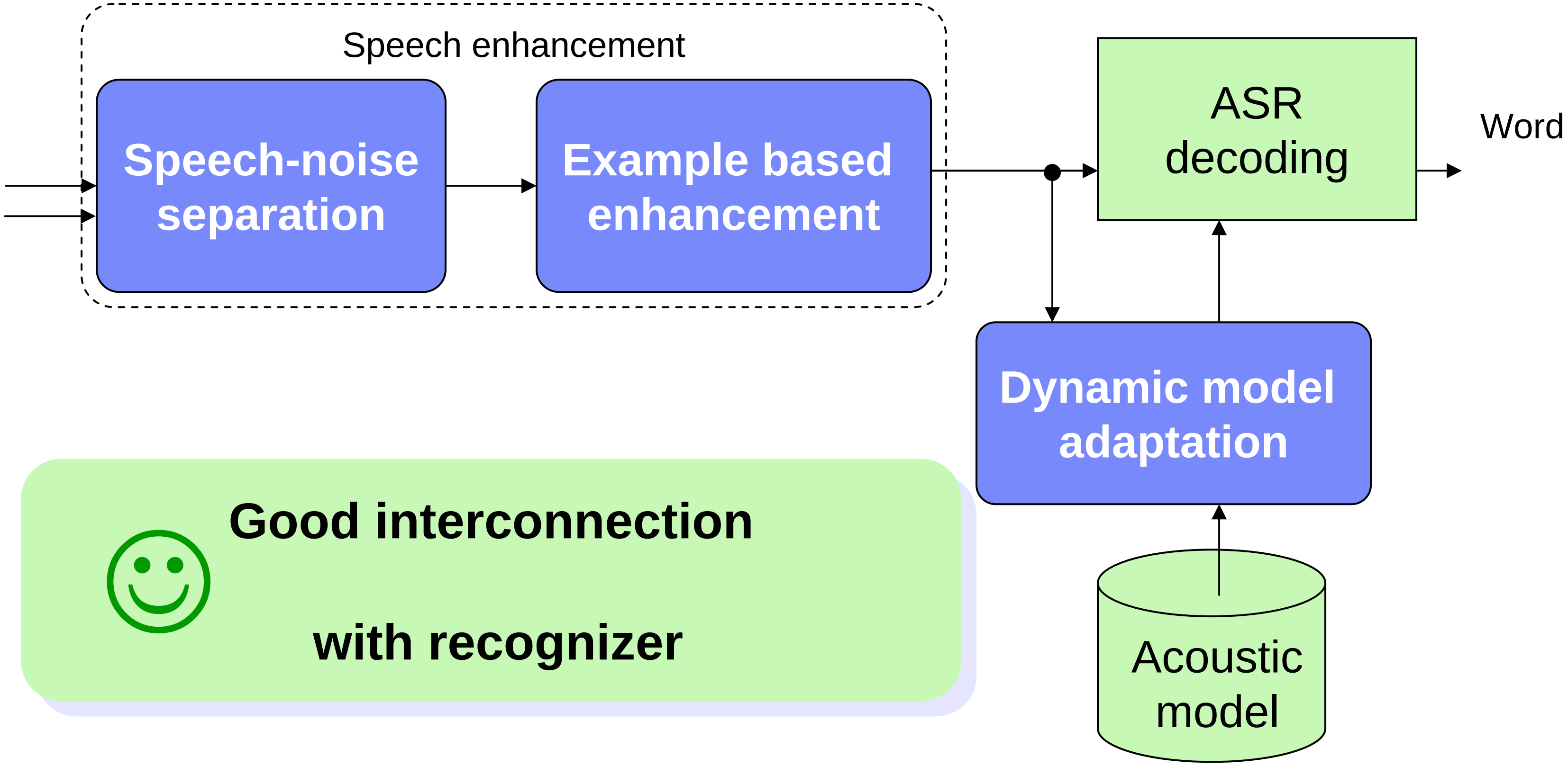
	Information used	Handling highly non-stationary noise	Interconnection w/ ASR
Acoustic model compensation, e.g. VTS	Spectral	☹	☺
Speech enhancement, e.g. BSS	Spatial/spectral/ temporal	😐	☹
Proposed			

System overview



Use spatial spectral and temporal information
→Enable removal of highly non-stationary noise

