Speaker and Noise Factorisation on the AURORA4 Task

Yongqiang Wang & Mark Gales

May. 2011

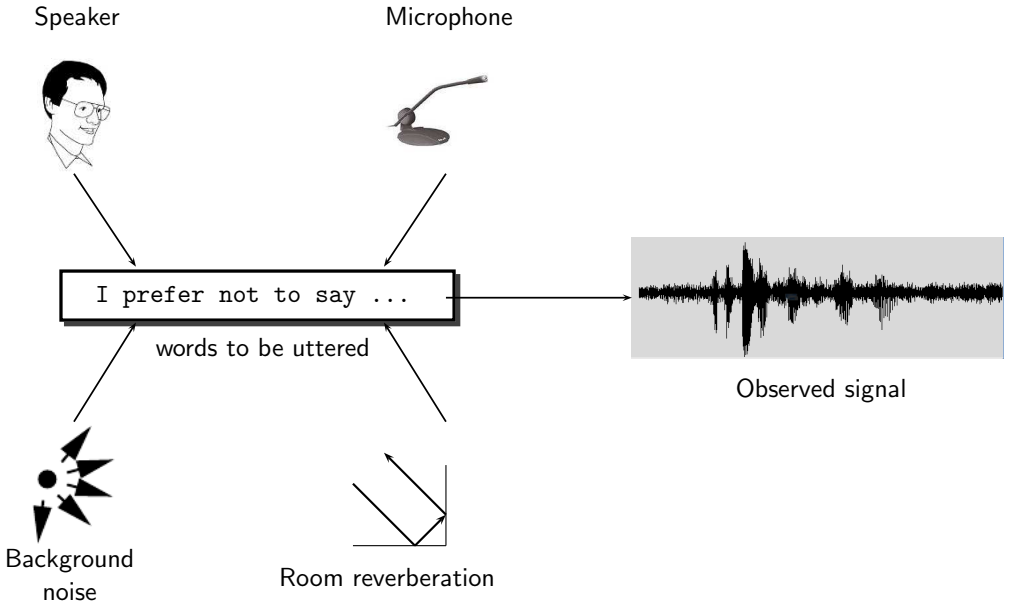
Cambridge University Engineering Department

Overview

- Model-based approaches to robust speech recognition
  - Acoustic environment
  - Speaker adaptation schemes
  - Noise robustness schemes
- Handling multiple acoustic factors
  - Acoustic factorisation
  - An example of acoustic factorisation – "Joint" (speaker and noise) adaptation
- Experiments
- Conclusion

Acoustic Environment

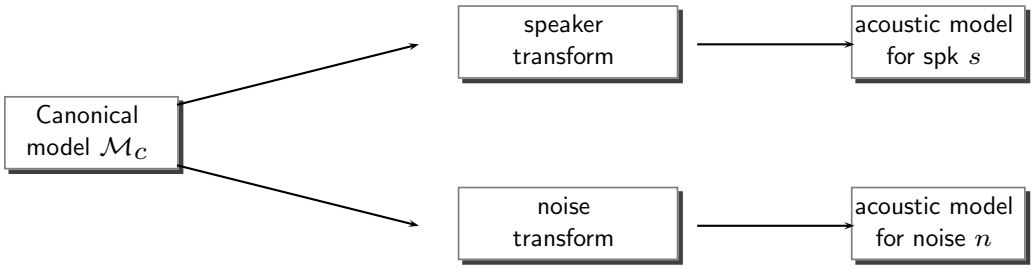
- Speech signals can be influenced by many factors
  - desired variations: words to be uttered
  - unwanted variations: speaker, noise , channel ...



