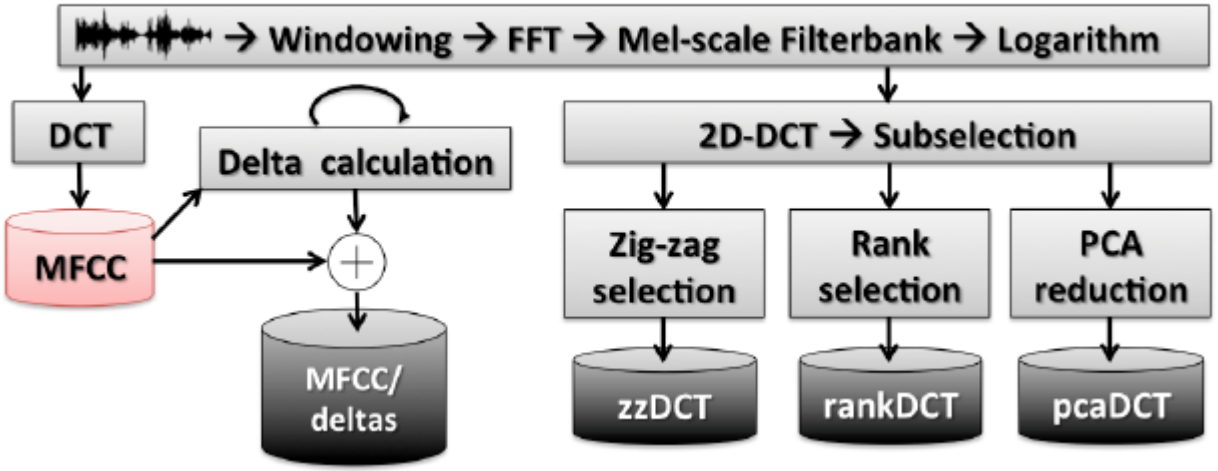
ASR and SID Research Frontier (I)

Dong Wang

DCT coefficients

- Add DCT coefficients (on Fbanks) in feature



Mitchell McLaren, Yun Lei, **IMPROVED SPEAKER RECOGNITION USING DCT COEFFICIENTS AS FEATURES, ICASSP 2015.**

Uncertainty Propagation

- How to use the covariance in i-vector extraction?
- How to deal with uncertainty caused by short utterances?
- Propagate covariance of i-vectors to PLDA

$$i_r = m + Vy + \epsilon_r \qquad i_r = m + U_r x_r + Vy + \epsilon_r.$$

- How to do length-normalization? Normalize T.

- Patrick Kenny, et al. “PLDA FOR SPEAKER VERIFICATION WITH UTTERANCES OF ARBITRARY DURATION”, ICASSP 2013
- Wei Rao, et al. “NORMALIZATION OF TOTAL VARIABILITY MATRIX FOR I-VECTOR/PLDA SPEAKER VERIFICATION”, ICASSP 2015

Multi-conditional training

- Add noise in model PLDA training and enrollment

Enroll MC	PLDA MC	Male	Female
No	No	2.25 (0.87)	2.23 (0.92)
Yes	No	1.58 (0.62)	2.06 (0.83)
No	Yes	1.10 (0.37)	1.26 (0.53)
Yes	Yes	1.06 (0.34)	1.3 (0.50)

Averaging	LN	Male	Female
No length normalization			
Avg. i-vec	None	1.58 (0.58)	2.08 (0.75)
Avg. score	None	1.71 (0.69)	2.22 (0.88)
With length normalization			
Avg. i-vec	Before	1.06 (0.34)	1.30 (0.50)
Avg. i-vec	After	1.11 (0.43)	1.43 (0.71)
Avg. score	Before	1.25 (0.48)	1.52 (0.66)

Minmax i-vector

- In order to improve performance with condition mismatch, good to minimize the max error.

- Ville Hautamaki et al. “Minimax i-vector extractor for short duration speaker verification”, Interspeech 2013

Non-Gaussian prior

$$\mathbf{m} = \mathbf{m_0} + \mathbf{T}\mathbf{w} + \mathbf{T}\mu_\mathbf{p}$$

$$\mathbf{w} \sim \mathcal{N}(0, \mathbf{\Sigma_p})$$

- Start out with an already trained T matrix.
- For each source, extract an informative prior N (μp, Σp) using the minimum divergence estimation.
- Re-center the first order statistics Fe around the relevant source-specific mean to give F,
- Extract i-vectors, by matching the now zero-mean informative prior N (0, Σp)for each source to the relevant re-centered first-order statistics

- Sven Ewan Shepstone et al., SOURCE-SPECIFIC INFORMATIVE PRIOR FOR I-VECTOR EXTRACTION, ICASSP 2015