System overview

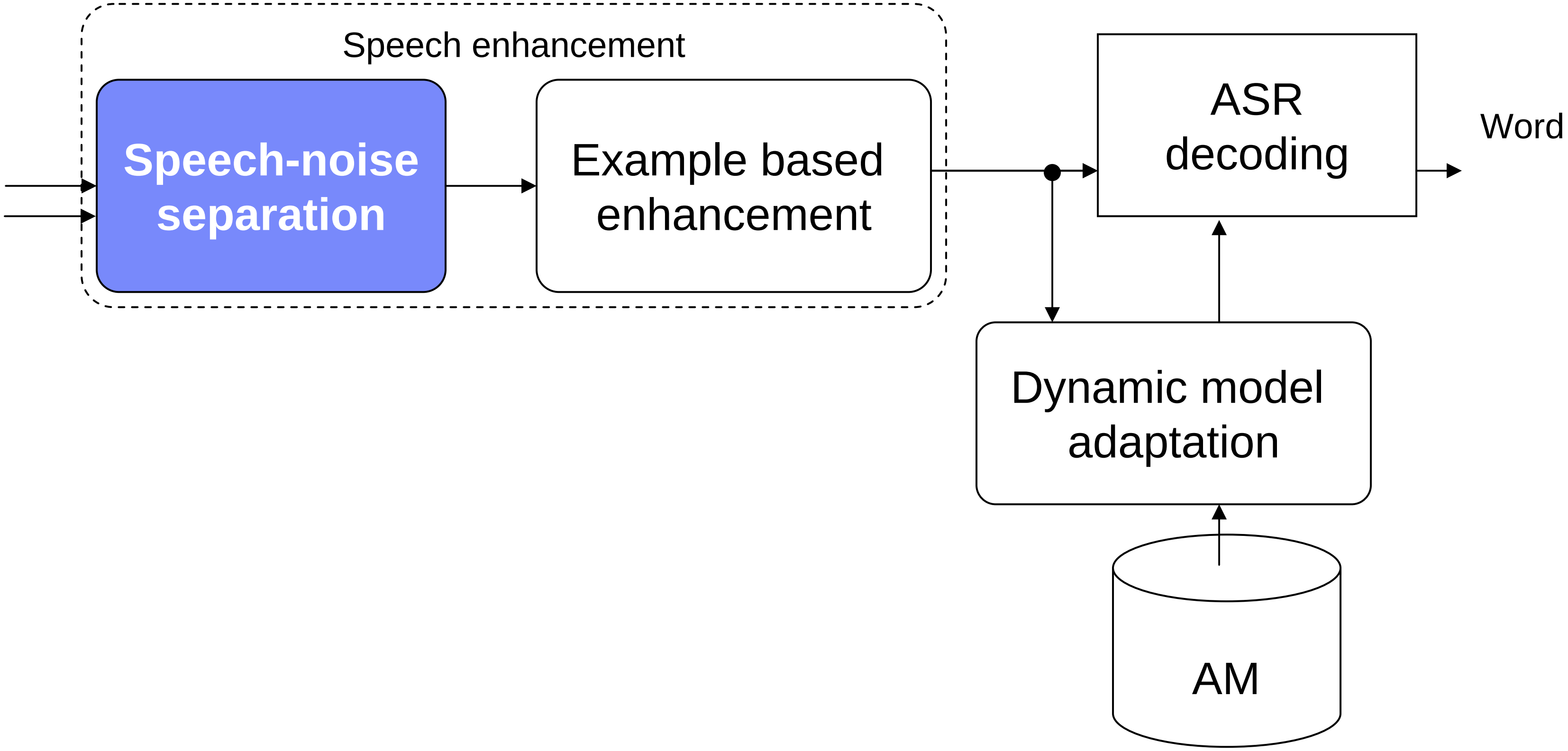


Good interconnection with recognizer

Approaches for noise robust ASR

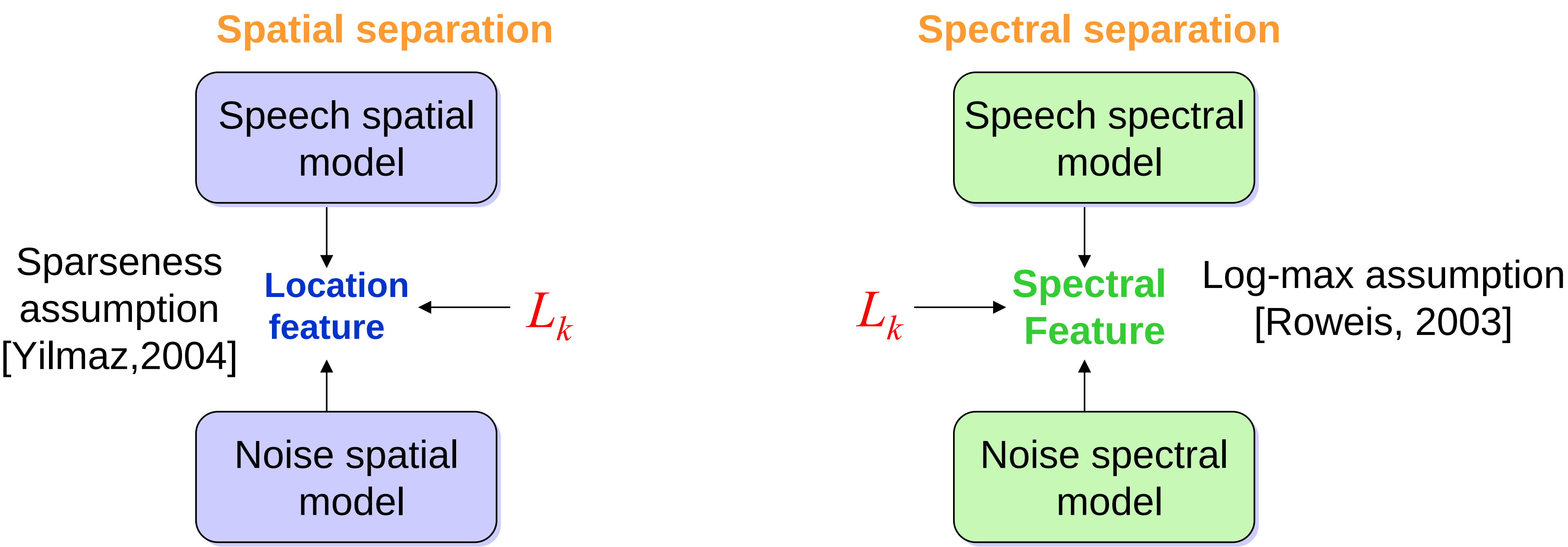
	Information used	Handling highly non-stationary noise	Interconnection w/ ASR
Acoustic model compensation, e.g. VTS	Spectral	☹	☺
Speech enhancement, e.g. BSS	Spatial/spectral/ temporal	😐	☹
Proposed	Spatial, spectral & temporal	☺	☺

System overview



Speech-noise separation [Nakatani, 2011]

- Integrate spatial-based and spectral-based separation in a single framework

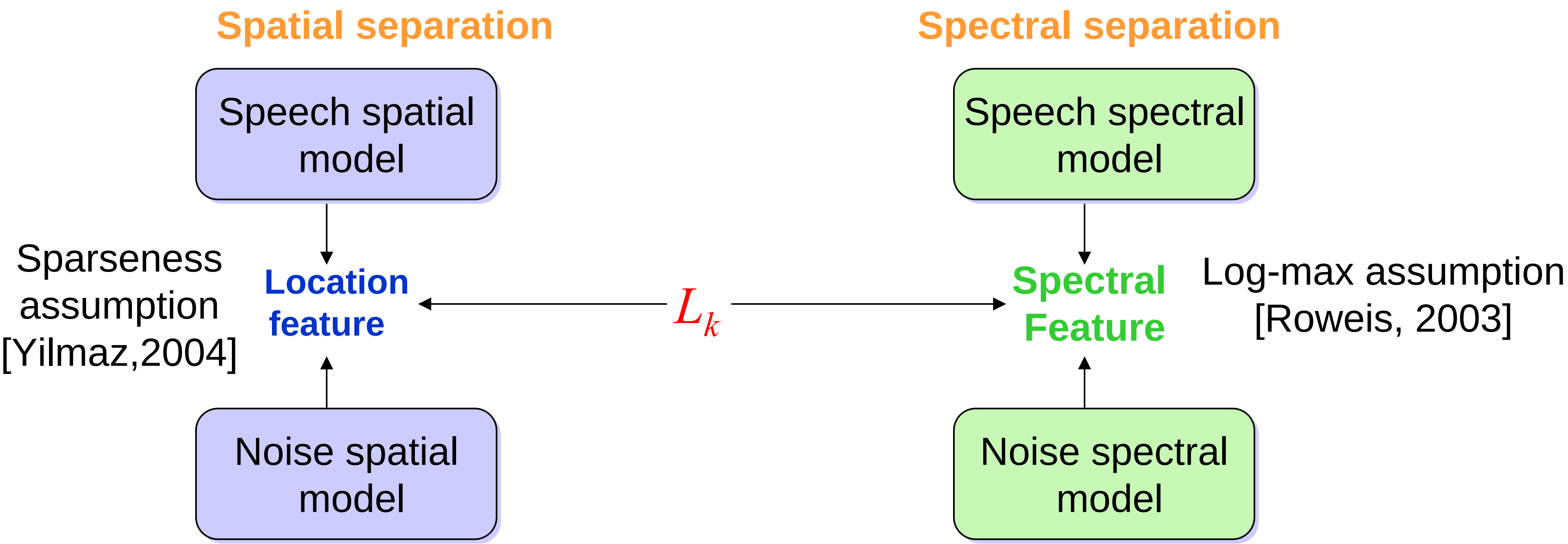


L_k **: dominant source index,**

i.e. indicates whether speech or noise is more dominant at each frequency k

Speech-noise separation [Nakatani, 2011]

