Controllable and Adaptable Statistical Parametric Speech Synthesis Systems

Mark Gales
work with colleagues at **Toshiba Cambridge Research Lab.** and CUED

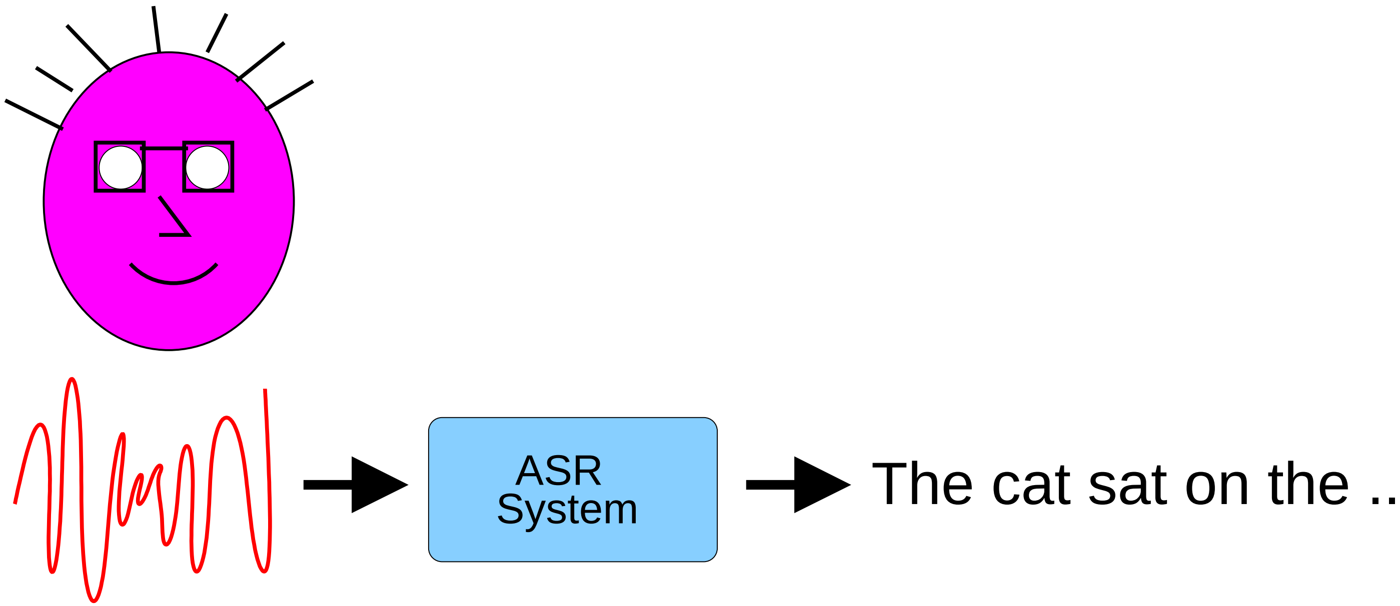
June 2014

Cambridge University Engineering Department

Overview

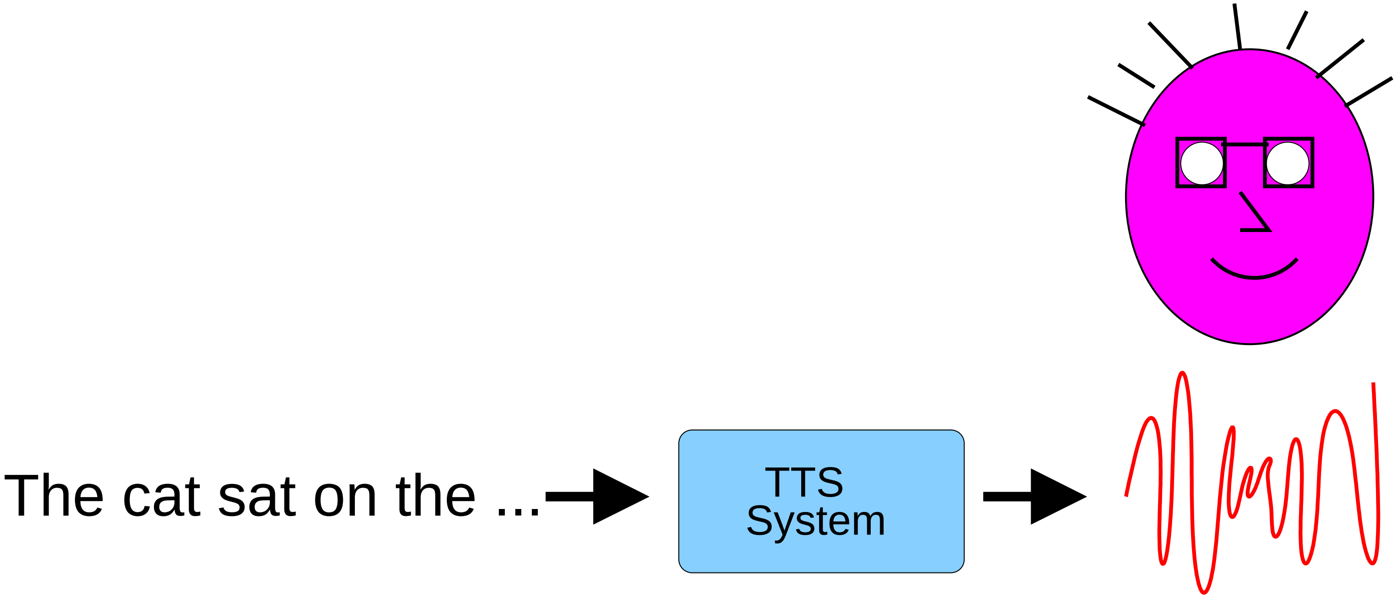
- Speech Synthesis
  - statistical parametric speech synthesis
- Adaptation and Adaptive Training Approaches
  - linear transform-based adaptation
  - cluster adaptive training
- Acoustic Factorisation
- Applications
  - polyglot synthesis
  - controllable speaker and expressive synthesis
  - e-book reading
  - expressive talking head synthesis - “Zoe”

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

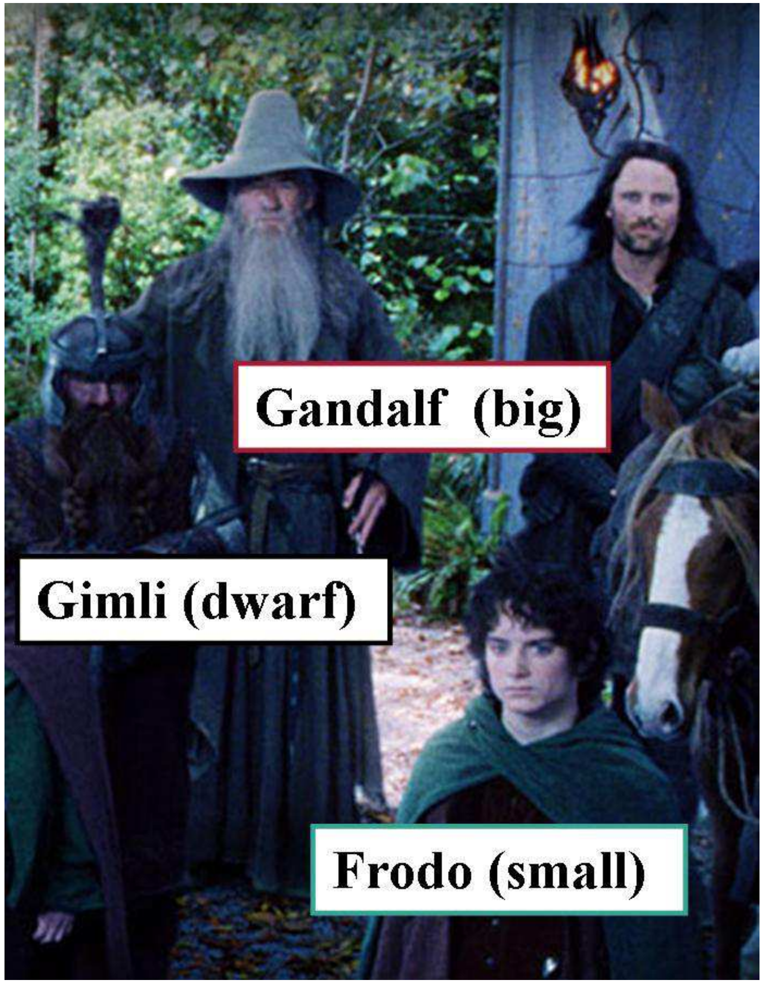
Speech Synthesis (TTS) as a Task



- 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

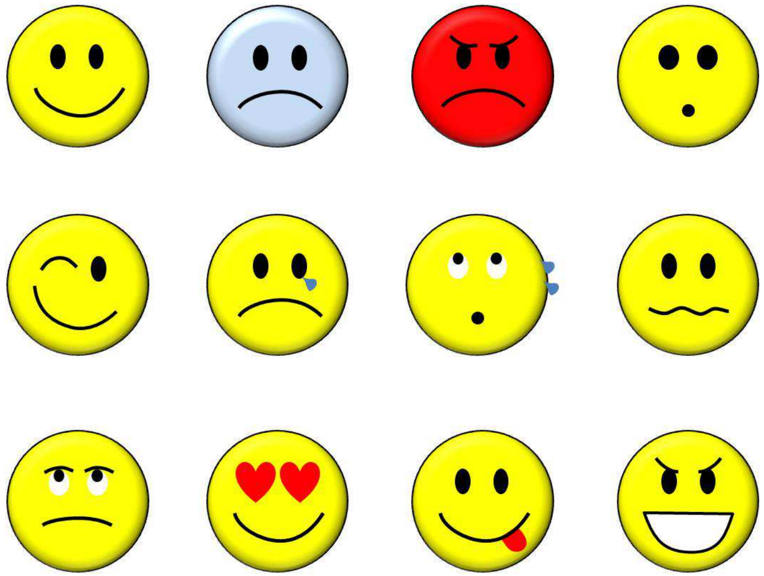
Speaker Differences

- 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



Expressive Speech

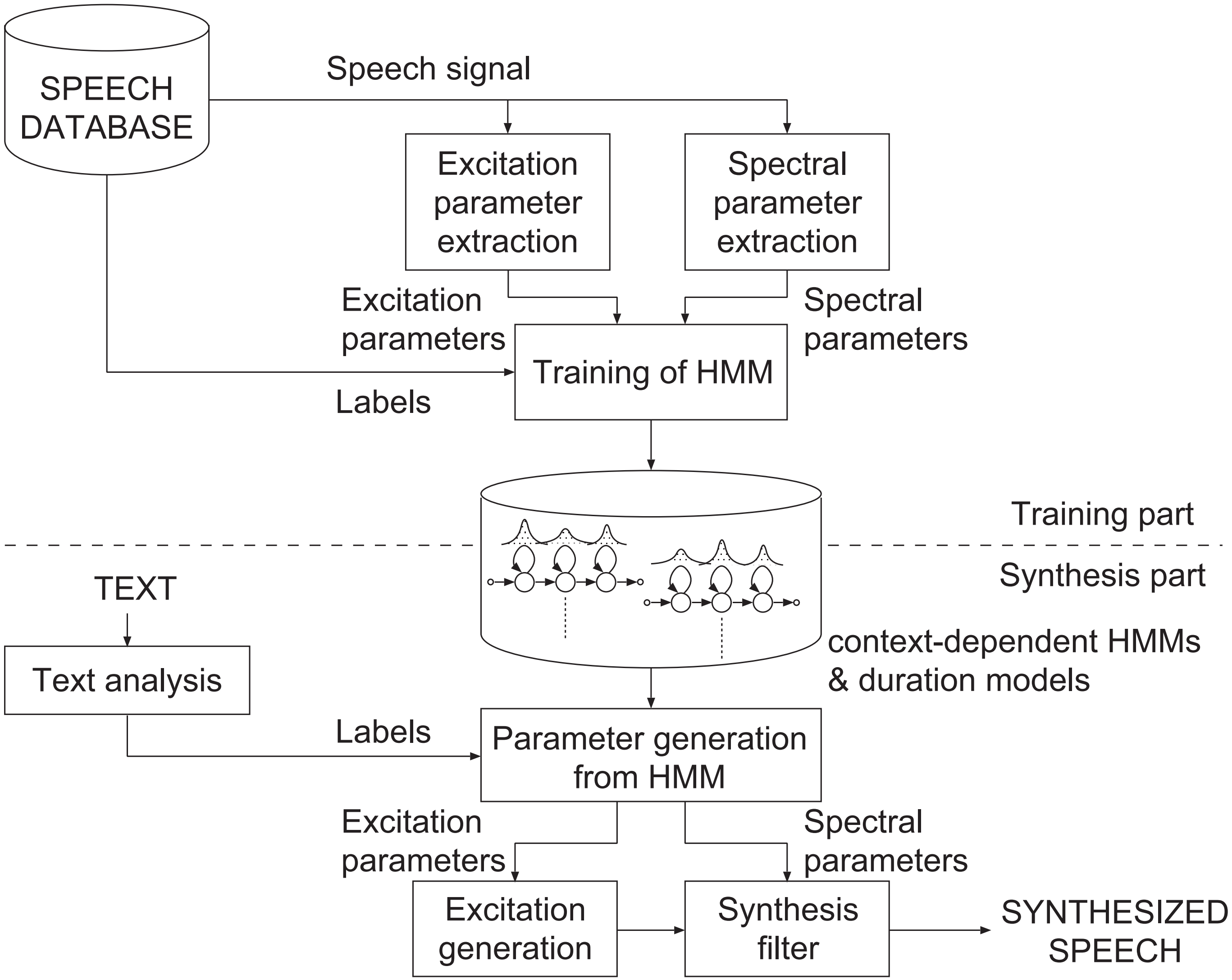
- Spoken communication is more than just conveying a sequence of words
  - expressive speech is essential for efficient communication
- Wide range of expressive states
- Emotional state e.g.
  - happy, sad, angry
- Contextual state e.g.
  - relationship to listener
  - interaction on a dialogue
- "Definition" vague
  - how to label expressive state?



