Generalization problems in ASR acoustic model training and adaptation

Sadaoki Furui
Department of Computer Science
Tokyo Institute of Technology
furui@cs.titech.ac.jp

Outline

1. Introduction
2. Model adaptation
3. Generalization problem
4. Constraining the degree of freedom by using a priori knowledge
5. Constraining the degree of freedom without using a priori knowledge
6. Combinations and extensions
7. Confidence measures
8. Special training methods for the models used for adaptation
9. Conclusion and future works

Outline

1. Introduction
2. Model adaptation
3. Generalization problem
4. Constraining the degree of freedom by using a priori knowledge
5. Constraining the degree of freedom without using a priori knowledge
6. Combinations and extensions
7. Confidence measures
8. Special training methods for the models used for adaptation
9. Conclusion and future works

ASR by statistical pattern recognition

- True joint distribution of a word sequence, W, and its corresponding acoustic vector sequence, X, is assumed to be modeled by a true parametric pdf:

$$P(W,X) = P_{\Lambda}(X|W)\ P_{\Gamma}(W) \qquad (1)$$

- The optimal decoder which achieves the expected minimum word error rate becomes

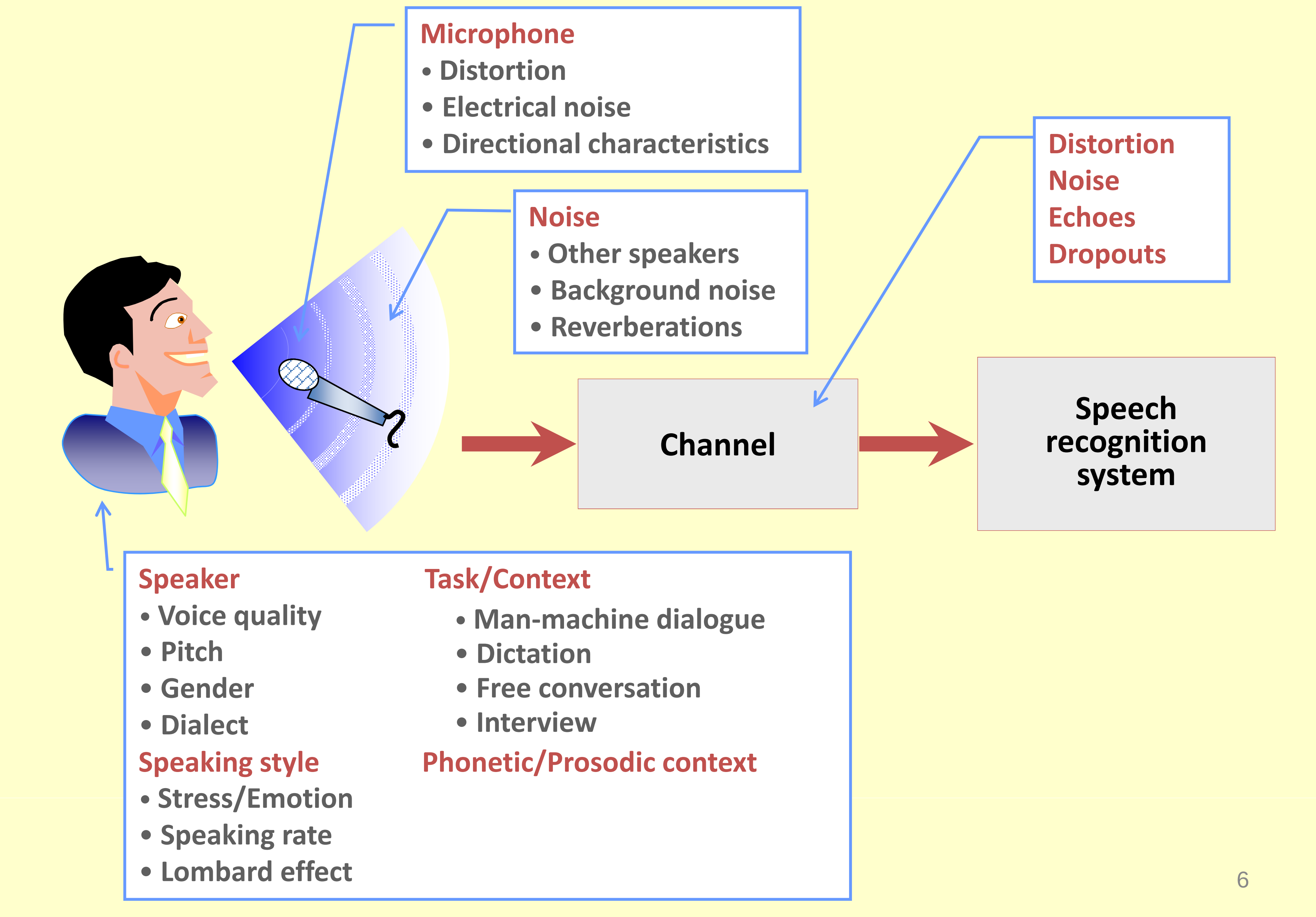
$$\hat{W} = \underset{W}{\mathrm{argmax}}\ P(W|X) = \underset{W}{\mathrm{argmax}}\ P_{\Lambda}(X|W)\ P_{\Gamma}(W) \qquad (2)$$

- Since we do not know the true parametric form of $P(W,X)$ nor true parameter values, they need to be estimated from a large set of labeled speech and text training data.

Data sparseness problem

- Speech has a large number of sources of variations.
- Mismatch between training and testing
- “There is no data like more data.”
- We always have a data sparseness problem.
- Generalization
  - Model training
  - Model adaptation

Main causes of acoustic variation in speech



Microphone
- Distortion
- Electrical noise
- Directional characteristics

Noise
- Other speakers
- Background noise
- Reverberations

Distortion
Noise
Echoes
Dropouts

Channel

Speech recognition system

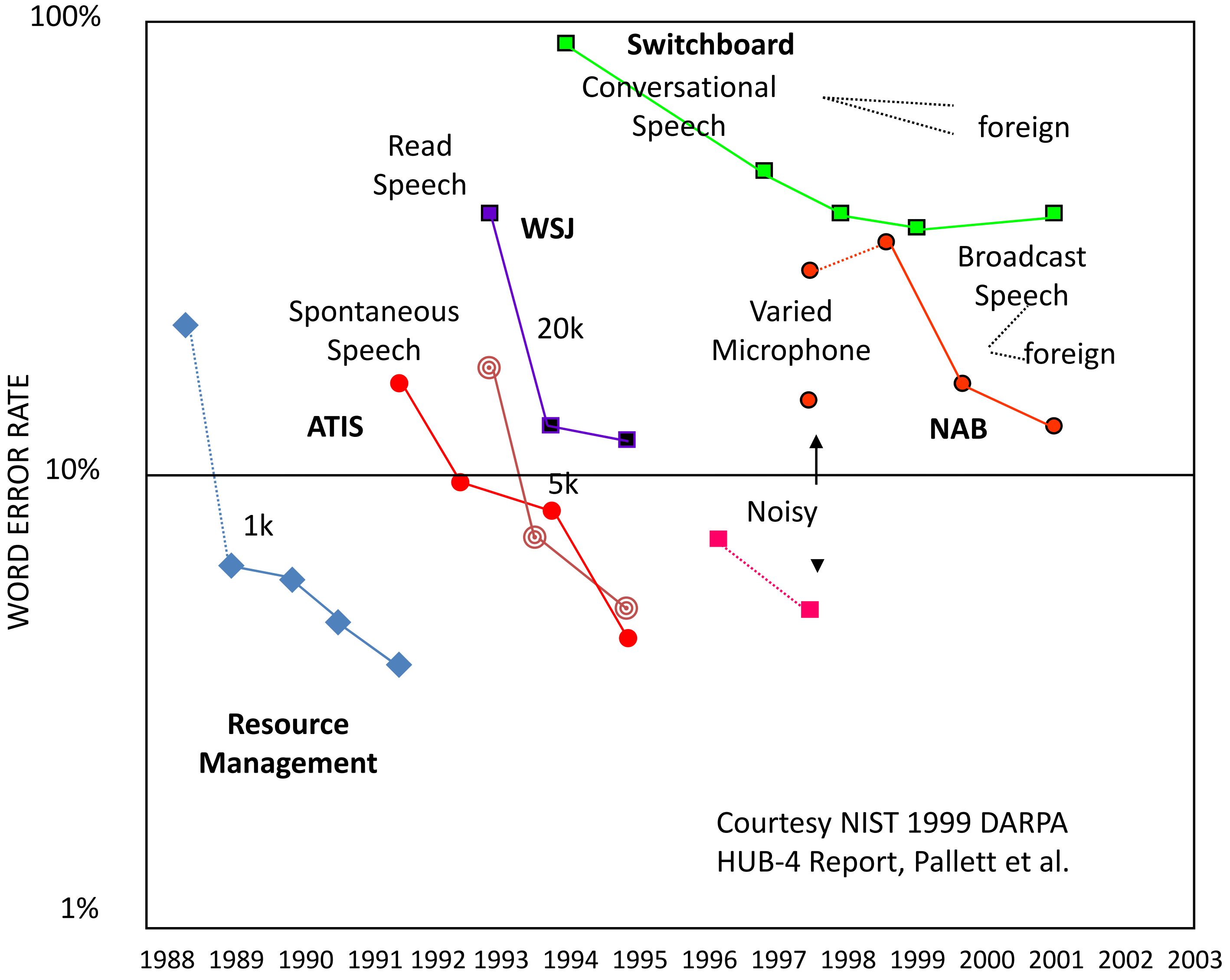
Speaker
- Voice quality
- Pitch
- Gender
- Dialect