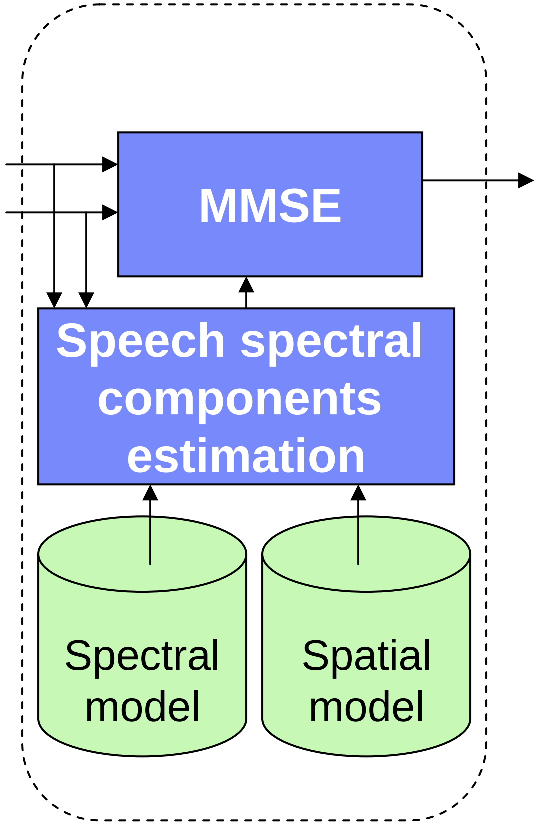
- Combined using dominant source index L_k



L_k **: dominant source index,**
i.e. indicates whether speech or noise is more dominant at each frequency k

Speech-noise separation [Nakatani, 2011]

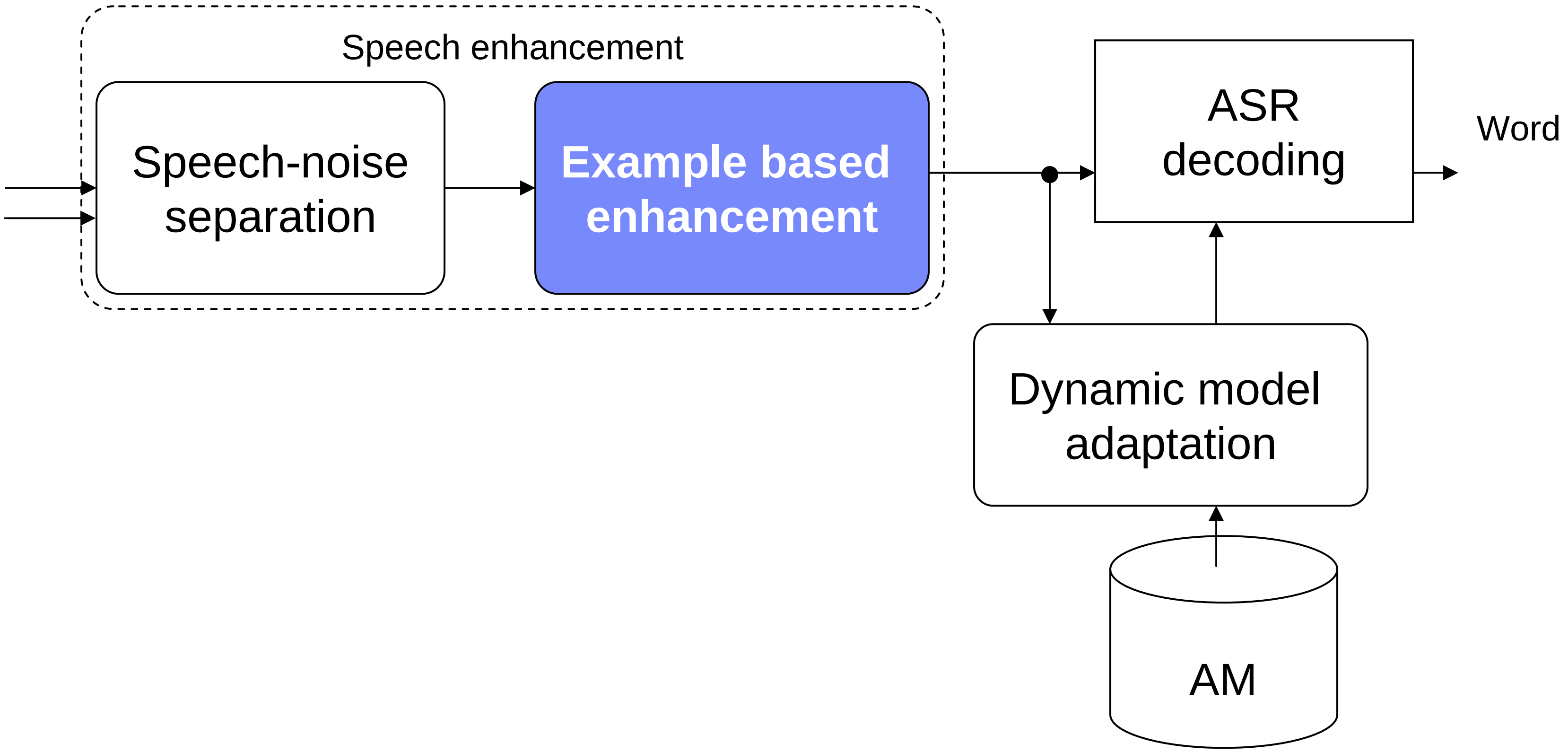
DOLPHIN *dominance based locational and power-spectral characteristics integration*



- Estimate speech spectral component sequence using EM algorithm
- Estimated speech obtained using MMSE

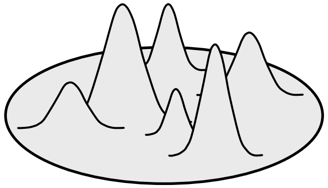
Integrate efficiently spatial and spectral information
to remove non-stationary noise

