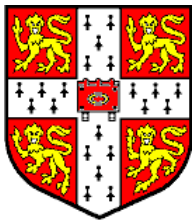
Acoustic Factorisation for Speech Recognition and Speech Synthesis

Mark Gales
work with Yongqiang Wang, Heiga Zen (Toshiba Research Europe Ltd)

March 2012

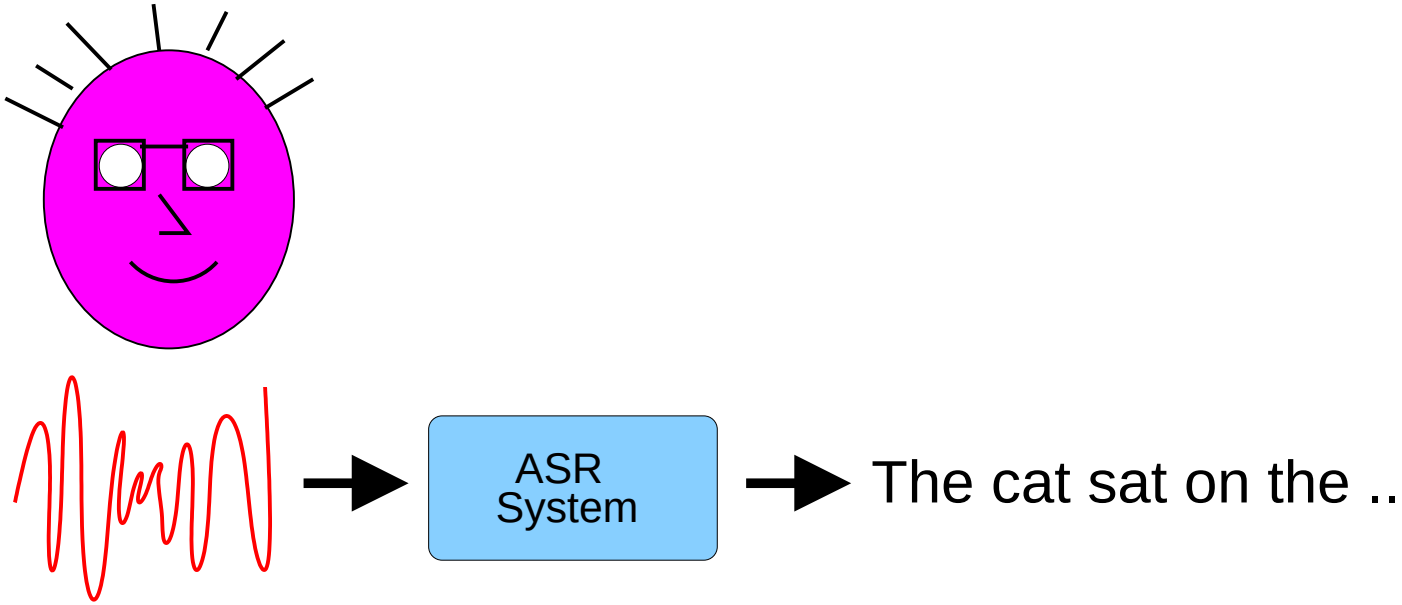


Cambridge University Engineering Department

Overview

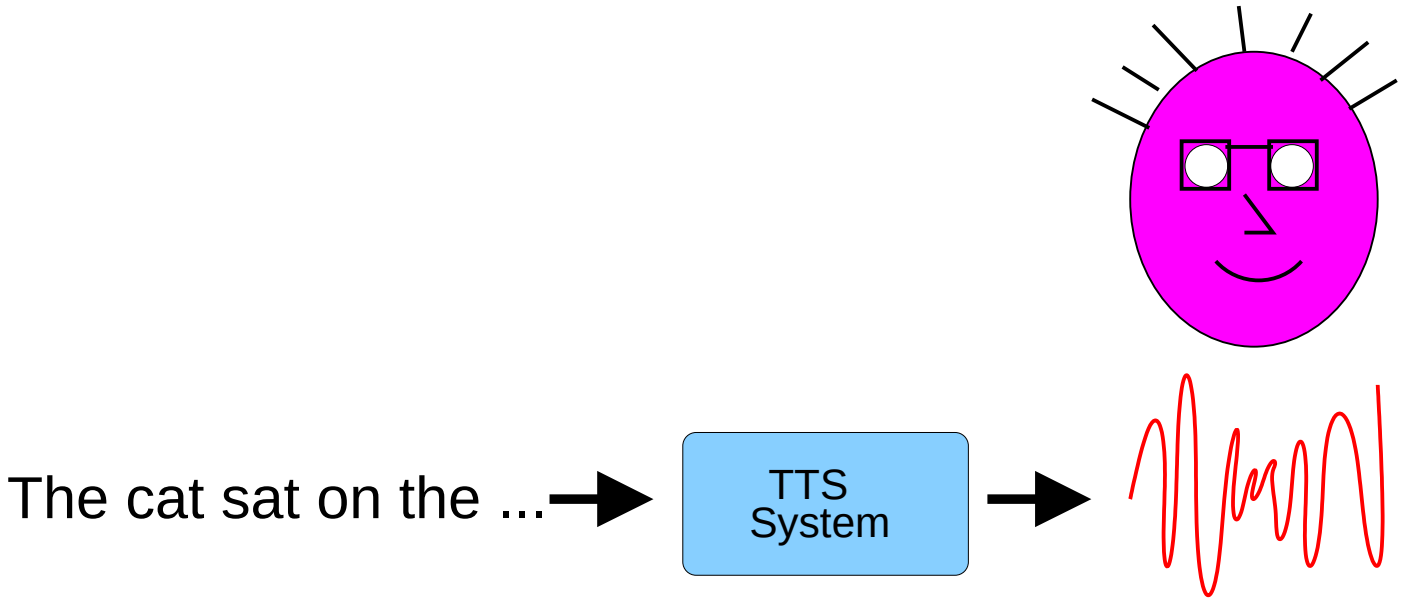
- Speech Recognition and Speech Synthesis Tasks
- Adaptation Approaches
  - linear transform-based adaptation
  - cluster adaptive training
  - model-based noise robustness
  - adaptive training
- Acoustic Factorisation
  - advantages of factorised approaches
  - "orthogonality" of transformations
- Acoustic Factorisation Examples
  - speaker and noise factorisation for speech recognition
  - speaker and language factorisation for polyglot synthesis

Speech Recognition (ASR/STT) as a Task



- Convert (parametrised) acoustic waveform $\boldsymbol{Y}$ into words $\boldsymbol{w}$
  - same "task" for all domain - recognition of words
  - but realisation of words impacted by multiple factors: speaker, noise, task differences
  - need to remove impact of factors on "clean" speech
  - output sentences highly dependent on domain

Speech Synthesis (TTS) as a Task



- Convert word sequence $\boldsymbol{w}$ into (raw) waveform $\boldsymbol{Y}$
  - highly specific task - synthesis of a particular voice
  - but realisation of words impacted by multiple factors: speaker, language, context, expressiveness
  - need to add impact of factors on "clean" speech

Speaker Differences

- Large differences between speakers
- Linguistic Differences e.g.
  - Accents
    tomato in RP/American English
  - Speaker idiosyncrasies
    either in English
  - non-native speaker
- Physiological Differences e.g.
  - physical attributes - gender,
    length of vocal tract
  - transitory effects
    cold/stress/public speaking

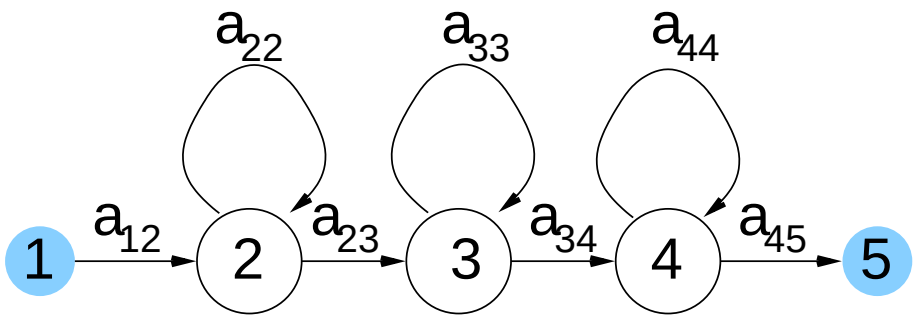
Environment Differences

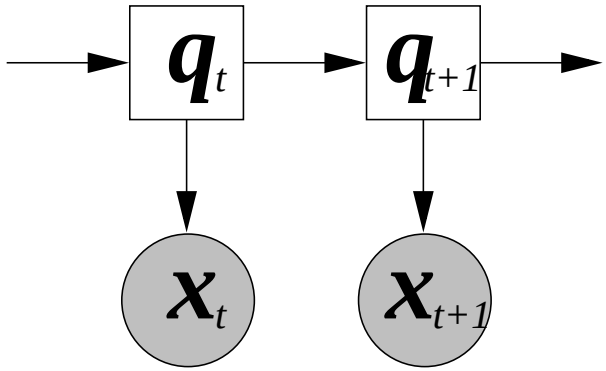
Limited Links with Machine Learning

- Speech researchers have examined robustness/adaptation for many years
  - key issue for achieving high performance speech recognition
  - increasingly important for high quality controllable speech synthesis
- Mismatches in training/test scenarios not unique to speech
  - general issue for a range of tasks, domain adaptation
- Speech research has generally adopted speech-specific approaches
  - take advantage of parametric nature of acoustic models
  - take advantage of physical attributes of acoustic factors

Interesting question whether general approaches will help?

