Combining Eigenvoice Speaker Modeling and VTS-based Environment Compensation for Robust Speech Recognition

Zhijian Ou, Kan Deng

ozj@tsinghua.edu.cn, dengk11@mais.tsinghua.edu.cn
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

Motivation to combine eigenvoice and VTS

- **Speaker differences** and **environmental variations** are two major random factors in speech.
- They always **coexist** in real-world speech.
- Beneficial to consider joint handling of these two random factors.

Why choose eigenvoice and VTS

Previous studies :

1	Acoustic factorization	Use MLLR as speaker transform and cluster adaptive training as noise transform, both of which are linear transforms. This is not optimal if considering the nonlinear nature of the mismatch function relating the clean speech and the noisy speech.
2	VTS+MLLR	Conduct MLLR on top of the standard VTS. We can hardly interpret the MLLR used in this scheme as modeling the speaker variation.
3	Joint VTS MLLR	Replace the clean speech model used in the VTS with a speaker-adapted clean speech model by MLLR transform. The speaker's MLLR transform estimated from the noisy speech still carries information about current noise characteristics.

Our study :

- Consider how to do better **speaker and noise factorization.**
- Speaker and environmental variations have different characteristics.
  - For speaker variation, the a priori information could be obtained → eigenvoice
  - Noise is hard to be modeled a priori → VTS

How to combine eigenvoice and VTS

Replace the clean speech model used in the VTS
with a speaker-adapted clean speech model by eigenvoice.

$$y = x + h + C\ln\left(1+\exp\left(C^{-1}\left(n-x-h\right)\right)\right) \triangleq g(x,n,h)$$

$$\mu_x = e_0 + \sum_{r=1}^{R} w_r e_r$$

μ_h
Unknown constant

$$N\left(\mu_n, \Sigma_n\right)$$

Experimental results

Table 1: Recognition accuracies for per-utterance unsupervised eigenvoice adaptation under the clean condition

Eigenvoice Num	SI	1	2	3	**4**	5	6	7	8	9
Clean Acc. (%)	99	98.98	99.1	99.1	**99.1**	99.11	99.09	99.08	99.1	99.09

For each utterance, initialize the noise mean and variance μ_n, Σ_n using the first and last several frames that are assumed to be speech-free, and set $\mu_h = 0$.
Update μ_y, Σ_y with w=0, and **do one pass recognition.**

$$\begin{cases} \mu_{y,jk}^{(u)} = g\left(e_{0,jk} + \sum_{r=1}^{R} w_r e_{r,jk}, \mu_n^{(u)}, \mu_h^{(u)}\right) \\ \Sigma_{y,jk}^{(u)} = G_{x,jk}^{(u)} \Sigma_{x,jk} \left(G_{x,jk}^{(u)}\right)^T \\ \qquad + (I - G_{x,jk}^{(u)}) \Sigma_n^{(u)} (I - G_{x,jk}^{(u)})^T \end{cases}$$

re-estimate μ_n, Σ_n
Update μ_y, Σ_y with w=0, and **do one pass recognition**

$$\hat{\mu}_n = \mu_n + \left[\sum_{j,k} \gamma_{jk} G_{n,jk}^T \Sigma_{y,jk}^{-1} G_{n,jk}\right]^{-1} \sum_{jk} G_{n,jk}^T \Sigma_{y,jk}^{-1} c_{y,jk}$$

Table 2: Average recognition accuracies for per-utterance unsupervised adaptation under noisy conditions by various schemes

Scheme	SetA	SetB	SetC	Avg. Acc.
Baseline	59.33	56.19	66.26	59.46
VTS Init	87.65	88.38	88.11	88.04
VTS with SI model	90.03	90.39	90.30	90.23
VTS with 4 eigenvoices	90.58	91.15	90.43	90.78

Based on current estimates of noisy speech parameters μ_y, Σ_y and eigenvoice coefficients w, re-estimate the eigenvoice coefficients and update the speaker adapted mean.
Update μ_y, Σ_y with new w, and **do one pass recognition.**

Table 3: Per-utterance unsupervised adaptation experimental results for recognizing the clean utterance, using the clean speaker model estimated from the noisy utterance under the "VTS with 4 eigenvoices" scheme.

The "clean" represents the standard unsupervised eigenvoice adaptation scheme (i.e. **using the clean speaker model estimated from the clean utterance itself**)

SNR	**clean**	20dB	15dB	10dB	5dB	0dB
Acc (%)	**99.10**	99.10	99.11	99.12	99.08	99.00

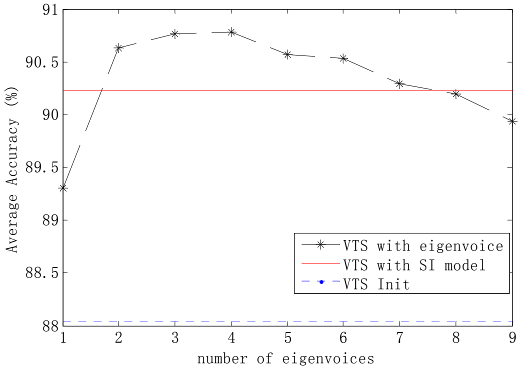

Fig. 1: Average recognition accuracies for per-utterance unsupervised adaptation under noisy conditions by various schemes