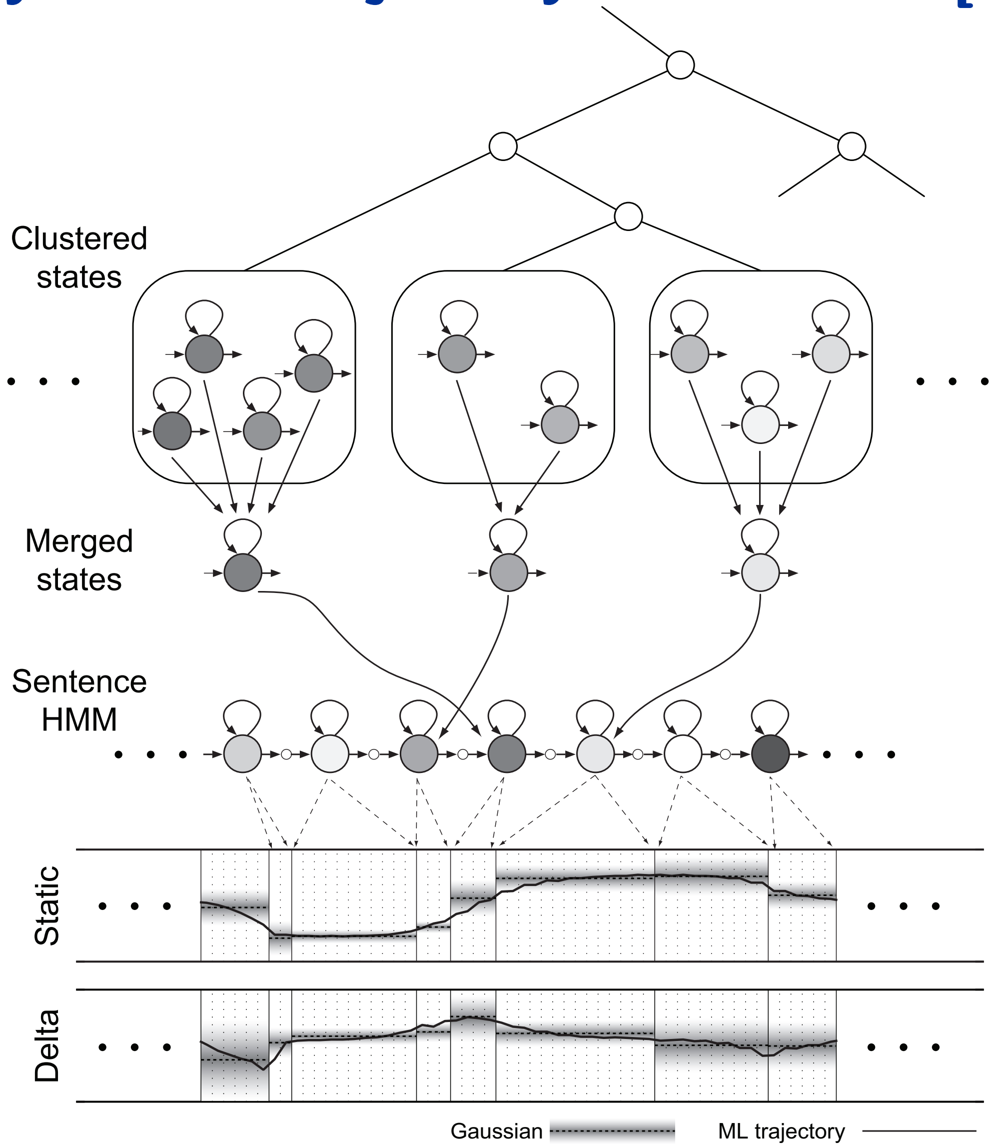
Environment Differences

Statistical Speech Synthesis

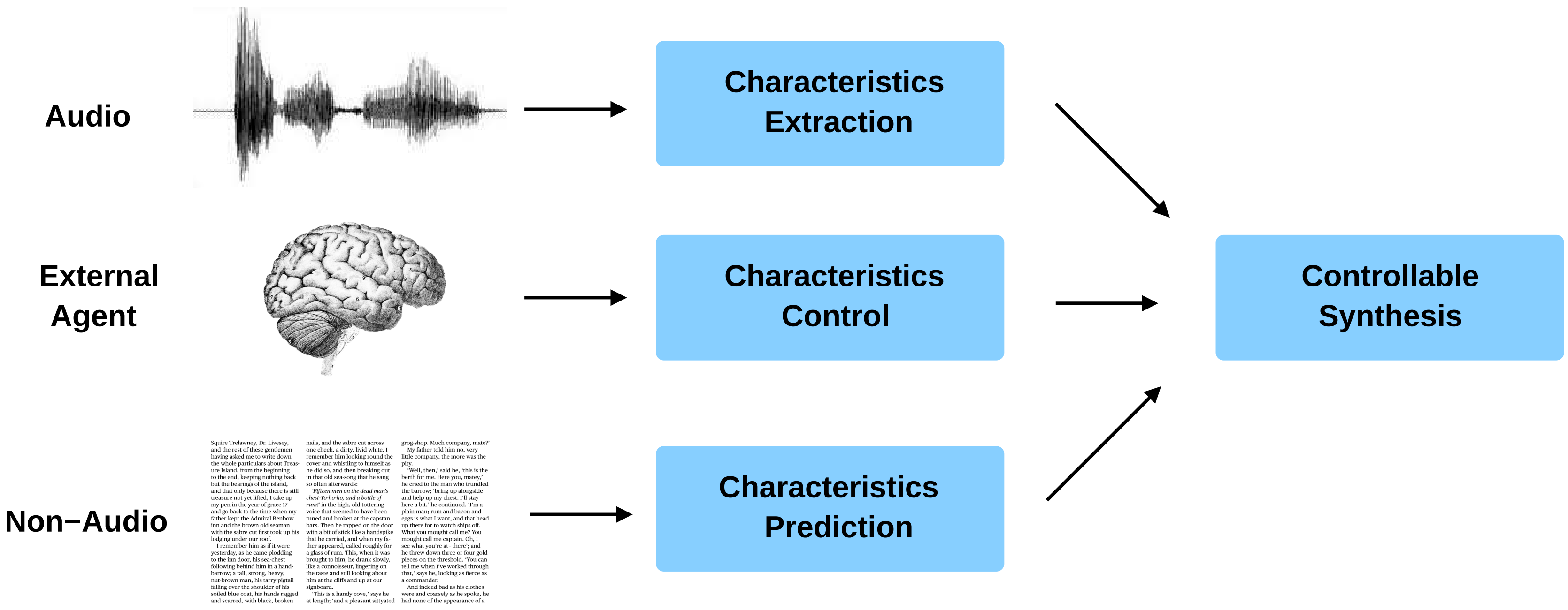
Training and Synthesis [1]



Synthesis Trajectory Generation [1]



Controllable/Adaptable Speech Synthesis

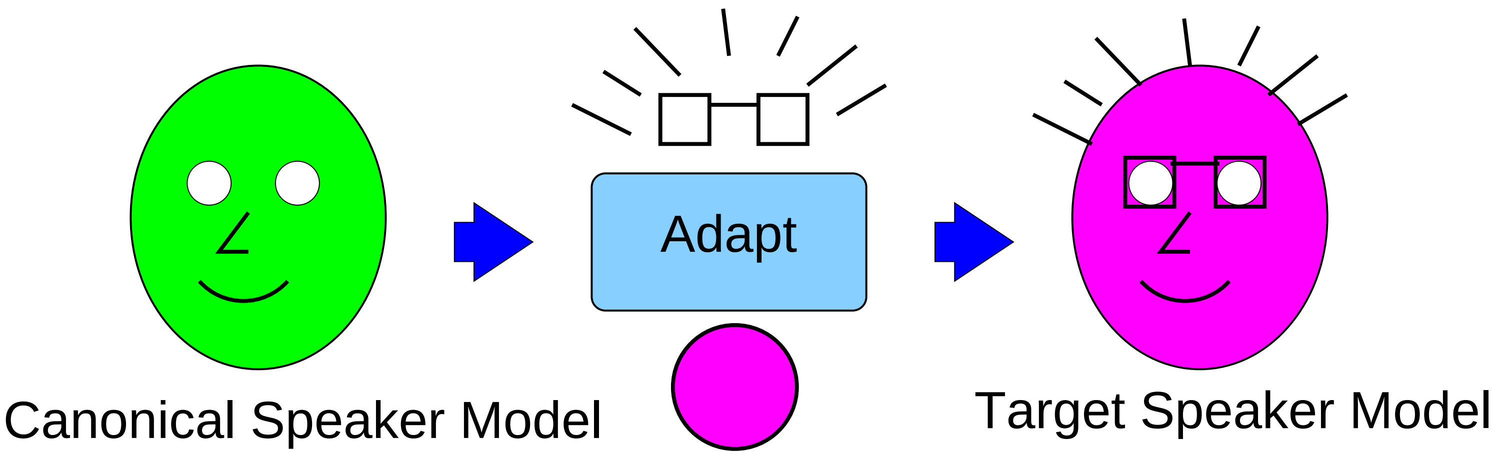


- Different characteristics required depending on information source
  - external agent: meaningful labels required for synthesis system
  - audio/non-audio: consistent labels required for synthesis system
- Influences training and synthesis requirements

Acoustic Model Adaptation

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

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

- 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
  - can be used to derive a prior for adaptive schemes

