Automatic Speaker Recognition

于嘉威

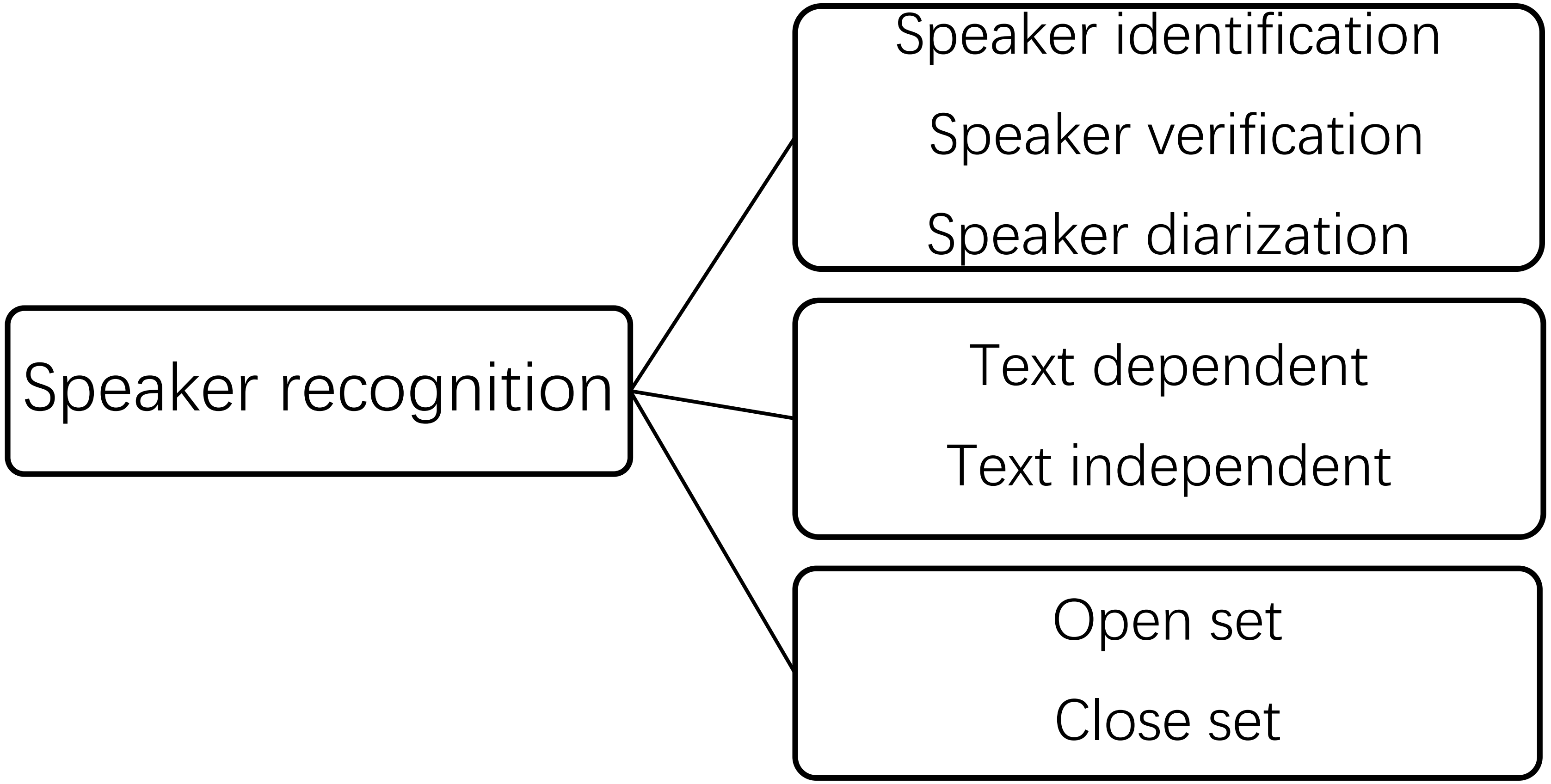
2018/8/13

Outline

- Introduction
- The i-vector methodology of speaker recognition
- The d-vector methodology of speaker recognition
- The end-to-end methodology of speaker recognition
- Inter-speaker variability in speaker recognition
- Variations in speaker recognition

