Speaker-Phonetic Vector Estimation for Short Duration Speaker Verification

Jianbo Ma[1], Vidhyasaharan Sethu[1], Eliathamby Ambikairajah[1,2], Kong Aik Lee[3]

[1] The School of Electrical Engineering and Telecommunications, UNSW, Sydney NSW 2052, Australia

[2] DATA61, CSIRO, Australia, Sydney NSW 2015, Australia

[3]Data Science Research Laboratories, NEC Corporation, Japan

jianbo.ma@unsw.edu.au

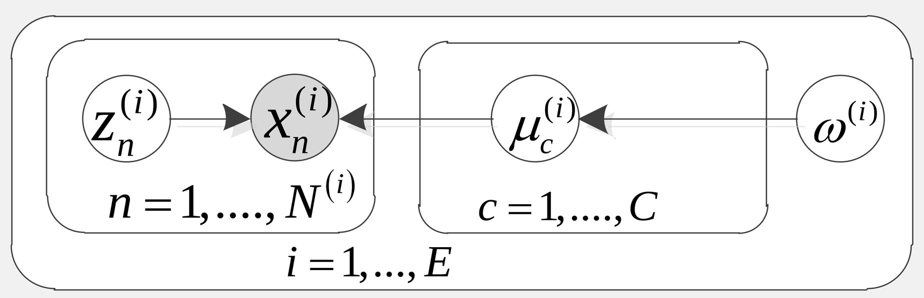
1. Introduction

- State-of-art text-independent system includes i-vector representation.
- Gaussian distribution is conventionally used to model distributions of latent variable for deriving i-vector representations.
- Relaxing the Gaussian assumption can form vector representations with both phonetic and speaker meaning for each utterance.
- These representations is able to perform content matching that is beneficial for short duration speaker verification.

2. Total Variability Model

- i-vector generative model

$$\mu_c{}^{(i)} = \mu_{c0}{}^{(i)} + T_c\omega^{(i)}$$



- Prior distribution of latent variable ω

$$p(\omega) = \mathcal{N}(0, I)$$

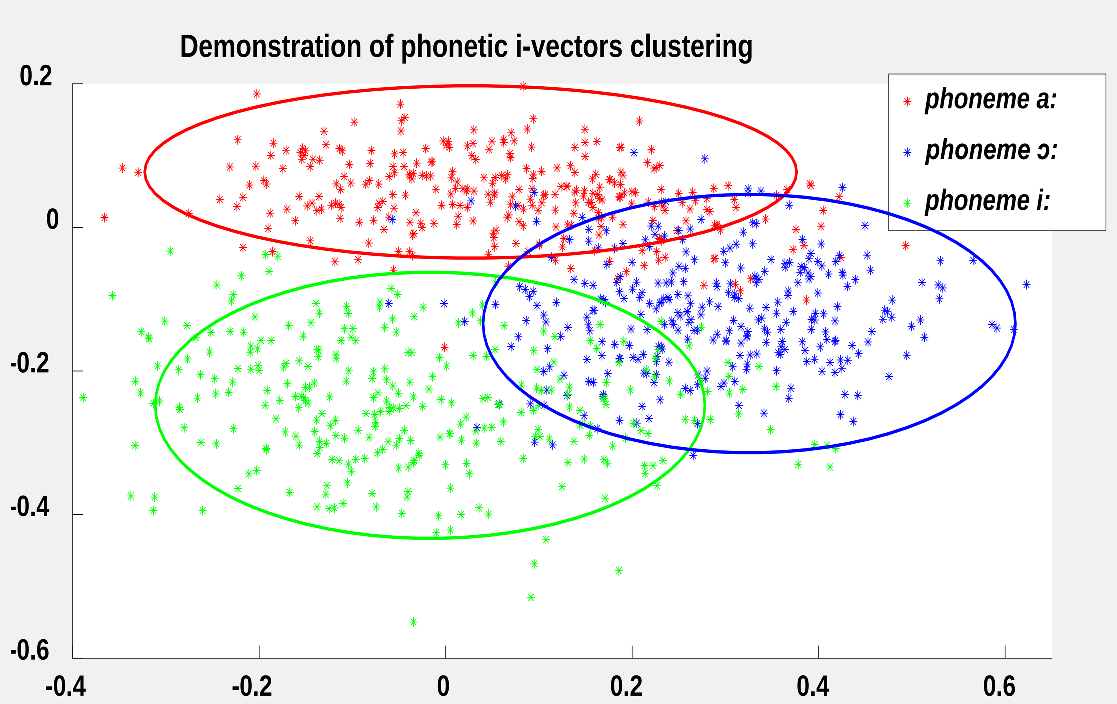
- Latent variable x and corresponding supervectors (M_i) are assumed to have Gaussian distributions.
- Inference of i-vector

$$p(\omega|X) \propto p(X|\omega)p(\omega)$$

z - labeling variables;
x - feature frames;
μ - means of the supervectors;
ω - latent variable;.
i – utterance index;
c - mixture component in UBM;
n - feature frame index.