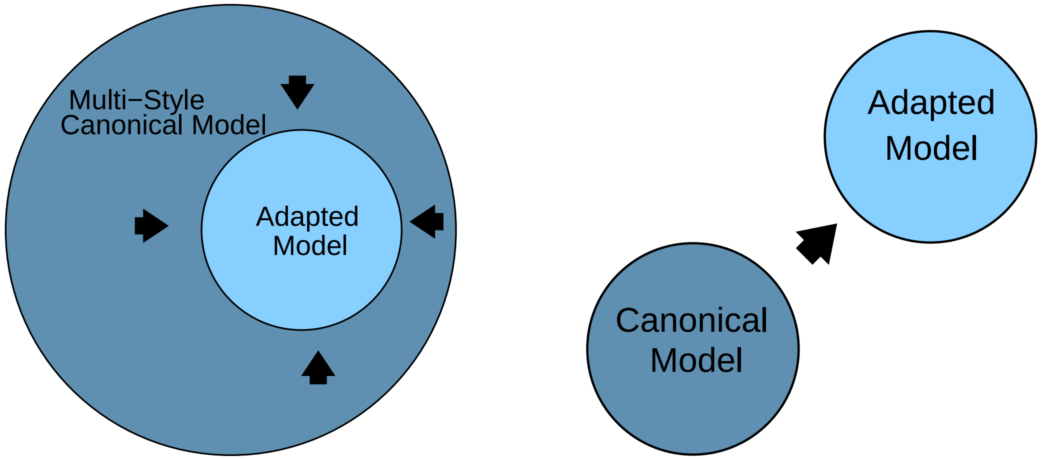
Adaptation Examples

- Adaptive Approaches examples:
  - Maximum A-Posteriori MAP [2] adaptation: general "robust" estimation
  - Cluster Selection: Gender-dependent (GD) models
  - Cluster Interpolation: combine multiple cluster parameters
    EigenVoices[3], CAT [4] 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

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 [5, 6, 7]

Form of the Adaptation Transform

- Dominant form for LVCSR are ML-based linear transformations
  - MLLR adaptation of the means [8]

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

  - MLLR adaptation of the covariance matrices [9, 6]

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

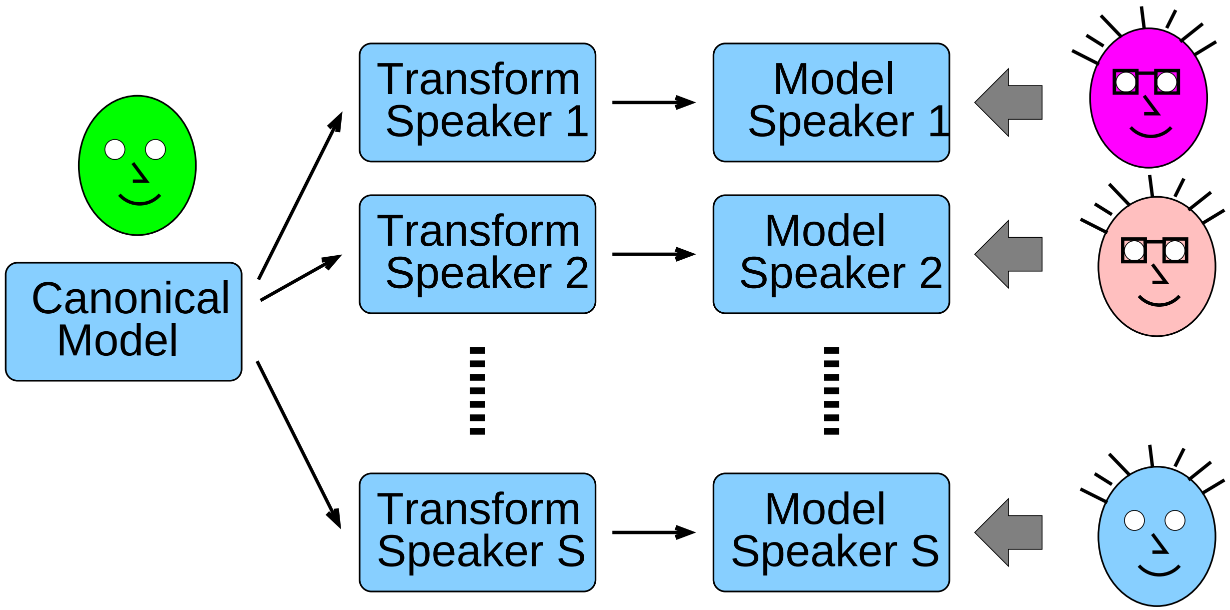
  - Constrained MLLR adaptation [6]

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

- Forms may be combined into a hierarchy [10] e.g.

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

Speaker Adaptive Training (AVM) [5, 6, 11]



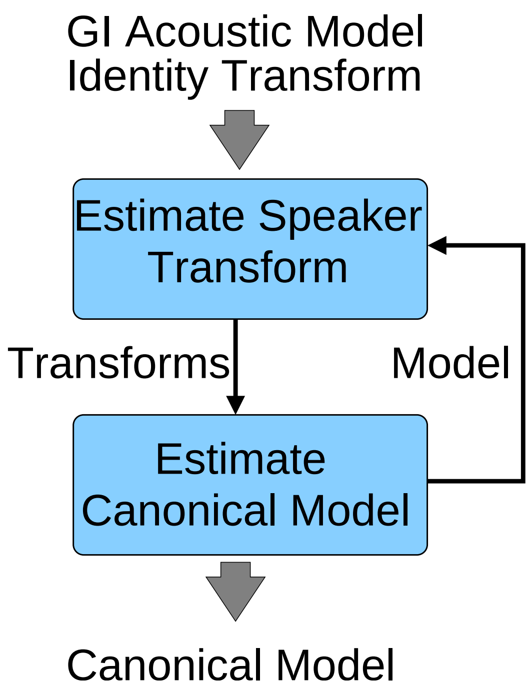
- In adaptive training the training corpus is split into "homogeneous" blocks
  - use adaptation transforms to represent unwanted acoustic factors
  - canonical model only represents desired variability
- All forms of linear transform can be used for adaptive training
  - CMLLR adaptive training highly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

CMLLR Adaptive Training

- The CMLLR likelihood may be expressed as [6]:

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

same as feature normalisation - simply train model in transformed space



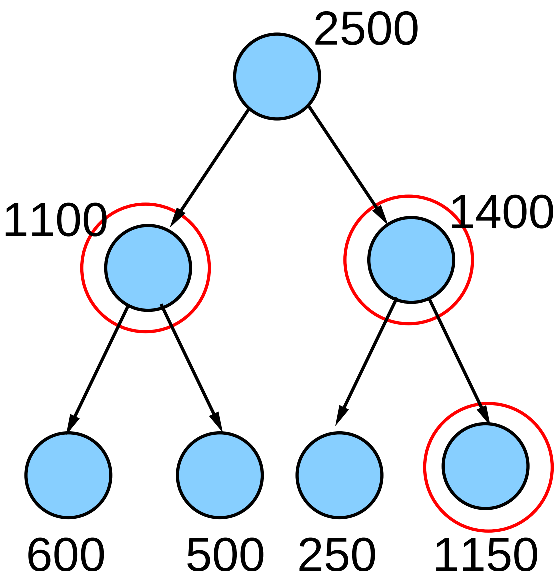
- Interleave Model and transform estimation
- Update formulae for mean

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

Adaptation Transform Complexity

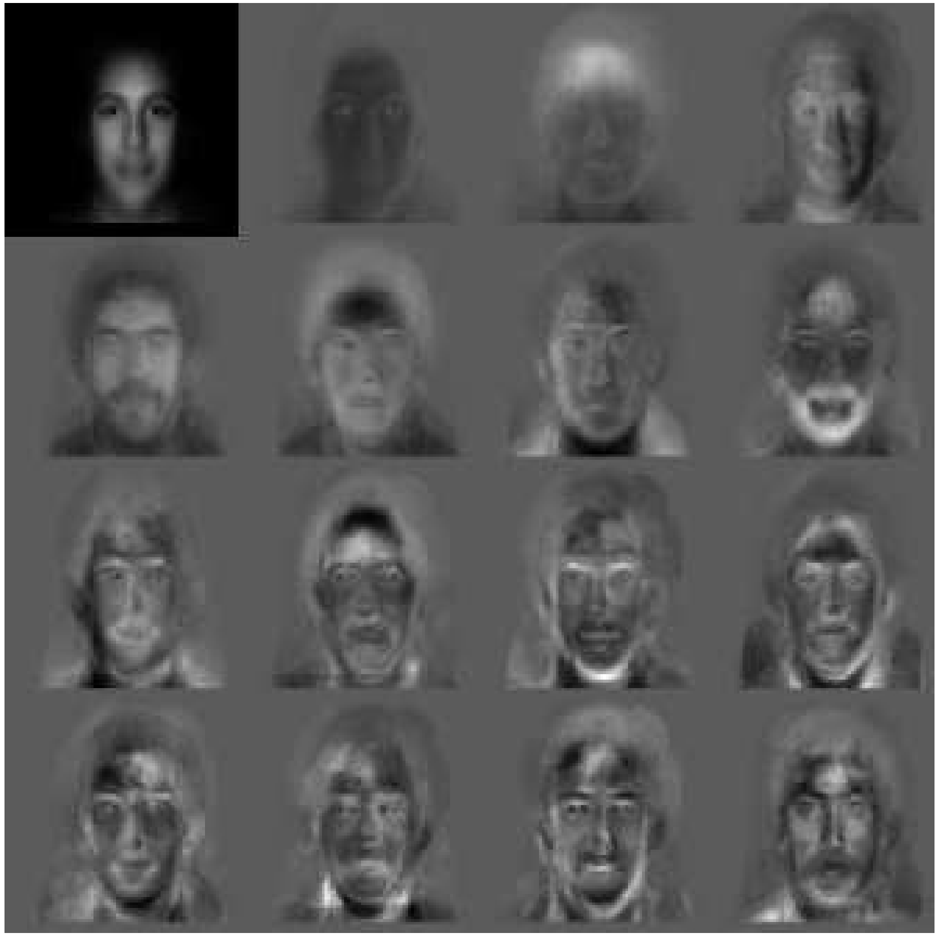
- Two aspects of transform complexity can be controlled:
  - structure of the transform: full, block, diagonal
  - number of transforms

  The structure is normally determined by an "expert"



- Regression Class trees often used [12] to determine number of transforms
- Example with a threshold of 1000 shown:
- Also able to incorporate a prior
  - CSMAPLR [13]