Clean i-vector Estimation

- Estimate clean i-vectors from noisy ones

$$N = Y - X$$

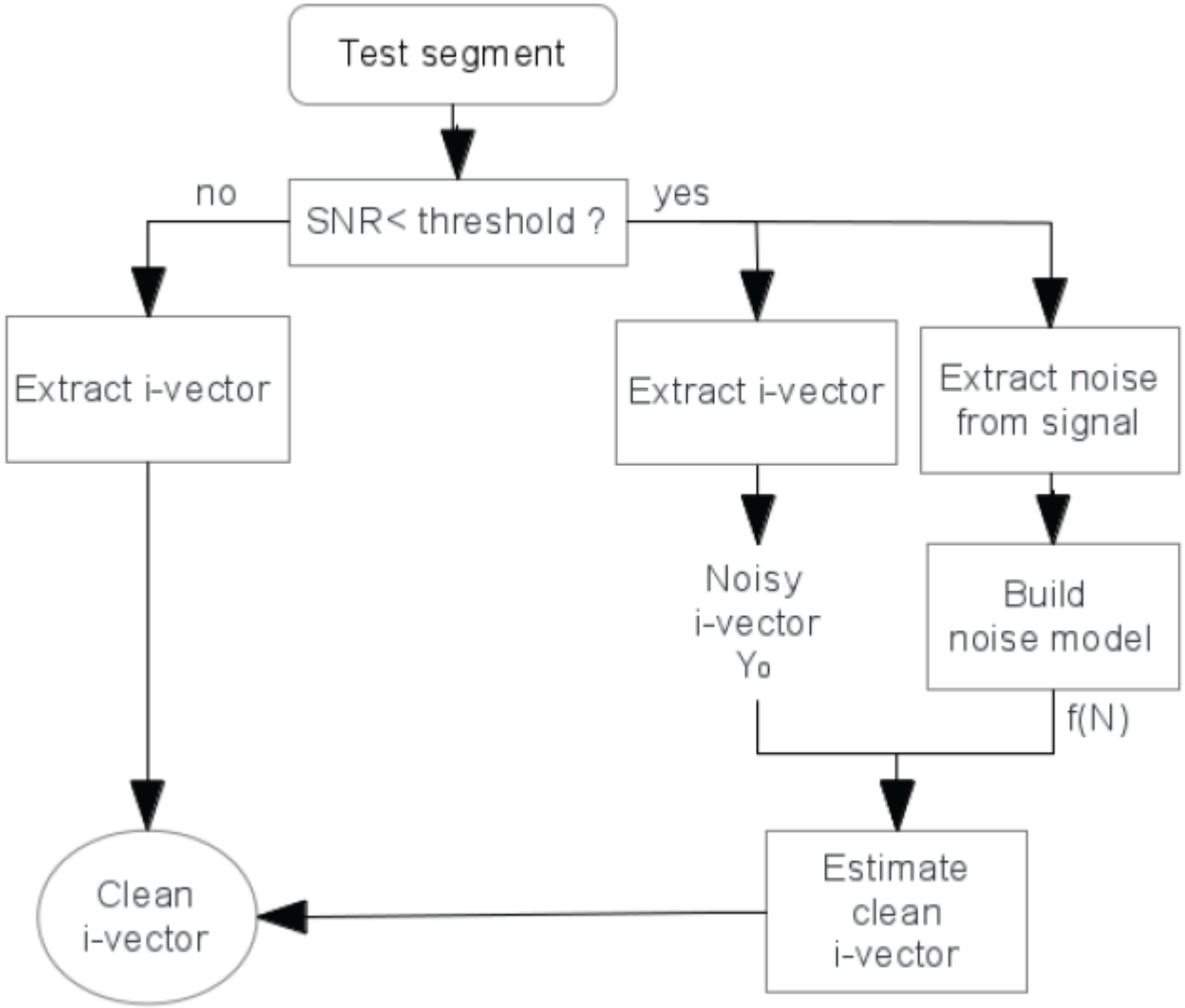
$$f(X) = \mathcal{N}(\mu_X, \Sigma_X) \qquad f(N) = \mathcal{N}(\mu_N, \Sigma_N)$$

$$f(Y_0|X) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_N|^{\frac{1}{2}}} exp^{-\frac{1}{2}(Y_0 - X - \mu_N)^t \Sigma_N^{-1}(Y_0 - X - \mu_N)}$$

$$\hat{X}_0 = \underset{X}{\operatorname{argmax}}\{\ln f(X/Y_0)\} \qquad \hat{X}_0 = \underset{X}{\operatorname{argmax}}\{\ln f(Y_0/X) f(X)\}$$

$$\hat{X}_0 = (\Sigma_N^{-1} + \Sigma_X^{-1})^{-1}(\Sigma_N^{-1}(Y_0 - \mu_N) + \Sigma_X^{-1}\mu_X)$$

Waad Ben Kheder, ADDITIVE NOISE COMPENSATION IN THE I-VECTOR SPACE FOR SPEAKER RECOGNITION, ICASSP 2015.


Test segment
SNR< threshold ?
no
yes
Extract i-vector
Extract i-vector
Extract noise from signal
Noisy i-vector Y0
Build noise model
f(N)
Estimate clean i-vector
Clean i-vector

Gaussian Mixture for i-vector Residual

- In DNN-based i-vector, the residual is not necessarily Gaussian. Therefore a GMM is better.

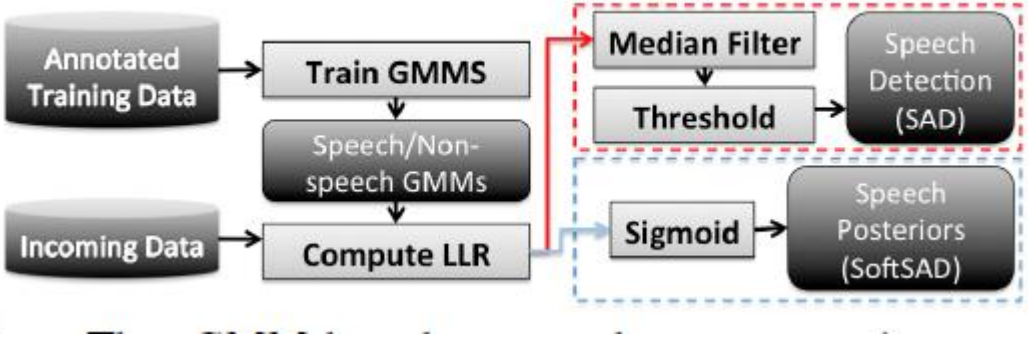
$$\tilde{F}_{ij} = \mathbf{T_i x_j} + \epsilon_{ij}$$

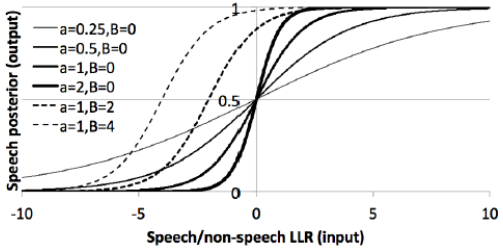
$$P(\mathbf{x_j}) = \mathcal{N}(\mathbf{0}, I), \qquad P(\tilde{F}_{ij}|\mathbf{x_j}) = \mathcal{N}(\mathbf{T_i x_j}, \frac{{\sigma_i}^2}{N_{ij}})$$

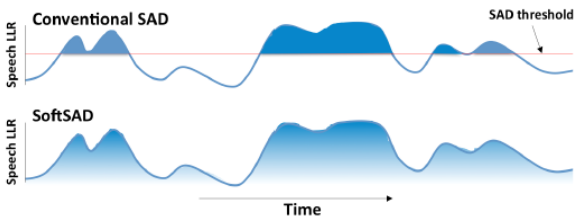
$$P(\mathbf{x_j}) = \mathcal{N}(\mathbf{0}, I), \qquad P(\tilde{F}_{ij}|\mathbf{x_j}) = \sum_{k=1}^{K} \pi_{ik} \mathcal{N}(\mathbf{T_i x_j}, \frac{{\sigma_{ik}}^2}{N_{ij}})$$

Frame Confidence in Frame Selection

- Weight In statistic domain by frame confidence
- Weight In score domain? Submitted to interspeech 2015.