Speaking style
- Stress/Emotion
- Speaking rate
- Lombard effect

Task/Context
- Man-machine dialogue
- Dictation
- Free conversation
- Interview

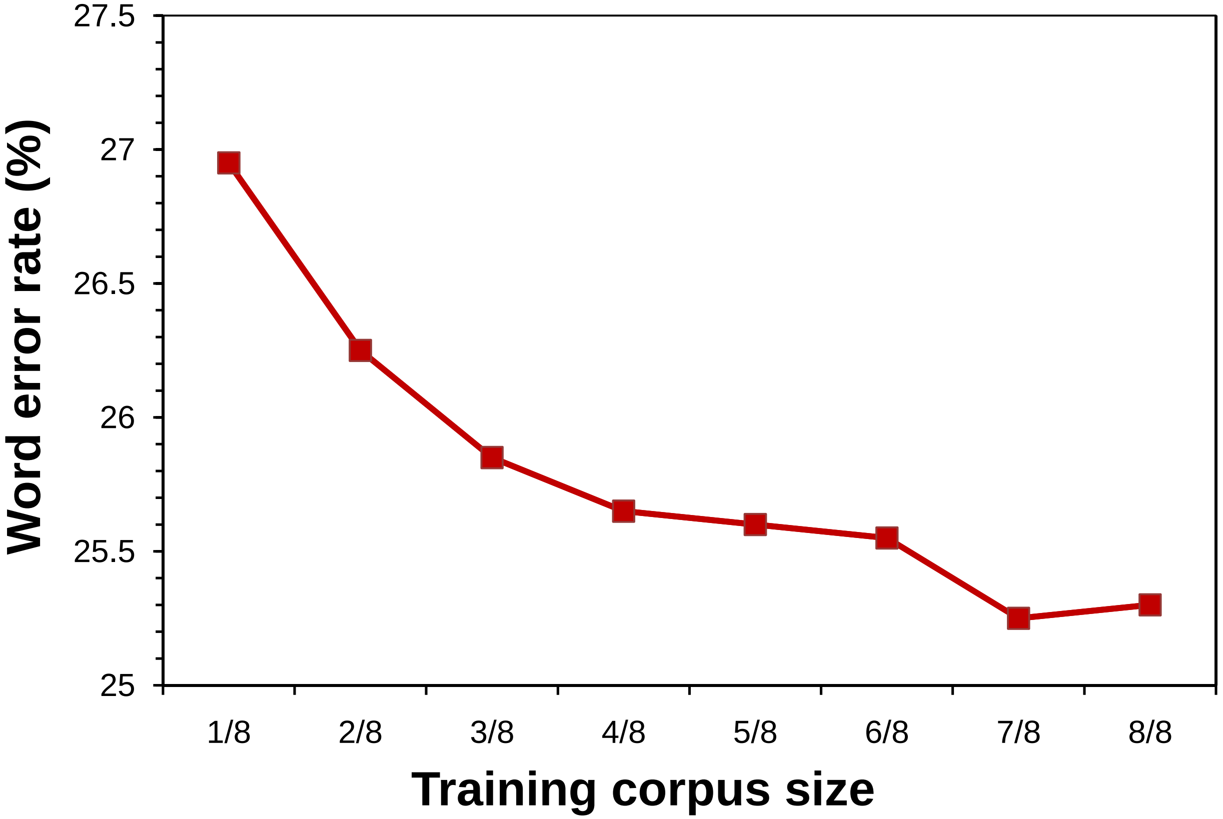
Phonetic/Prosodic context

History of DARPA speech recognition benchmark tests



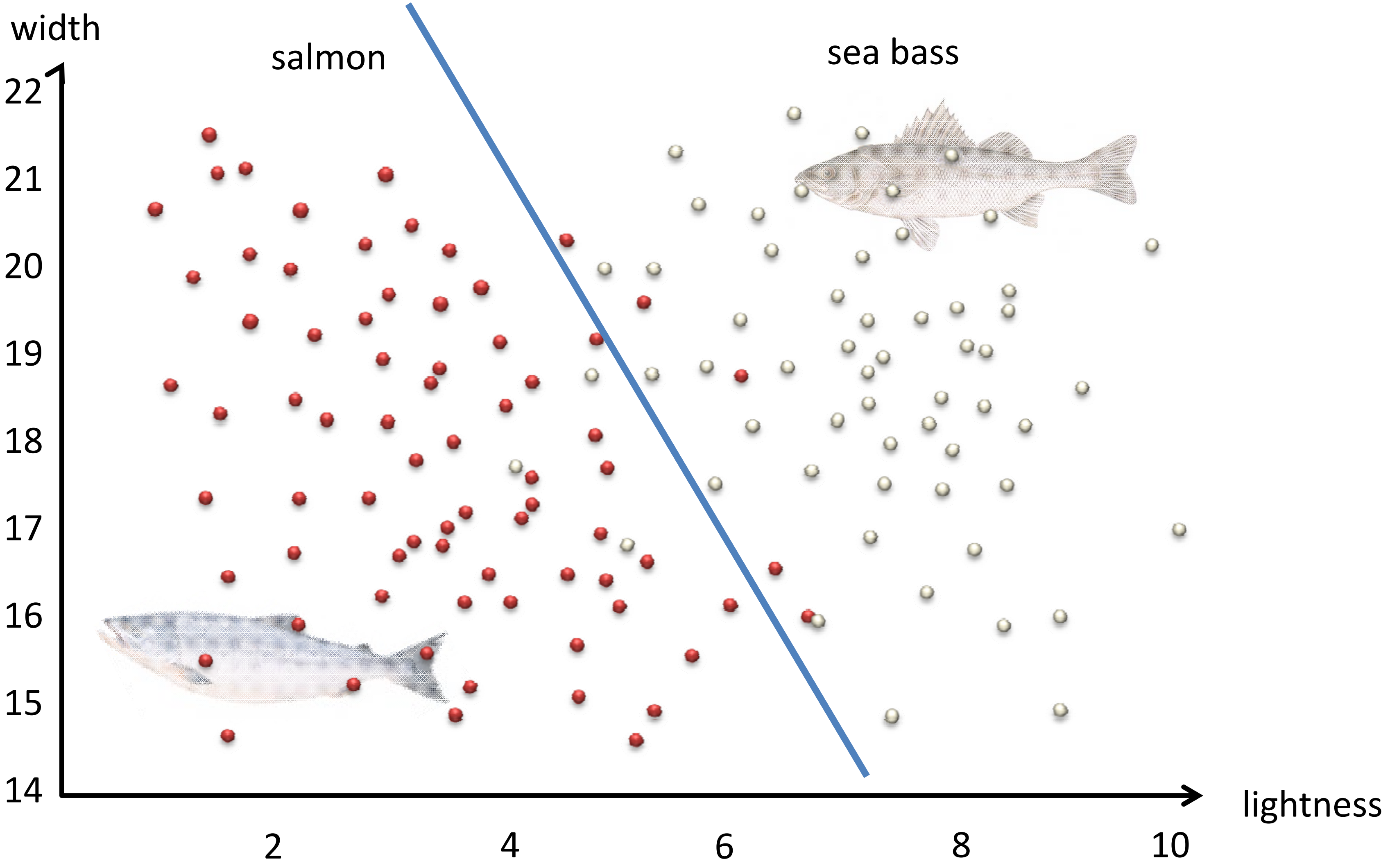
Courtesy NIST 1999 DARPA HUB-4 Report, Pallett et al.

Word error rate (WER) as a function of the size of acoustic model training data (8/8 = 510 hours)



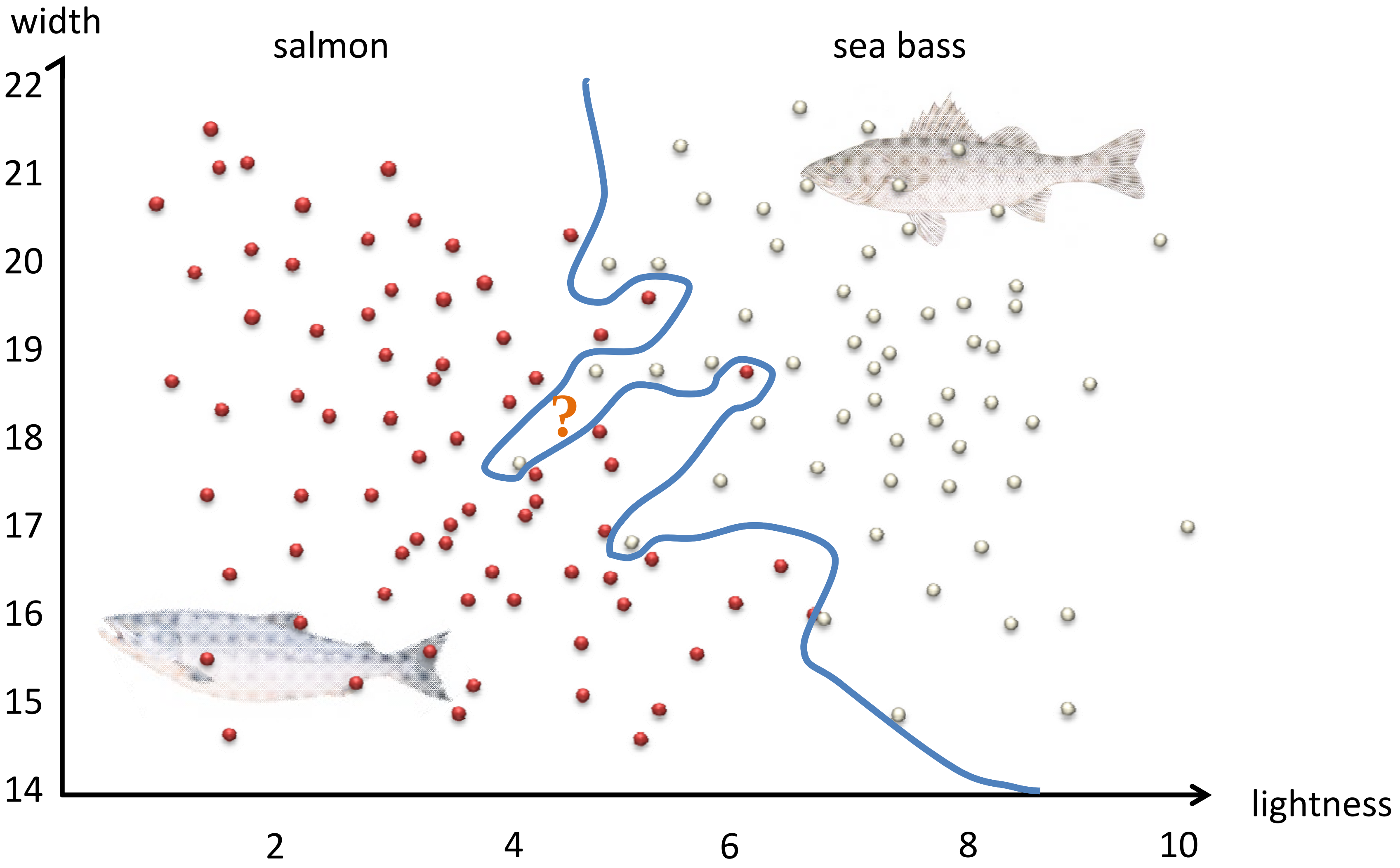
Salmon/sea bass classification using a straight line

(Too simple)

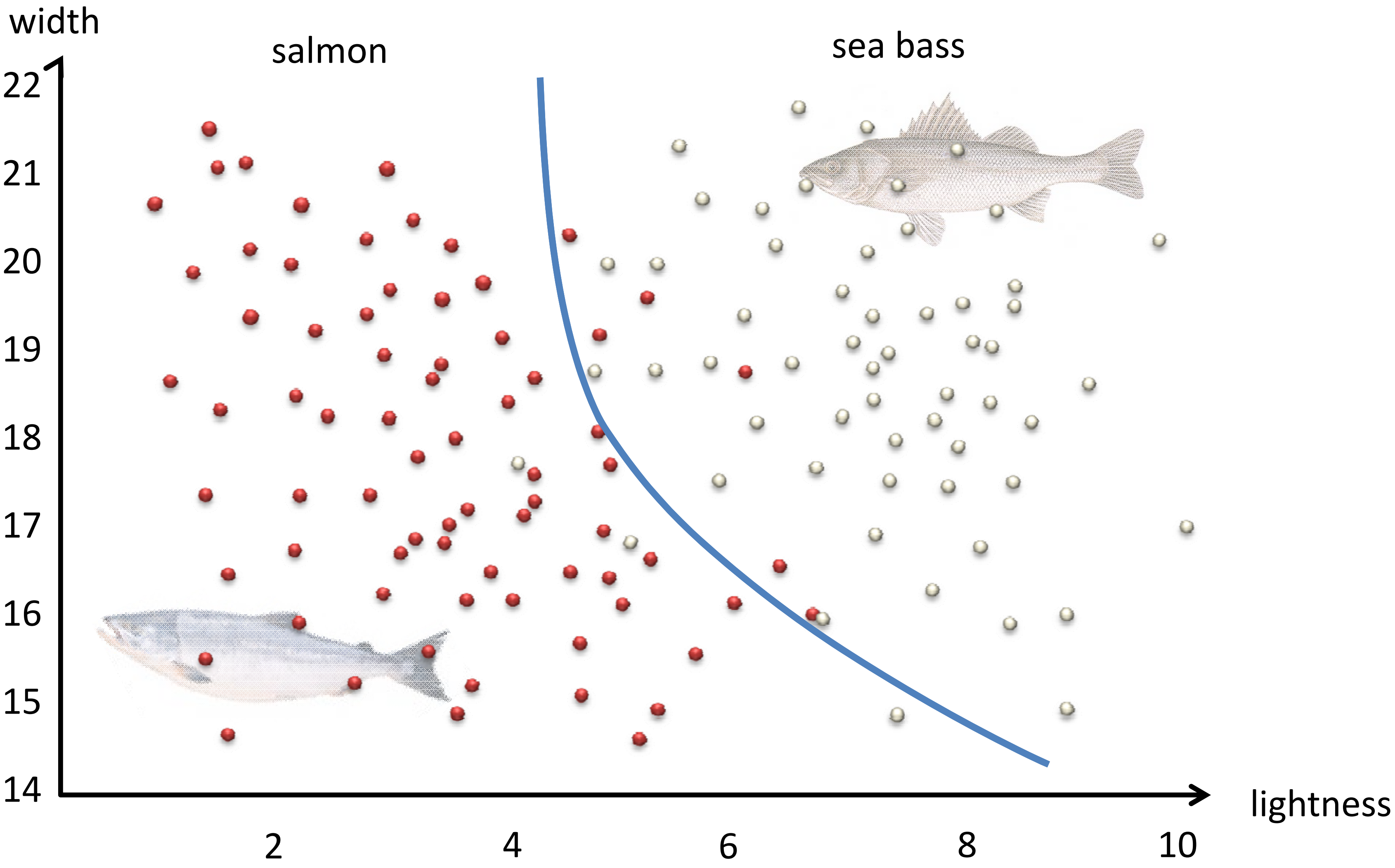


Salmon/sea bass classification using complex models

(Over-tuning)