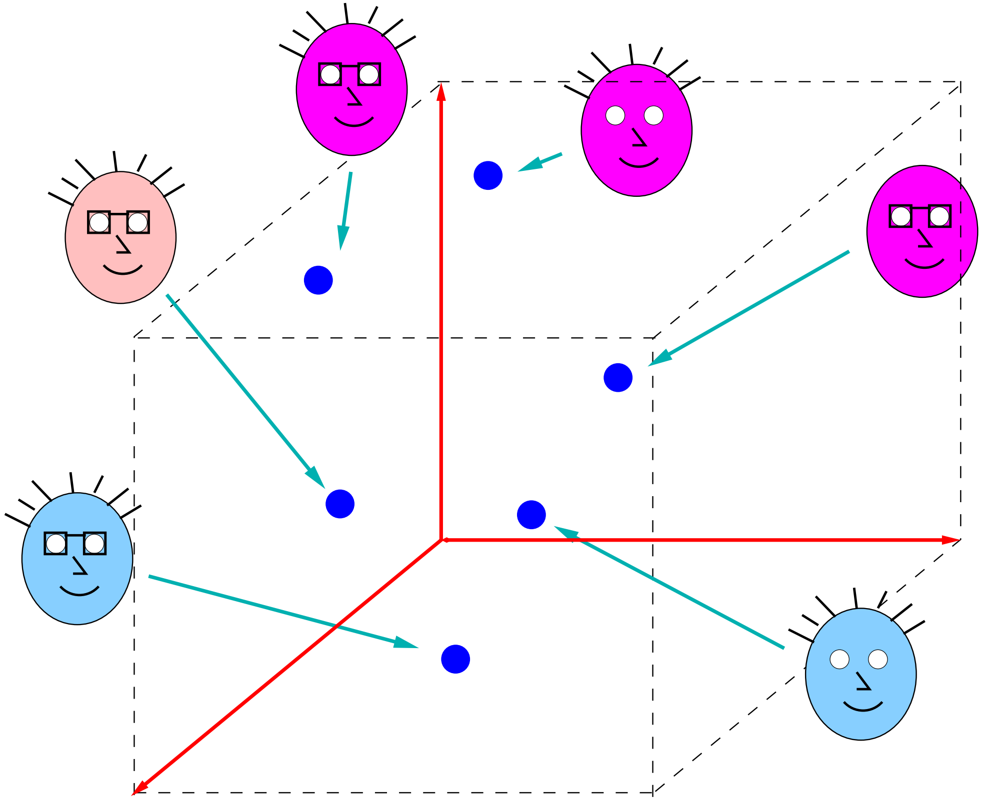
EigenFaces



- 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

  100K (`comp`) $\times$ 39 (`dim`) $=$ 3.9M

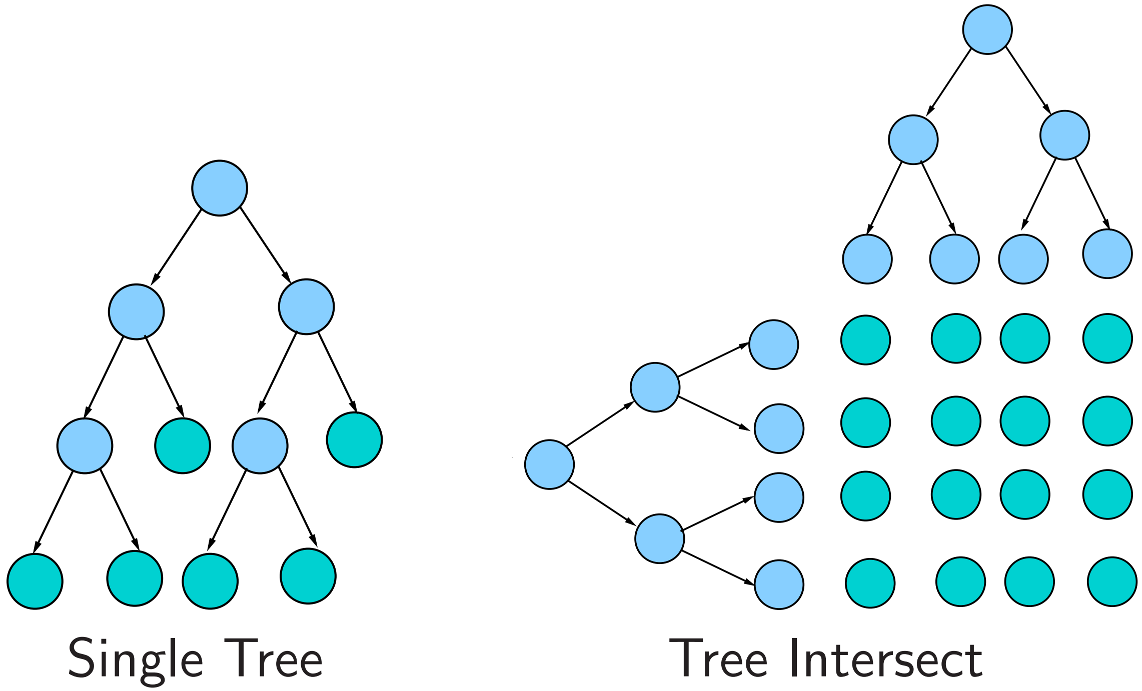
- Low-dimensional (3-10) subspace
- Each speaker represented by a point $\boldsymbol{\lambda}$ in the subspace
  - the speaker specific mean is

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

- CAT yields complete ML estimation (PCA for original EigenVoices) [3, 4]

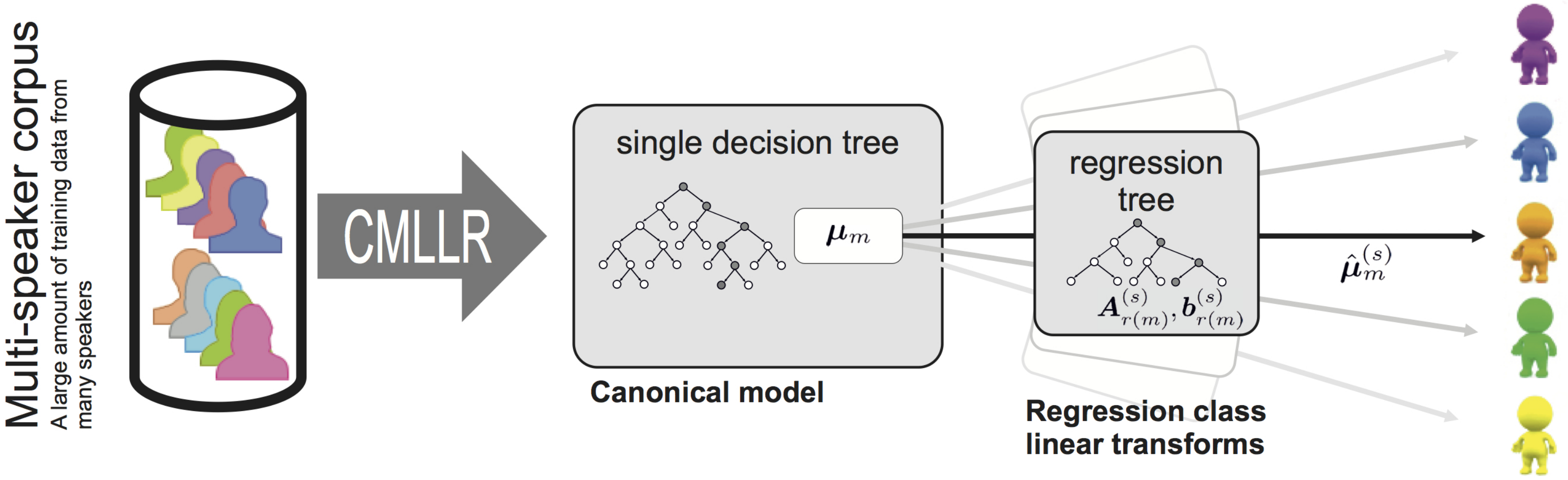
Each speaker specified by only 3-10 parameters!

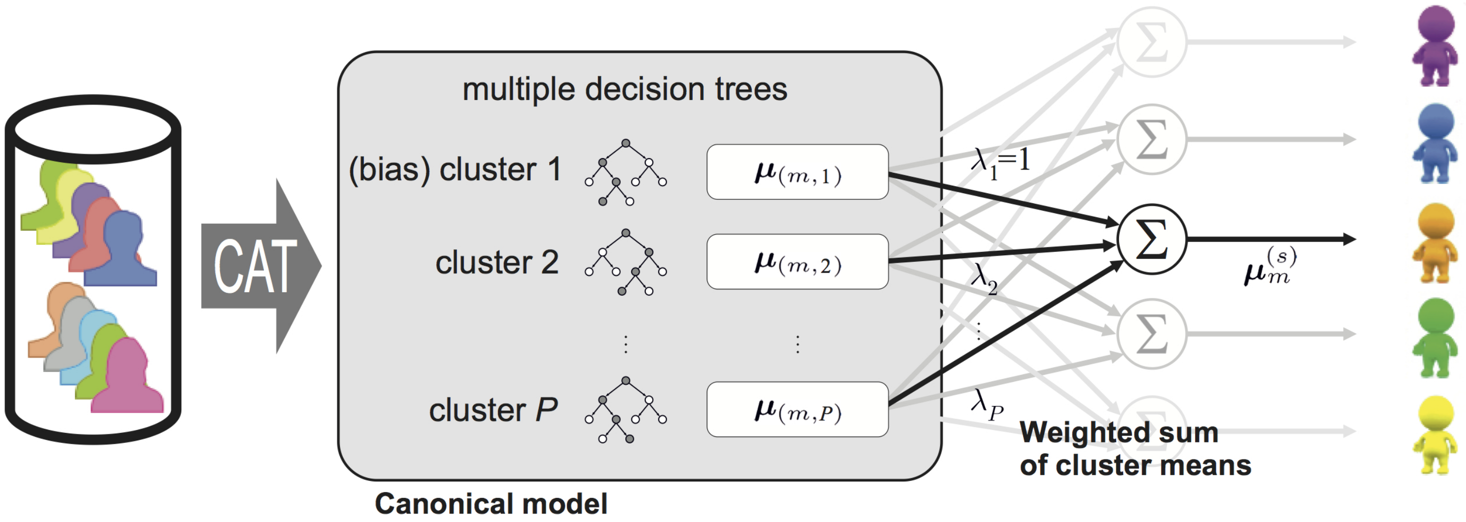
Multiple Decision Trees [14]



- An important aspect of the acoustic model is the decision tree
  - each leaf (context group) usually modelled by a single Gaussian
- Tree interact models yield a compact way of representing many contexts
  - associate a separate tree with each cluster