Acoustic Model - Hidden Markov Model

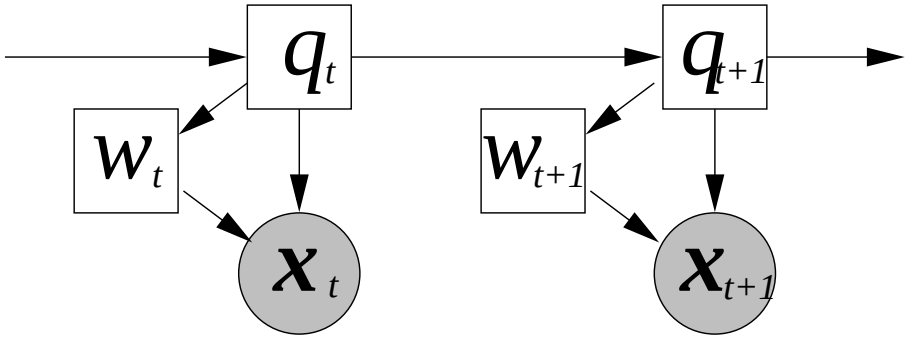




- HMMs dominate state-of-the-art speech recognition
  - "clean" speech, $\boldsymbol{x}_t$ generated by HMM (standard assumption)
  - not ideal (understatement) but good enough (sometimes)
  - form considered in this talk
- Statistical parametric synthesis uses closely related model
  - static "trajectory" obtained using delta and delta-delta parameters

Same Approaches can be Applied for Both

(Standard) ASR State Output Distribution



- Observations conditionally independent of other observations given state.

$$p(\boldsymbol{x}_t|q_t, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}) = p(\boldsymbol{x}_t|q_t) = \sum_{m=1}^{M} c_{\mathtt{x}}^{(m)} \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}\right)$$

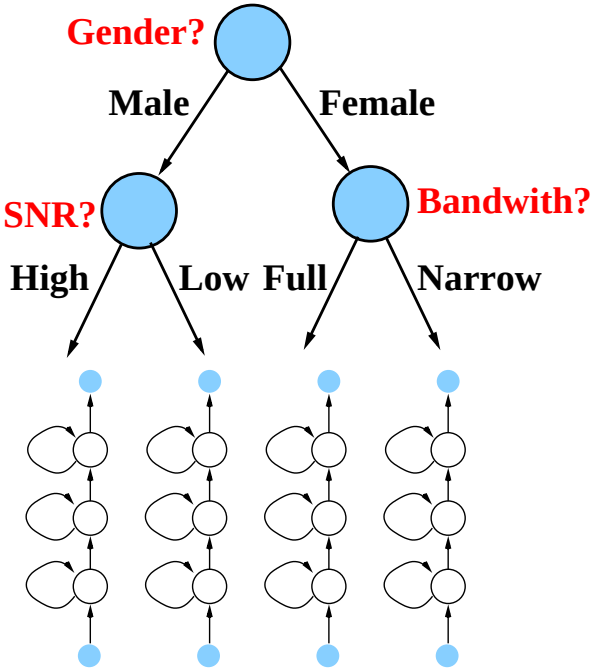
- State-of-the-art systems typically very large - for ASR typically
  - 39-dimensional feature space (PLP/MFCC + delta/delta-deltas/HLDA/STC)
  - state-clustered triphone acoustic models (6K-10K distinct states)
  - 36 components/distinct state - **upto 28 million parameters**

and Clean Speech Never Observed in Training

Training Multiple Models

- ASR (and increasingly TTS) systems trained on thousands of hours of data
  - data collected from multiple speakers/environments/bandwiths
  - single model trained on all data too "broad"

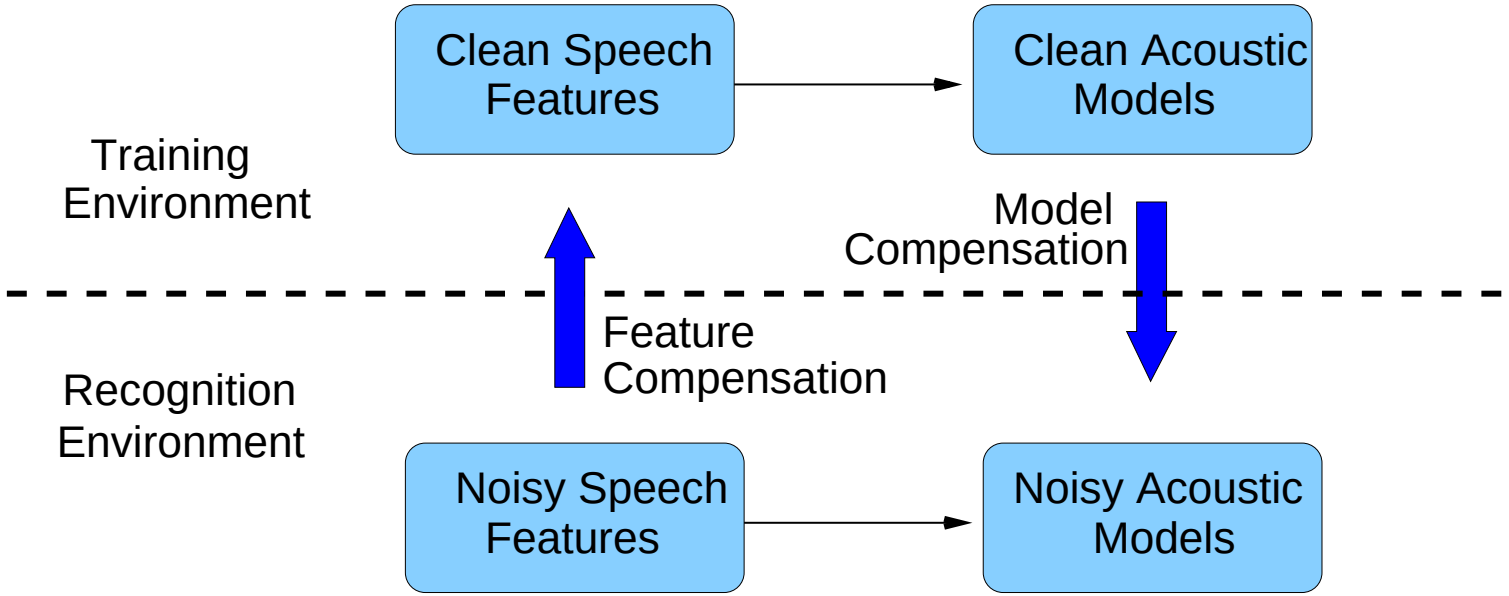
Simply train many acoustic models!



- Split data into blocks
  - train separate models for each block
- How many blocks?
  - more blocks better "resolution"
  - BUT systems have millions of parameters
  - sufficient data for robust estimates
  - still large spread of data/block

Adaptation Approaches

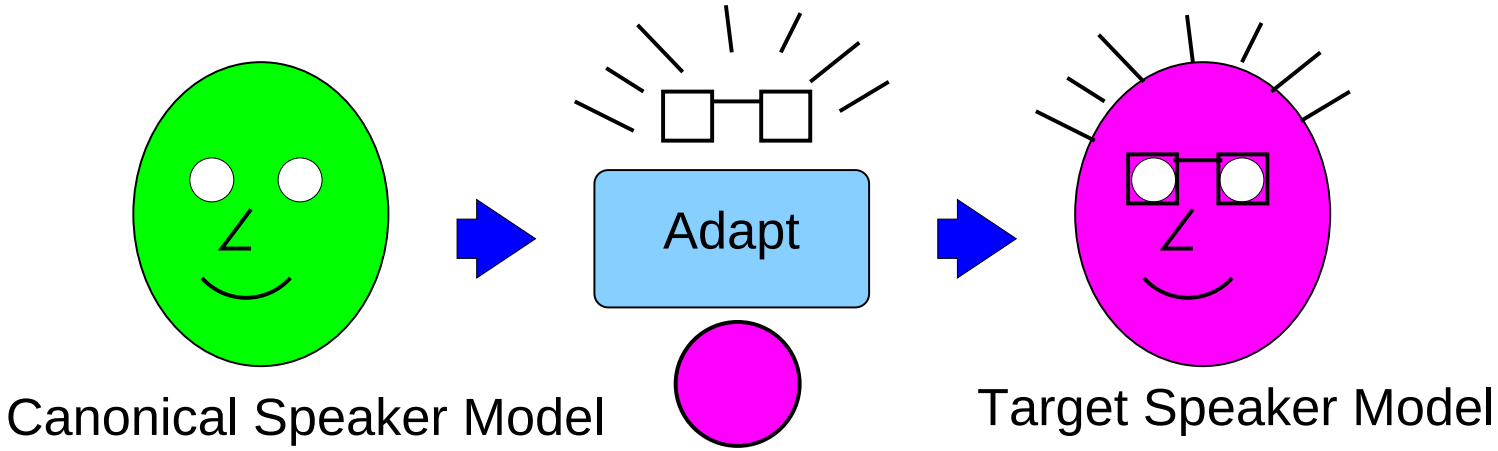
Adaptation Approaches - Robust ASR



- Two main approaches to adaptation:
  - feature compensation: "clean" the noisy features
  - model compensation: "corrupt" the clean models
- This work concentrates on model compensation approaches

Model Adaptation Process

- Aim: modify a "canonical" model to represent a target speaker or domain
  - should require minimal data from the target scenario
  - should accurately represent target scenario speech



- Need to determine
  - nature (and complexity) of the adaptation transformation
  - how to train the "canonical" model that is adapted

Forms of Model Adaptation

$$\sum_{m=1}^{M} c_{\mathrm{x}}^{(m)} \mathcal{N}(\boldsymbol{\mu}_{\mathrm{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathrm{x}}^{(m)}) \rightarrow \sum_{n=1}^{N} c_{\mathrm{y}}^{(n)} \mathcal{N}(\boldsymbol{\mu}_{\mathrm{y}}^{(n)}, \boldsymbol{\Sigma}_{\mathrm{y}}^{(n)})$$

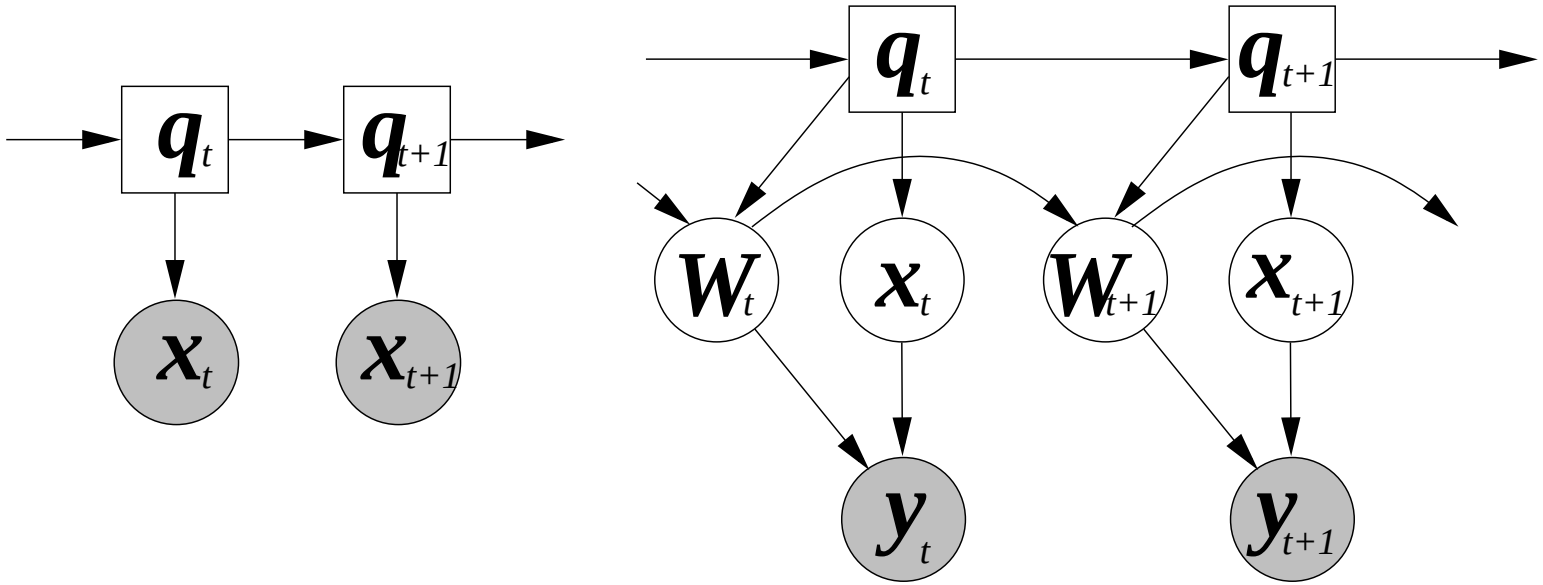
- Adaptive Compensation: general transform
  - not specifically related to particular form of distortion
  - often limited to (piecewise) linear transformation
  - typically requires large number of model parameters
- Predictive Compensation: use "model" of acoustic factor
  - impact of distortions explicitly represented
  - requires definition of (approximate) mismatch function
  - often non-linear in nature
  - typically very low dimensional representation