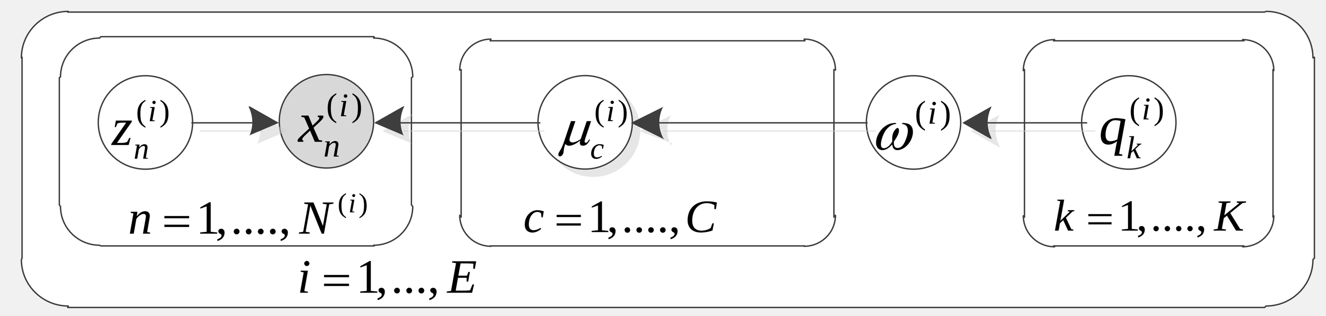
3. Phonetic i-vectors Analysis

- Phonetic i-vector clustering
  - Phonetic i-vectors are estimated by using features belong to same phonetic class.
  - Phonetic i-vector projected by PCA.
  - Different distributions found for different phonetic i-vectors.



- *For long duration utterances, it is not a problem due to sufficient information for each phoneme.*
- *For short duration utterances, i-vector biased toward some dominant groups and differ from one to another, resulting in larger within-class covariance.*