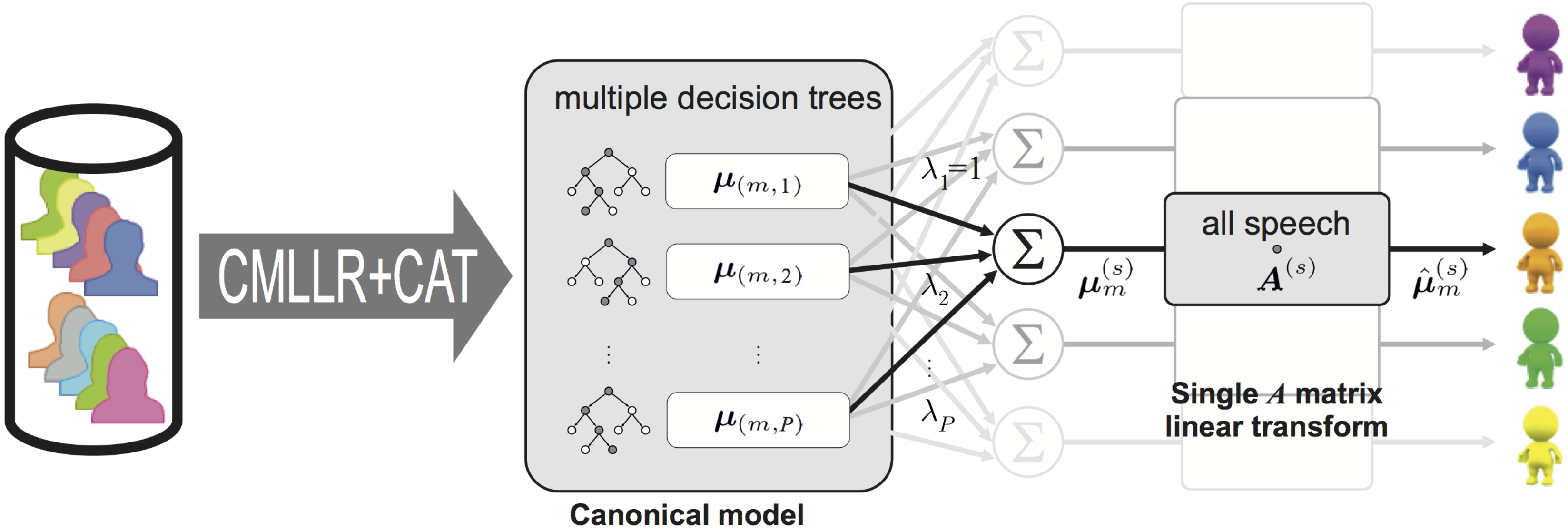
CAT [15] and AVM [11] Representations





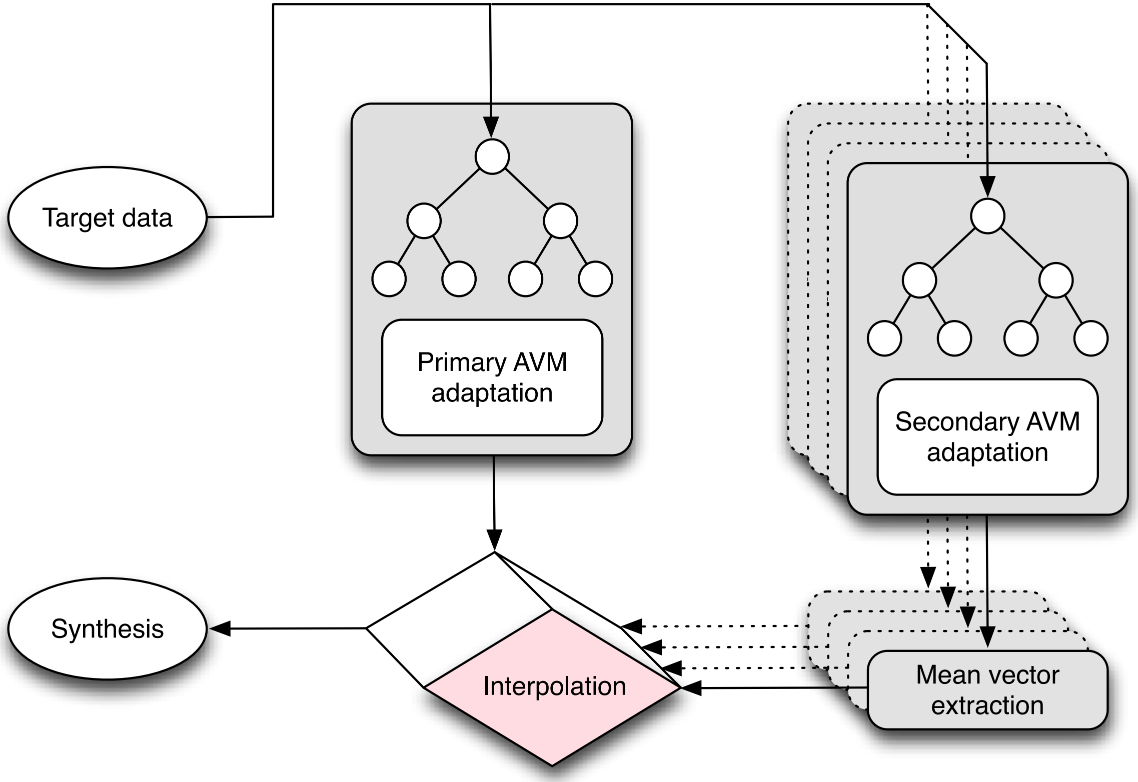
Combined Schemes - AVM plus CAT [16]

- AVM and CAT have complementary attributes
  - CAT: very fast, good quality, average similarity
  - AVM: slow(er), average/good quality, good similarity



- Use CAT as the canonical model the AVM
  - additional transform improves similarity

Combined Schemes - Multiple AVMs [17]

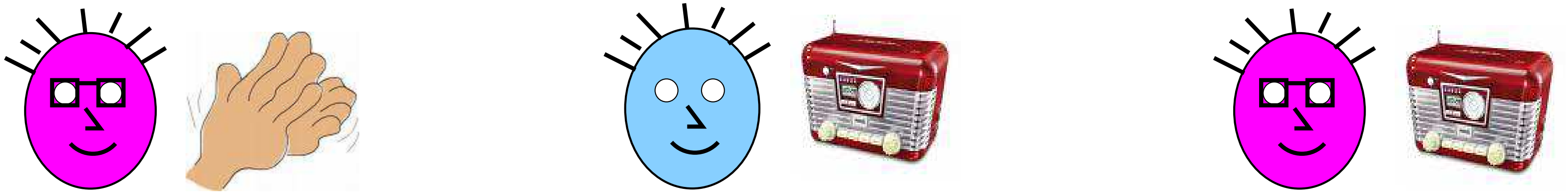


- Possible to combine multiple AVMs together
  - effectively tunes CAT bases to the target speaker
  - only the means are interpolated (simplifies the maths)

Acoustic Factorisation

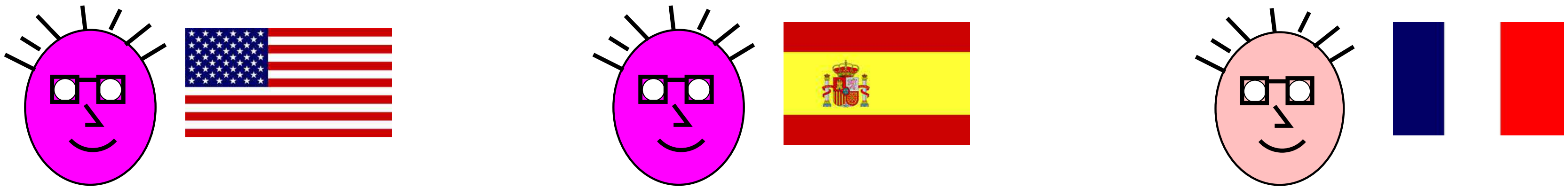
Multiple Acoustic Factors

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



the same speaker may be observed in multiple noise conditions

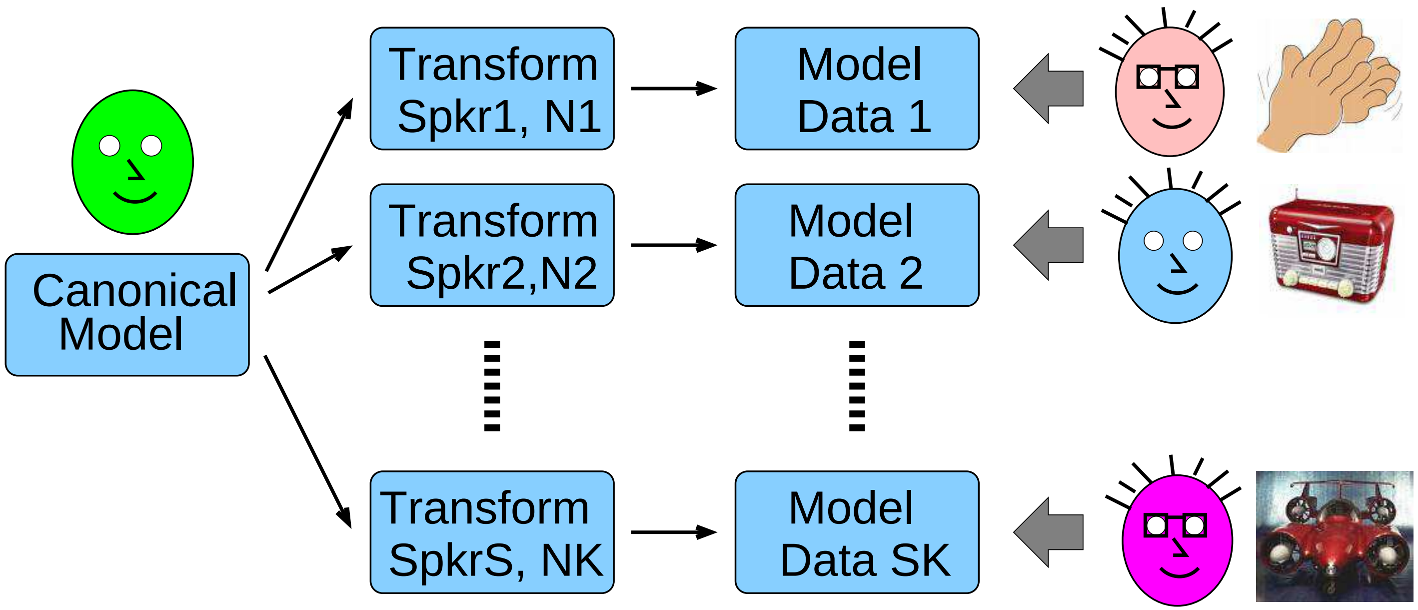
  - speaker and language:



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

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

- BUT ignores aspects of speaker/noise relationships

How to Incorporate this Information?

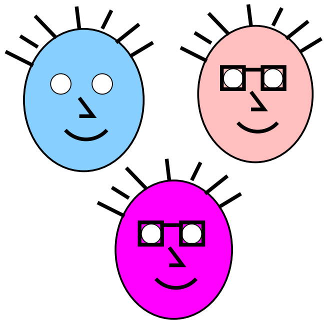
Acoustic Factorisation

- Conceptually the process is very easy [18]

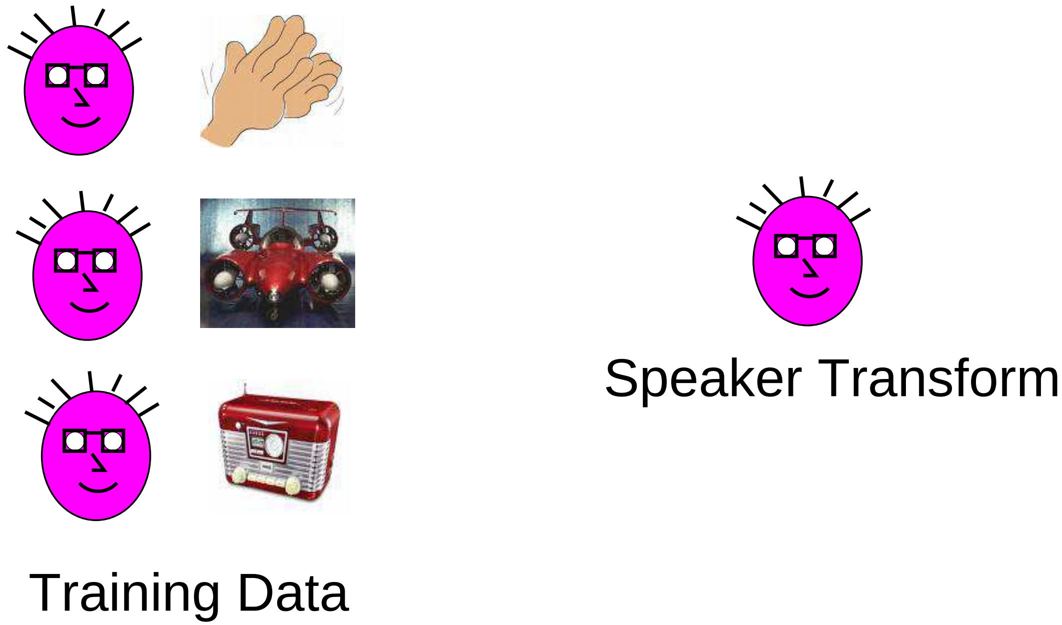
$$\mathcal{M}_{\mathtt{f}}^{(sn)} = \mathcal{M}_{\mathtt{s}}^{(s)} \otimes \mathcal{M}_{\mathtt{n}}^{(n)}$$