- This work will consider model-based approach to robust speech recognition

Model-based Framework to Robust Speech Recognition

- *Canonical model* is built to model the desired variations
- *Transforms* are used to adapt canonical model to different acoustic conditions
  - Different transforms are developed to handle specific acoustic factors
    - Speaker adaptation
    - Noise compensation



  - Transforms can be combined to handle multiple acoustic factors

Speaker Adaptation

- Linear transforms are widely used to adapt the HMM parameters
  - Mean transform
$$\boldsymbol{\mu}^{(s)} = \mathbf{A}^{(s)}\boldsymbol{\mu} + \boldsymbol{b}^{(s)}$$
  - Linear transform $\mathbf{W}^{(s)} = [\mathbf{A}^{(s)}, \boldsymbol{b}^{(s)}]$ represents speaker s's characteristic
  - Large number of parameters to be estimated
    ↳not possible to estimate transforms robustly from single utterances
    ↳not suitable for very rapid adaptation

- Originally designed for speaker adaptation

- Also used as general linear transform for environmental adaptation

Noise Compensation Schemes

- A mismatch function is defined for the impact of environment
  - In the cepstral domain, clean speech $\boldsymbol{x}$ and corrupted speech $\boldsymbol{y}$ are related:

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

$\boldsymbol{h}$ channel distortion
$\boldsymbol{n}$ noise cepstral

- Model parameters are compensated using nonlinear transform:
  - VTS approximation

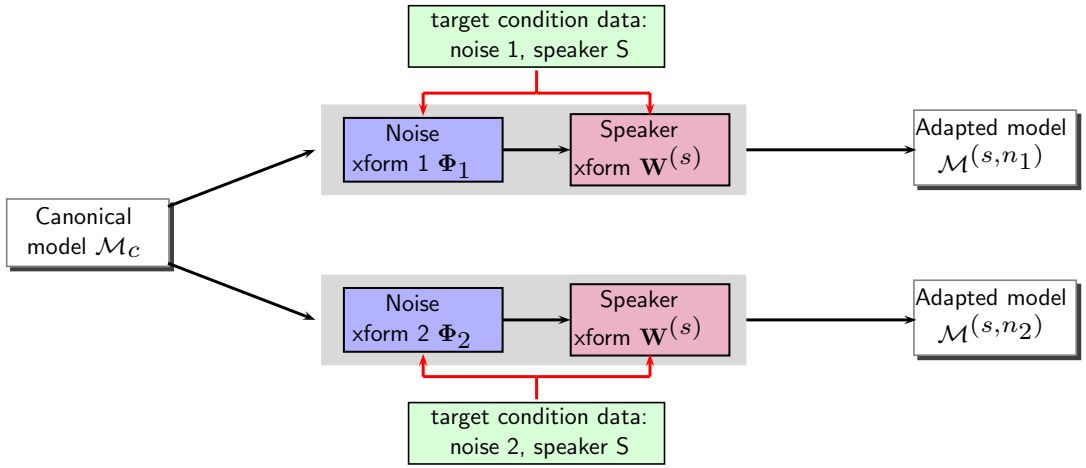
$$\begin{aligned}\boldsymbol{\mu}_{\mathrm{y}} &\approx \mathbf{f}(\boldsymbol{\mu}_{\mathrm{x}}, \boldsymbol{\mu}_{\mathrm{h}}, \boldsymbol{\mu}_{\mathrm{n}}) \\ \boldsymbol{\Sigma}_{\mathrm{y}} &\approx \mathrm{diag}\left(\mathbf{J}_{\mathrm{x}}\boldsymbol{\Sigma}_{\mathrm{x}}\mathbf{J}_{\mathrm{x}}^{\mathsf{T}} + \mathbf{J}_{\mathrm{n}}\boldsymbol{\Sigma}_{\mathrm{n}}\mathbf{J}_{\mathrm{n}}^{\mathsf{T}}\right)\end{aligned}$$

$\boldsymbol{\Phi} = (\boldsymbol{\mu}_{\mathrm{h}}, \boldsymbol{\mu}_{\mathrm{n}}, \boldsymbol{\Sigma}_{\mathrm{n}})$
noise transform parameter

  - Noise transform $\boldsymbol{\Phi}$ has a small number of parameters
    ↳noise transform parameter can be estimated at per utterance basis
    ↳ suitable for very rapid adaptation

Speaker and Noise Compensation: Batch Mode

- In real environment, there exists multiple acoustic attributes, e.g., speaker and noise
  - Transforms need to be combined to adapt to the target condition
- Batch mode combination
  - An example: VTS combined with MLLR ("VTS+MLLR")