Adaptation Examples

- Adaptive Approaches examples:
  - Maximum A-Posteriori MAP [1] adaptation: general "robust" estimation
  - Cluster Selection: Gender-dependent (GD) models
  - Cluster Interpolation: combine multiple cluster parameters
    EigenVoices[2], CAT [3] more complex (interesting) forms.
  - Linear Transform Adaptation: dominant form for LVCSR
    linear transform comprises: transformation $\mathbf{A}^{(s)}$ and bias $\mathbf{b}^{(s)}$

- Predictive Approaches examples:
  - Vocal Tract Length Normalisation: motivated from physiological perspective
  - Vector Taylor Series Compensation: model-based environment compensation

Linear Transform DBN



- Only "corrupted" speech $\boldsymbol{y}_t$ observed
  - unable to train/use standard model on left as "clean" speech $\boldsymbol{x}_t$ unknown
- Introduce dependence on linear transform $\mathbf{W}_t$ (adaptive HMM [4])
  - observed data of form: $\boldsymbol{y}_t = \boldsymbol{f}(\boldsymbol{x}_t, \mathbf{W}_t)$
  - transform same for each homogeneous block ($\mathbf{W}_t = \mathbf{W}_{t+1}$)

Form of the Adaptation Transform

- Dominant form for LVCSR are ML-based linear transformations
  - MLLR adaptation of the means [5]

$$\boldsymbol{\mu}_{\mathrm{y}}^{(s)} = \mathbf{A}\boldsymbol{\mu}_{\mathrm{x}} + \mathbf{b}^{(s)}$$

  - MLLR adaptation of the covariance matrices [6, 7]

$$\boldsymbol{\Sigma}_{\mathrm{y}}^{(s)} = \mathbf{H}^{(s)}\boldsymbol{\Sigma}_{\mathrm{x}}\mathbf{H}^{(s)\mathsf{T}}$$

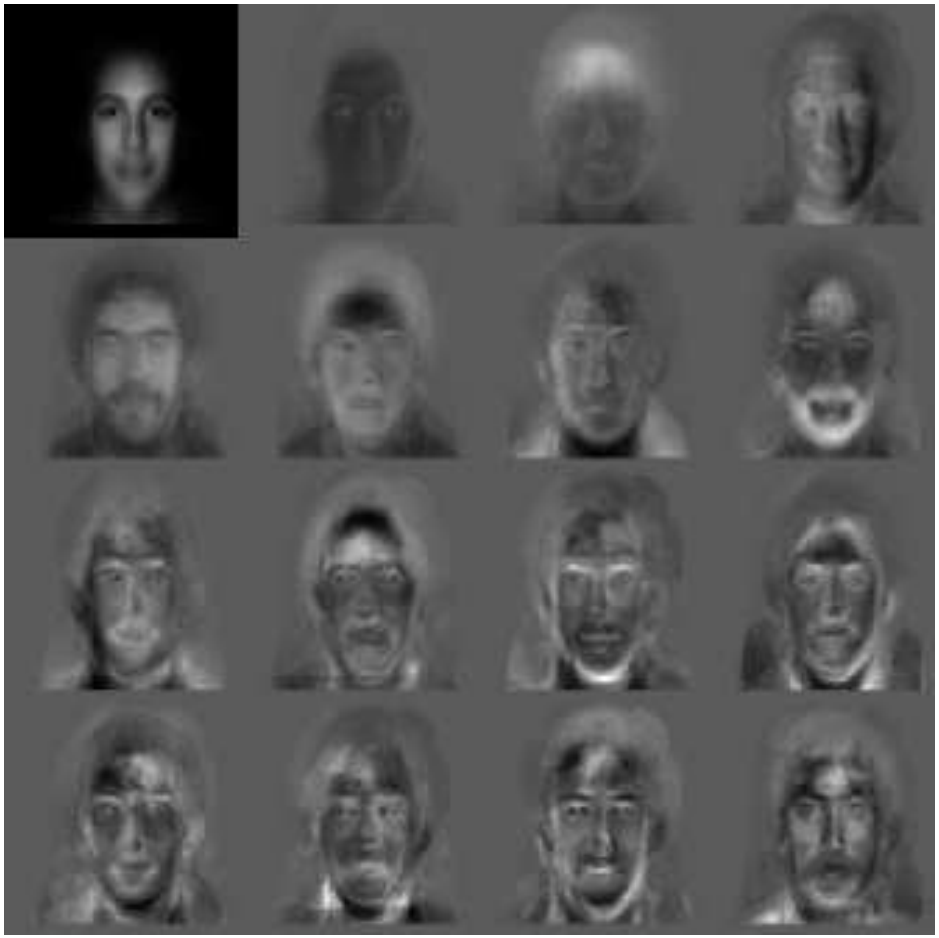
  - Constrained MLLR adaptation [7]

$$\boldsymbol{\mu}_{\mathrm{y}}^{(s)} = \mathbf{A}^{(s)}\boldsymbol{\mu}_{\mathrm{x}} + \mathbf{b}^{(s)}; \quad \boldsymbol{\Sigma}_{\mathrm{y}}^{(s)} = \mathbf{A}^{(s)}\boldsymbol{\Sigma}_{\mathrm{x}}\mathbf{A}^{(s)\mathsf{T}}$$

- Forms may be combined into a hierarchy [8] e.g.

$$\texttt{CMLLR} \rightarrow \texttt{MLLRMEAN}$$

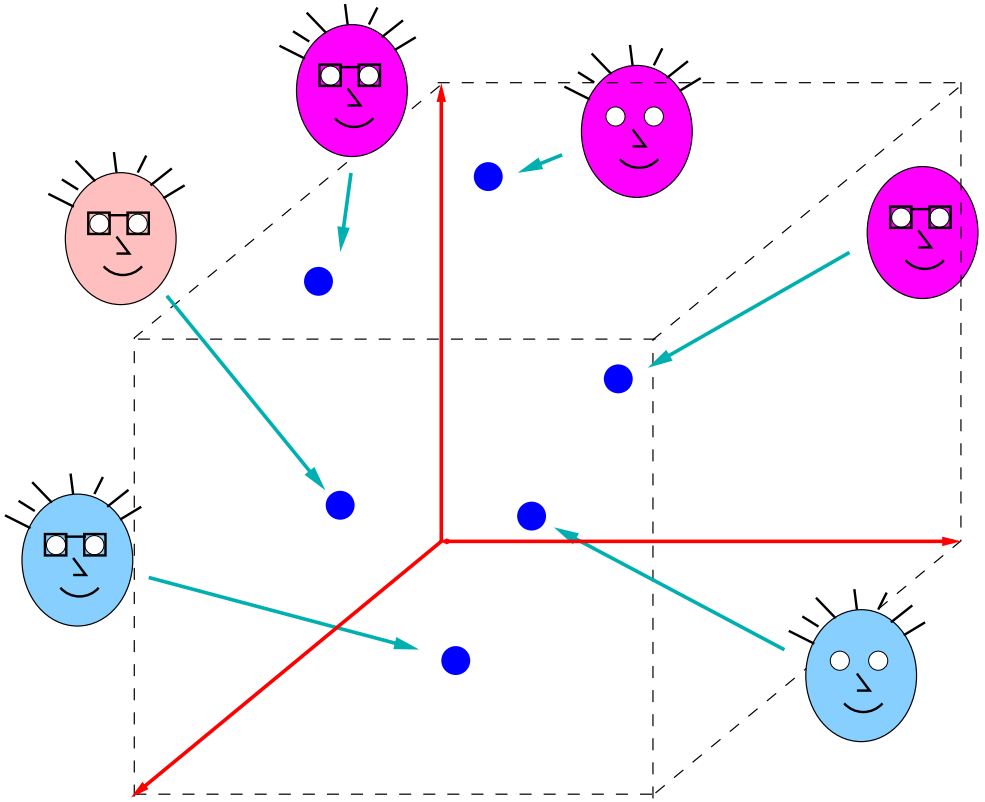
EigenFaces



- Developed for face recognition
- Estimate the average face
- Dimensions yield "face" variability
  - combine dimensions to yield a face
  - any face represented as a point

Apply same concept to speaker adaptation

Cluster Adaptive Training (EigenVoices)



- The dimensionality of the space is

  $$100\text{K}\ (\texttt{comp}) \times 39\ (\texttt{dim}) = 3.9\text{M}$$

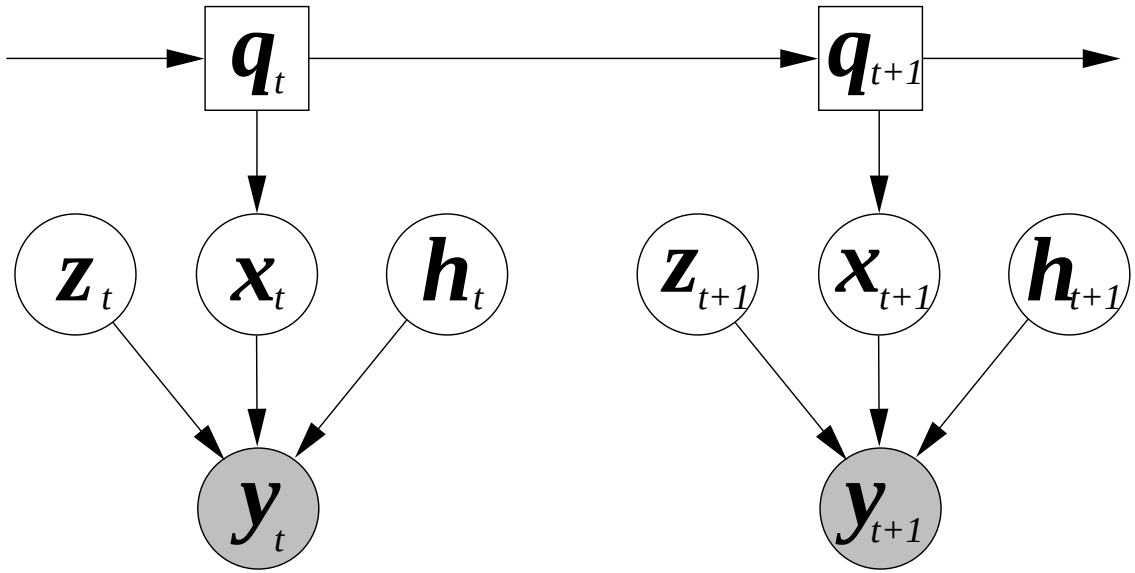
- Low-dimensional (3-10) subspace
- Each speaker represented by a point $\boldsymbol{\lambda}$ in the subspace
  - the speaker specific mean is

$$\boldsymbol{\mu}_{\mathsf{y}}^{(s)} = \boldsymbol{\mu}_{\mathsf{b}} + \sum_{i=1}^{P} \lambda_i^{(s)} \mathbf{c}_i$$

- CAT yields complete ML estimation (PCA for original EigenVoices) [2, 3]

Each speaker specified by only 3-10 parameters!