  - 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

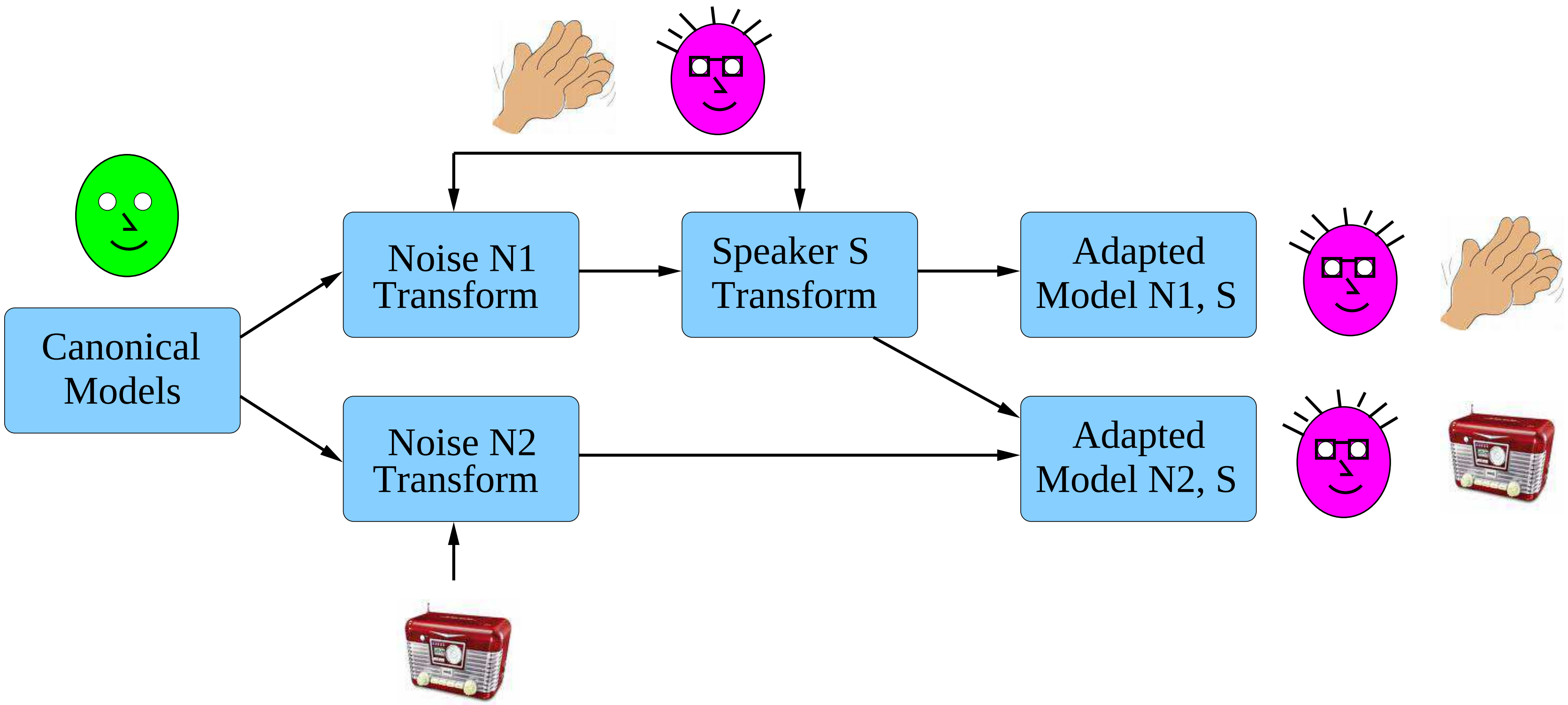



Example (1) - "Practical" Speaker Enrolment



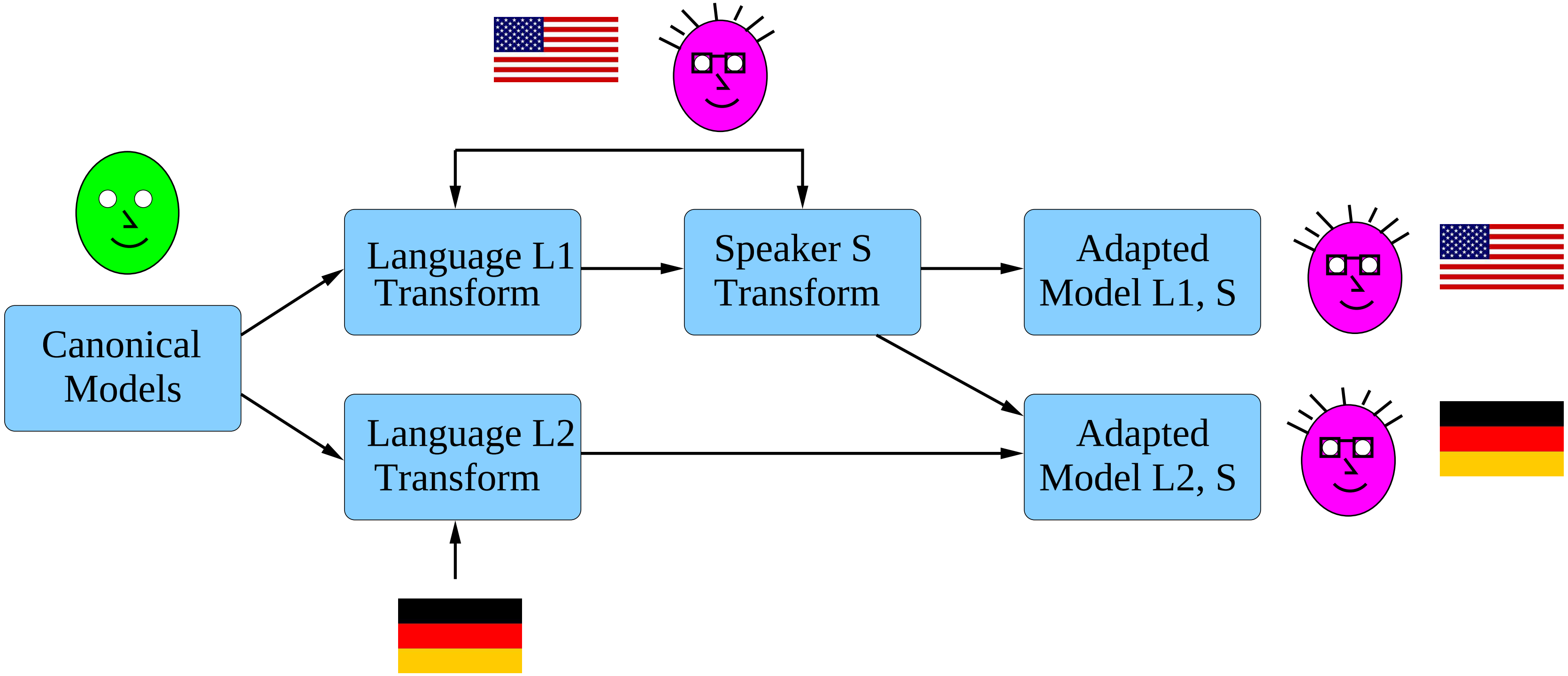
- Often only see data from a speaker with varying acoustic condition
  - consider in-car navigation system/ recording session variability
- Canonical speaker transform required
  - recognition in different environments [19]/speaker identification [20]

Example (2) - Rapid Adaptation



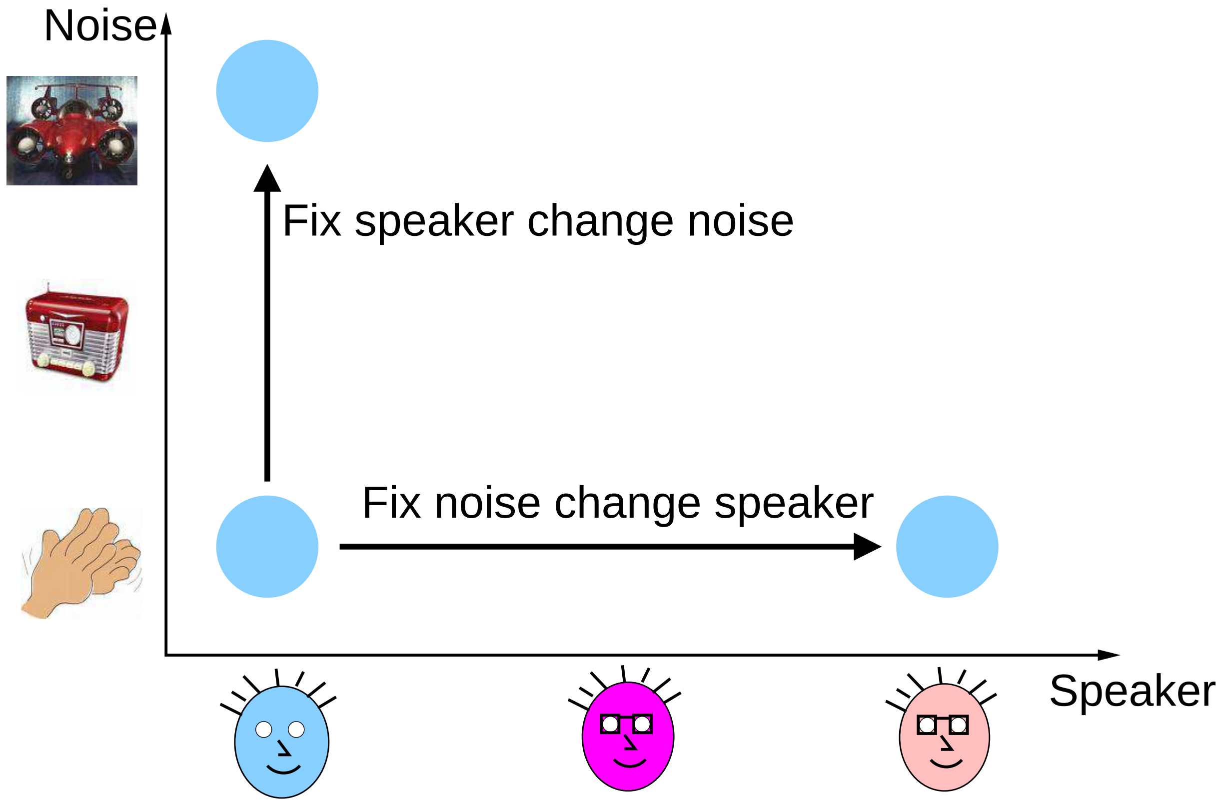
- Consider the above condition for speaker and noise:
  - general speaker transform requires ≈ 1500 frames for robust estimate
  - VTS environment model requires ≈ 100 frames for robust estimate

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

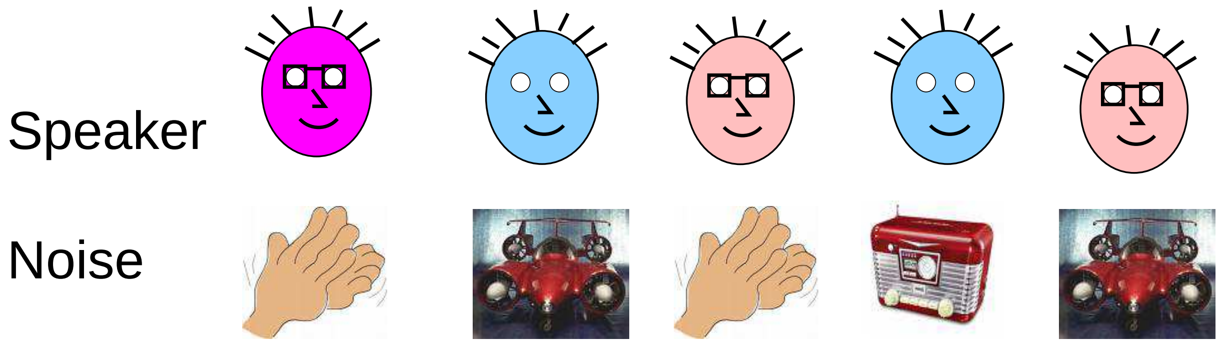
Multiple Linear Transforms

- Consider the case of using linear transforms for both speaker and noise [21]

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

  - there's no orthogonality - transform structure the same

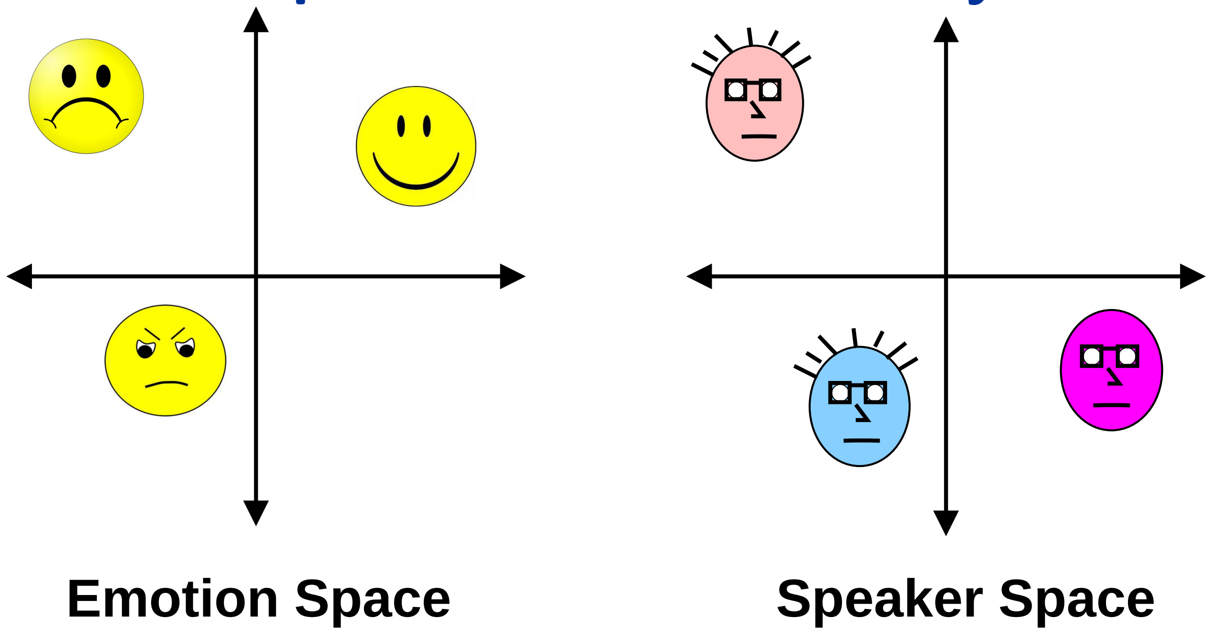
- Simplest solution is to ensure speaker/noise overlaps