SOFTSAD: INTEGRATED FRAME-BASED SPEECH CONFIDENCE FOR SPEAKER RECOGNITION, ICASSP 2015.

Adaptation: Subspace GMM

- Using SGMM for i-vector adaptation

$$p(\mathbf{x} \mid j) = \sum_{i=1}^{M_j} w_{ji} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_{ji}),$$

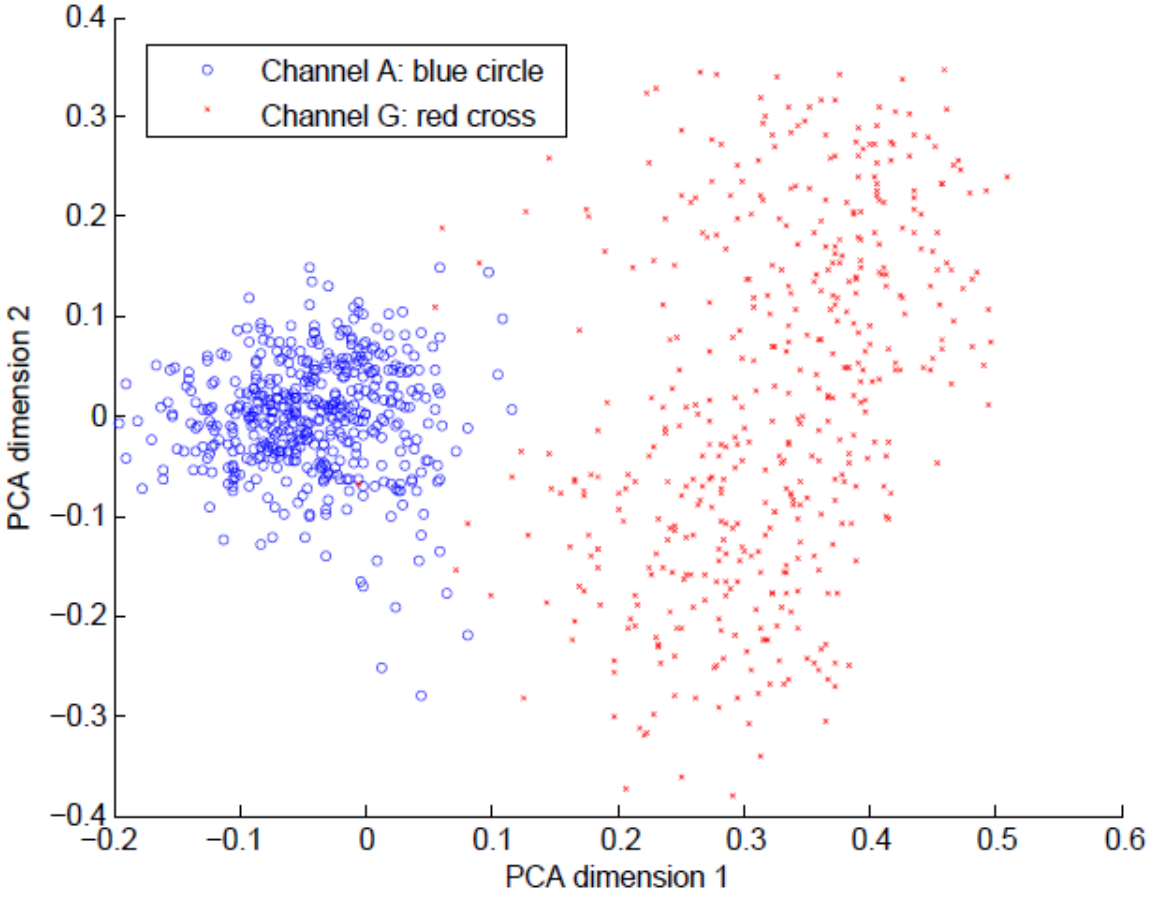
$$p(\mathbf{x} \mid j, s) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{jmi}^{(s)}, \boldsymbol{\Sigma}_i)$$

$$\boldsymbol{\mu}_{jmi}^{(s)} = \mathbf{M}_i \mathbf{v}_{jm} + \mathbf{N}_i \mathbf{v}^{(s)}$$

$$w_{jmi} = \frac{\exp \mathbf{w}_i^{\mathrm{T}} \mathbf{v}_{jm}}{\sum_{l=1}^{I} \exp \mathbf{w}_l^{\mathrm{T}} \mathbf{v}_{jm}},$$

Petr Motlicek, et al. **EMPLOYMENT OF SUBSPACE GAUSSIAN MIXTURE MODELS IN SPEAKER RECOGNITION, ICASSP 2015.**

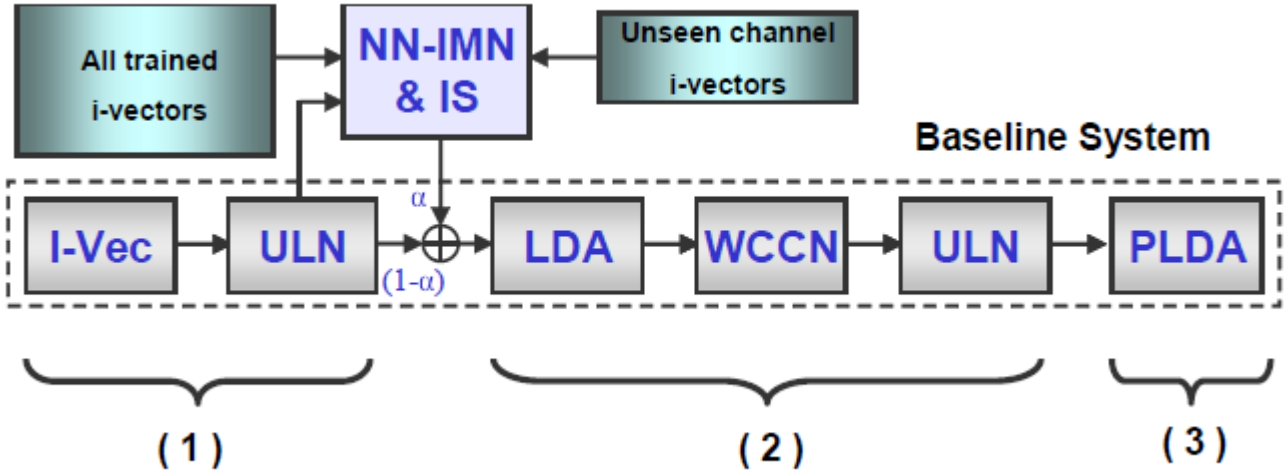
I-vector Normalization for Unseen Channels