Predictive Acoustic Environment Corrupted DBN



- "Clean" speech distorted by:
  - z_t: additive noise
  - h_t: convolutional distortion (reverberant effects can also be considered)
- Noise terms assumed to have no temporal structure (not a requirement)

Mismatch Functions

- Standard assumption about impact of noise on "clean" speech [9]:

$$\boldsymbol{y}_t = \mathbf{C} \log \left(\exp(\mathbf{C}^{-1}\boldsymbol{x}_t + \mathbf{C}^{-1}\boldsymbol{h}_t) + \exp(\mathbf{C}^{-1}\boldsymbol{z}_t) \right) = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{n}_t, \boldsymbol{h}_t)$$

  $\mathbf{C}$ is the DCT, magnitude-based Cepstra

  - non-linear relationship between the clean speech, noise and corrupted speech
  - not possible to get simple expression for all parametrisation

- This has assumed sufficient smoothing to remove all "cross" terms

  - some sites use interaction likelihoods or phase-sensitive functions [10, 11]
  - given $\boldsymbol{x}_t, \boldsymbol{h}_t$ and $\boldsymbol{n}_t$ there is a distribution

$$\boldsymbol{y}_t \sim \mathcal{N}\left(\boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{n}_t, \boldsymbol{h}_t), \boldsymbol{\Phi}\right)$$

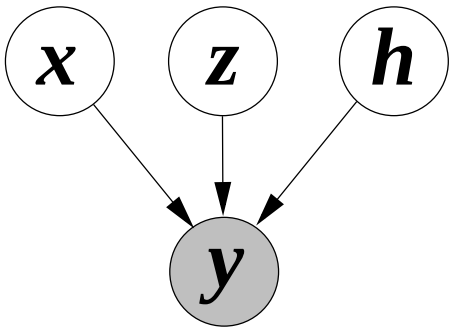
Linearised Mismatch Functions

- To simplify estimation linearise function about an expansion point, $\mathcal{T}^{(m)}$, [12]

$$\boldsymbol{y}|m = \boldsymbol{\Lambda}_{\mathsf{x}}^{(m)}\boldsymbol{x} + \boldsymbol{\Lambda}_{\mathsf{z}}^{(m)}\boldsymbol{z} + \boldsymbol{\Lambda}_{\mathsf{h}}^{(m)}\boldsymbol{h} + \boldsymbol{\epsilon}$$

parameters of the Gaussian noise, $\boldsymbol{\epsilon}$, determined by expansion point

$$\boldsymbol{x} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{\mathsf{x}}^{(m)}, \hat{\boldsymbol{\Sigma}}_{\mathsf{x}}^{(m)}), \quad \boldsymbol{z} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{\mathsf{z}}, \hat{\boldsymbol{\Sigma}}_{\mathsf{z}}), \quad \boldsymbol{h} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{\mathsf{h}}, \hat{\boldsymbol{\Sigma}}_{\mathsf{h}})$$

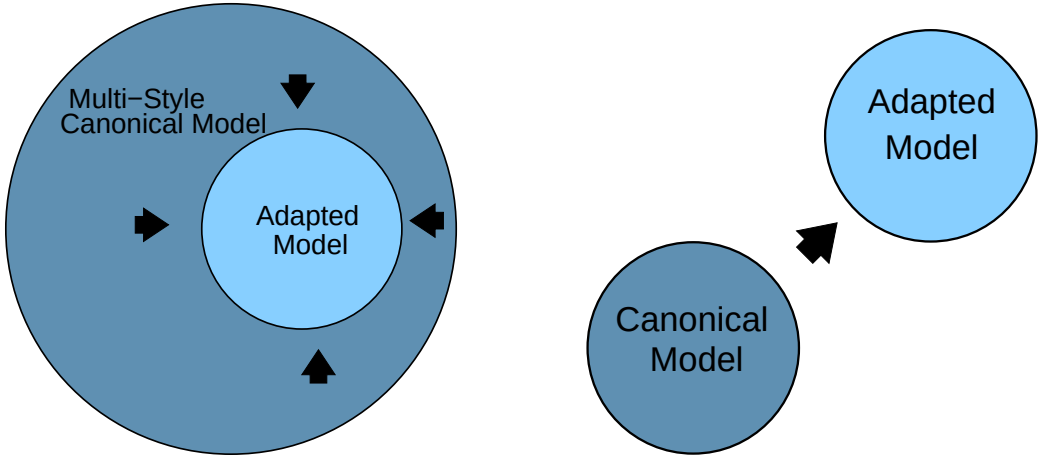


- Looks like generalised factor analysis [13]
  - loading matrices have the form:

$$\boldsymbol{\Lambda}_{\mathsf{x}}^{(m)} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}^{\mathsf{T}}}\right|_{\mathcal{T}^{(m)}}, \quad \boldsymbol{\Lambda}_{\mathsf{z}}^{(m)} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}^{\mathsf{T}}}\right|_{\mathcal{T}^{(m)}} = \mathbf{I} - \mathbf{J}_{\mathsf{x}}^{(m)}$$

Training a "Good" Canonical Model

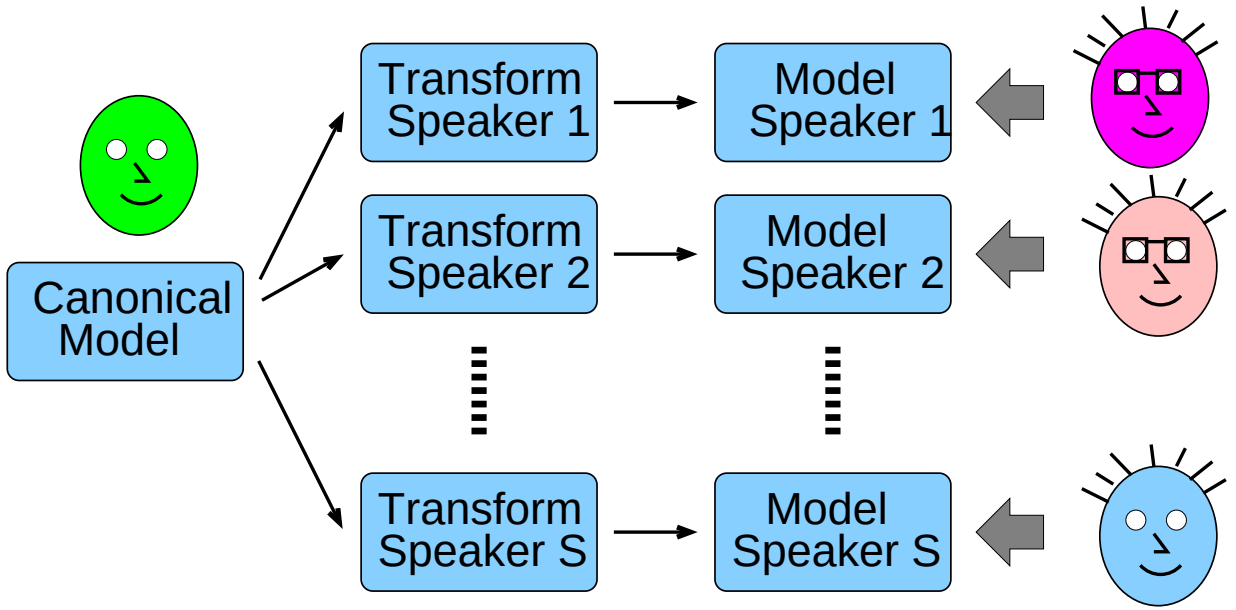
- Need to estimate model parameters for the clean speech
  - for both general and environment conditions, clean speech, $\boldsymbol{x}_t$, unobserved
  - how to estimate the clean speech models $\mathcal{M}_{\mathrm{x}}$



Two different forms of canonical model:

- Multi-Style: treat observed data $\boldsymbol{y}_t$ as the clean speech
- Adaptive: attempt to estimated underlying clean model [14, 7, 15]

Speaker Adaptive Training



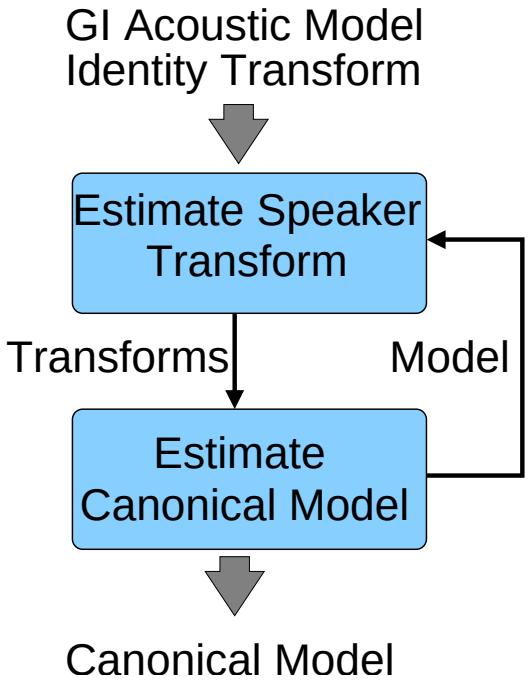
- In adaptive training the training corpus is split into "homogeneous" blocks
  - use adaptation transforms to represent unwanted acoustic factors
  - canonical model only represents desired variability
- All forms of linear transform can be used for adaptive training
  - CMLLR adaptive training highly efficient

CMLLR Adaptive Training

- The CMLLR likelihood may be expressed as [7]:

$$p(\boldsymbol{y}_t^{(s)}|\hat{\mathcal{M}}_{\mathtt{x}}, \mathcal{M}_{\mathtt{s}}^{(s)}, m) = |\mathbf{A}^{(s)}|\mathcal{N}(\mathbf{A}^{(s)}\boldsymbol{y}_t + \mathbf{b}^{(s)}; \hat{\boldsymbol{\mu}}_{\mathtt{x}}^{(s)}, \hat{\boldsymbol{\Sigma}}_{\mathtt{x}}^{(m)})$$