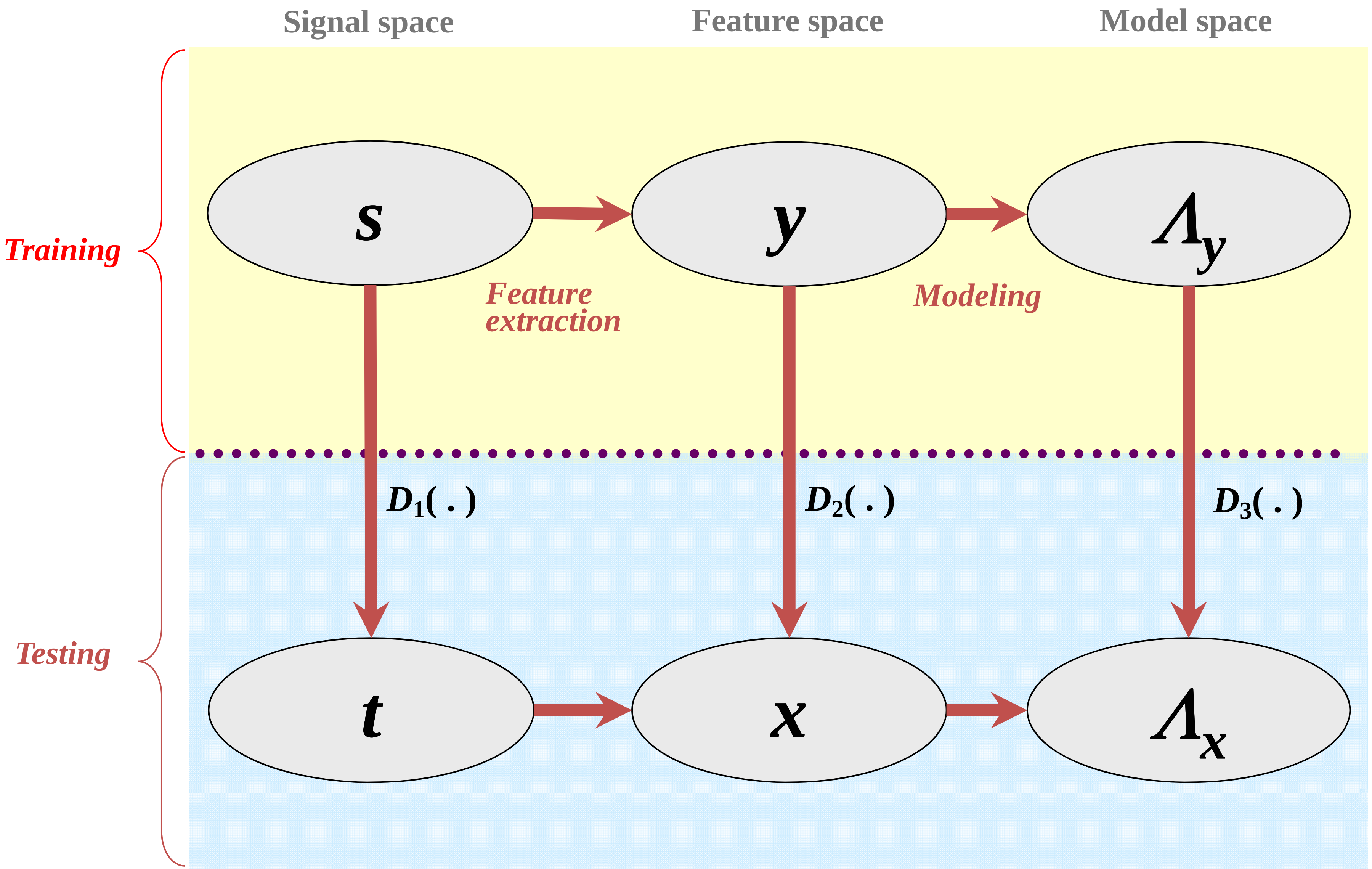
Salmon/sea bass classification using a simple curve



Outline

1. Introduction
2. Model adaptation
3. Generalization problem
4. Constraining the degree of freedom by using a priori knowledge
5. Constraining the degree of freedom without using a priori knowledge
6. Combinations and extensions
7. Confidence measures
8. Special training methods for the models used for adaptation
9. Conclusion and future works

Mismatch between training and testing



Model adaptation

- Supervised adaptation
- Unsupervised adaptation
  - Recognition hypotheses are used as supervision information
  - On-line/off-line adaptation
  - Instantaneous/batch adaptation
  - Iterative adaptation: recognition errors are reinforced during the iteration
- ML, MAP or discriminative estimation

Discriminative training

- Maximum mutual information (MMI) (Normandin, 1996)
  - The mutual information between data and their corresponding labels/symbols is maximized
- Minimum classification error (MCE) (Juang & Katagiri, 1992)
  - The recognition error rate of the classifier is embedded in a smooth function form, and the expected loss of the classifier is minimized
- Minimum phone/word error (MPE/MWE) (Povey, 2003)
  - Performance is optimized at the substring pattern level
- Discriminative training techniques are more heavily biased towards the supervision hypothesis

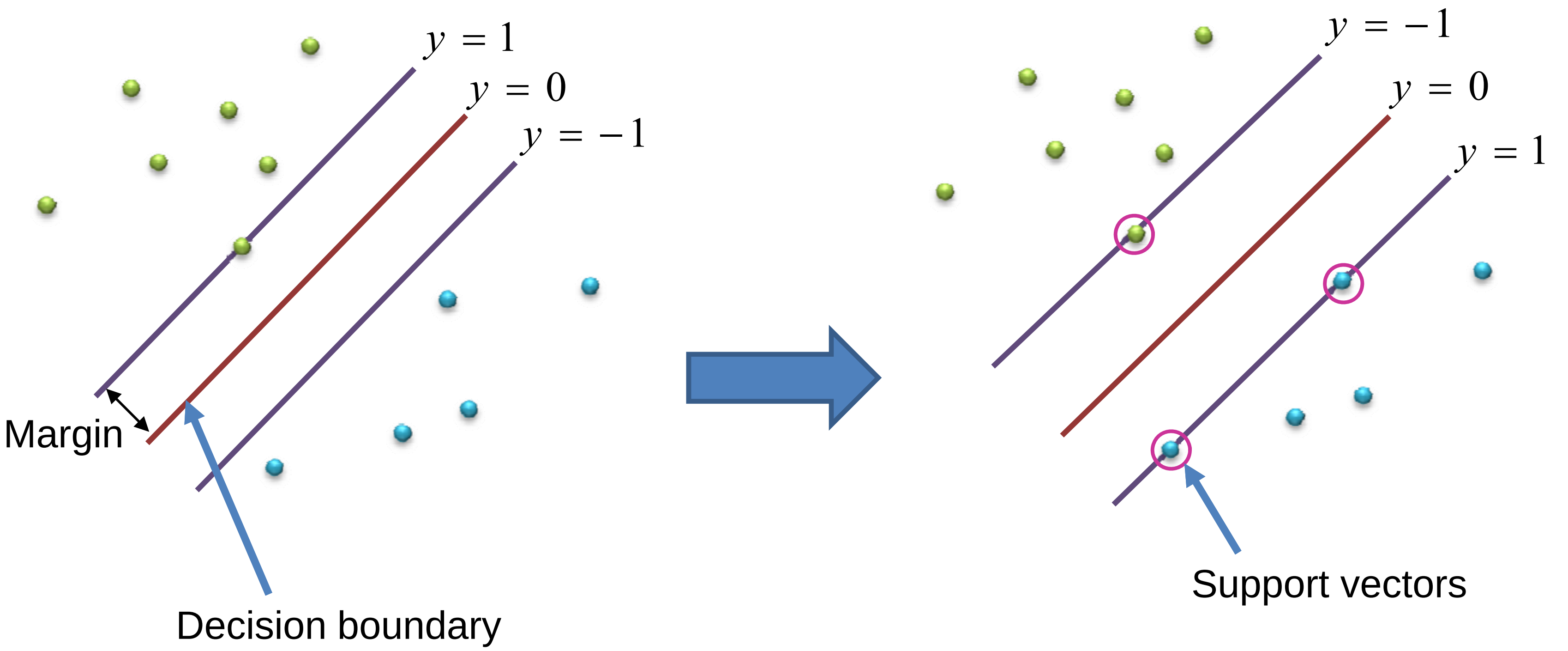
Outline

1. Introduction
2. Model adaptation
3. Generalization problem
4. Constraining the degree of freedom by using a priori knowledge
5. Constraining the degree of freedom without using a priori knowledge
6. Combinations and extensions
7. Confidence measures
8. Special training methods for the models used for adaptation
9. Conclusion and future works

Generalization problem

- How to reduce the effect of hypothesis bias and allow robust estimates using a limited amount of data (no theoretical solution)
- By controlling the degree of freedom of the model/smoothing (Occam’s razor)
  - Trade-off between a complicated model and a constrained model is optimized
  - With/without using a priori knowledge
  - A priori knowledge is obtained from our speech knowledge or from training data
- By maximizing “margins” between training samples and decision boundaries

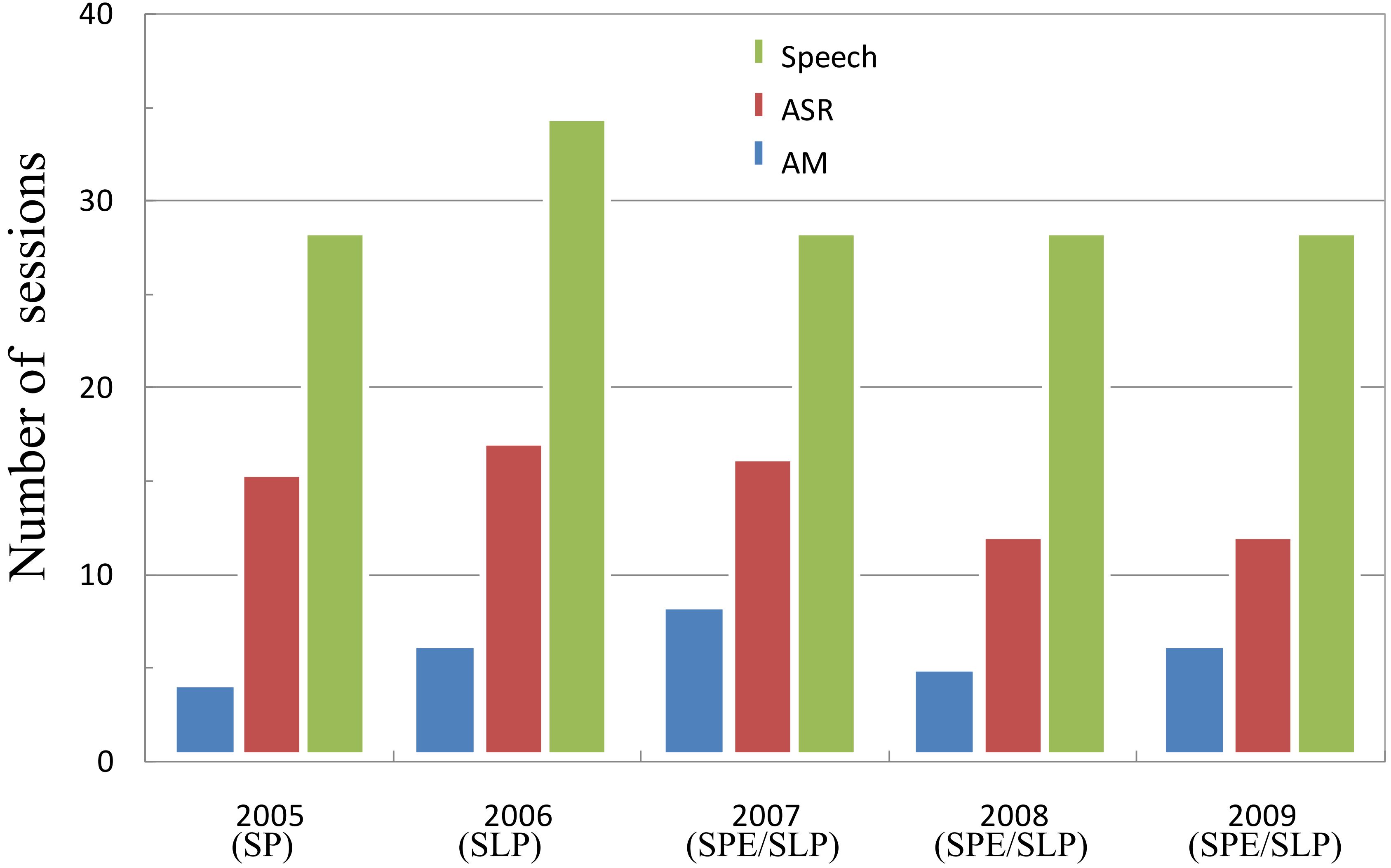
Margin maximization



“Margin” is defined as the perpendicular distance between the decision boundary and the closest data points. Maximizing the margin leads to a particular choice of decision boundary determined by a subset of the data points, “support vectors”.