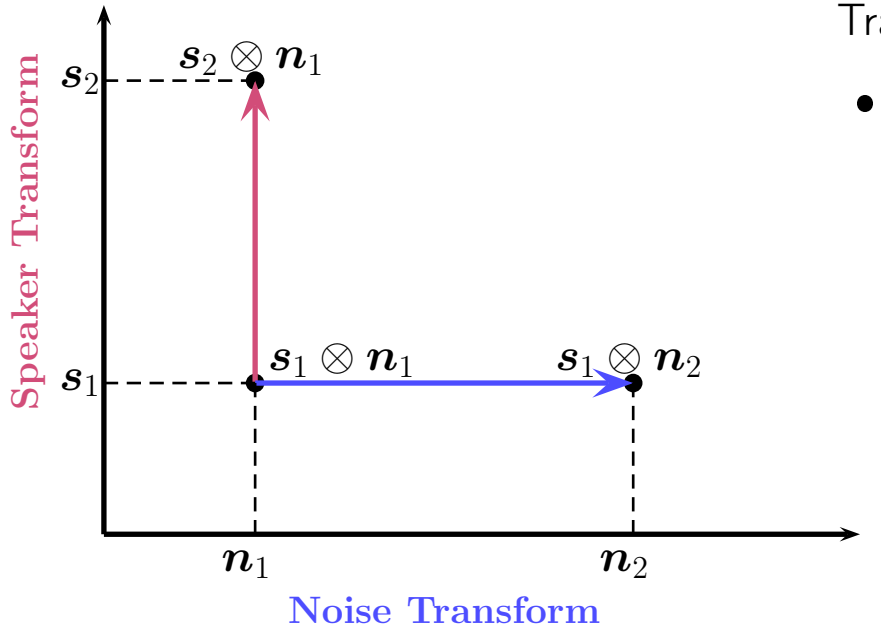


$$\boldsymbol{\mu}_{\mathrm{y}}^{(s)} = \mathbf{A}^{(s)}\mathbf{f}(\boldsymbol{\mu}_{\mathrm{x}}, \boldsymbol{\mu}_{\mathrm{h}}, \boldsymbol{\mu}_{\mathrm{n}}) + \boldsymbol{b}^{(s)}$$

$$\boldsymbol{\Sigma}_{\mathrm{y}}^{(s)} = \boldsymbol{\Sigma}_{\mathrm{y}}$$

  - Transforms are estimated for each speaker and noise conditions
    - VTS are estimated per utterance
    - Linear transform $\mathbf{W}^{(s)}$ must be estimated using multiple utterances

Acoustic Factorisation



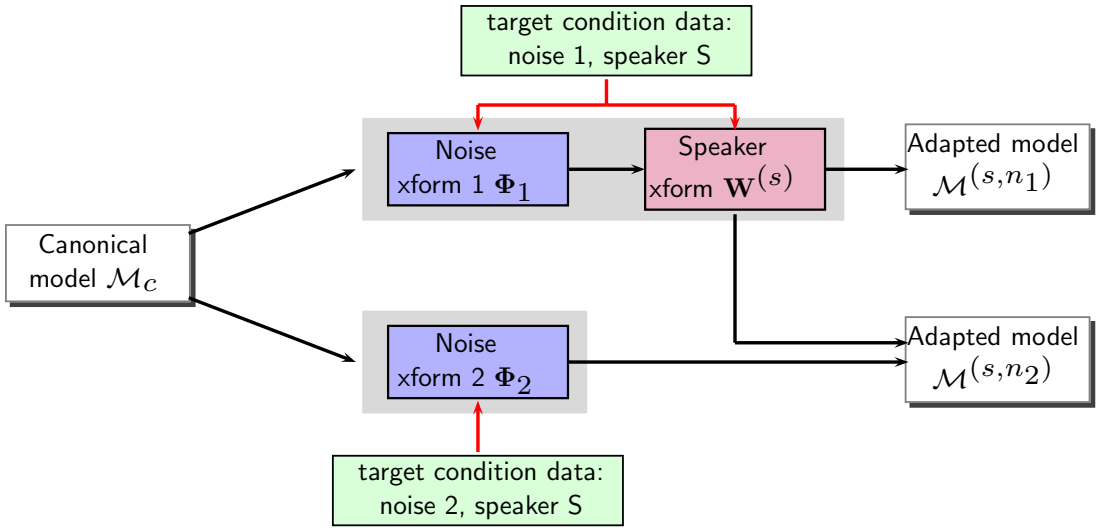
Transforms are "orthogonal":