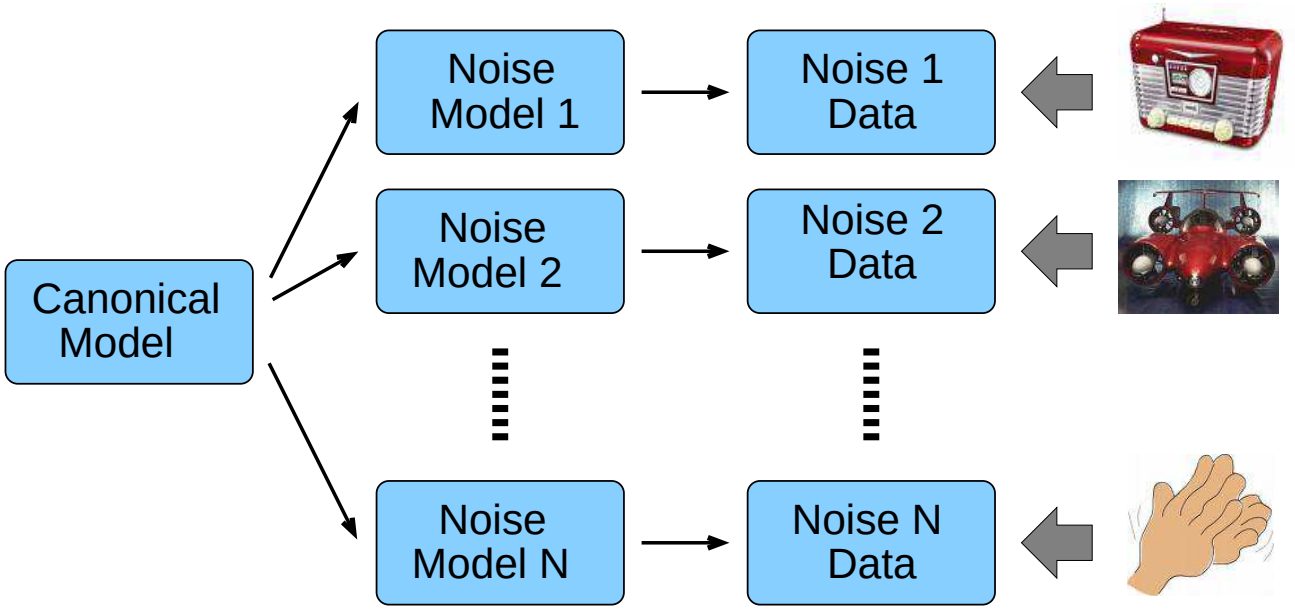
same as feature normalisation - simply train model in transformed space



- Interleave Model and transform estimation
- Update formulae for mean

$$\hat{\boldsymbol{\mu}}_{\mathtt{x}}^{(m)} = \frac{\sum_{s,t} \gamma_t^{(sm)} \left(\mathbf{A}^{(s)}\boldsymbol{y}_t + \mathbf{b}^{(s)}\right)}{\sum_{s,t} \gamma_t^{(sm)}}$$

Noise Adaptive Training

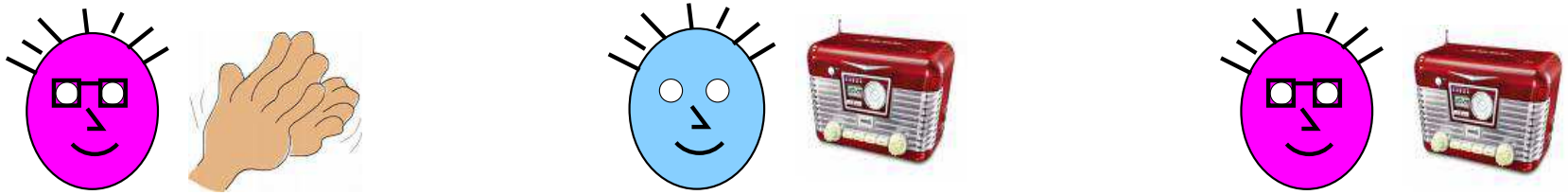


- Recently adaptive training to noise conditions also considered [16, 17, 18]
  - impact of noise differences (often) larger than speaker differences
  - harder due to the non-linear impacts of noise
- Linearised forms of mismatch functions used
  - generalised FA-style approaches, or gradient descent to train models

Acoustic Factorisation

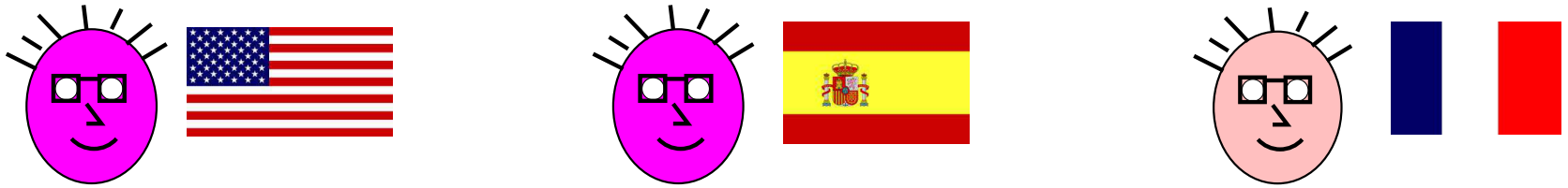
Multiple Acoustic Factors

- In most scenarios multiple acoustic factors impact the signal
  - speaker and noise:



the same speaker may be observed in multiple noise conditions

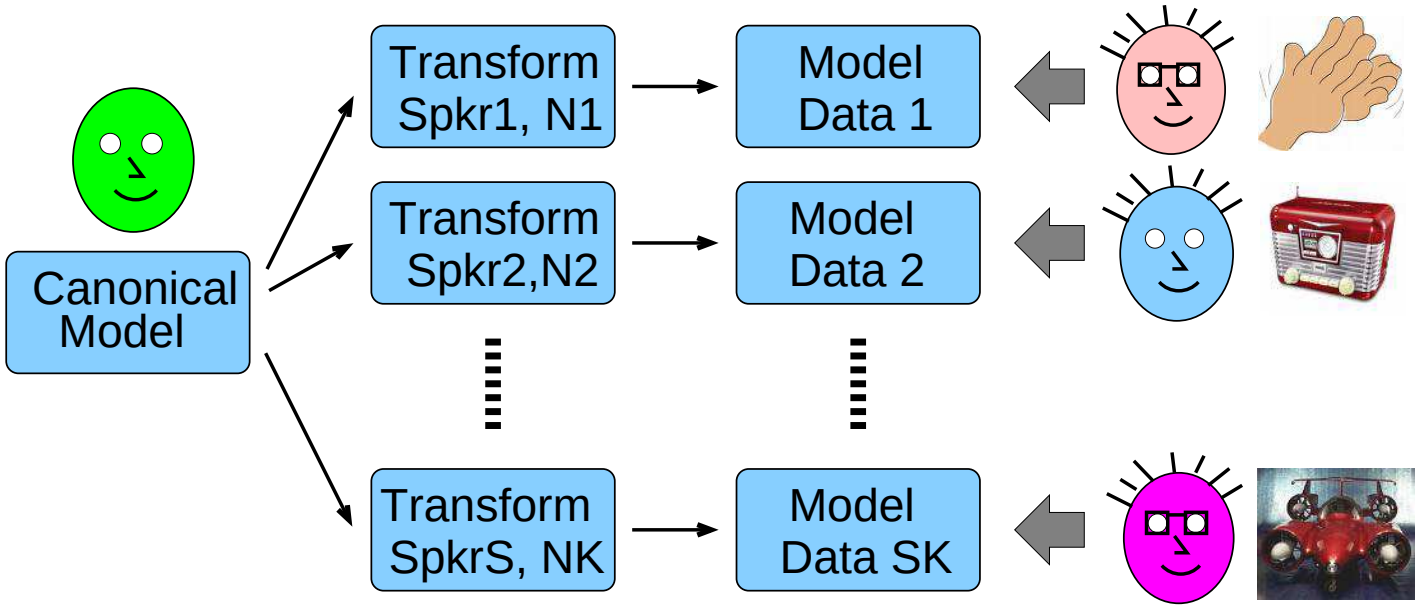
  - speaker and language:



the same speaker characteristics will be perceived irrespective of language

How to Use/Estimate Transforms in this Case?

Standard Approach



- The standard approach is estimating a transform for speaker/noise pairs

$$\mathcal{M}_{\mathtt{f}}^{(sn)} = \text{argmax}\left\{p(\boldsymbol{Y}^{(sn)}|\mathcal{H};\mathcal{M},\mathcal{M}_{\mathtt{x}})p(\mathcal{M})\right\}$$

- BUT ignores aspects of speaker/noise relationships

How to Incorporate this Information?

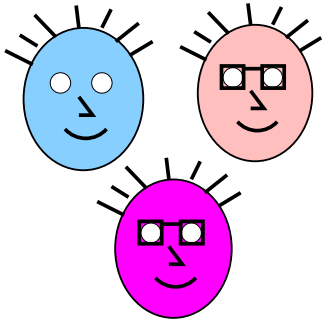
Acoustic Factorisation

- Conceptually the process is very easy [19]

$$\mathcal{M}_{\mathtt{f}}^{(sn)} = \mathcal{M}_{\mathtt{s}}^{(s)} \otimes \mathcal{M}_{\mathtt{n}}^{(n)}$$

  - form of transform for the speaker $\mathcal{M}_{\mathtt{s}}$
  - form of transform for the environment $\mathcal{M}_{\mathtt{n}}$

- Aim is to avoid exponential growth of number of transforms
  - transforms assigned to specific acoustic factors



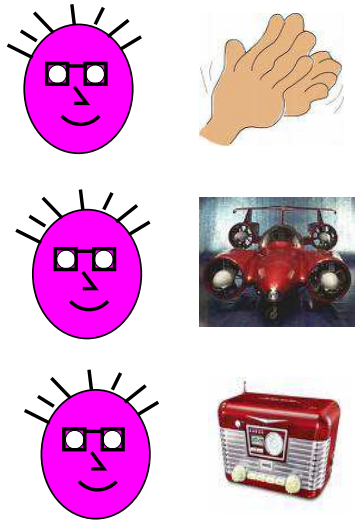

"Factoring-In" for ASR

- If individual acoustic factors modelled - highly flexible
  - choose which factors to specify
  - which factors to marginalise
- Consider simple case of speaker and noise acoustic factors
  - unknown speaker: need to be speaker-independent
  - known environment: need to be noise specific, $\mathcal{M}_{\mathrm{n}}^{(n)}$
- Simple to enforce with the factorised model:

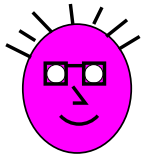
$$p(\boldsymbol{Y}|\mathcal{M}_{\mathrm{x}}, \mathcal{M}_{\mathrm{n}}^{(n)}) = \int p(\boldsymbol{Y}|\mathcal{M}_{\mathrm{x}}, \mathcal{M}_{\mathrm{s}} \otimes \mathcal{M}_{\mathrm{n}}^{(n)}) p(\mathcal{M}_{\mathrm{s}}) d\mathcal{M}_{\mathrm{s}}$$

  - (posterior) distribution $p(\mathcal{M}_{\mathrm{s}})$ obtained from training data
  - Bayesian approaches required - challenging for this data/models!