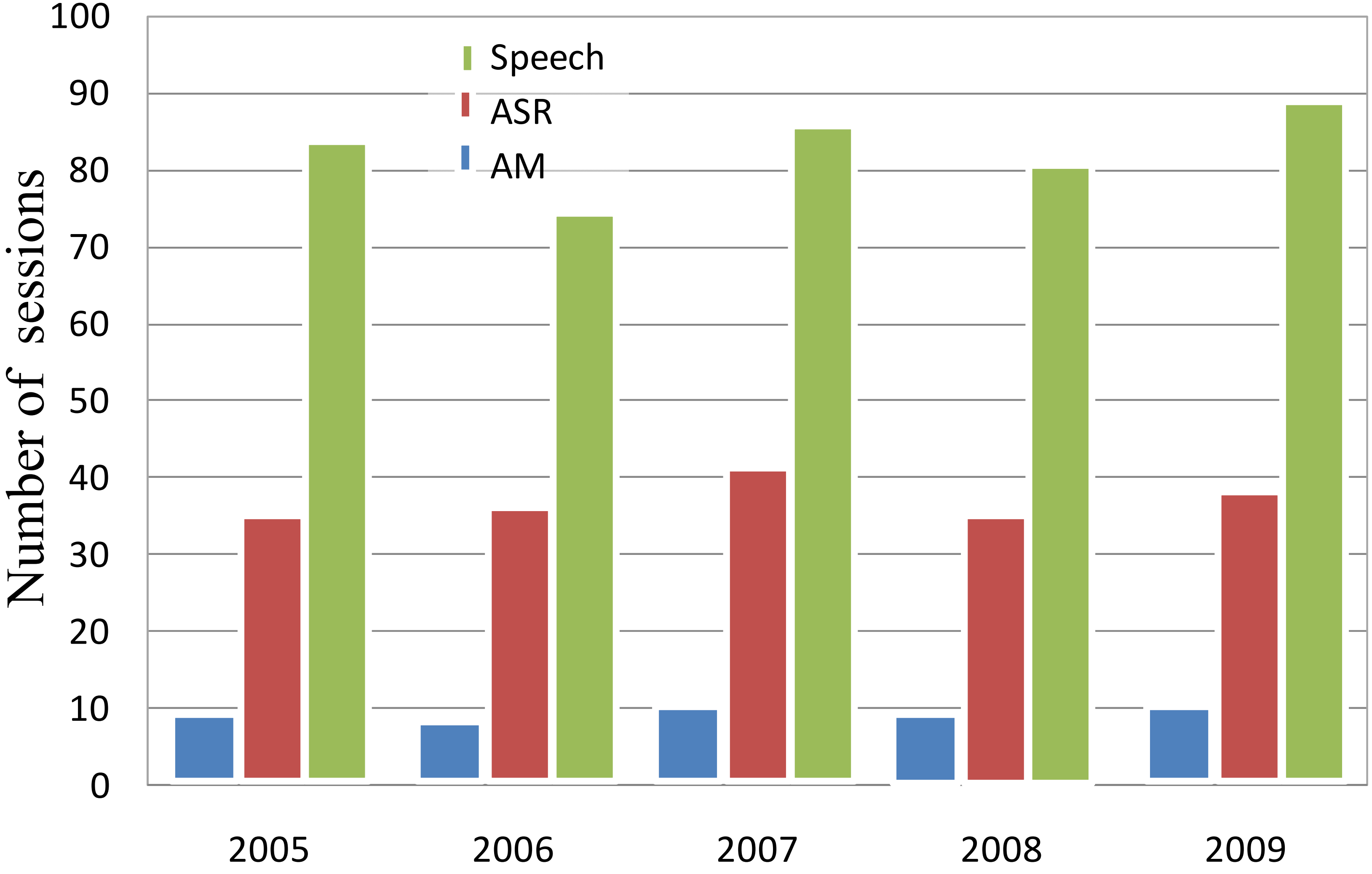
ICASSP speech sessions

(AM: acoustic model training and adaptation)

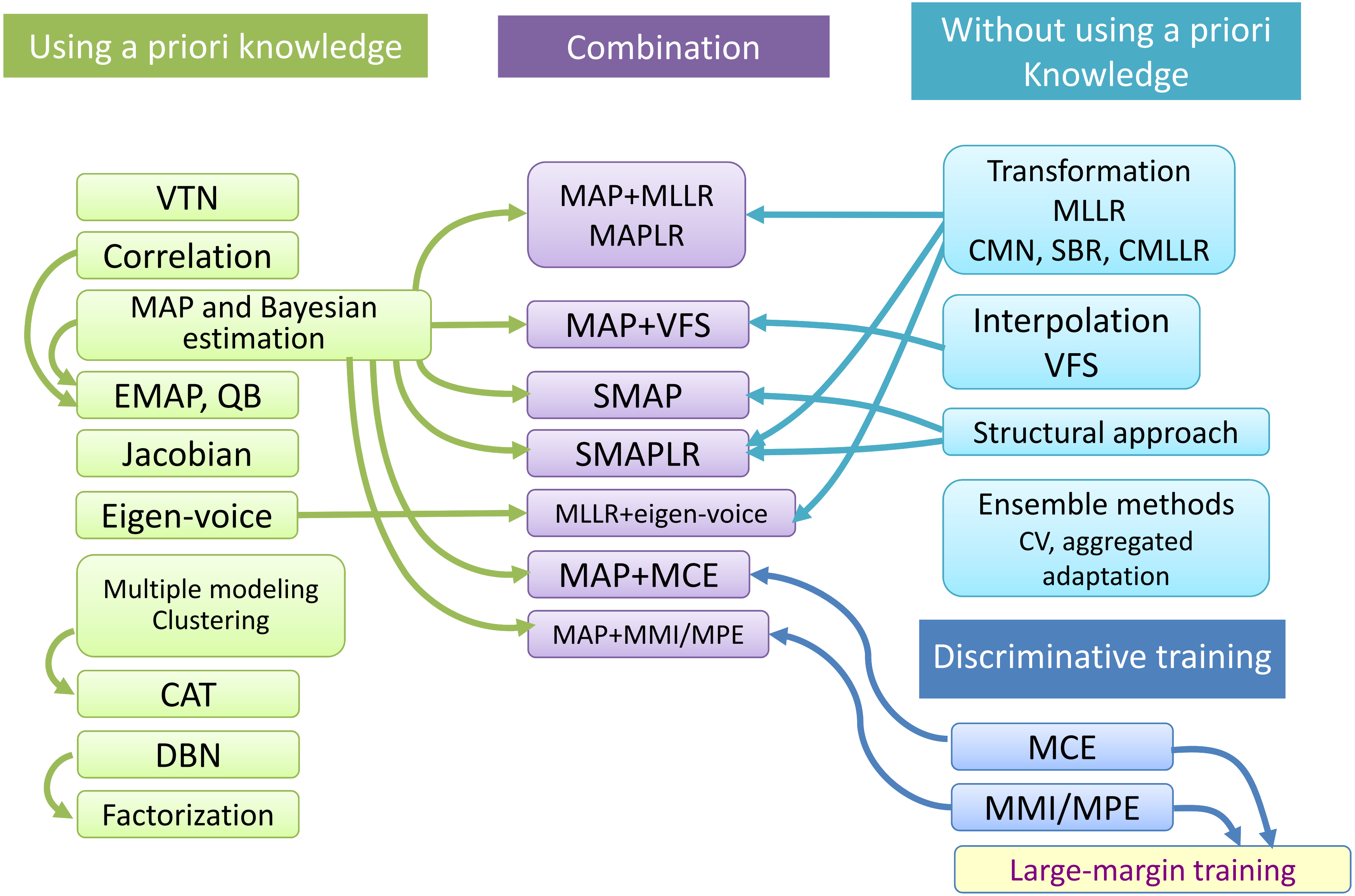


Interspeech speech sessions

(AM: acoustic model training and adaptation)



Constraining the degree of freedom with/without using a priori knowledge



Using a priori knowledge

Combination

Without using a priori Knowledge

VTN

Correlation

MAP and Bayesian estimation

EMAP, QB

Jacobian

Eigen-voice

Multiple modeling Clustering

CAT

DBN

Factorization

MAP+MLLR MAPLR

MAP+VFS

SMAP

SMAPLR

MLLR+eigen-voice

MAP+MCE

MAP+MMI/MPE

Transformation MLLR CMN, SBR, CMLLR

Interpolation VFS

Structural approach

Ensemble methods CV, aggregated adaptation

Discriminative training

MCE

MMI/MPE

Large-margin training

Outline

1. Introduction
2. Model adaptation
3. Generalization problem
4. Constraining the degree of freedom by using a priori knowledge
5. Constraining the degree of freedom without using a priori knowledge
6. Combinations and extensions
7. Confidence measures
8. Special training methods for the models used for adaptation
9. Conclusion and future works

Constraining the degree of freedom by using a priori knowledge

- Vocal tract length normalization (VTN)
- Correlation
- MAP and Bayesian estimation
- EMAP and quasi-Bayes (QB) methods
- Jacobian approach
- Eigen-voice
- Multiple modeling (multi-style training)
  - Cluster-based model selection
- Cluster adaptive training (CAT)
- Bayesian networks

VTN

- Simple constrained models
- Piecewise linear or bilinear warping in the frequency domain (Wakita, 1977)
- Speaker-specific Bark/Mel scale warping (Lee & Rose, 1996)

Correlation

- Estimation of model parameters for units, including those not included in the adaptation data, based on pair-wise unit correlation (Furui, 1980)

$$\hat{U}_i = \left(r \Big/ \sum_j \omega_{ij} \right) \sum_j \omega_{ij} \Phi_{ij} V_j + (1 - r) \bar{U}_i$$

$$\hat{U}_i = (u_{i1}, u_{i2}, \cdots, u_{iN})'$$

Model vector of phoneme i obtained by the whole vocabulary

$$V_j = (v_{j1}, v_{j2}, \cdots, v_{jN})'$$

Model vector of phoneme j obtained by a fraction of the vocabulary

Φ_{ij} : N x N matrix estimated by multiple regression analysis using training data

ω_{ij} : Weighting coefficient based on multiple correlation coefficient

r : $0 < r < 1$

- Extended to the Quasi-Bayes technique (Huo and Lee, 1998)

MAP and Bayesian estimation

- Maximum a posteriori (MAP) estimation (Gauvain and Lee, 1994)

$$\Lambda_{MAP} = \underset{\Lambda}{\text{argmax}}\; P(\Lambda|X,W) = \underset{\Lambda}{\text{argmax}}\; P(X|\Lambda,W)\; P(\Lambda)$$

- Bayesian estimation

$$P(\Lambda|X,W) = P(X|\Lambda,W)\, P(\Lambda) \Big/ \int P(X|\Lambda,W) P(\Lambda)\, d\Lambda$$

- Mean values are equivalent, but variances are different
- Bayesian estimation is more robust and more effective when the adaptation data is limited
- Bayesian estimation is computationally expensive

EMAP and QB methods

- Extension of MAP approach for incremental adaptive learning of HMM parameters
- Using the correlation between the mean values of different speech units
- EMAP (extended MAP) (Zavaliagkos et al., 1995)
  - Explicitly introducing correlations
- QB (quasi-Bayes) (Huo and Lee, 1998)
  - Based on the assumption that all mean vectors have a joint prior distribution

Jacobian approach

- An analytic approach to adapting models under an initial condition to a target condition (Sagayama et al., 1997, 2001), assuming that
  - the variation can be analytically modeled, and
  - the difference between the two conditions is relatively small.
- Changes are related by Jacobian matrices and the adaptation is performed by simple matrix arithmetic.

Eigen-voice

- Speaker-dependent models from many speakers are created, and PCA is carried out for model parameters for all the speakers.
- The lower order eigen-vectors are selected as eigen-voices.
- For a new speaker, weights for each eigen-voice are estimated in a maximum likelihood estimation to be used for model adaptation (Kuhn et al., 1998, Nguyen et al., 2002)

Multiple modeling (multi-style training)

- Ensemble of condition-specific models (gender, age, speaking rate, spontaneity, etc.) are trained and used within a selection, competition or combination framework (Nanjo & Kawahara, 2002).
- Dynamic Bayesian networks (DBN) handles model dependencies with respect to auxiliary variables or hidden factors.
- Cluster-based model selection