I-vector Normalization for Unseen Channels

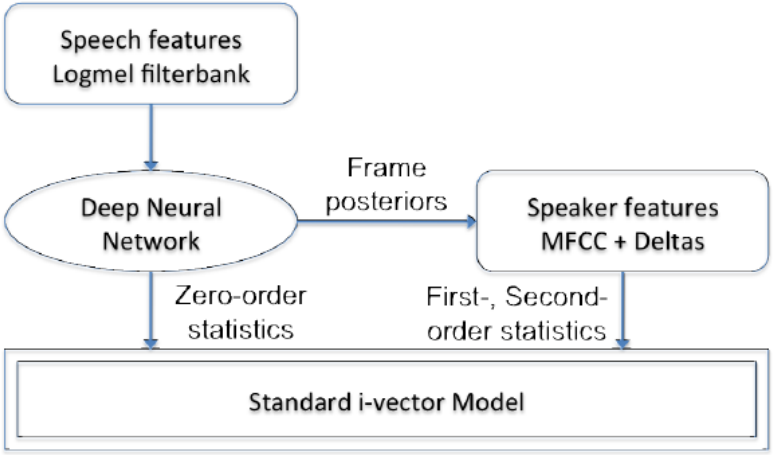
$$\mathbf{w}_c = \mathbf{w} - \frac{1}{K}\sum_{k=1}^{K} NN_k(\mathbf{w}).$$

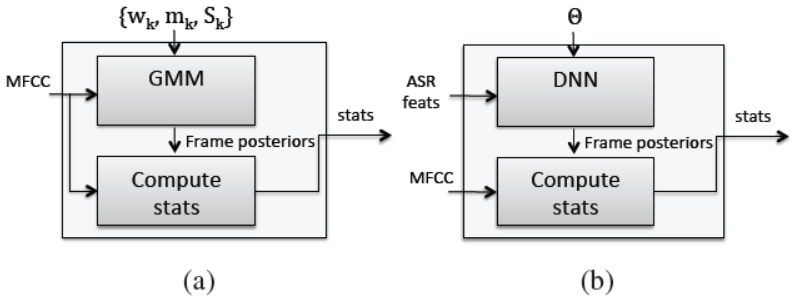
$$\begin{aligned}\widetilde{\mathbf{w}}_c &= \alpha\mathbf{w}_c + (1-\alpha)\mathbf{w} \\ &= \mathbf{w} - \frac{\alpha}{K}\sum_{k=1}^{K} NN_k(\mathbf{w}),\end{aligned}$$



Weizhong Zhu et al. NEAREST NEIGHBOR BASED I-VECTOR NORMALIZATION FOR ROBUST SPEAKER RECOGNITION UNDER UNSEEN CHANNEL CONDITIONS , ICASSP 2015

DNN for SID





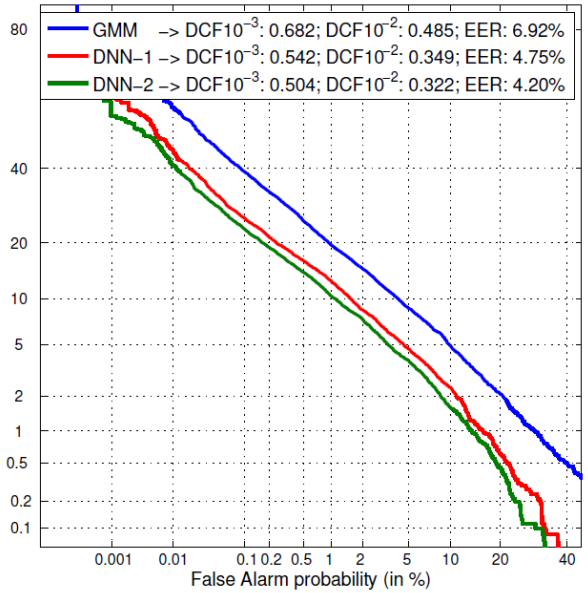
Y Lei, N. Scheffer, L. Ferrer and M. McLaren"A novel scheme for speaker recognition using phonetically-aware deep neural net," in Proc. of ICASSP, Florence, Italy, 2014.

Patrick Kenny, Deep Neural Networks for extracting Baum-Welch statistics for Speaker Recognition, Speaker Odyssey 2014

Ming Li and Wenbo Liu, "Speaker verification and spoken language identification using a generalized i-vector framework with phonetic tokenizations and tandem features," in Proc. INTERSPEECH, 2014.

- Adaptation for DNN-based SID

Task	System	$DCF10^{-3}$	$DCF10^{-2}$	EER(%)
OOD	GMM	0.682	0.485	6.92
	DNN-2	0.504	0.322	4.20
UA-LN	GMM	0.627	0.425	5.55
	DNN-2	0.431	0.273	3.31
UA-LN-PLDA	GMM	0.445	0.264	2.72
	DNN-2	0.271	0.172	2.09
IND	GMM	0.399	0.235	2.32
	DNN-2	0.260	0.164	1.82



D. Garcia-Romero, X. Zhang, A. McCree, and D. Povey, “Improving Speaker Recognition Performance in the Domain Adaptation Challenge using Deep Neural Networks”, in Proc. of SLT, USA, 2014.