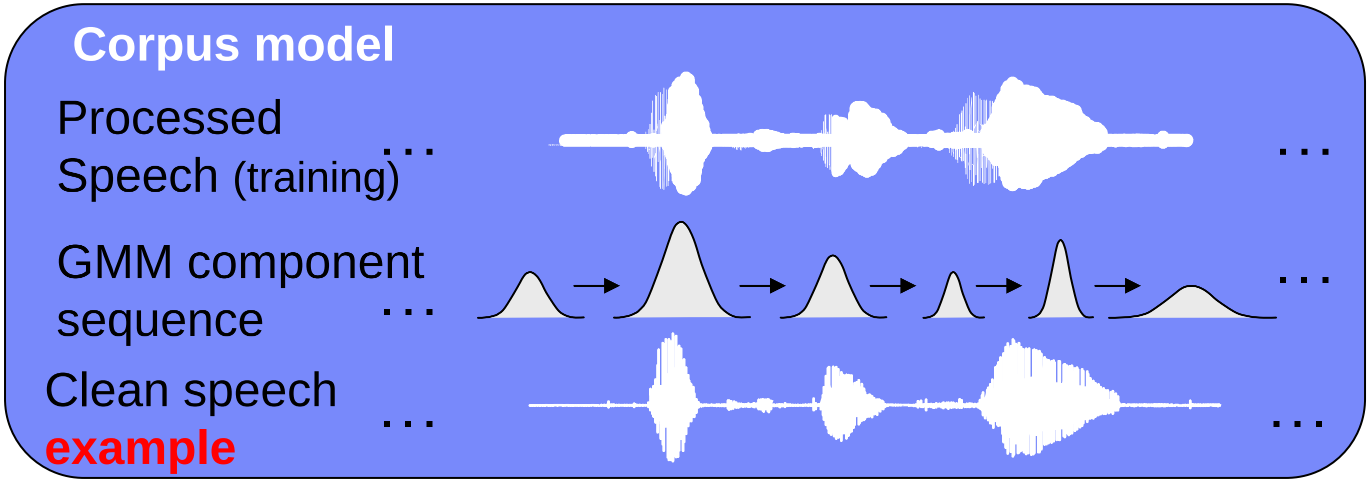
System overview



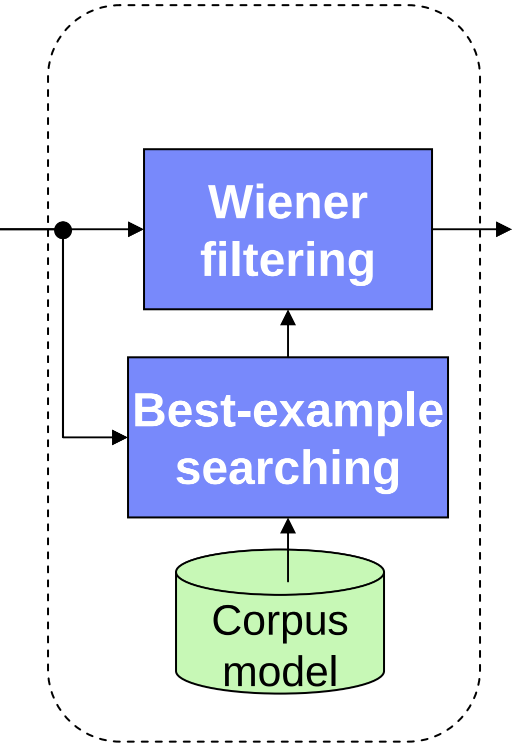
Example-based enhancement [Kinoshita,2011]

- **Use a parallel corpus model (clean and processed speech) that represents the fine spectral and temporal structure of speech**
  - Train a GMM from multi-condition training data processed with DOLPHIN
  - Generate corpus model

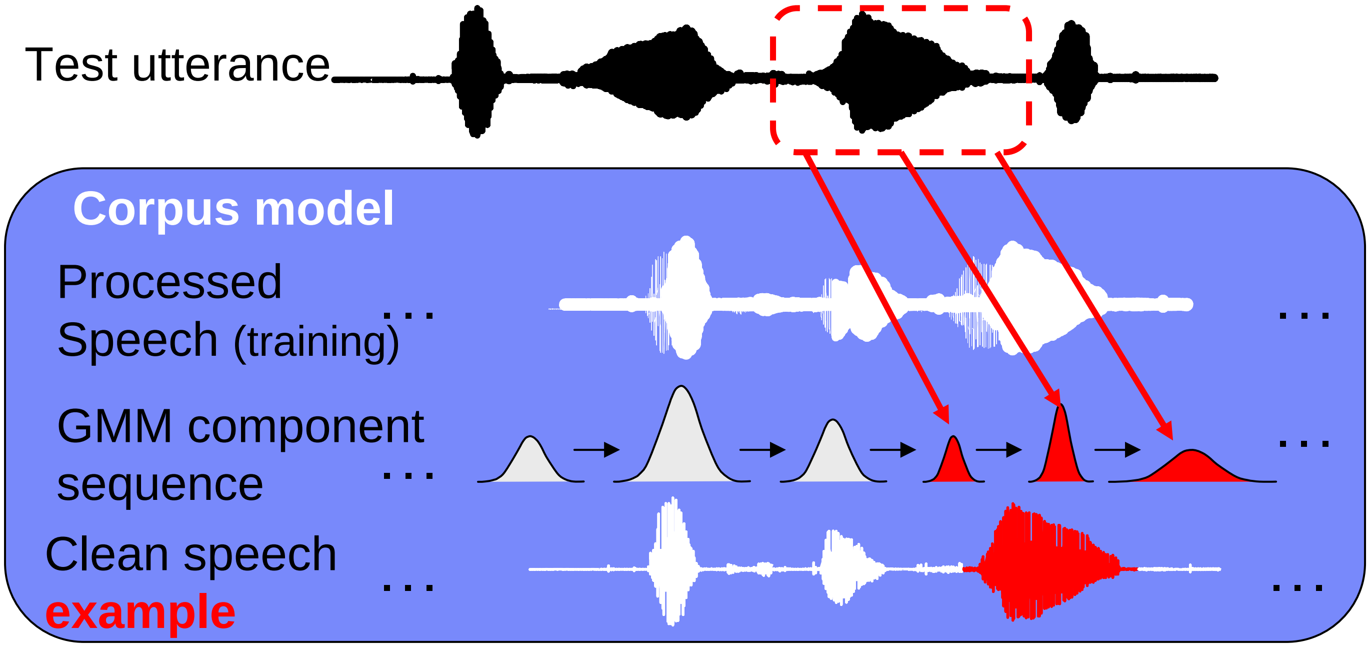




Example-based enhancement [Kinoshita,2011]