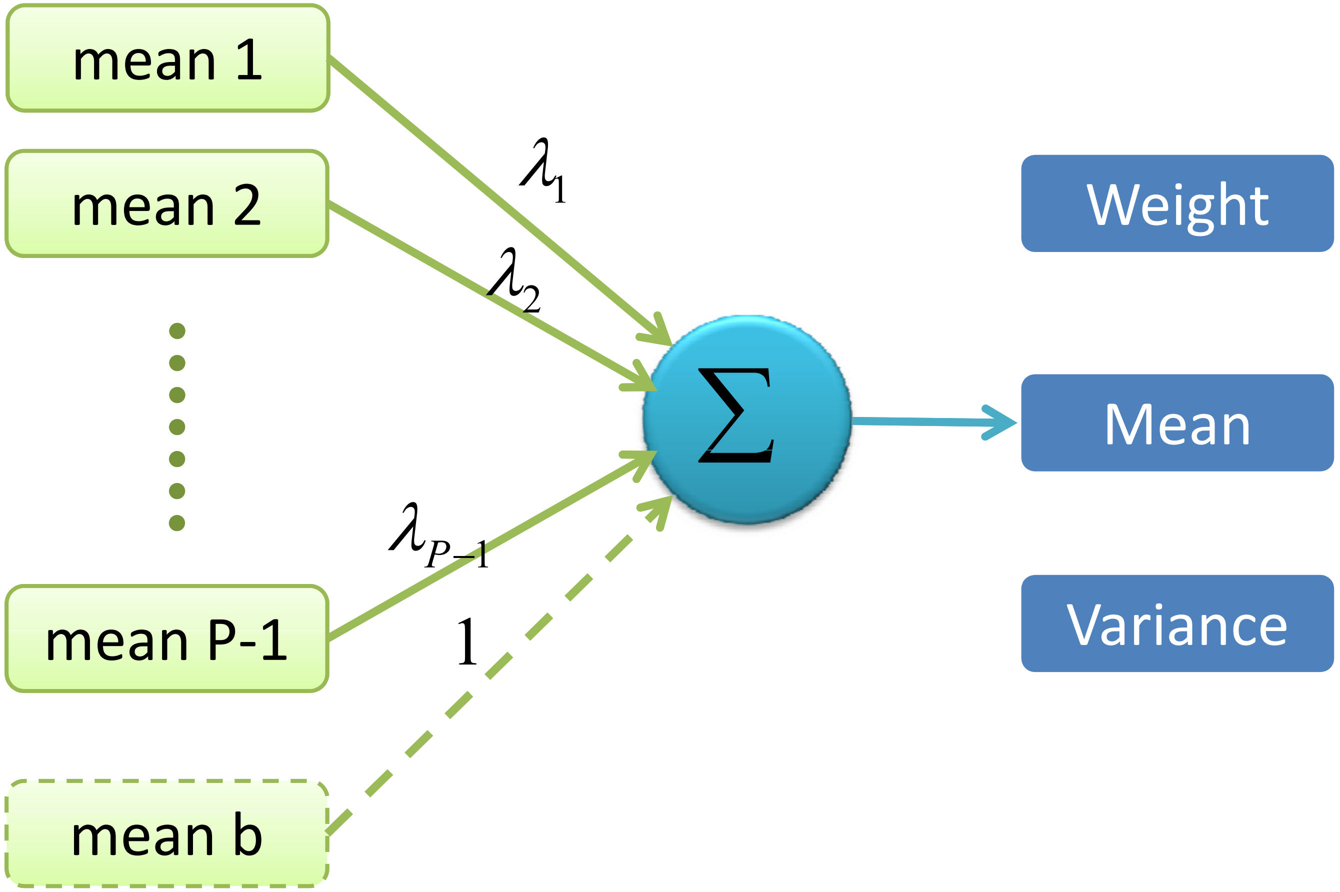
Cluster-based model selection

- Multiple models are prepared using a clustering technique.
- The optimum model for input speech is selected to maximize likelihood (Kosaka et al., 1994; Padmanabhan et al., 1998).
- Clustering training data at the utterance level provides better performance than that at the lecture level (Shinozaki & Furui, 2004).

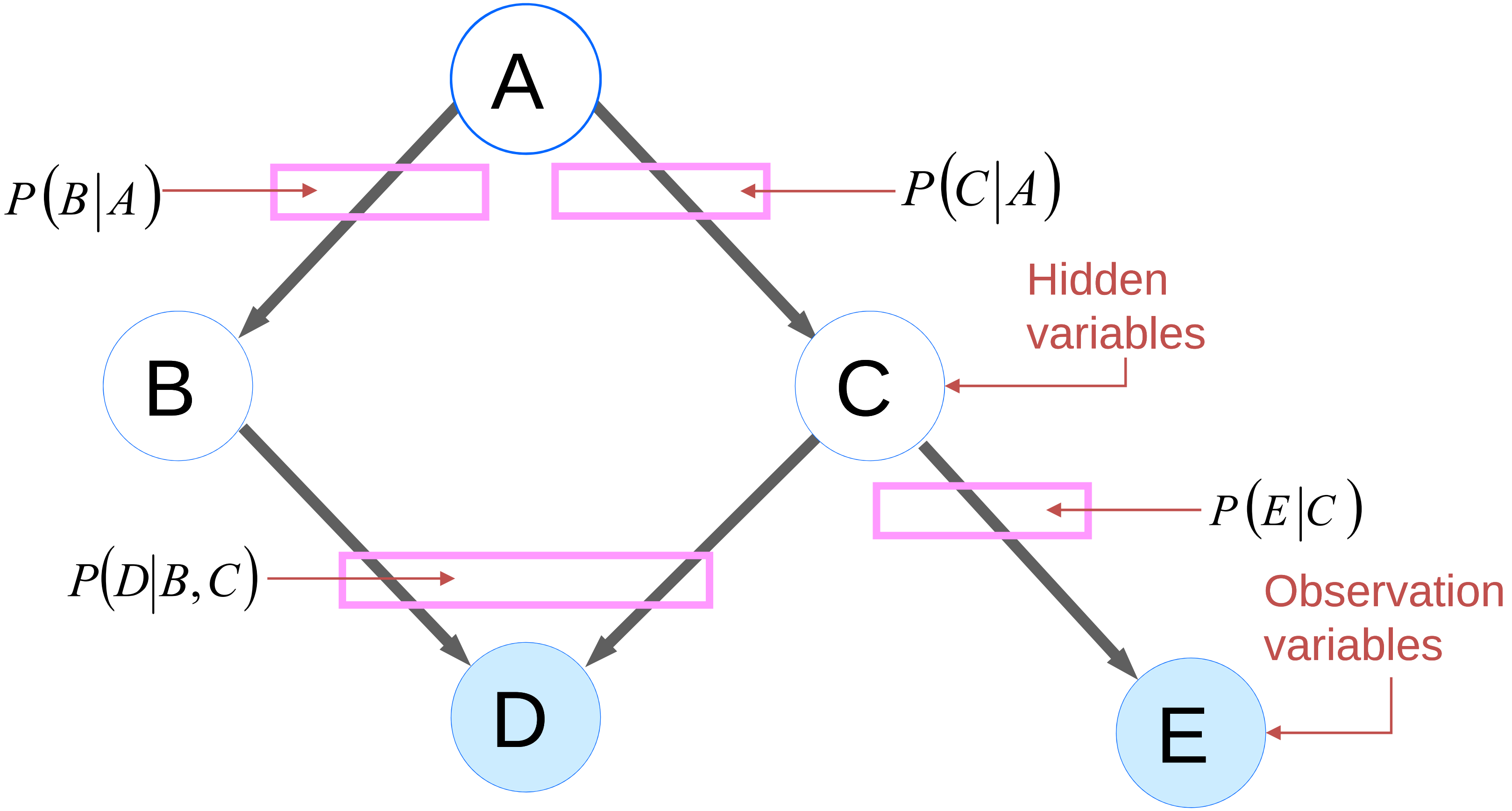
Cluster adaptive training (CAT)

- A linear interpolation of all the clusters is used (Gales, 1998).
- HMM component weights and variances are tied over all the speaker clusters, and a set of interpolation values, the weight vector, is estimated.
- An explicit set of means or cluster dependent MLLR transforms of some canonical model are used.
- CAT and eigen-voice methods are mathematically similar: linear combinations of some basis vectors representing “prototypical” speakers.

Cluster adaptive training



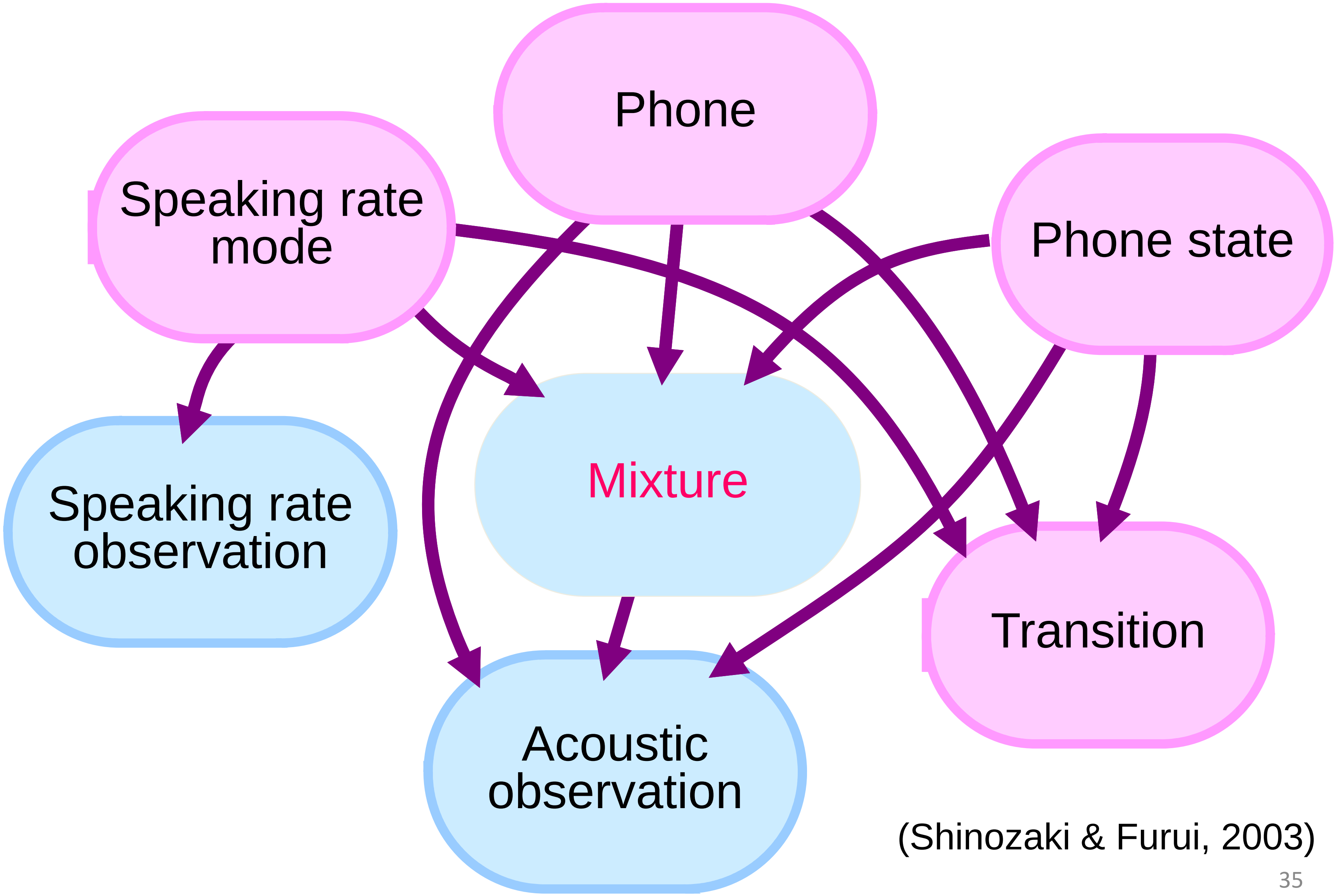
A Bayesian network with five variables



Joint distribution: $\boldsymbol{P(A,B,C,D,E) = P(A)P(B|A)P(C|A)P(D|B,C)P(E|C)}$

Variables with known values are shaded. Conditional probability functions (indicated by boxes) are associated with each variable and used to return numerical values for conditional probabilities.

Bayesian network representation of HMM incorporating speaking rate variations



(Shinozaki & Furui, 2003)

Outline

1. Introduction
2. Model adaptation
3. Generalization problem
4. Constraining the degree of freedom by using a priori knowledge
5. Constraining the degree of freedom without using a priori knowledge
6. Combinations and extensions
7. Confidence measures
8. Special training methods for the models used for adaptation
9. Conclusion and future works