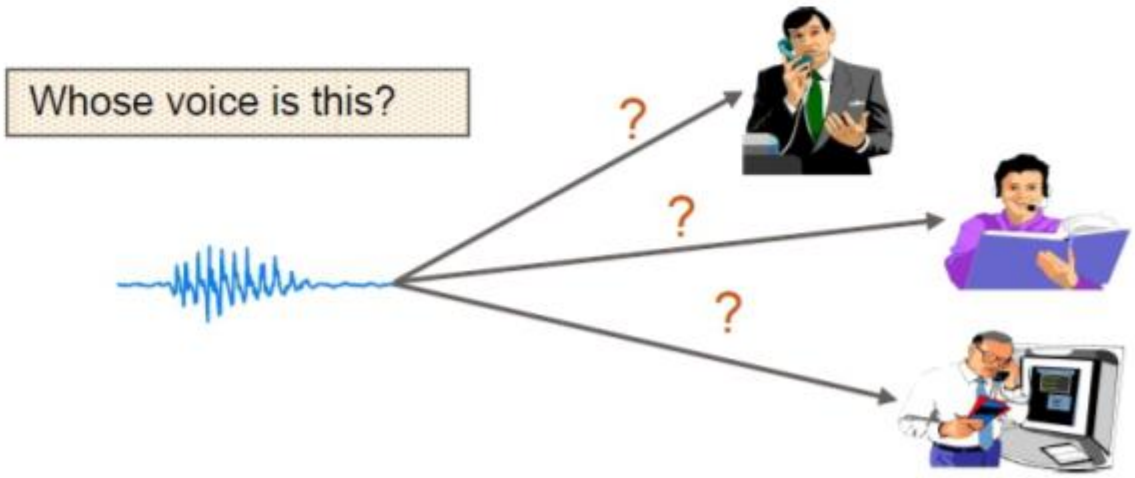
Introduction

- Definition: It is the method of recognizing a person based on his voice



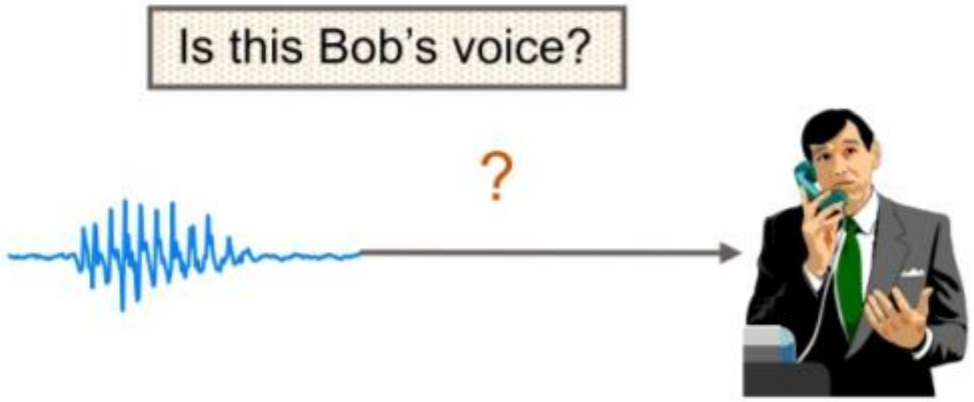
Speaker Identification

- Definition: Determine whether unknown speaker matches one of a set known speakers
- One-to-many mapping
- Often assumed that unknown voice must come from a set of known speakers – referred to as close-set identification
- Adding “none of the above” option to closed-set identification gives open-set identification



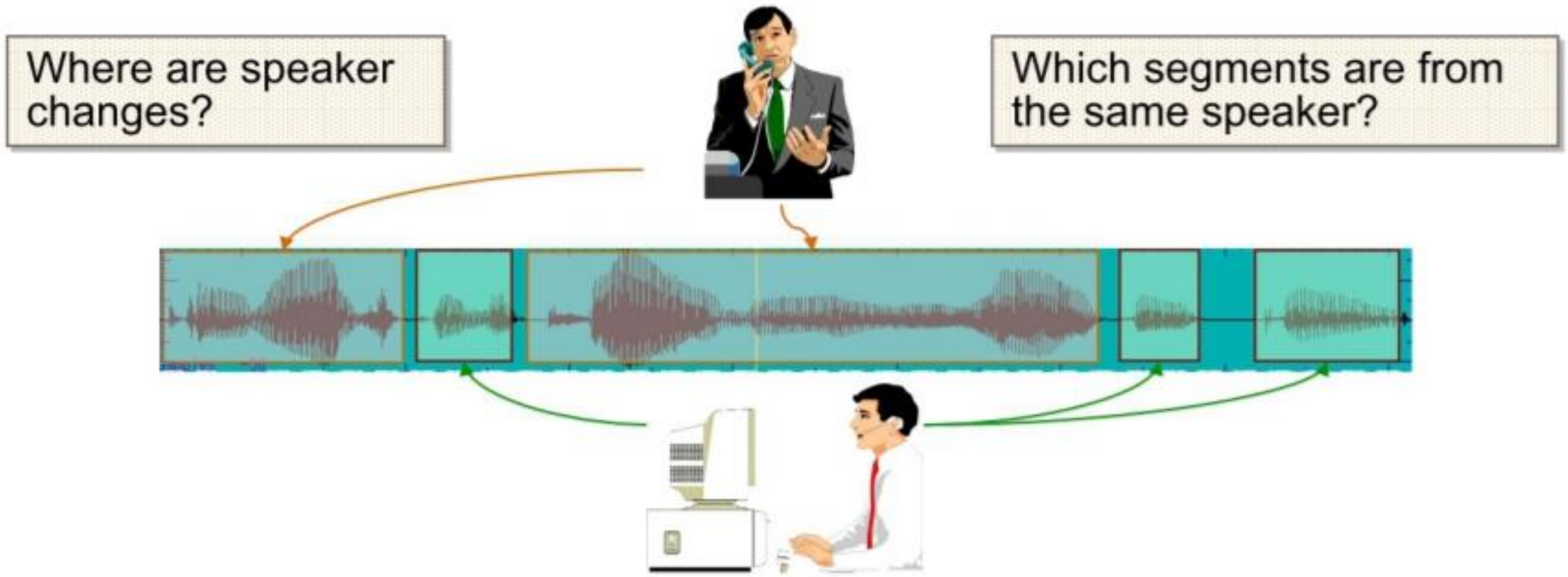
Speaker Verification

- Determine whether unknown speaker matches a specific speaker
- One-to-one mapping
- Close-set verification: The population of clients is fixed
- Open-set verification: New clients can be added without having to redesign the system.