Change Point Detection

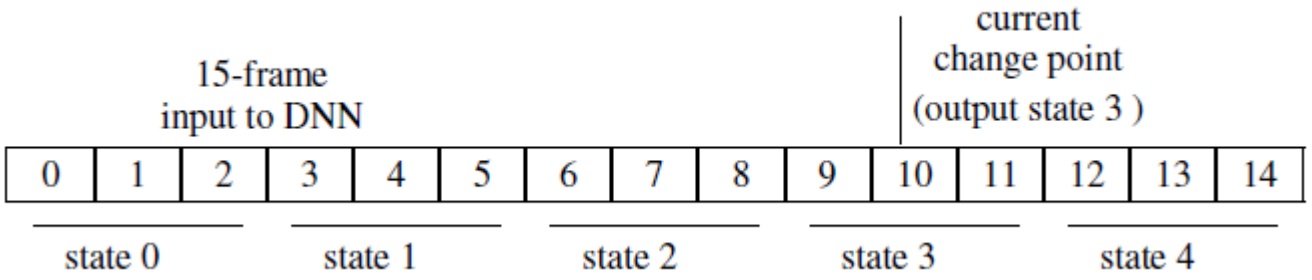
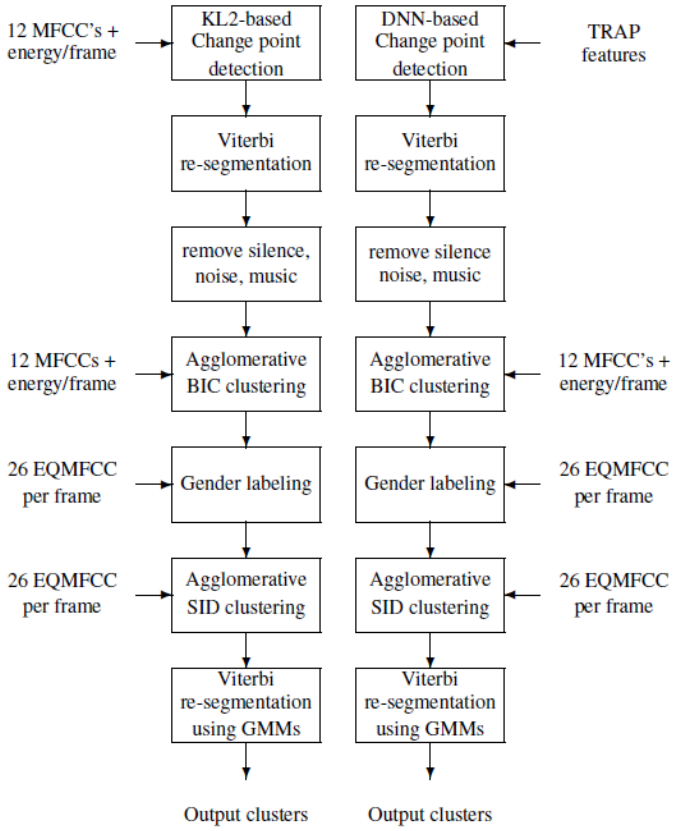


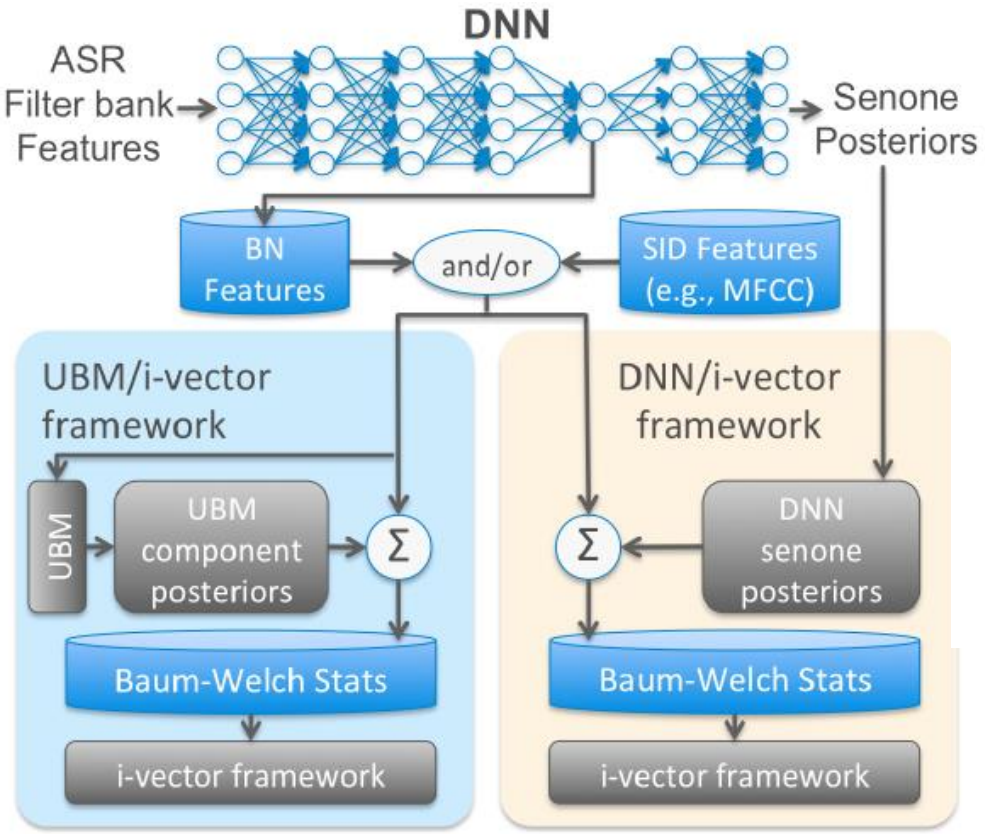
Fig. 1. *15-frame input to DNN illustrating how output state is assigned to this DNN input during DNN training. In this example, the speaker change point corresponds to frame 10, so the output state is 3. If there was no speaker change point inside these 15 frames, then the output state assigned would be state 5.*

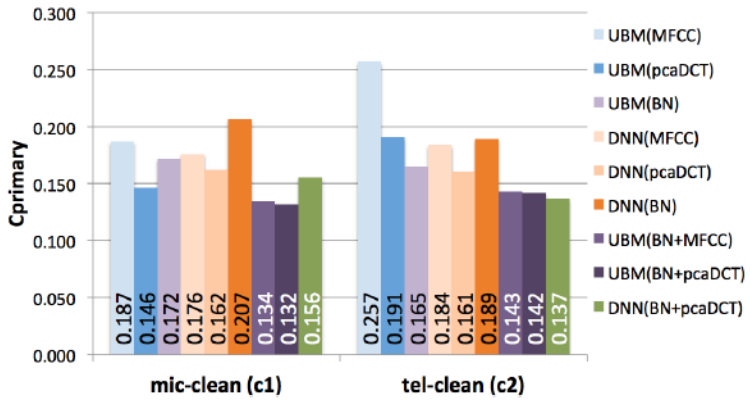
- Vishwa Gupta, SPEAKER CHANGE POINT DETECTION USING DEEP NEURAL NETS, ICASSP 2015.