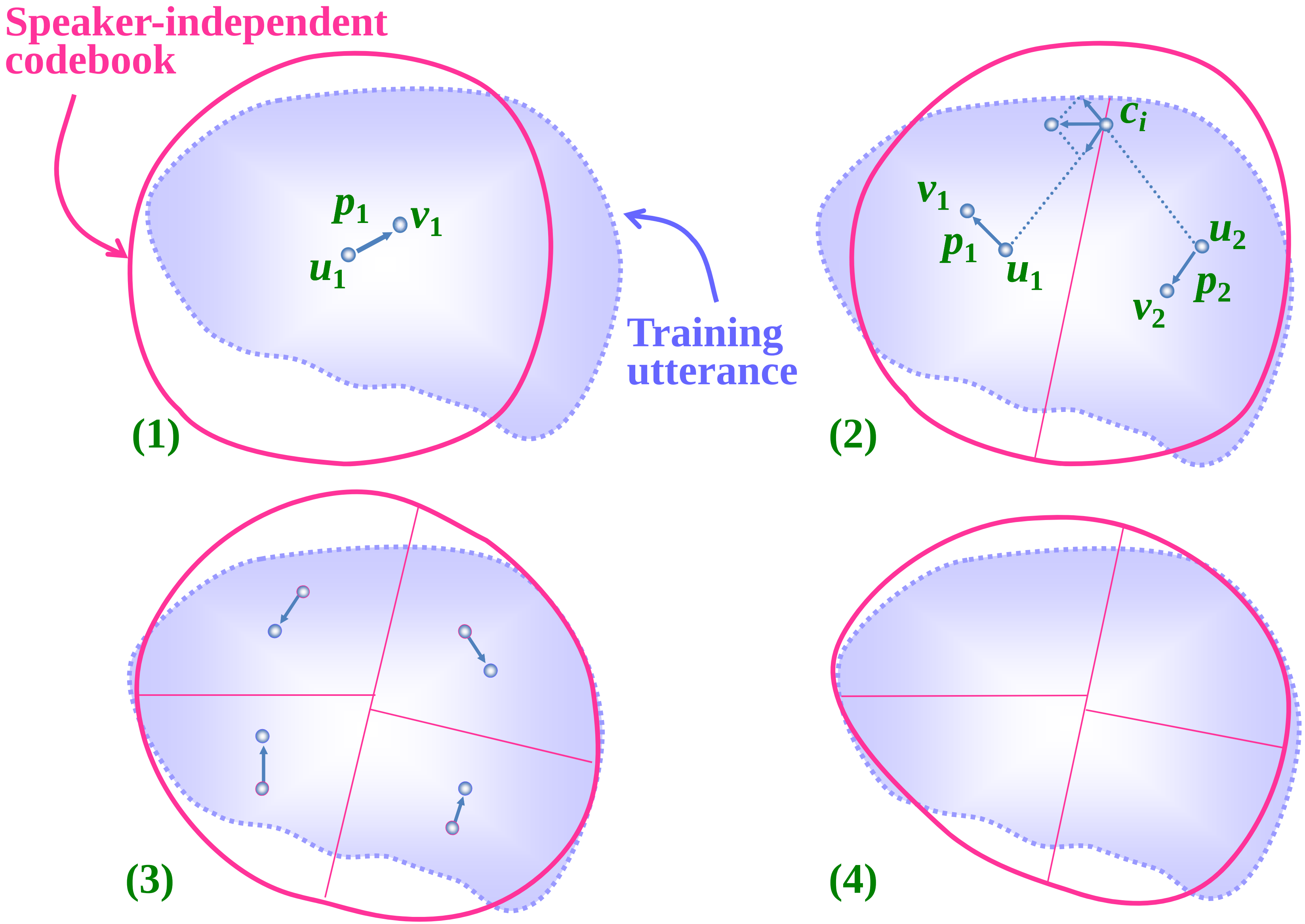
Constraining the degree of freedom without using a priori knowledge

- Structural approach
- Transformation-based approaches
  - Cepstral mean normalization (CMN)
  - MLLR
  - Signal bias removal (SBR)
  - CMLLR
- Interpolation
- Vector field smoothing (VFS)
- Ensemble methods
  - Cross validation adaptation
  - aggregated adaptation

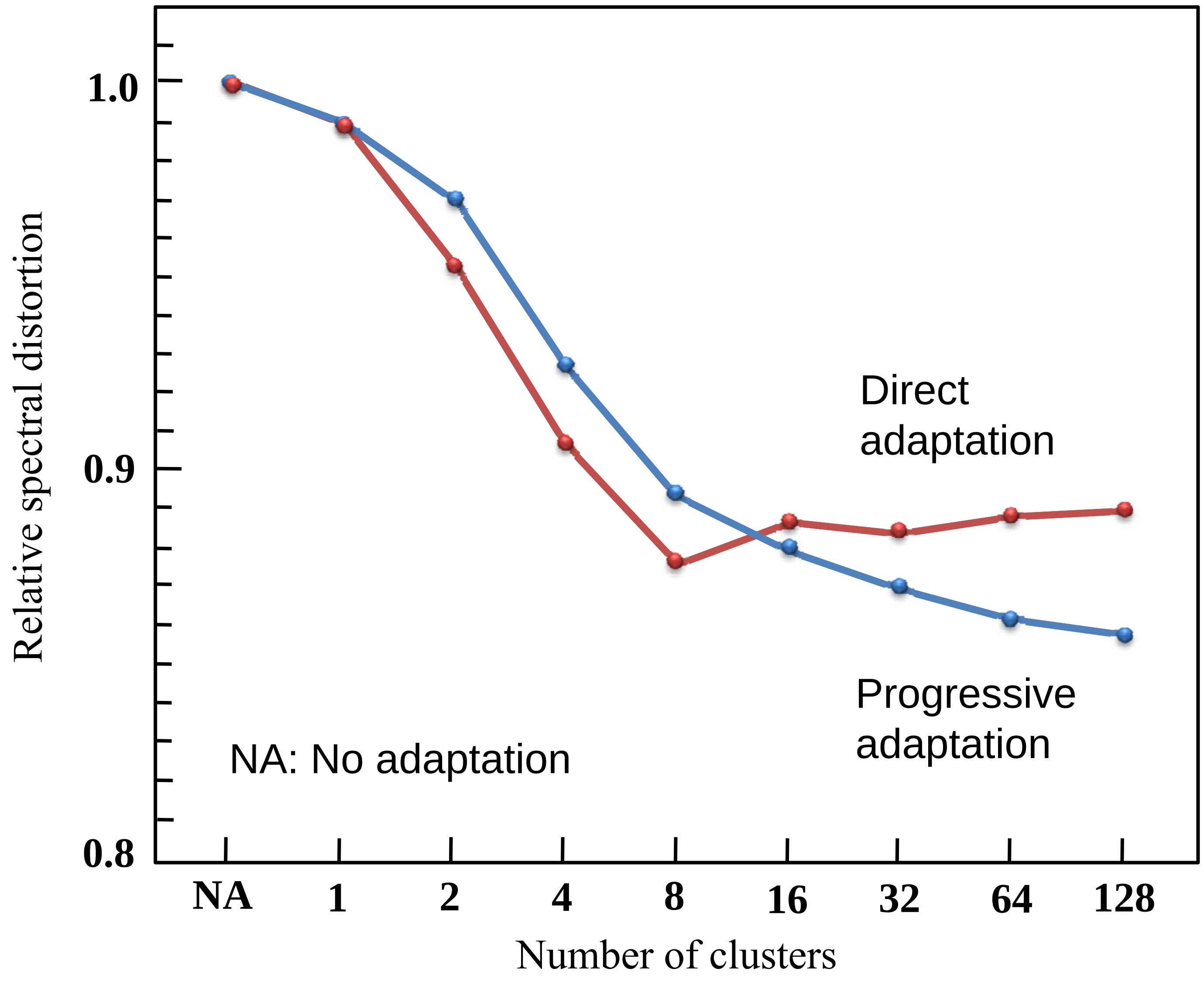
Structural approach

- Hierarchical codebook adaptation algorithm (Shiraki & Honda, 1987; Furui, 1989): a set of spectra in adaptation speech and the reference codebook elements are clustered hierarchically by increasing the number of clusters.
- Adaptation is performed hierarchically from the global variation characteristics down to the local ones.
- The method was extended to continuous HMMs (Matsui & Furui, 1998; Shinoda & Watanabe, 1995).
- Also applied to SMAP and SMAPLR methods.

Hierarchical codebook adaptation algorithm maintaining continuity between adjacent clusters



Cepstral distortion between input speech and reference templates resulted from hierarchical code-word adaptation



Transformation-based approaches

- The number of free parameters is limited by tying the HMM parameters or by applying some constraints on the parameters.
  - Cepstral mean normalization (CMN)
  - Maximum likelihood linear regression (MLLR)
  - Signal bias removal (SBR)
  - Constrained MLLR (CMLLR)
  - Interpolation
  - Vector field smoothing (VFS)

MLLR

- Most widely used transformation-based approach
- Originally the mean vectors in HMMs were modeled using an affine transformation (Leggetter & Woodland, 1995).
- Extended to update variances (Gales & Woodland, 1996).
- Gaussian distributions in HMM are clustered into phone classes, and transformation is shared.
- Regression class trees for robust clustering
- Signal bias removal (SBR) (Rahim and Juang, 1996) corresponds to a special case of MLLR.

MLLR (maximum likelihood linear regression) for adaptation of continuous density HMMs

$\hat{\mu} = \Gamma\zeta$

$\zeta = [\omega, \mu_1, \ldots \mu_n]'$: ($n$ +1)-dimensional extended mean vector

μ : n-dimensional mean vector

ω : offset term

$$\begin{cases} \omega = 1 : \text{include an offset in the regression} \\ \omega = 0 : \text{ignore offsets} \end{cases}$$

$\hat{\mu}$: adapted mean vector

Γ : n × (n +1) transformation matrix maximizing the likelihood of the adaptation data

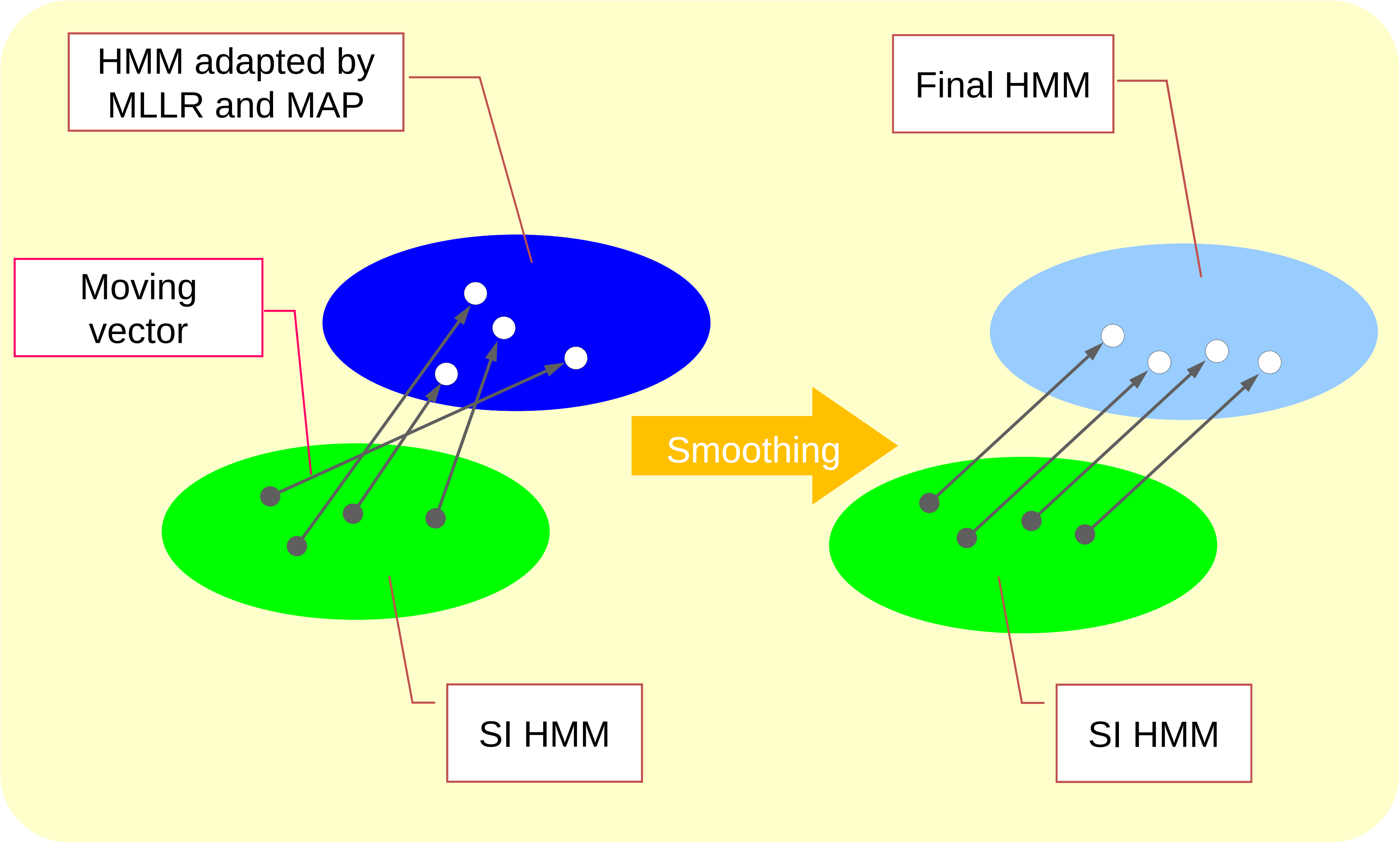
Constrained MLLR (CMLLR)

- The same transformation matrix is used for the covariance matrices and the mean vectors of HMMs (Digalakis et al., 1995; Gales, 1998).
- This method can be used not only for model adaptation but also as a feature adaptation technique.
- Usually implemented for diagonal covariance, continuous density HMMs, due to computational reasons.

Interpolation and VFS

- The bias of a parameter having no adaptation data was estimated by interpolating the biases of nearby parameters (Shinoda et al., 1991)
- Vector field smoothing (VFS): Correspondence of feature vectors between different speakers are viewed as a smooth vector field (Ohkura et al., 1992).
  - Interpolation and smoothing of the correspondence are introduced into the adaptation process to reduce "observation errors".