Example (1) - "Practical" Speaker Enrolment



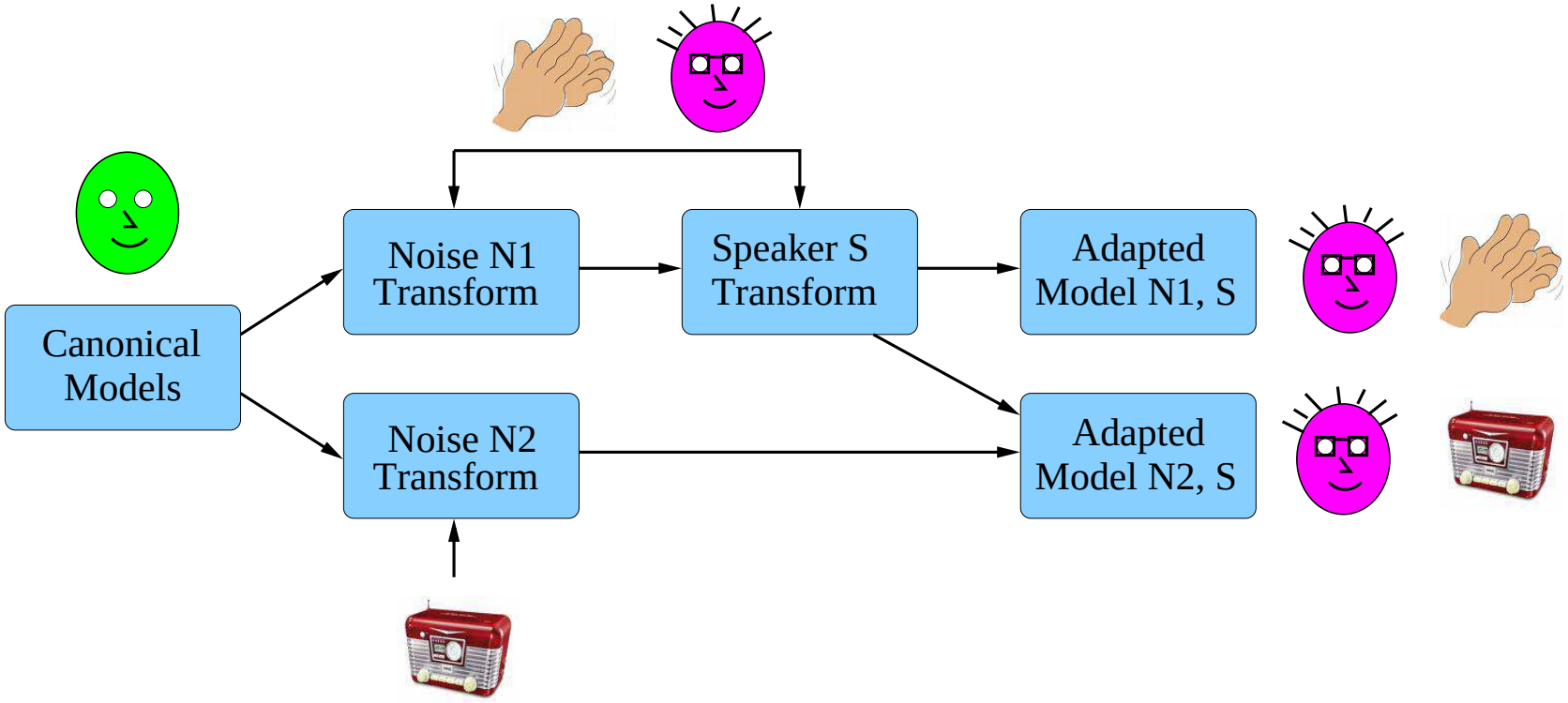
Training Data



Speaker Transform

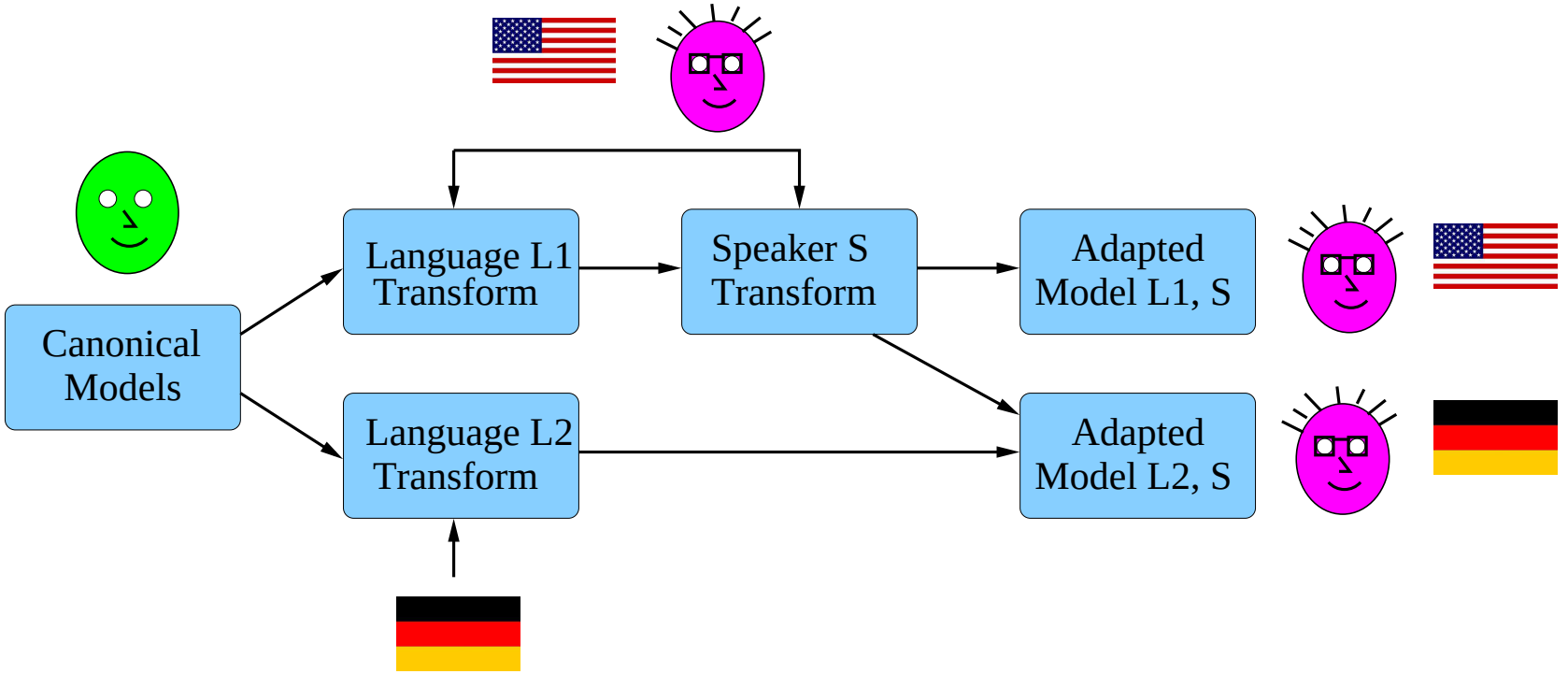
- Often only see data from a speaker with varying acoustic condition
  - consider in-car navigation system/ recording session variability
- Canonical speaker transform required
  - recognition in different environments [20]/speaker identification [21]

Example (2) - Rapid Adaptation



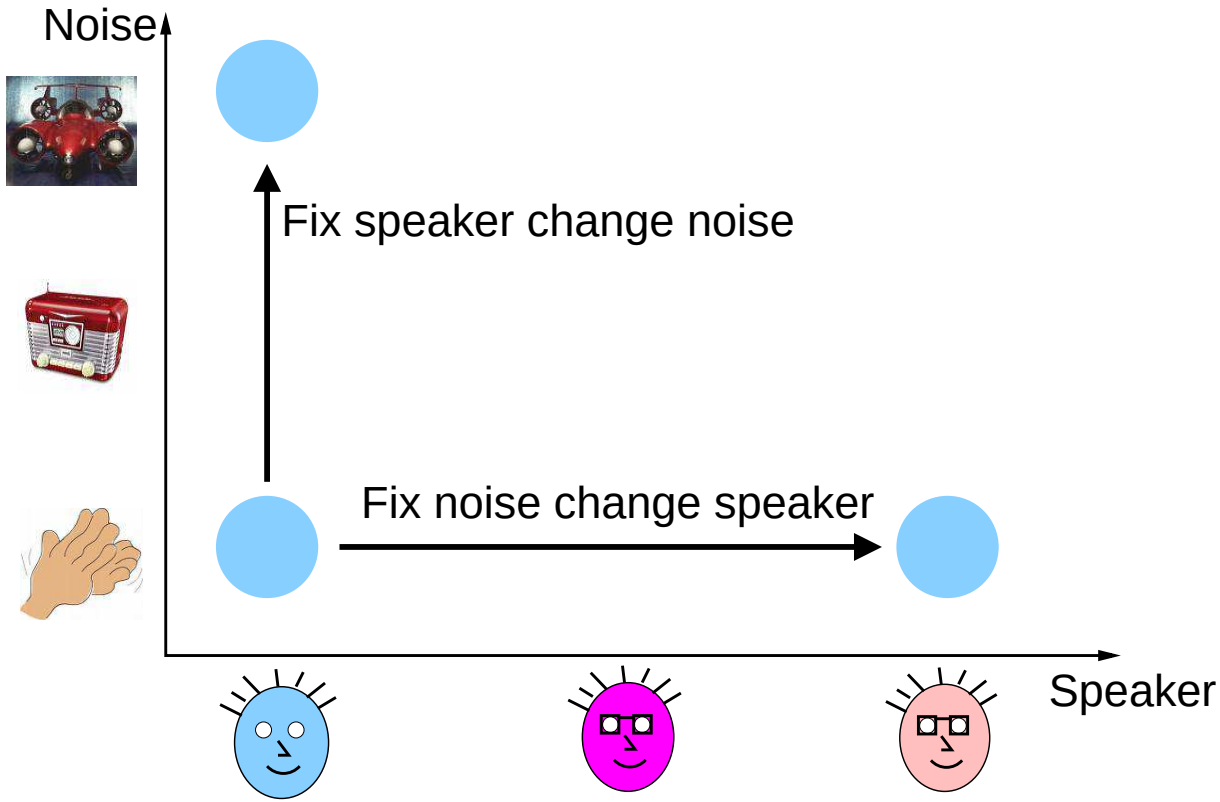
- Consider the above condition for speaker and noise:
  - general speaker transform requires ≈ 1500 frames for robust estimate
  - VTS environment model requires ≈ 100 frames for robust estimate