- Not always possible to control nature of the data
  - possible to impose explicit orthogonality constraints [22]

Example Applications

Controllable Speaker and Emotion Synthesis [23]



- Use CAT to define both speaker and emotion spaces
  - train so that spaces are orthogonal to one another
  - enables separate control over speaker and emotion characteristics
- System trained on a range of emotions and speakers
  - enables appropriate spaces to be automatically generated

Controllable Speaker and Emotion Synthesis

Video

Polyglot Synthesis [14]

- 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

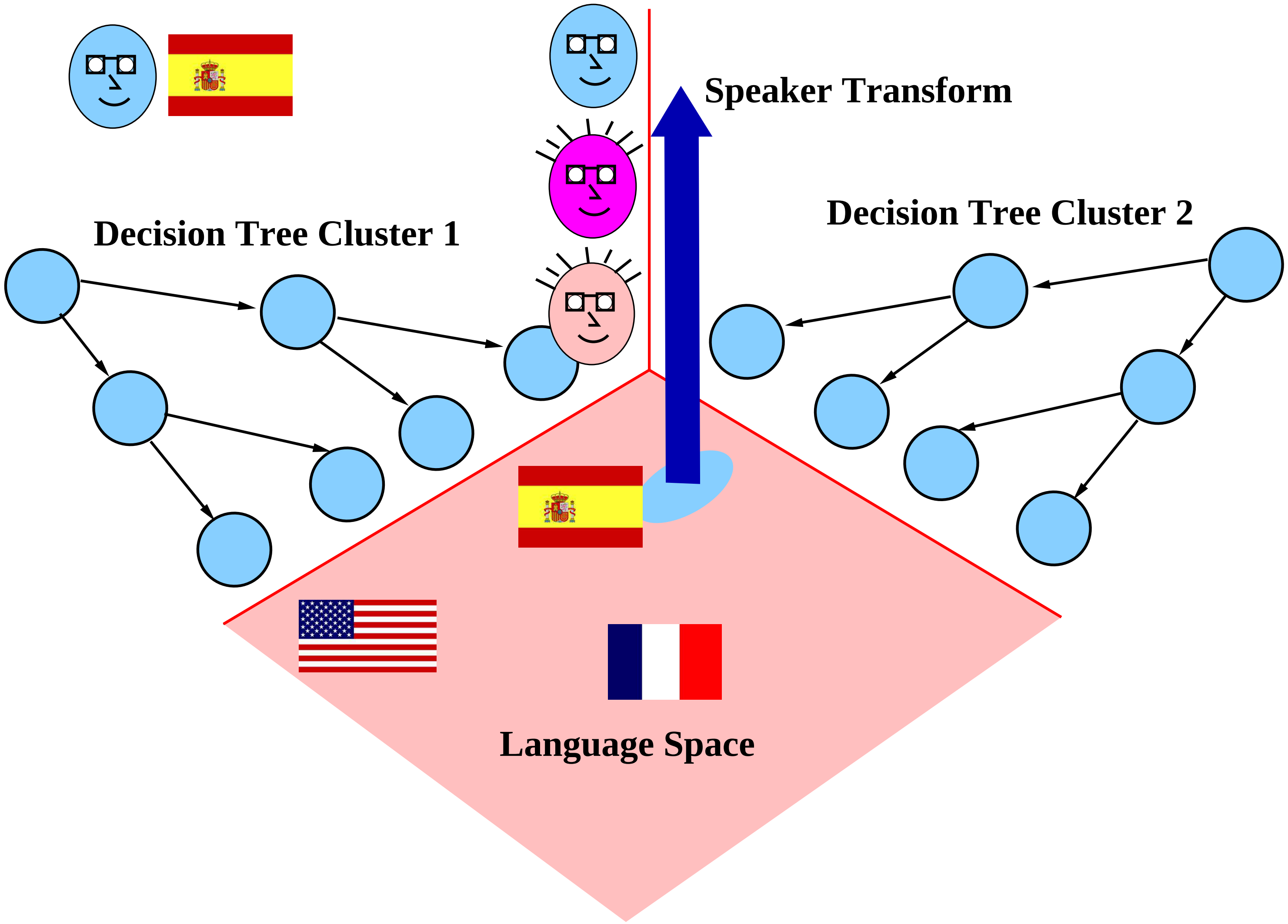
- Parametric statistical speech synthesis is attractive for this task

  – 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
- CAT to multiple decision trees [14]
  - a CAT specified language space for language attributes
  - use CMLLR to represent the speaker attributes

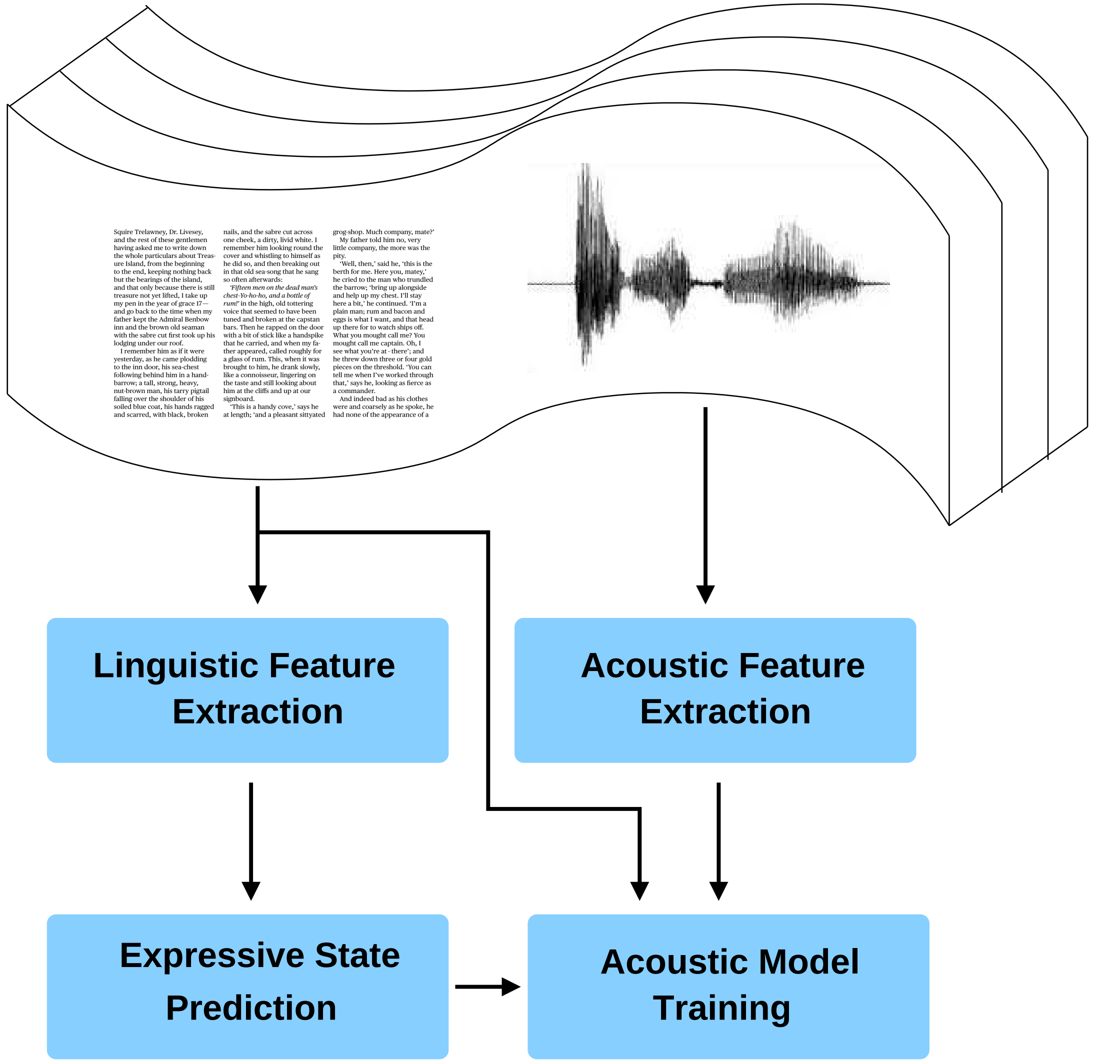
Speaker and Language Transformations



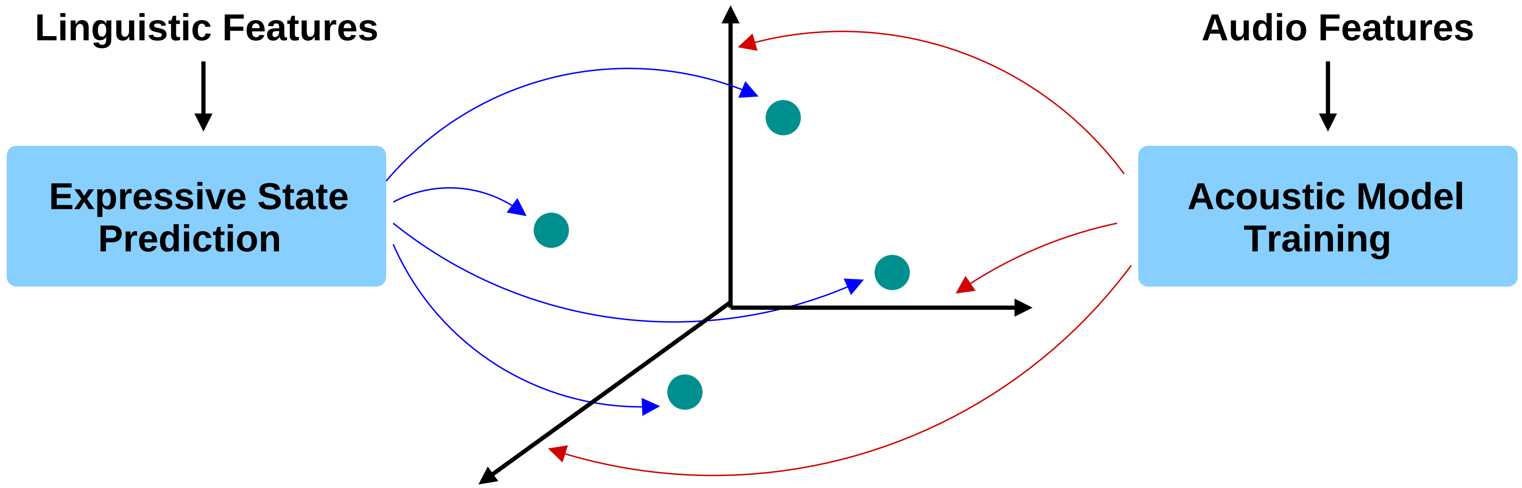
E-Book Reading

- E-books (Kindle, Kobo etc) increasingly popular
  - audio-books useful extension - “eyes-free” listening
  - is it possible to automatically generate an audio-book from text?
- Highly challenging task:
  - paragraph-level (not sentence) synthesis
  - high level of “listenability” required
  - expressive synthesis often used in reading
  - character voices sometimes used
- Consider expressive synthesis aspect:
  - normally need label the expressive state of every utterance
  - very hard to get consistency (what is an expressive state?)
- Jointly train extraction and synthesis to maximise likelihood

Integrated Expressive Speech Training [24]



Expressive Space Representation



- Automatically initialise a set of expressive states [25]
- Update complete system to maximise likelihood of audio data
  1. estimate acoustic models given prediction model
     determines bases of expressive space
  2. estimate prediction model given acoustic model
     determines position in expressive space
- Neural network based predictor used in initial work