Vector Field Smoothing (VFS)

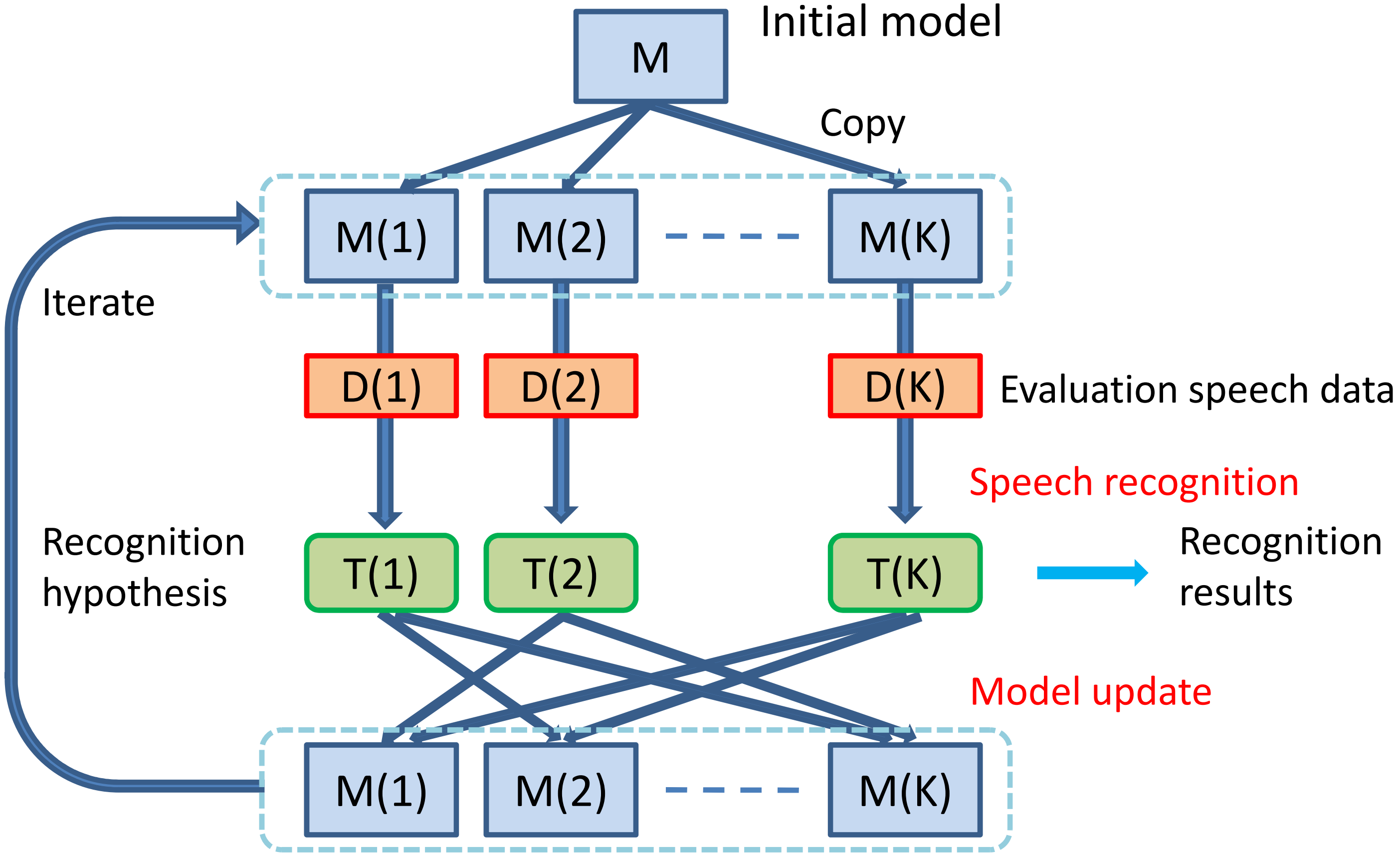


(Ohkura et al., 1992)

Cross validation (CV) and aggregated adaptation methods

- In both methods, adaptation utterances are split into K exclusive subsets, each with roughly the same size, to suppress the negative effects of recognition errors (Shinozaki et al., 2009).
- CV adaptation: Adaptation utterances used in the decoding step and those used in the model updating step are separated based on the K-fold CV technique.
- Aggregated adaptation: Each adaptation utterance set is decoded N times using separate models, based on the idea of the bagging approach.

Unsupervised cross-validation (CV) adaptation



- Reducing the influence of recognition errors by separating the data used for the decoding step and the model update step

Outline

1. Introduction
2. Model adaptation
3. Generalization problem
4. Constraining the degree of freedom by using a priori knowledge
5. Constraining the degree of freedom without using a priori knowledge
6. Combinations and extensions
7. Confidence measures
8. Special training methods for the models used for adaptation
9. Conclusion and future works

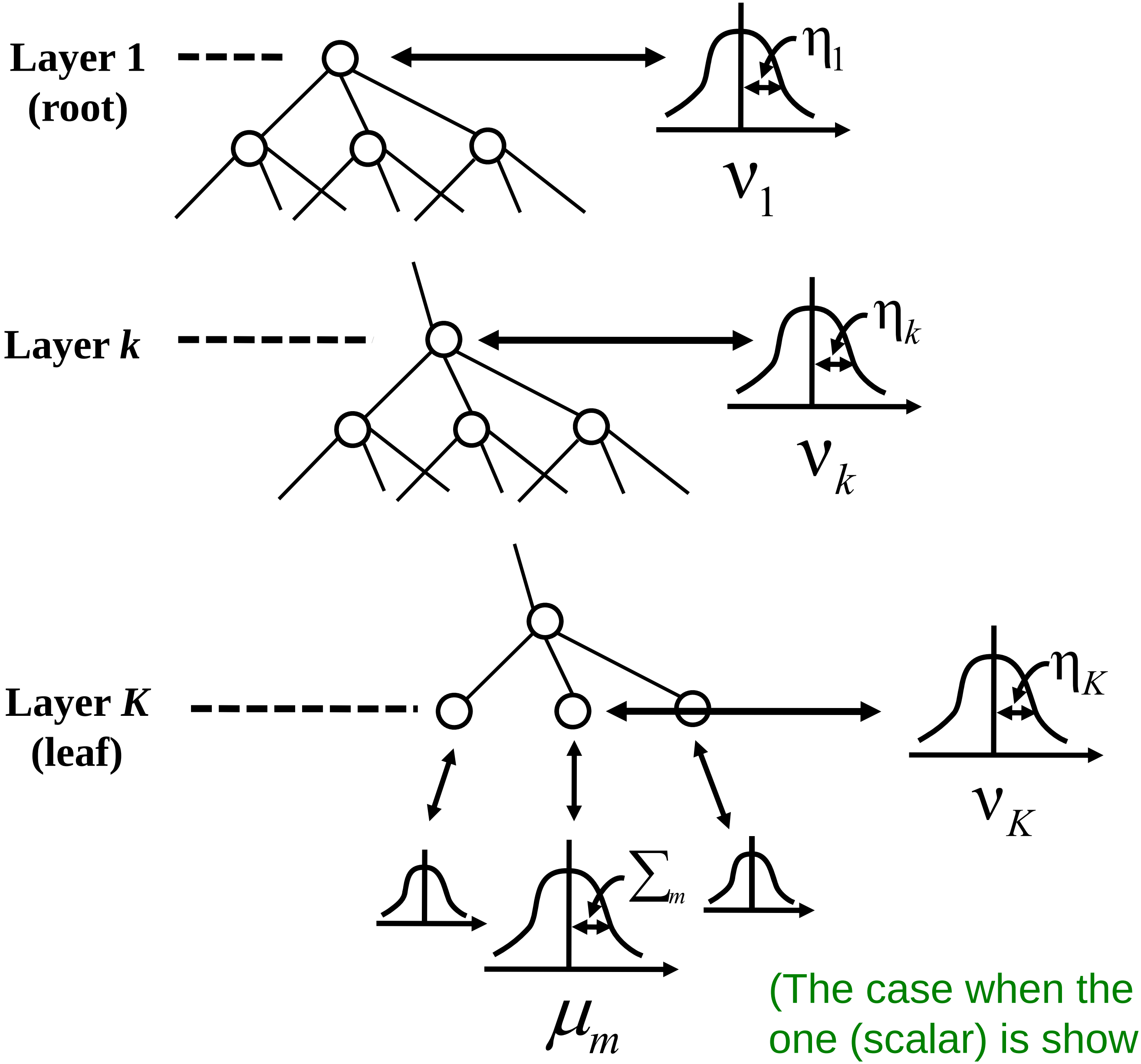
Combinations and extensions

- ML-based combinations
  - MAP + MLLR (MAPLR, etc)
  - MAP + structural method (SMAP)
  - MAP + VFS
  - MAP + affine transform + structural method (SMAPLR)
  - MLLR + eigen-voice method
- Discriminative approach based combinations
  - MAP + MCE
  - MAP + MMI/MPE

Structural MAP (SMAP)

- Combination of MAP estimation and the flexible parameter tying strategy (Shinoda & Lee, 2001).
- A hierarchical structure in the model space is made.
- Priors corresponding to child nodes in the tree are derived from the parent node.
- All the priors for all the HMM parameters are specified to perform efficient and effective adaptation.

Tree structure for Gaussian pdfs in continuous density HMMs used in the SMAP method



(The case when the dimension is one (scalar) is shown for simplicity)

N-best based method

- To reduce the effects of recognition errors in the hypothesis
- N-best list based instantaneous unsupervised adaptation using MAP method (Matsui & Furui, 1998)
- Smooth estimation and utterance verification are combined
- N-best list based Bayesian framework for MLLR (Yu & Gales, 2008)

Discriminative approach based combination and extension

- MAP+MCE (Matsui & Furui, 1995)
- Bayesian discriminative unsupervised adaptation (Discriminative MAP estimation) (Raut & Gales, 2009)
- I-smoothing (Interpolation between MLE and a discriminative objective function (MMI)) (Povey & Woodland, 2002)
- Large-margin discriminative training: equivalent to Support Vector Machines (SVM)

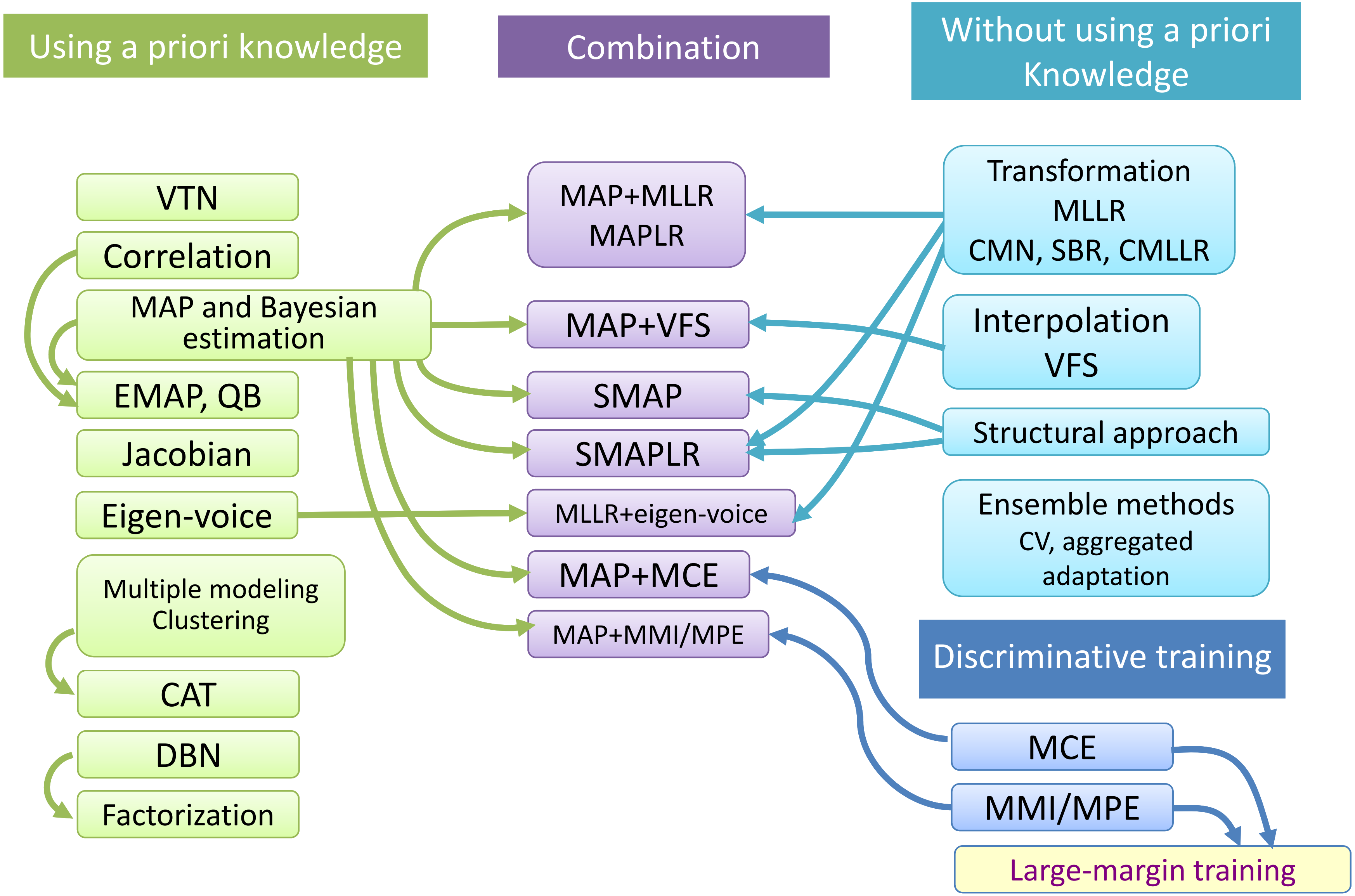
Bayesian discriminative adaptation (Discriminative MAP estimation)

- MAP Bayesian approach for discriminative unsupervised adaptation (Raut & Gales, 2009)
- Bayesian framework reduces the hypothesis bias and makes the discriminative adaptation less sensitive to hypothesis errors.
- Allows robust estimation of discriminative transforms.