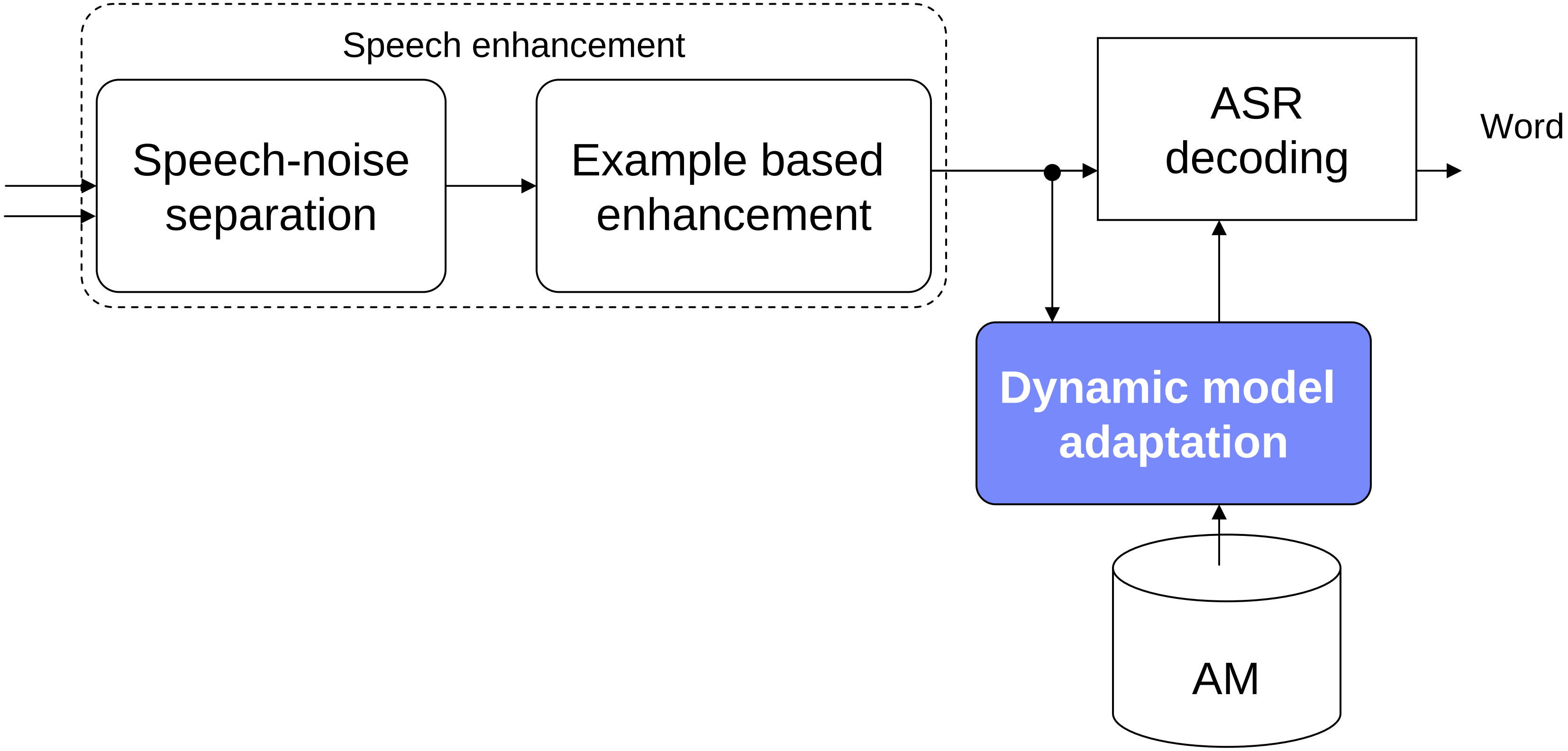
- Look for the longest example segments



- Use the corresponding clean speech example for Wiener filtering

Using precise model of temporal structure of speech
→ remove remaining highly non-stationary noise
→ recover precisely speech

System overview

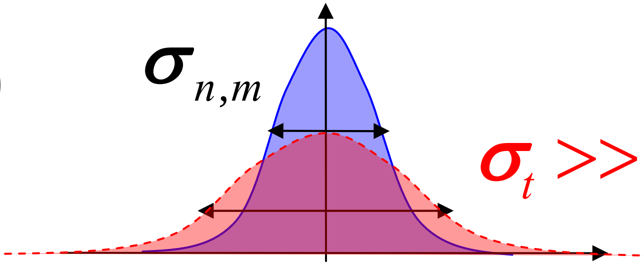


Dynamic model adaptation [Delcroix, 2009]

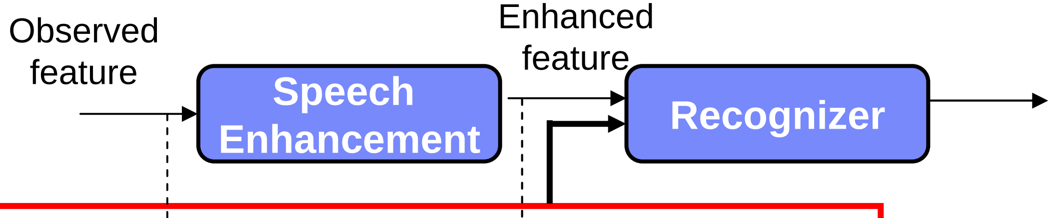
- **Compensate mismatch between enhanced speech and acoustic model**
  - Non-stationary noise & frame by frame processing
  - → Mismatch changes frame by frame (dynamic)
  - → *Conventional acoustic model compensation techniques (MLLR) not sufficient*
- **Dynamic variance compensation (Uncertainty decoding) [Deng, 2005]**
  - Mitigate the mismatch frame by frame by considering feature variance

$$p(y_t \mid n) = \sum_m p(m) N(y_t; \mu_{n,m}, \sigma_{n,m} + \boxed{\sigma_t})$$

Enhanced speech feature



Dynamic feature variance model



$$\sigma_t(\alpha) = (u_t - y_t)^2 \alpha$$

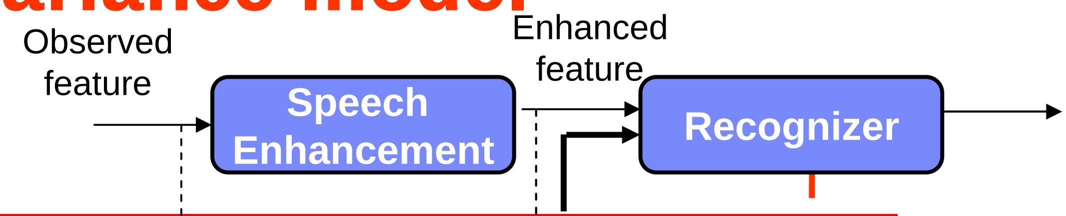
For each feature dimension

- Assumption

Feature variance (Feature uncertainty)	$\propto$	Amount of noise reduction	$\propto$	Amount of Noise