4. Proposed Speaker–Phonetic Vector



- **Introduce mixture of Gaussians as priors**

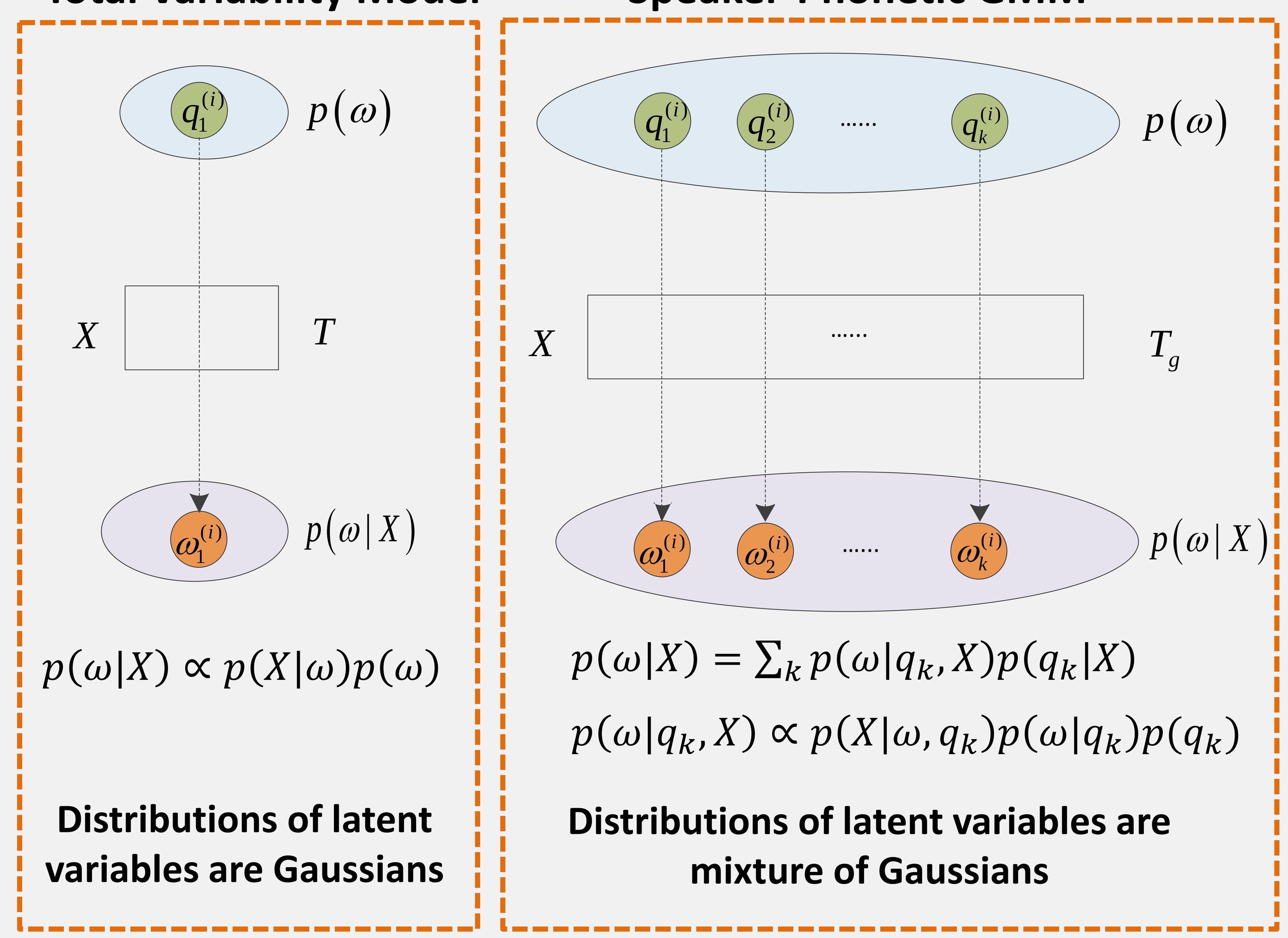
$$p(\omega) = \sum_k p(\omega|q_k)p(q_k)$$

$$p(\omega|q_k) = \mathcal{N}(m_k, B_k) \qquad p(q_k) = 1/K$$

q - state variables
k - state index

- Latent variables and supervectors are distributed as mixture of Gaussians

Total Variability Model | **Speaker-Phonetic GMM**



- **$E[\omega|q, X]$ is the phonetic-speaker vector, ω_k**
- A bank of GPLDAs are used to compare phonetic-speaker vectors. Scores are combined as:

$$Score(\omega_e, \omega_t) = \sum_k \gamma_k Score(\omega_{ik}, \omega_{tk})$$

where $\gamma_k = \frac{N_{tk}}{\sum_k N_{tk}}$, N_{tk} is the zeroth-order statistics of state k.

5. Experimental Results

- The BUT group's phoneme decoder of Hungarian language is used to obtain phonetic posterior probabilities $p(q_k|X)$
- Similar phonemes are grouped to form 14 phonetic groups
- One Gaussian $\mathcal{N}(m_k, B_k)$ is then assigned to each phonetic group to fit the phonetic vectors

Table 1. Experimental results (EER %) of NIST SRE' 2010 8CONV-10SEC

EER % results NIST SRE' 2010 8CONV-10SEC							
		Male			**Female**		
	System	10s	5s	3s	10s	5s	3s
1	Baseline	5.12	10.61	17.43	6.16	12.43	18.90
2	Proposed	5.34	10.26	14.26	6.68	11.54	16.52
4	Fusion 1+2	3.82	8.10	12.19	4.94	8.90	14.15
5	LV system*	4.40	8.99	14.06	5.92	11.24	15.31

- Proposed phonetic-speaker vector representation outperformed i-vector baseline for shorter conditions.
- Substantial improvements are obtained by fusing phonetic-speaker vector and i-vector systems in score level, showing complementary behaviour.
- The proposed method is compared with local acoustic variability model. Phonetic-speaker vector outperformed it in both single and fused systems.

* J. Ma, V. Sethu, E. Ambikairajah, and K. A. Lee, "Incorporating Local Acoustic Variability Information into Short Duration Speaker Verification," Proc. Interspeech 2017, pp. 1502-1506, 2017

6. Conclusion

- i-vectors of different phonemes are not identically distributed. This leads to i-vector representation having larger within-class covariance for short duration utterances.
- The proposed phonetic-speaker vector representation is derived by introducing mixture of Gaussians to model distributions of latent variables.
- The proposed method is able to perform soft content matching and outperformed i-vector representation system in short condition.