Example (3) - Polyglot Synthesis



- Consider the above condition for speaker and language:
  - synthesis speaker characteristics in a different language

Transform Orthogonality



- Need to be able to apply transforms independently - transform orthogonality

How to ensure this orthogonality/attributable to factors

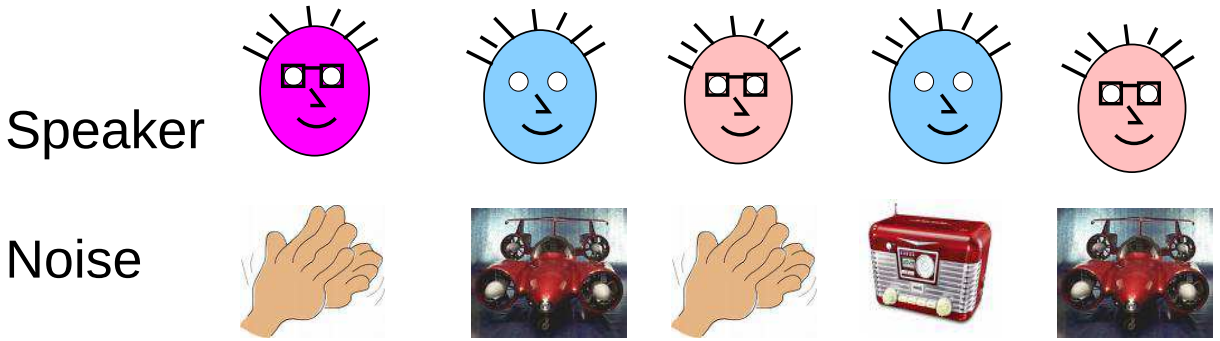
Multiple Linear Transforms

- Consider the case of using linear transforms for both speaker and noise [22]

$$\mathbf{A}^{(s)}(\mathbf{A}^{(n)}\boldsymbol{y}_t + \mathbf{b}^{(n)}) + \mathbf{b}^{(s)} = \mathbf{A}^{(sn)}\boldsymbol{y}_t + \mathbf{b}^{(sn)}$$

  - there's no orthogonality - transform structure the same

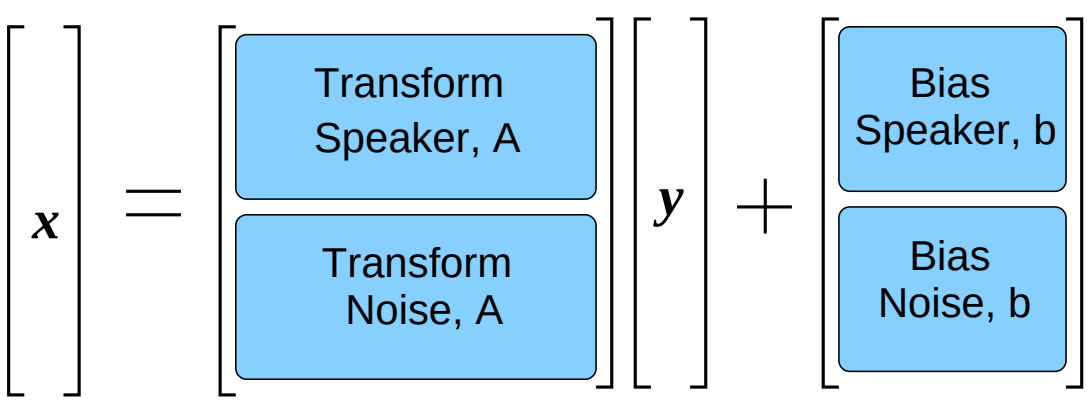
- Simplest solution is to ensure speaker/noise overlaps



- Not always possible to control nature of the data
  - how to ensure factorisation without overlaps?

"Orthogonal" Linear Subspaces

- Alternative approach is to have speaker and noise "sub-spaces"



$$\mathbf{A}^{(sn)} = \begin{bmatrix} \mathbf{A}^{(s)}_{[p\times d]} \\ \mathbf{A}^{(n)}_{[\overline{p}\times d]} \end{bmatrix} \qquad \mathbf{b}^{(sn)} = \begin{bmatrix} \mathbf{b}^{(s)}_{[p]} \\ \mathbf{b}^{(n)}_{[\overline{p}]} \end{bmatrix}$$

  - p-dimensional speaker-space, $\overline{p}(= d - p)$-dimensional noise-space
  - can use standard row-by-row CMLLR/MLLR updates

    additional constraints for CMLLR - $\mathbf{A}^{(sn)} = \begin{bmatrix} \mathbf{A}^{(s)} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{(n)} \end{bmatrix} \mathbf{A}$
  - limit power of transforms to represent speaker/noise
- Similar to (noise-less) ICA, but on varying transforms
  - possible to add noise to transform (Noisy CMLLR)

Orthogonal Transforms

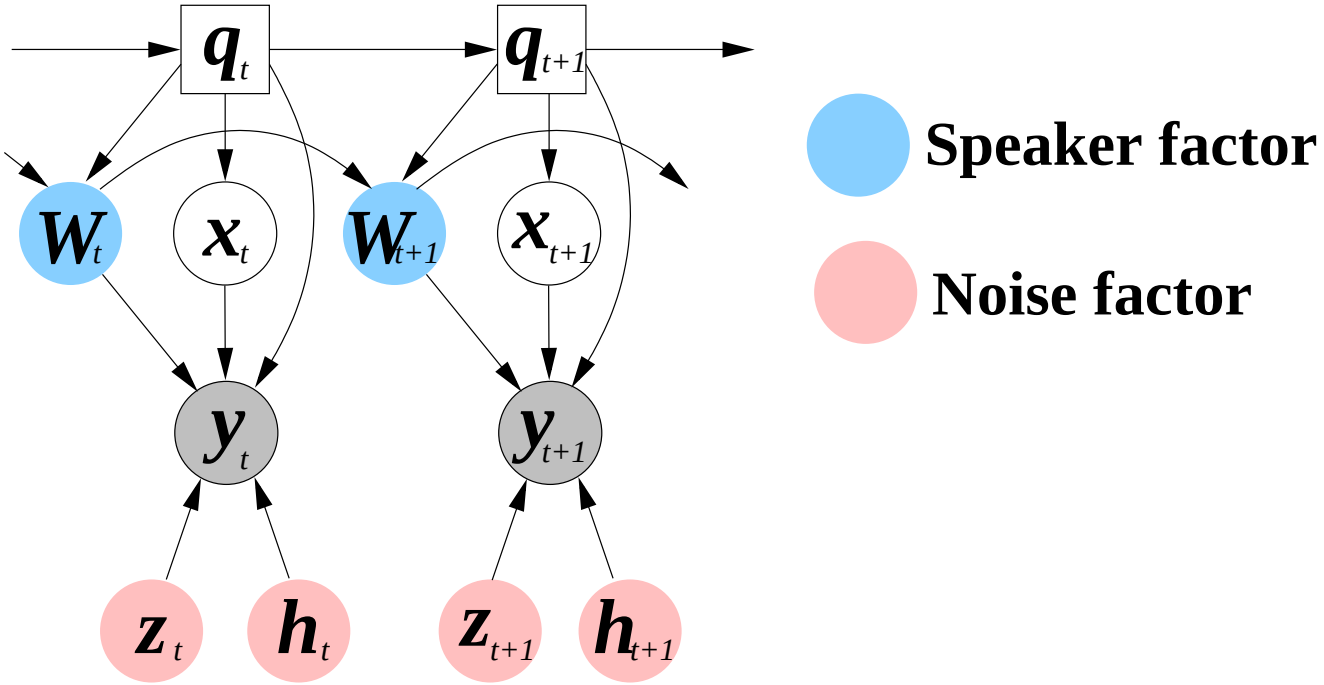
- If the attributes of the transforms are very different
  - each transform "tuned" to a specific factor
  - optimal transforms will be expected to model specific factors
- Again consider speaker (CMLLR) and noise (VTS) representations

$$\begin{aligned}\boldsymbol{y}_t^{(sn)}|m = &\boldsymbol{f}\left(\boldsymbol{A}\boldsymbol{\mu}_{\mathtt{x}}^{(m)} + \mathbf{b}^{(s)}, \boldsymbol{\mu}_{\mathtt{z}}^{(n)}, \boldsymbol{\mu}_{\mathtt{h}}^{(n)}\right) + \boldsymbol{\Lambda}_{\mathtt{x}}^{(m)}(\mathbf{A}^{(s)}(\boldsymbol{x}_t - \boldsymbol{\mu}_{\mathtt{x}}^{(m)}) + \mathbf{b}^{(s)}) \\ &+ \boldsymbol{\Lambda}_{\mathtt{n}}^{(m)}(\boldsymbol{z}_t - \boldsymbol{\mu}_{\mathtt{z}}^{(n)}) + \boldsymbol{\Lambda}_{\mathtt{h}}^{(m)}(\boldsymbol{h}_t - \boldsymbol{\mu}_{\mathtt{h}}^{(n)})\end{aligned}$$

  - again linear transforms $\mathbf{A}^{(s)}$, $\boldsymbol{\Lambda}_{\mathtt{x}}^{(m)}$, etc
  - BUT some global $\mathbf{A}^{(s)}$, some component specific $\boldsymbol{\Lambda}_{\mathtt{x}}^{(m)}$
- Automatically yields orthogonality for appropriately "tuned" transforms
  - again subspaces, but typically non-linear and linked to transforms

Speaker and Noise Factorisation

Speaker and Noise Factors DBN



- Need the impacts of the speech and noise transforms to act independently [19, 20]

$$\boldsymbol{y}_t = \boldsymbol{f}\left(\mathbf{A}_t\boldsymbol{x}_t + \mathbf{b}_t, \boldsymbol{z}_t, \boldsymbol{h}_t\right)$$

  - implies CMLLR for the speaker, VTS for the noise condition

Speaker and Noise Configuration

- Investigated form comprises:

$$p(\boldsymbol{y}_t|\mathcal{M}_{\mathtt{x}}, \mathcal{M}_{\mathtt{s}}^{(s)} \otimes \mathcal{M}_{\mathtt{n}}^{(n)}, m) = \mathcal{N}\left(\boldsymbol{y}_t; \boldsymbol{\mu}_{\mathtt{y}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{y}}^{(m)}\right)$$

  where MLLR and VTS are combined

$$\boldsymbol{\mu}_{\mathtt{y}}^{(m)} = \boldsymbol{f}\left(\mathbf{A}^{(s)}\boldsymbol{\mu}_{\mathtt{x}}^{(m)} + \mathbf{b}^{(s)}, \boldsymbol{\mu}_{\mathtt{h}}^{(n)}, \boldsymbol{\mu}_{\mathtt{n}}^{(n)}\right)$$

$$\boldsymbol{\Sigma}_{\mathtt{y}}^{(m)} = \mathrm{diag}\left(\boldsymbol{\Lambda}_{\mathtt{x}}^{(m)}\boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}\boldsymbol{\Lambda}_{\mathtt{x}}^{(m)\mathsf{T}} + \boldsymbol{\Lambda}_{\mathtt{n}}^{(m)}\boldsymbol{\Sigma}_{\mathtt{n}}^{(n)}\boldsymbol{\Lambda}_{\mathtt{n}}^{(m)\mathsf{T}}\right)$$

- MLLR is applied to the "clean" speech parameters
  - transformation should be independent of the noise
  - however it will compensate for any limitations of the mismatch function.