Architecture



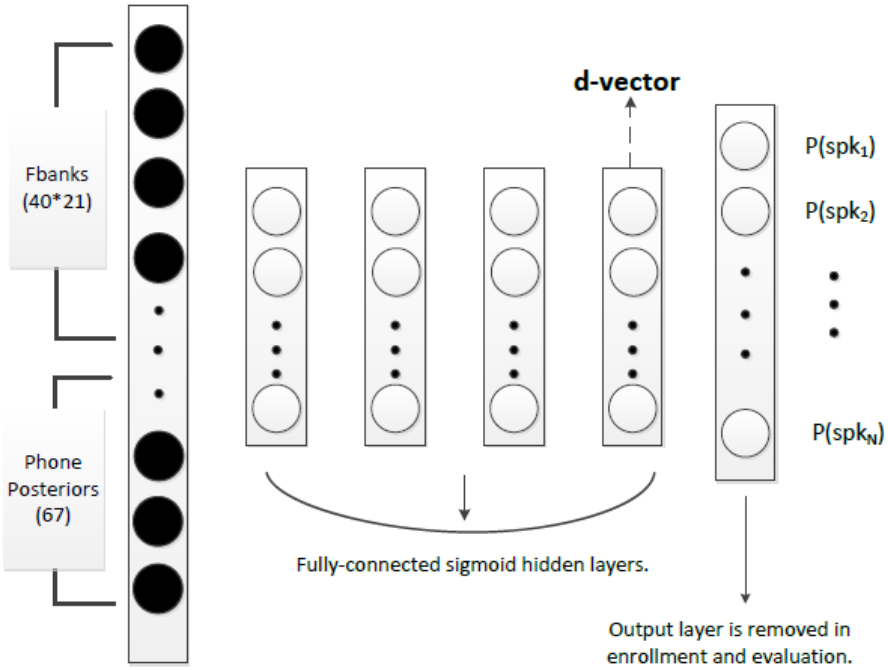
Feature-based DNN SID





M. McLaren, Y. Lei, and L. Ferrer , Advances in deep neural network frameworks for speaker recognition, ICASSP 2015

D-vector

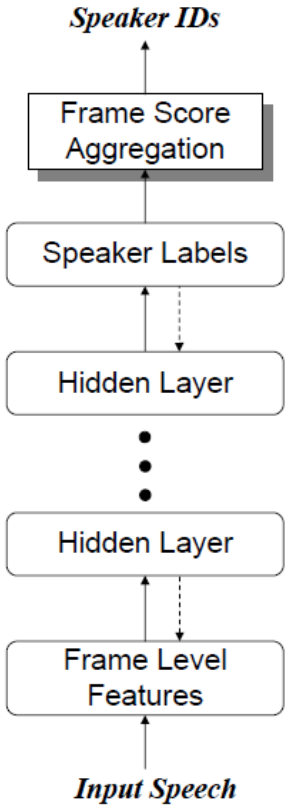


	cosine	LDA	PLDA	NLDR
i-vector	19.32	11.09	8.70	-
d-vector	13.58	13.07	15.45	12.79

V. Ehsan, L. Xin, M. Erik, L. M. Ignacio, and G.-D. Javier, “Deep neural networks for small footprint text-dependent speaker verification,” IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), vol. 28, no. 4, pp. 357–366, 2014.

DNN-based cochannel SID

- Speaker posterior probability aggregation



Xiaojia Zhao, **DEEP NEURAL NETWORKS FOR COCHANNEL SPEAKER IDENTIFICATION, ICASSP2015.**

RBM vectors

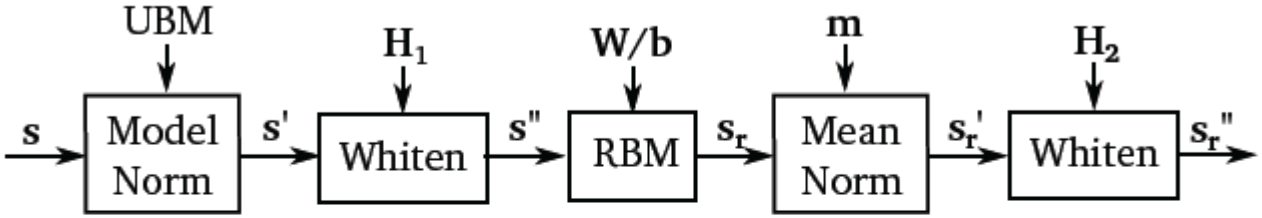
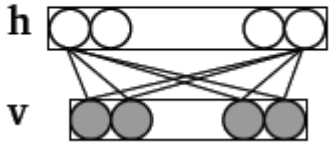


Fig. 1: Block-diagram of the process of transformation of raw GMM supervectors ($\mathbf{s}$) to the proposed RBM supervectors ($\mathbf{s}_r''$). $\mathbf{H}_1$ and $\mathbf{H}_2$ are whitening matrices, $\mathbf{W}$ and $\mathbf{b}$ are RBM parameters obtained on the development data



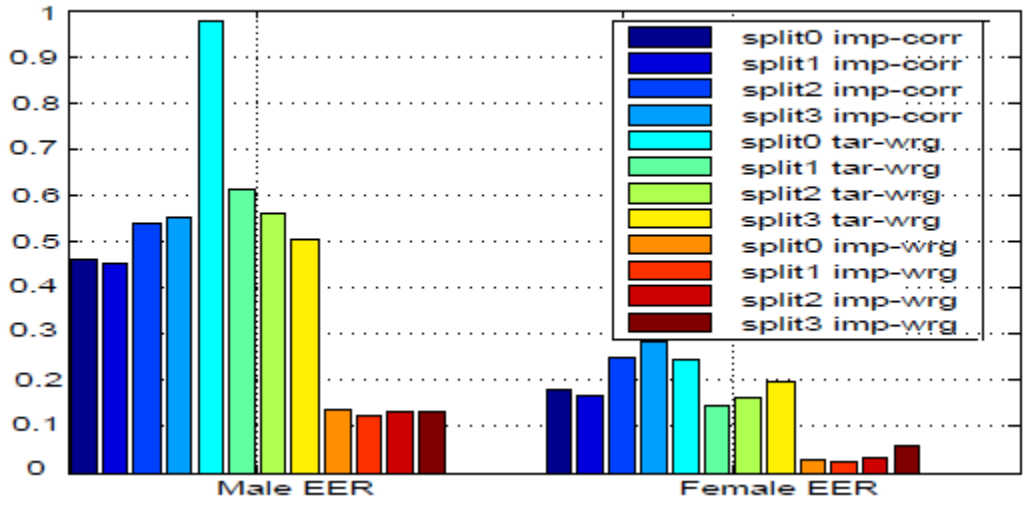
- Omid Ghahabi, **RESTRICTED BOLTZMANN MACHINE SUPERVECTORS FOR SPEAKER RECOGNITION, ICASSP 2015.**