The more we process the signal, the more we introduce uncertainty

Dynamic feature variance model



$$\boxed{\sigma_t(\alpha) = (u_t - y_t)^2 \alpha}$$

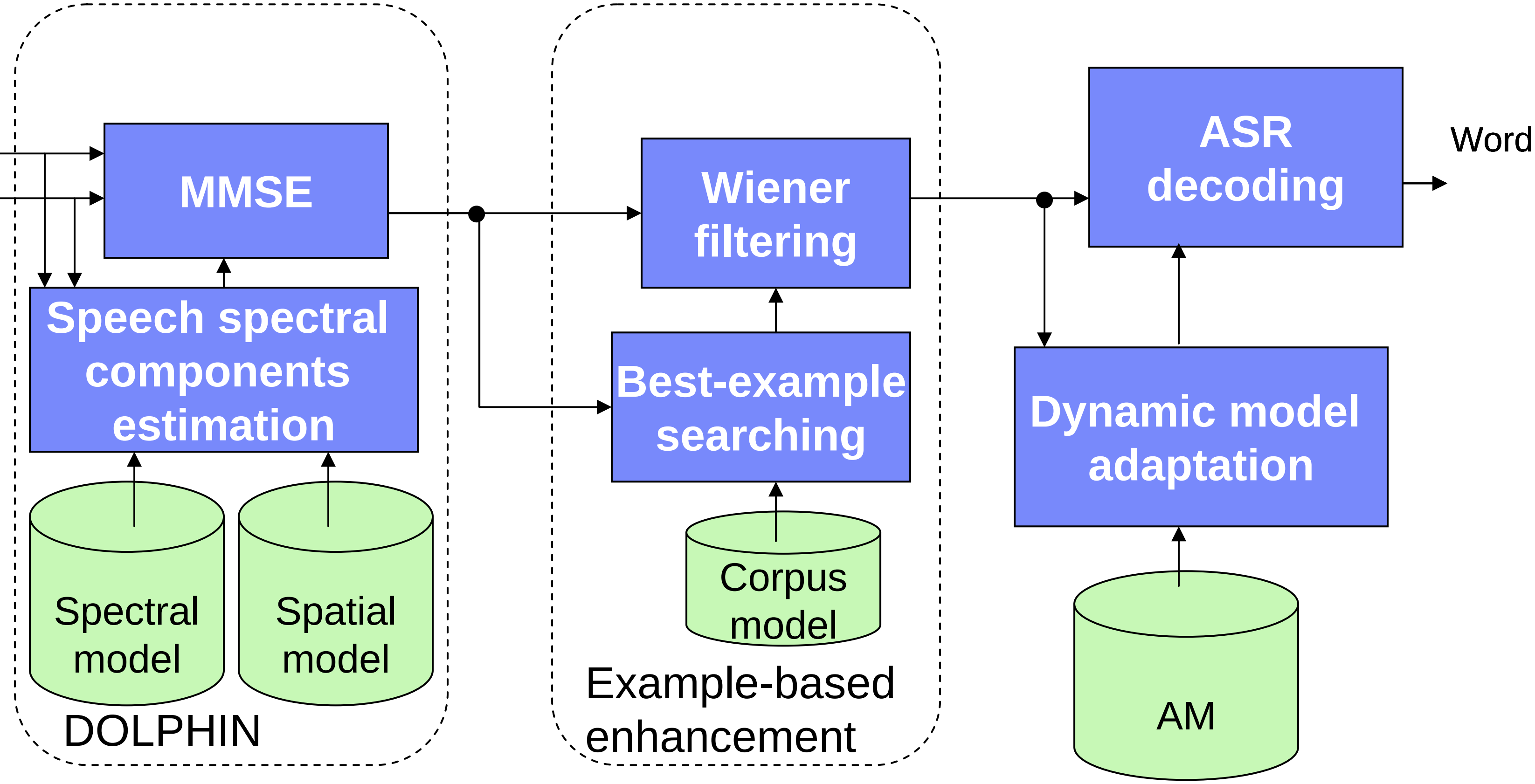
For each feature dimension

- Optimized for recognition with ML criterion using adaptation data (Dynamic variance adaptation – DVA)
- Can be combined with MLLR for static adaptation of the acoustic model mean parameters