Large-margin discriminative training

- Margin: distance between the well-classified samples and the decision boundary
  - Margin is directly maximized, or
  - Some form of combined scores of the margin and the empirical error rate is optimized.
- Sigmoid bias in the MCE training is interpreted as a soft margin and optimized (Yu et al., 2008).
- Standard MPE and MMI training has been extended to large-margin based methods (Heigold et al., 2008; Saon et al., 2008).

Constraining the degree of freedom with/without using a priori knowledge



Using a priori knowledge

Combination

Without using a priori Knowledge

VTN

Correlation

MAP and Bayesian estimation

EMAP, QB

Jacobian

Eigen-voice

Multiple modeling Clustering

CAT

DBN

Factorization

MAP+MLLR MAPLR

MAP+VFS

SMAP

SMAPLR

MLLR+eigen-voice

MAP+MCE

MAP+MMI/MPE

Transformation MLLR CMN, SBR, CMLLR

Interpolation VFS

Structural approach

Ensemble methods CV, aggregated adaptation

Discriminative training

MCE

MMI/MPE

Large-margin training

Outline

1. Introduction
2. Model adaptation
3. Generalization problem
4. Constraining the degree of freedom by using a priori knowledge
5. Constraining the degree of freedom without using a priori knowledge
6. Combinations and extensions
7. Confidence measures
8. Special training methods for the models used for adaptation
9. Conclusion and future works

Confidence measures (CMs)

- CMs are widely used in unsupervised adaptation to select more reliable speech segments (words or utterances).
- Posterior probability in the standard MAP decision rule is widely used.
- It is hard to precisely estimate the normalization term in the denominator.
- CM problem is sometimes formulated as a statistical hypothesis testing problem (likelihood ratio testing (LRT)) : utterance verification framework.
- Major difficulty with LRT is how to model the alternative hypothesis (general background model, hypothesis-specific anti-model, a set of competing models, etc.)

Outline

1. Introduction
2. Model adaptation
3. Generalization problem
4. Constraining the degree of freedom by using a priori knowledge
5. Constraining the degree of freedom without using a priori knowledge
6. Combinations and extensions
7. Confidence measures
8. Special training methods for the models used for adaptation
9. Conclusion and future works

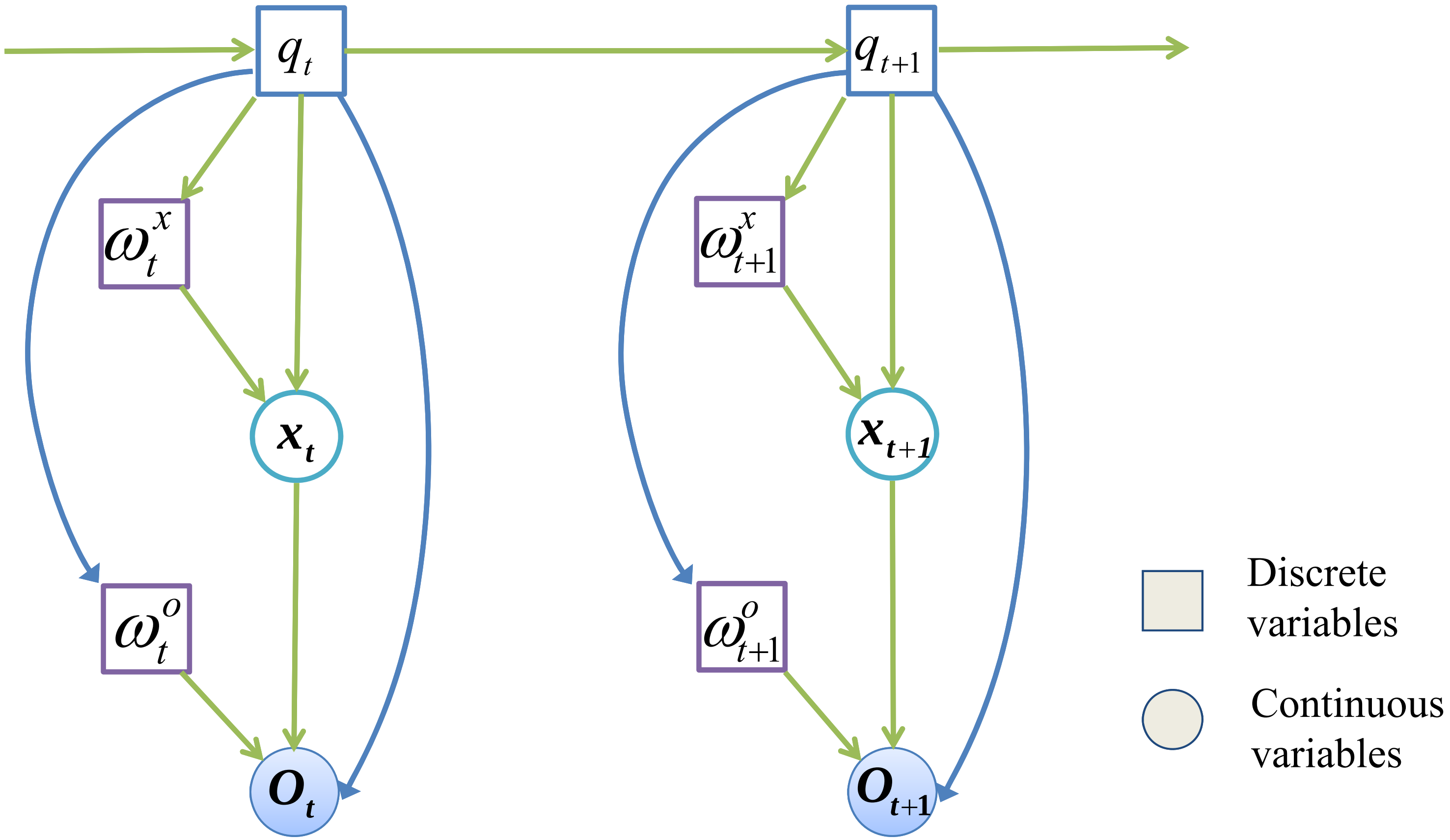
Special training methods for adaptation

- To make the adaptation process more effective or to keep the consistency between training and testing, special training methods have been investigated.
- Speaker adaptive training (SAT) (Anastasakos et al., 1996; Pye and Woodland, 1997)
  1. Mapping from each individual model to initial model is estimated.
  2. The mapping is applied to the data for each speaker.
  3. Mapped data is used to train the speaker-dependent model.
  4. This process is iterated until convergence.
  5. Canonical models represent only variability from individual speakers.

Acoustic factorization

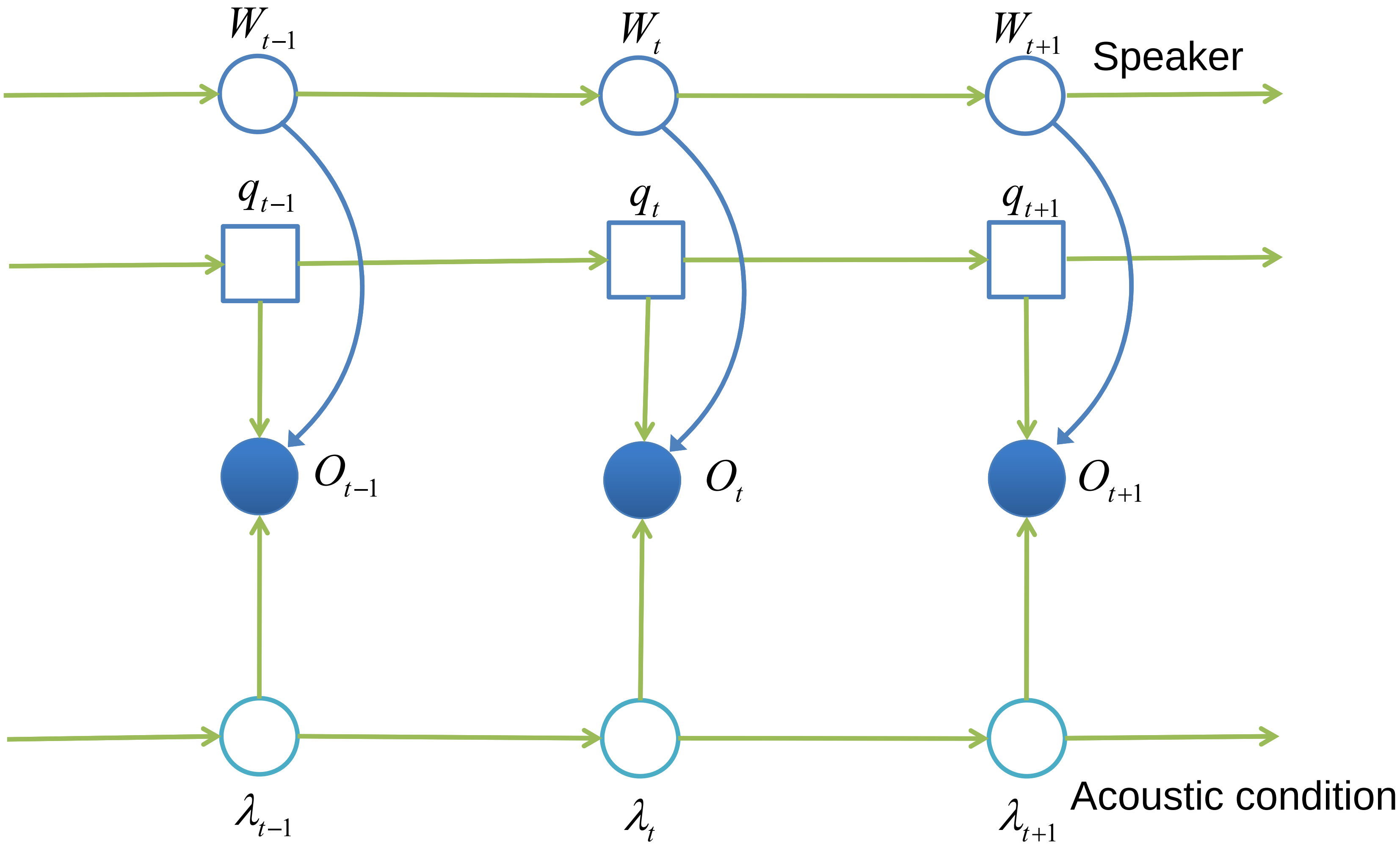
- Explicitly model all the factors affecting the acoustic signal (Rosti and Gales, 2002).
- The trained model set is expected to be used more flexibly than in standard SAT.
- It is possible to factor-in only those factors appropriate to a particular target domain.
- The target domain specific factors are simply estimated from limited target specific data.
- MLLR as the speaker transform and CAT as the noise (acoustic condition) transform (Gales, 2001).

DBN representing a factor analyzed HMM



x_t : state vector, O_t : observation vector, q_t : HMM state,
ω_t^x, ω_t^o : mixture indicator

DBN for acoustic factorization



Outline

1. Introduction
2. Model adaptation
3. Generalization problem
4. Constraining the degree of freedom by using a priori knowledge
5. Constraining the degree of freedom without using a priori knowledge
6. Combinations and extensions
7. Confidence measures
8. Special training methods for the models used for adaptation
9. Conclusion and future works

Conclusion and future works

- Human subjects produce one to two orders of magnitude fewer errors than machines.
- Human subjects are far more flexible and adaptive than machines against various variations of speech.
- How to train and adapt AMs using limited amounts of data: generalization problem.
  - Controlling the degree of freedom with/without using a priori knowledge
  - Maximizing “margins”
- There is no universal method.
- We need to know more about human speech processing and natural speech variation.
- Future systems need to have an efficient way of representing, storing, and retrieving various knowledge resources.