Fast i-vector and Small Footprint i-vector

- O. Glembek, L. Burget, P. Matˇejka, M. Karafi´at, and P. Kenny, “Simplification and optimization of i–vector extraction,” in *Proceedings of ICASSP 2011*, pp.
- H. Aronowitz and O. Barkan, “Efficient approximated i–vector extraction,” in *Proceedings of ICASSP 2012*, pp. 4789–4792, 2012. 4516–4519, 2011.
- P. Kenny, “A small footprint i-vector extractor,” in *Proceedings of Odyssey 2012*, pp. 1–6, 2012.
- S. Cumani, P. Laface, and V. Vasilakakis, “Memory and computation effective approaches for i–vector extraction,” in *Proceedings of Odyssey 2012*, pp. 7–13, 2012.
- S. Cumani and P. Laface, “Memory and computation trade-offs for efficient i-vector extraction,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 934–944, 2013.
- S. Cumani and P. Laface, “Factorized sub-space estimation for fast and memory effective i-vector extraction,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 248–259, 2013.
- Sandro Cumani et al, **MEMORY-AWARE I-VECTOR EXTRACTION BY MEANS OF SUB-SPACE FACTORIZATION, ICASSP 2015.**

Resurging of text-dependent SID

- Work on RSR2015



- Hanwu SUN, Kong Aik LEE and Bin MA, A NEW STUDY OF GMM-SVM SYSTEM FOR TEXT-DEPENDENT SPEAKER RECOGNITION, ICASSP 2015.

JFA for Text-dependent Modeling

- Extend GMM to HMM, and use multiple hidden variables for the HMM states.

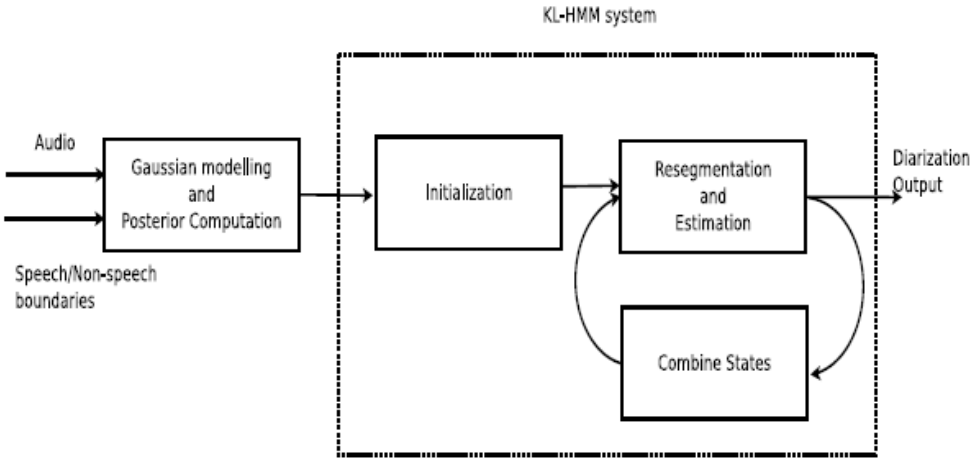
$$\boldsymbol{s}_r = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_r + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{D}\boldsymbol{z}$$

$$\frac{P_T(X_e, X_t)}{P_N(X_e, X_t)}$$

Patrick Kenny, **JFA MODELING WITH LEFT-TO-RIGHT STRUCTURE AND A NEW BACKEND FOR TEXT-DEPENDENT SPEAKER RECOGNITION, ICASSP 2015.**

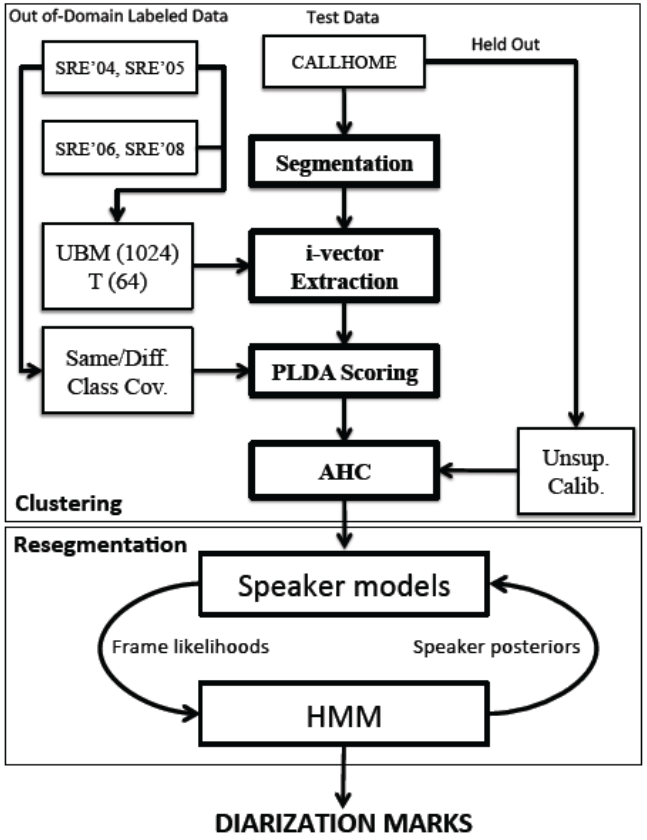
KL-HMM speaker diarization

- KL-HMM as a standalone method for diarization, like GMM-HMM
- An extension to IB



Srikanth Madikeri, Hervé Bourlard, **KL-HMM BASED SPEAKER DIARIZATION SYSTEM FOR MEETINGS, ICASSP 2015.**