Good interconnection

with recognizer

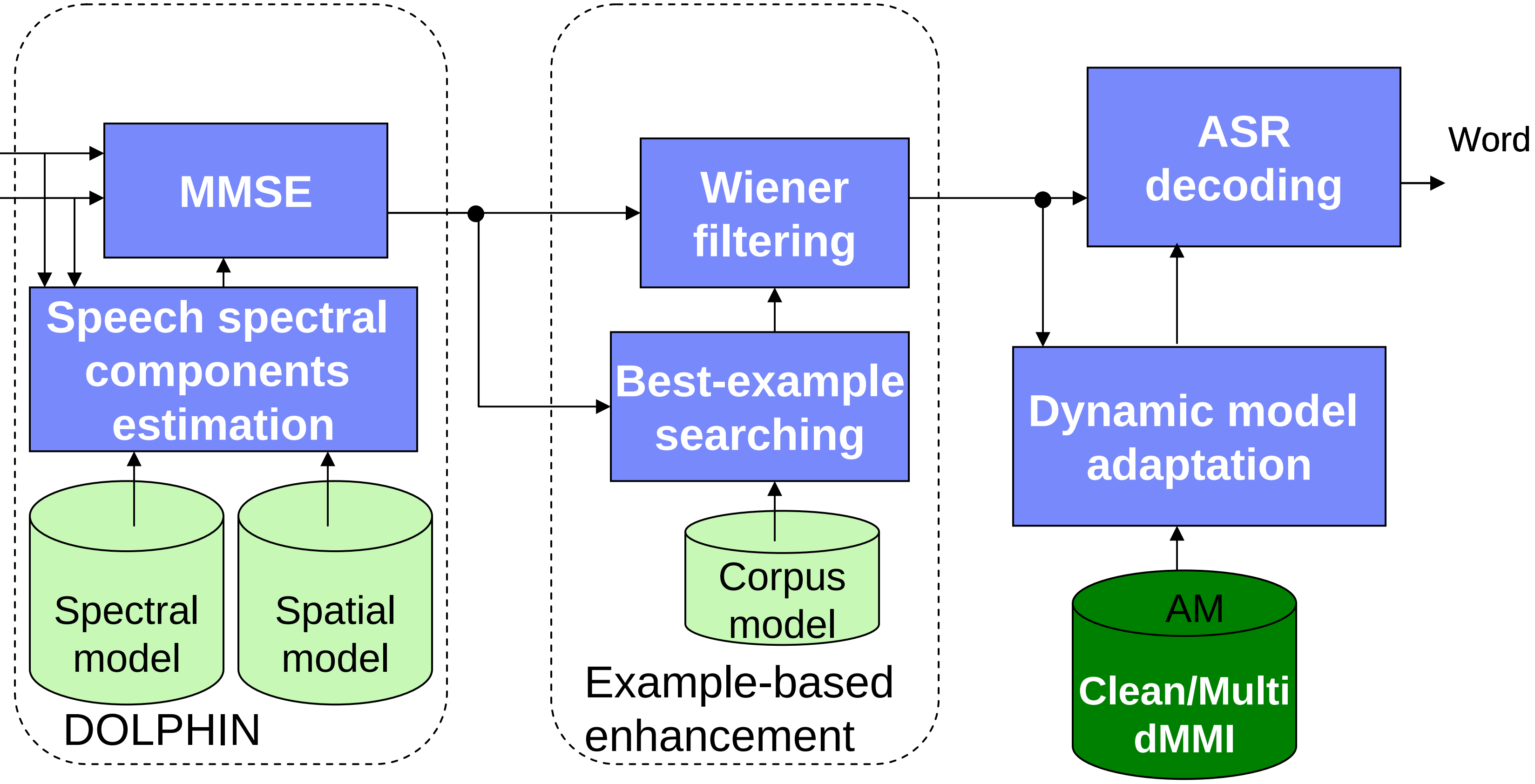
System overview



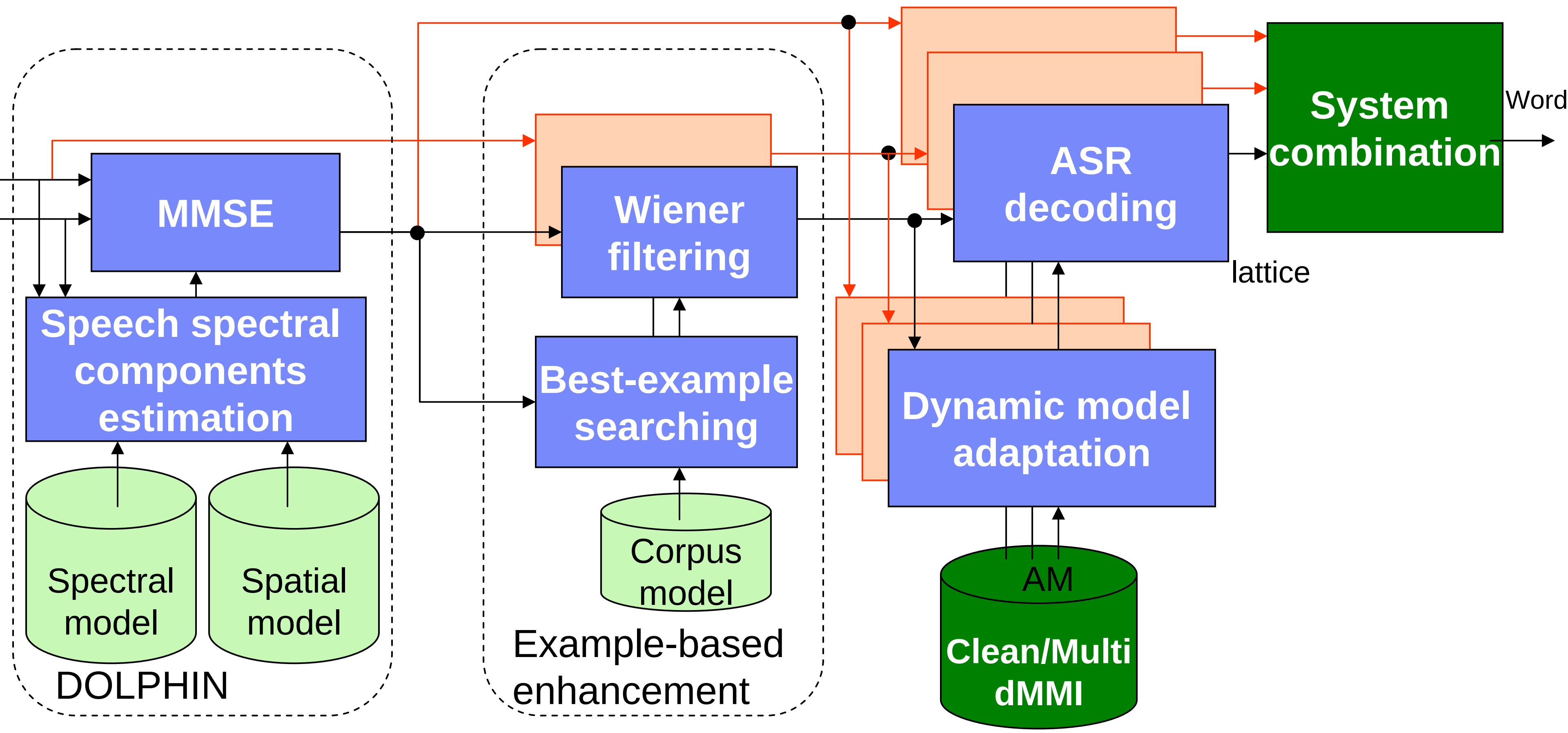
Multi-condition / discriminative training

Add background noise training samples to clean training data

dMMI : differenced maximum mutual information [McDermott, 2010]



System combination [Evermann, 2000]



Settings - Enhancement

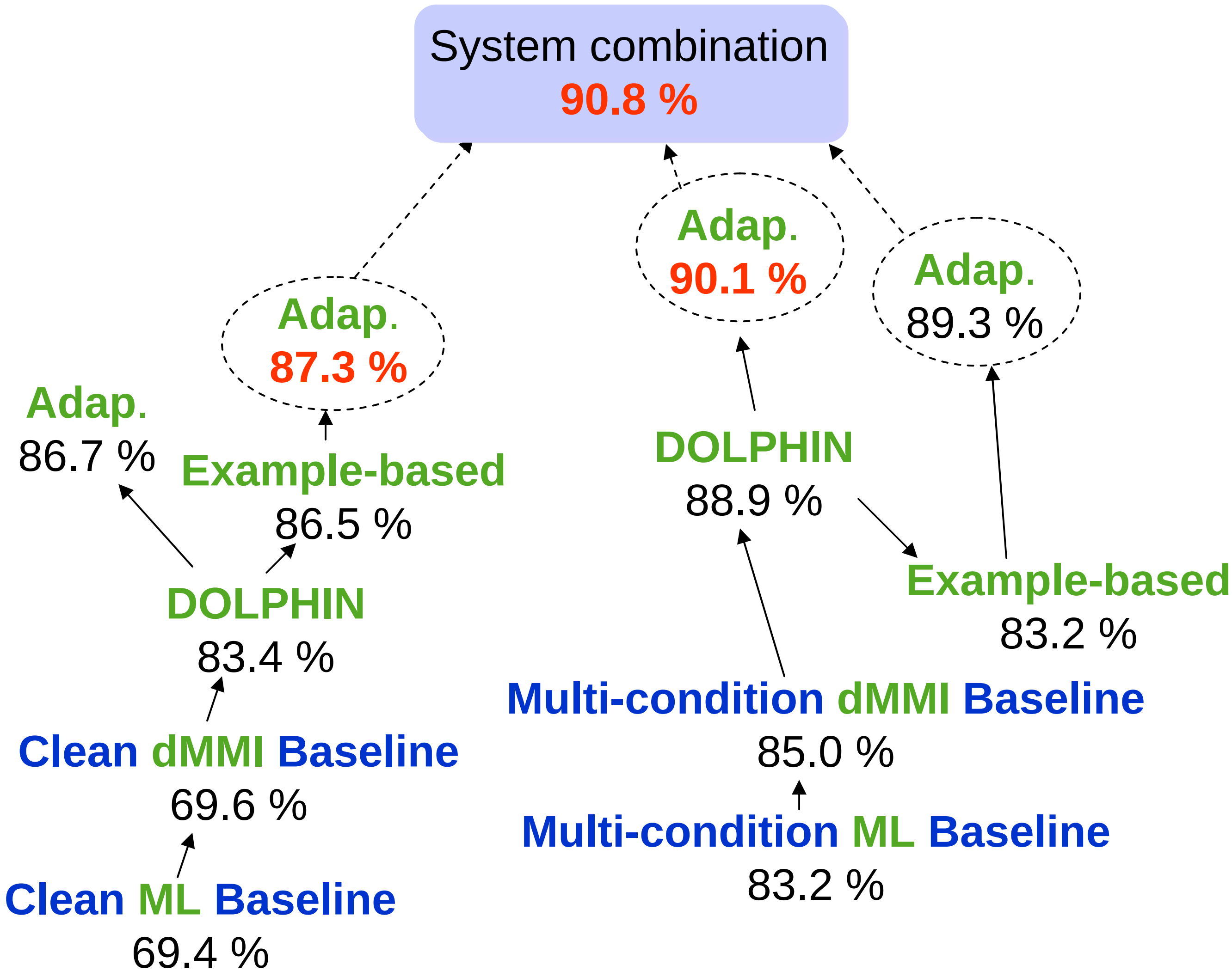
DOLPHIN	▪ Spatial model - 4 mixture components ▪ Spectral model - 256 mixture components - Speaker dependent model ▪ Models trained in advanced using the noise/speech training data ▪ Long windows (100 ms) to capture reverberation
Example-based	▪ Corpus model - GMM w/ 4096 mixture components - Trained on DOLPHIN processed speech - Features 60 order MFCC w/ log energy