- From (s_1, n_1) to condition (s_1, n_2):
  - Speaker transform is locked as s_1
  - Only need to update noise transform

- Flexible transforms:
  - Speaker transform can be used in a range of noise conditions
- Transforms are used in a factorised fashion, but are estimated jointly

Joint Speaker and Noise Compensation

- "Joint" transform: an example of acoustic factorisation



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

$$\boldsymbol{\Sigma}_{\mathrm{y}}^{(s)} = \mathrm{diag}(\mathbf{J}_{\mathrm{x}}\boldsymbol{\Sigma}_{\mathrm{x}}\mathbf{J}_{\mathrm{x}}^{\mathsf{T}} + \mathbf{J}_{\mathrm{n}}\boldsymbol{\Sigma}_{\mathrm{n}}\mathbf{J}_{\mathrm{n}}^{\mathsf{T}})$$

- Configuration investigated:
  - Estimate speaker transform in a particular noise condition ($\boldsymbol{n}_1$)
  - Use obtained speaker transform in all noise conditions ($\boldsymbol{n}_1$ and $\boldsymbol{n}_2$)
    ↳ Only noise transform need to be estimated for $\boldsymbol{n}_2$, very rapid adaptation

Experiments

- AURORA4 task
  - Derived from WSJ0 5k-word closed vocabulary dictation task
  - 7138 utterances from 83 speakers in training
  - 330 utterances from 8 speakers in testing
    - 14 test sets are defined

	Headset	Far-field micro.
No noise added	set A (test01)	set C (test08)
Noise added	set B (test02-07)	set D (test09-14)

- Acoustic model training:
  - Front-end: MFCC+C0+1^{st}, 2^{nd} derivatives
  - xwrd triphone, 3140 clustered states, 16-component system

Batch Mode Experiments

Schemes	set A	set B	set C	set D	Avg.
VTS	6.9	15.1	11.8	23.3	17.8
+MLLR	5.0	12.1	9.0	19.8	14.7
Joint	5.0	12.1	8.6	19.7	14.6

- speaker and noise transforms are estimated for each of the 14 test sets
  - Noise transforms varies from utterance to utterance
- Speaker and noise adaptation outperforms noise adaptation ("VTS") only
- Similar performances achieved by "Joint" and "VTS+MLLR" transform
- Batch mode adaptation requires a block of utterances to estimate transform
  ↳ Not very flexible to be used

Factorisation Experiments

Spk. Est.	Schemes	A	B	C	D	Avg.
—	VTS	6.9	15.1	11.8	23.3	17.8
01	+MLLR	5.0	20.2	16.5	28.0	22.2
	Joint	5.0	14.1	10.4	22.3	16.7
04	+MLLR	10.2	19.7	19.7	28.0	22.5
	Joint	7.0	12.5	11.0	20.4	15.4

- Speaker transform and noise transforms are estimated in a factorised mode
  - Speaker transform estimated from `test01/04` and fixed for all other sets
  - Noise tranform estimated per utterances
- Speaker transform estimated from `test04` improves average performance
- Performance of factorised adaptation (15.4%) is close to batch mode (14.6%)

Conclusion

- Handling multiple acoustic factors is important in realistic acoustic environment
- A powerful and flexible approach: acoustic factorisation
- An example of acoustic factorisation is presented: "Joint"
  - Allows very rapid speaker and noise adaptation
  - Speaker transforms are used across multiple noise conditions
  - The power and flexibility of model-based framework is demonstrated
    - "Joint" outperformed advanced front-end with MLLR for both batch and factorisation mode