Diarization with i-vectors



Gregory Sell, **DIARIZATION RESEGMENTATION IN THE FACTOR ANALYSIS SUBSPACE, ICASSP 2015.**

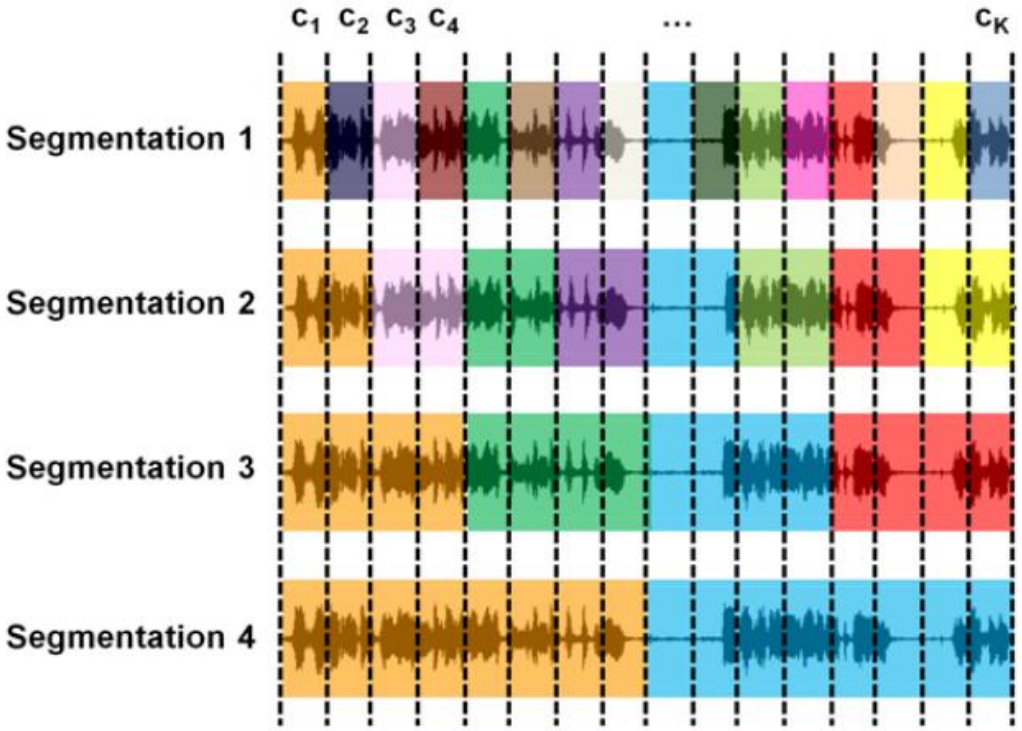
Diarization with PLDA

- PLDA is used to infer speaker vectors and speaker identity. VB is used for inference.
  - For each speaker s sample y_s, from $\mathcal{N}(y; \mu, \Lambda^{-1})$.
  - For each segment:
    - Sample i_m from the multinomial distribution $Mult(\Pi)$ where $\Pi = (\pi_1, ..., \pi_S)$. Let k be the index for which $i_{mk} = 1$, with all the other entries of i_m being 0.
    - Sample ϵ_m from $\mathcal{N}(\epsilon; \bar{0}, \mathcal{L}^{-1})$.
    - The observed segment i-vector is obtained as $\phi_m = y_k + \epsilon_m$.

Ahmet Emin Bulut, **PLDA-BASED DIARIZATION OF TELEPHONE CONVERSATIONS, ICASSP 2015.**

Non-resegmentation Diarization

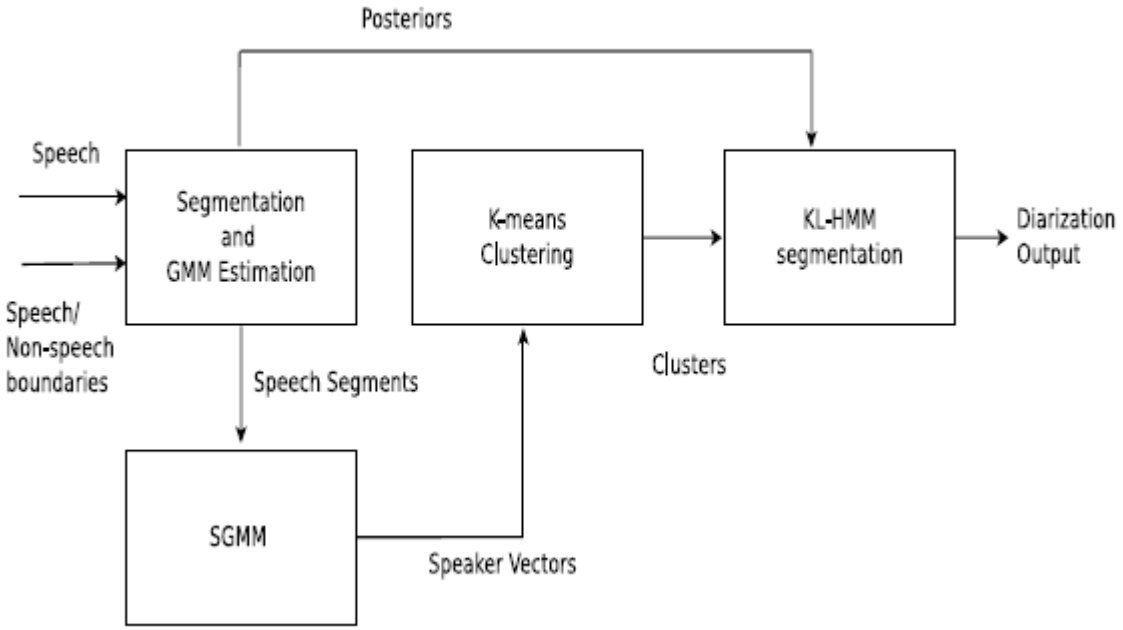
- Clustering on different segment length



Houman Ghaemmaghami, **A CLUSTER-VOTING APPROACH FOR SPEAKER DIARIZATION AND LINKING OF AUSTRALIAN BROADCAST NEWS RECORDINGS, ICASSP 2015.**

SGMM + KL-HMM Diarization

- SGMM for i-vector extraction, KL-HMM used for resegmentation



Srikanth Madikeri, **COMBINING SGMM SPEAKER VECTORS AND KL-HMM APPROACH FOR SPEAKER DIARIZATION, ICASSP 2015.**

Database and tools

- Spoofing
- https://wiki.inf.ed.ac.uk/CSTR/SASCorpus
- NAM TIMIT:
- http://homepages.inf.ed.ac.uk/jyamagis/page3/page57/page57.html
- VB diarization
- http://speech.fit.vutbr.cz/software/vb-diarization-eigenvoice-and-hmm-priors