Expressive Audio-Visual Synthesis

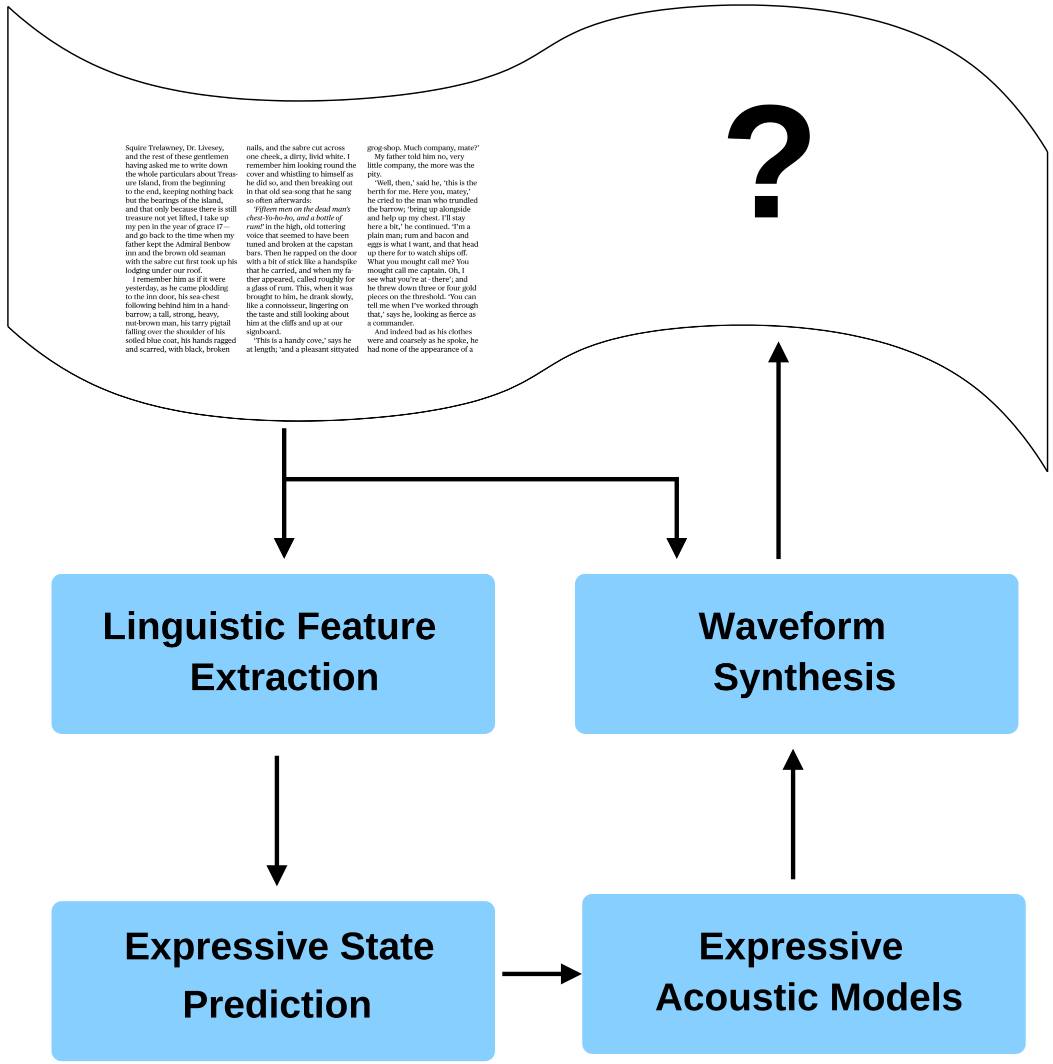
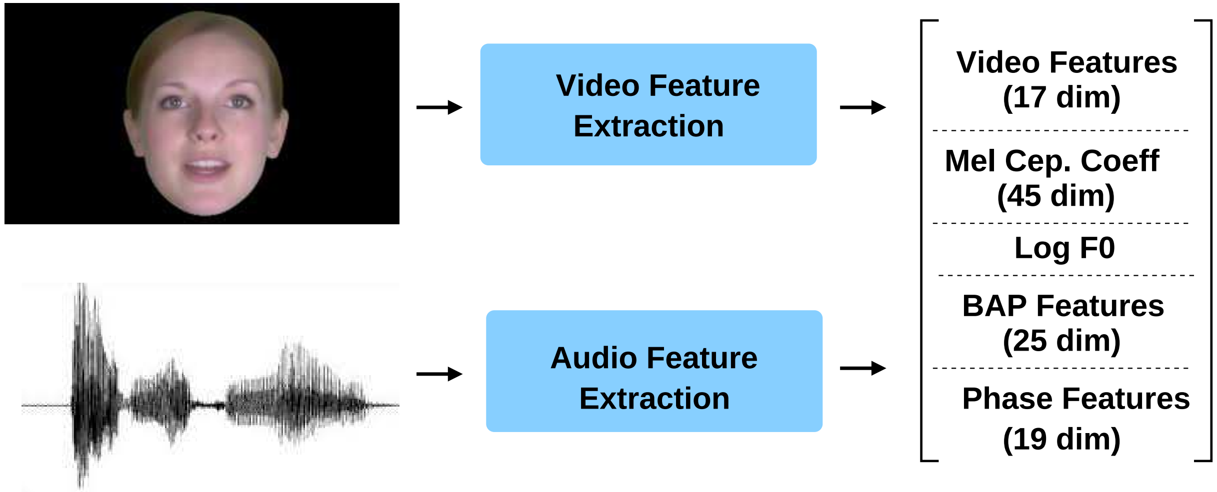
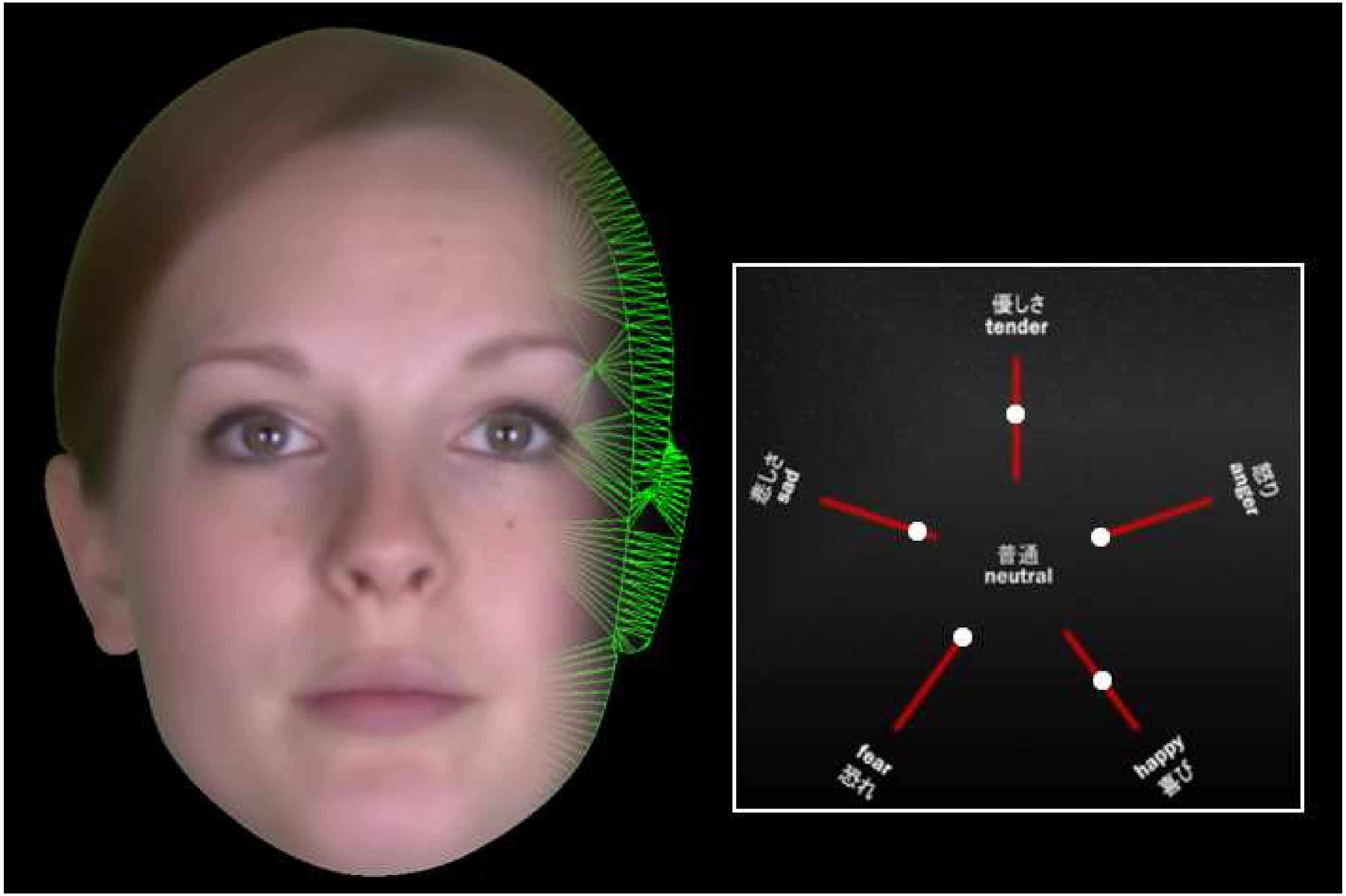


Photo-Realistic Expressive Speech Synthesis [26]



- Collect a high quality expressive audio-visual corpus
  - extract synchronised audio and video (upsampled) features
- Construct an expressive CAT speech synthesis system
  - add the video features (with delta and delta-deltas) as additional stream
- Synthesise audio and video features for selected emotion!

Photo-Realistic Expressive Speech Synthesis



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)
- Acoustic factorisation highly flexible/controllable adaptation
  - essential for controllable speech synthesis
  - improves efficiency/portability for speech recognition
- CAT and AVM useful approaches for diverse data
  - can be operated in adaptable or controllable modes

Acknowledgements

- This work has been funded from the following sources:

TOSHIBA
Leading Innovation

EPSRC
Pioneering research and skills

  - Toshiba Research Europe Ltd, Cambridge Research Lab
  - EPSRC - Natural Speech Technology

References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039 — 1064, 2009.

[2] 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.

[3] 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.

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

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.

[12] C. J. Leggetter and P. C. Woodland, "Flexible speaker adaptation for large vocabulary speech recognition," in *Proceedings Eurospeech*, 1995, pp. 1155–1158.

[13] Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis." in *INTERSPEECH*, 2006.

[14] 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.

[15] V. Wan, J. Latorre, K. Yanagisawa, N. Braunschweiler, L. Chen, M. J. Gales, and M. Akamine, "Building HMM-TTS voices on diverse data," *Journal of Selected Topics Signal Processing*, 2014.

[16] V. Wan, J. Latorre, K. Yanagisawa, M. J. Gales, and Y. Stylianou, "Cluster adaptive training of average voice models," in *ICASSP*, 2014.

[17] P. Lanchantin, M. Gales, S. King, and J. Yamagishi, "Multiple-average-voice-based speech synthesis," in *ICASSP*, 2014.

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

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

[20] 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.

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

[22] Y. Wang and M. Gales, "An explicit independence constraint for factorised adaptation in speech recognition." in *INTERSPEECH*, 2013, pp. 1233–1237.

[23] J. Latorre, V. Wan, M. J. Gales, L. Chen, K. Chin, K. Knill, and M. Akamine, "Speech factorization for HMM-TTS based on cluster adaptive training." in *INTERSPEECH*, 2012.

[24] L. Chen, M. Gales, N. Braunschweiler, M. Akamine, and K. Knill, "Integrated expression prediction and speech synthesis from text," *Journal of Selected Topics Signal Processing*, 2014.

[25] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M. J. Gales, and K. Knill, "Unsupervised clustering of emotion and voice styles for expressive TTS," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4009–4012.

[26] V. Wan, R. Anderson, A. Blokland, N. Braunschweiler, L. Chen, B. Kolluru, J. Latorre, R. Maia, B. Stenger, K. Yanagisawa, *et al.*, "Photo-realistic expressive text to talking head synthesis." in *INTERSPEECH*, 2013, pp. 2667–2669.