Settings - Recognition

Recognizer	▪ SOLON [Hori, 2007]
Acoustic Model	▪ Trained with SOLON (ML & discriminative (dMMI)) ▪ Clean - HMM w/ 254 states (include silent state) - HMM state modeled by GMM with 7 components ▪ Multi-condition - 20 components per HMM state - No silent model
Multi-condition data	▪ Added background noise samples to clean training data ▪ 7 noise environment x 6 SNR conditions
Adaptation	▪ Unsupervised/speaker dependent ▪ use all test data for a given speaker

Development

	m6dB	m3dB	0dB	3dB	6dB	9dB	MEAN
Baseline	49.75	52.58	64.25	75.08	84.25	90.58	69.42
Proposed	84.33	88.58	90.17	92.33	94.50	95.00	90.82

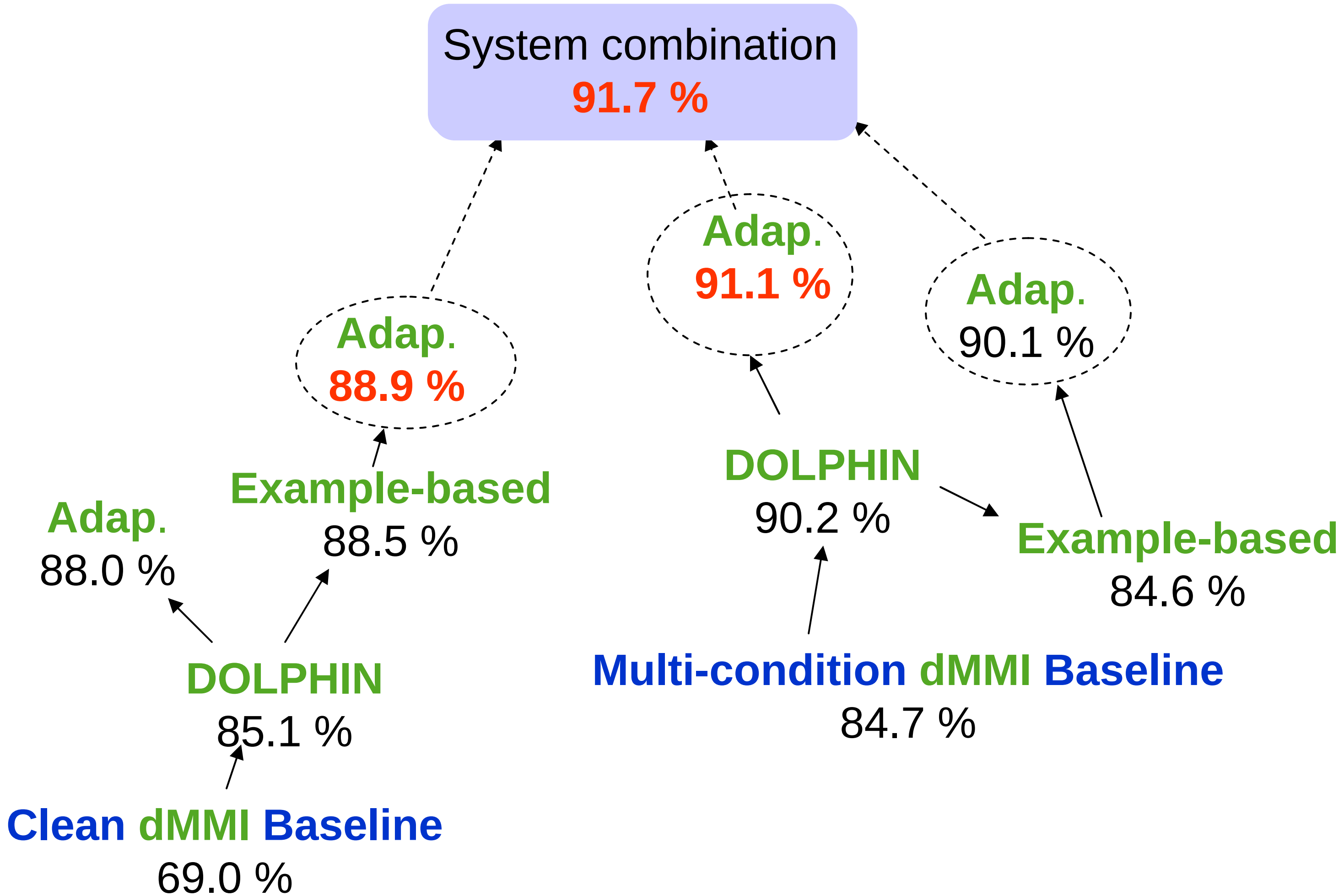


Relative improvement

Multi-cond	51%
Dolphin	**45%**
HTK baseline → SOLON	21%
Adap.	**20%**
Ex. based	**18%**
dMMI	10%
Sys comb	7%

Evaluation

	m6dB	m3dB	0dB	3dB	6dB	9dB	MEAN
Baseline	45.67	52.67	65.25	75.42	83.33	91.67	69.00
Proposed	85.58	88.33	92.33	93.67	94.17	95.83	91.65



Conclusion

- General approach
  - Fully use spatial, spectral and temporal information
  - Good interconnection with recognizer

  → Achieve great reduction of highly non-stationary noise

  → Improve ASR performance

  → Improve also audible quality

  (http://www.kecl.ntt.co.jp/icl/signal/kinoshita/publications/CHiME_demo/index.html)

- Remaining issues
  - Apply to more complex tasks
    - spontaneous speech
    - Unknown speaker location

Thank you!