Results on AURORA-4

- Medium vocabulary (WSJ-based) noise corrupted speech recognition task

Scheme	Spk. Est.	A	B	C	D	Avg.
VAT	—	8.5	13.7	11.8	20.1	15.9
`std`	`block`	5.6	11.0	8.8	17.8	13.4
`afact`	`noise04`	6.9	11.5	10.4	18.5	14.1

- Standard form (`std`) estimate VTS+MLLR transform for each block
  - works well, BUT requires sufficient data for speaker transform estimation
- Factored form (`afact`) speaker only estimated on one noise condition
  - as fast as VTS for all other noise conditions (single utterance)
  - acoustic factorisation has occurred (to some extent)

Speaker and Language Factorisation

Polyglot Synthesis

- An interesting challenge is

  How to have a speaker talk in a different language?

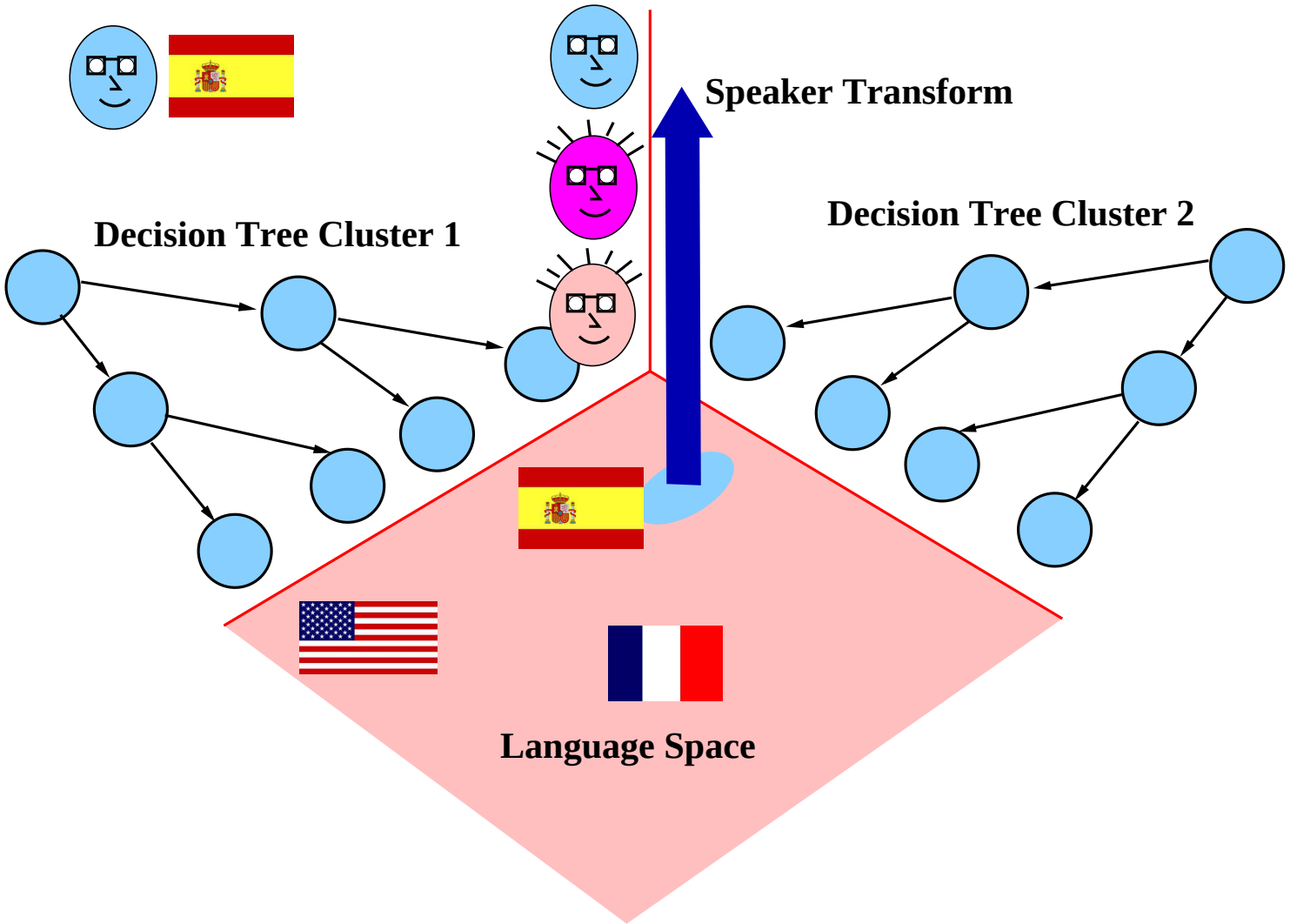
  - need to maintain the same speaker characteristics
  - need to change the language

- Not normally possible to get multi-lingual speakers to record corpora
  - would dramatically limit the size of corpus that could be used
- Parametric statistical speech synthesis is attractive for this task
  - based on graphical models (HMMs-like)
  - standard adaptation approaches can be applied
  - factorisation should be possible

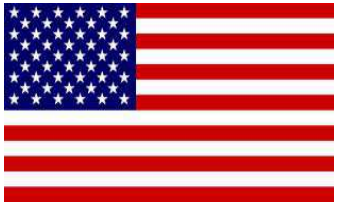
Multi-Lingual Synthesis

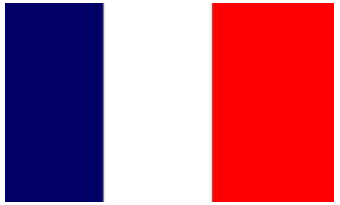
- Major problem with multi-lingual systems is variations in phonetic information
  - phone sets may differ between languages
  - contextual importance may differ between languages
  - some contextual/acoustic attributes shared (common physical system)
- Some of these attributes are reflected in the decision trees
  - a single decision tree will not be sufficient
  - multiple decision trees one option - yields a tree intersect style model
- Generalised CAT to multiple decision trees [23]
  - a CAT specified language space for language attributes
  - use CMLLR to represent the speaker attributes

Speaker and Language Transformations



Example Synthesis





Example Synthesis

- Highly challenging (too challenging ...)
- SLF system trained on limited data
  - 5 languages (US English, UK English, Spanish, French German)
  - 8 speakers per language
  - only 7 hours in total
- Target speaker data very limited
  - only 10 minutes of Barack Obama data
  - taken from State of the Nation address (not high quality)

Conclusions

Conclusions

- Speech is an incredibly rich signal
  - words only part of the speech information signal
  - signal has speaker/environment/channel/language distortions
- Makes speech recognition/synthesis interesting (and challenging)
- Adaptation and compensation an essential part of systems
  - need to train systems on vast quantities of data
  - wide-range of speakers/environments/channels/languages
- Acoustic factorisation highly flexible/controllable adaptation
  - essential for controllable speech synthesis
  - improves efficiency/portability for speech recognition

Acknowledgements

- This work has been funded from the following sources:
  - Cambridge Research Lab, Toshiba Research Europe Ltd
  - Google Research Award
  - DARPA RATS Programme
  - EPSRC Natural Speech Technology Programme Grant

References

[1] J. L. Gauvain and C.-H. Lee, "Maximum a-posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[2] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," in *Proceedings ICSLP*, 1998, pp. 1771–1774.

[3] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Transactions Speech and Audio Processing*, vol. 8, pp. 417–428, 2000.

[4] K. Yu and M. Gales, "Bayesian adaptive inference and adaptive training," *IEEE Transactions Speech and Audio Processing*, vol. 15, no. 6, pp. 1932–1943, August 2007.

[5] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, vol. 9, pp. 171–186, 1995.

[6] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Languages*, vol. 10, pp. 249–264, 1996.

[7] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[8] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, X. Liu, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[9] A. Acero, "Acoustical and environmental robustness in automatic speech recognition," Ph.D. dissertation, Carnegie Mellon University, 1990.

[10] T. Kristjansson, "Speech recognition in adverse environments: a probabilistic approach," Ph.D. dissertation, University of Waterloo, Waterloo, Canada, 2002.

[11] L. Deng, J. Droppo, and A. Acero, "Enhancement of log mel power spectra of speech using a phase sensitive model the acoustic environemnt and sequential estimation of the corrupting noise," *Proc. IEEE Transactions on Speech and Audio Processing*, 2004.

[12] P. Moreno, "Speech Recognition in Noisy Environments," Ph.D. dissertation, Carnegie Mellon University, 1996.

[13] F. Flego and M. J. F. Gales, "Discriminative adaptive training with VTS and JUD," in *Proc. ASRU*, 2009.

[14] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proceedings ICSLP*, 1996, pp. 1137–1140.

[15] J. Yamagishi, T.Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation method," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, pp. 66–83, 2009.

[16] Y. Hu and Q. Huo, "Irrelevant variability normalization based HMM training using VTS approximation of an explicit model of environmental distortions," in *Proc. InterSPeech*, 2007.

[17] H. Liao and M. Gales, "Adaptive training with joint uncertainty decoding for robust recognition of noisy data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007.

[18] O. Kalinli, M. Seltzer, and A. Acero, "Noise adaptive training using a vector taylor series approach for noise robust automatic speech recognition," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 3825–3828.

[19] M. J. F. Gales, "Acoustic factorisation," in *Proc. ASRU*, 2001.

[20] Y. Wang and M. J. F. Gales, "Speaker and noise factorisation on the AURORA4 task," in *Proc. ICASSP*, 2011.

[21] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions Audio Speech and Language Processing*, 2007.

[22] M. Seltzer and A. Acero, "Factored adaptation for separable compensation of speaker and environmental variability," in *Proc ASRU*, 2011.

[23] H. Zen, N. Braunschweiler, S. Buchholz, M. Gales, K. Knill, S. Krstulovic, and J. Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE Transactions Audio Speech and Language Processing*, 2012.