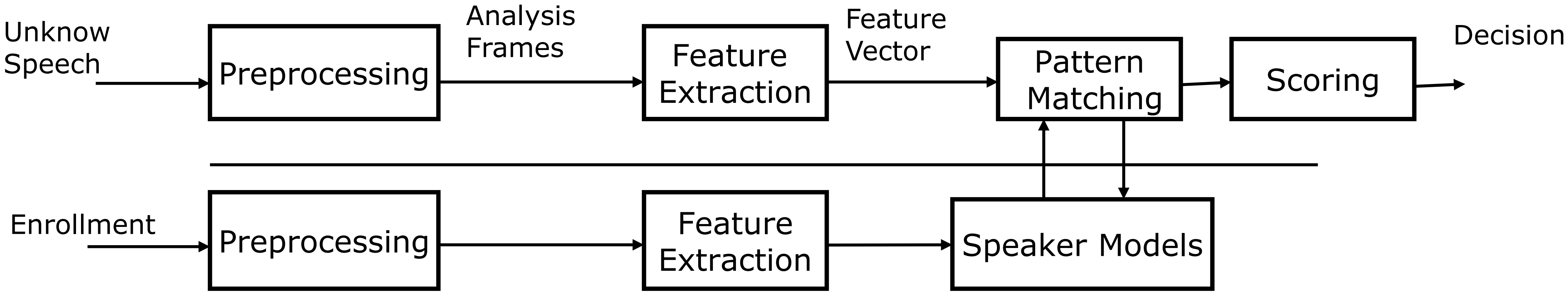
Speaker diarization

- Determine when a speaker change has occurred in speech signal (segmentation)
- Group together speech segments corresponding to the same speaker (clustering)
- Prior speaker information may or may not be available



Introduction: Generic Speaker Recognition System

- Basic structure of a speaker recognition system

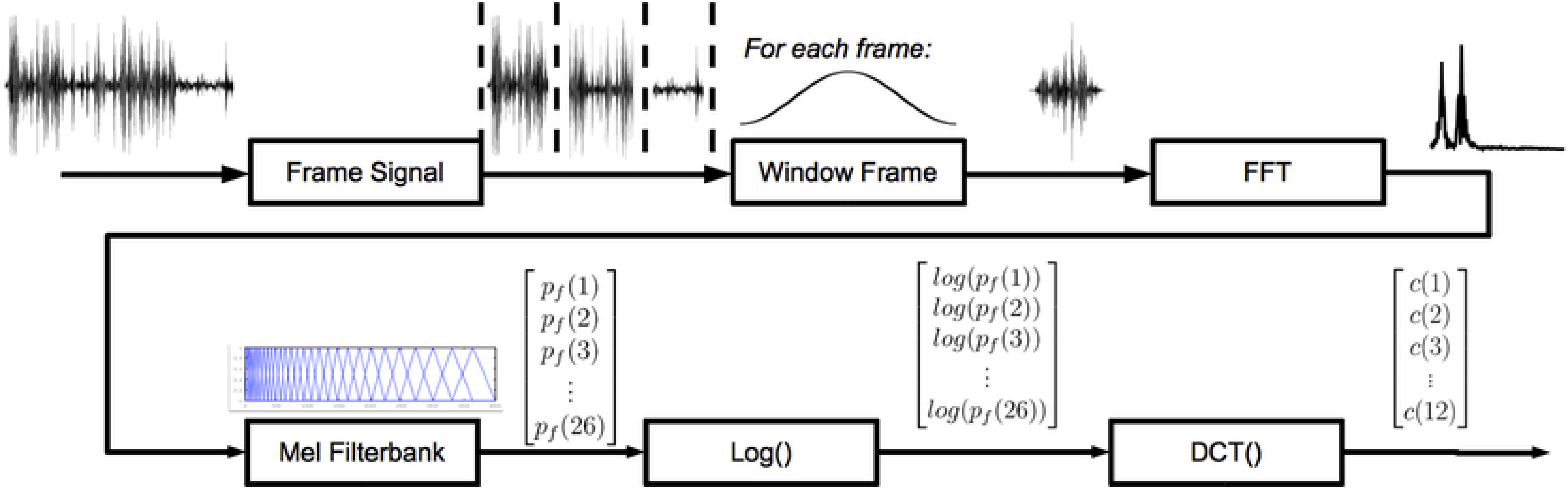


Introduction: Main Research Fields on SRE

- Feature Extraction
- Pattern matching
- Scoring method

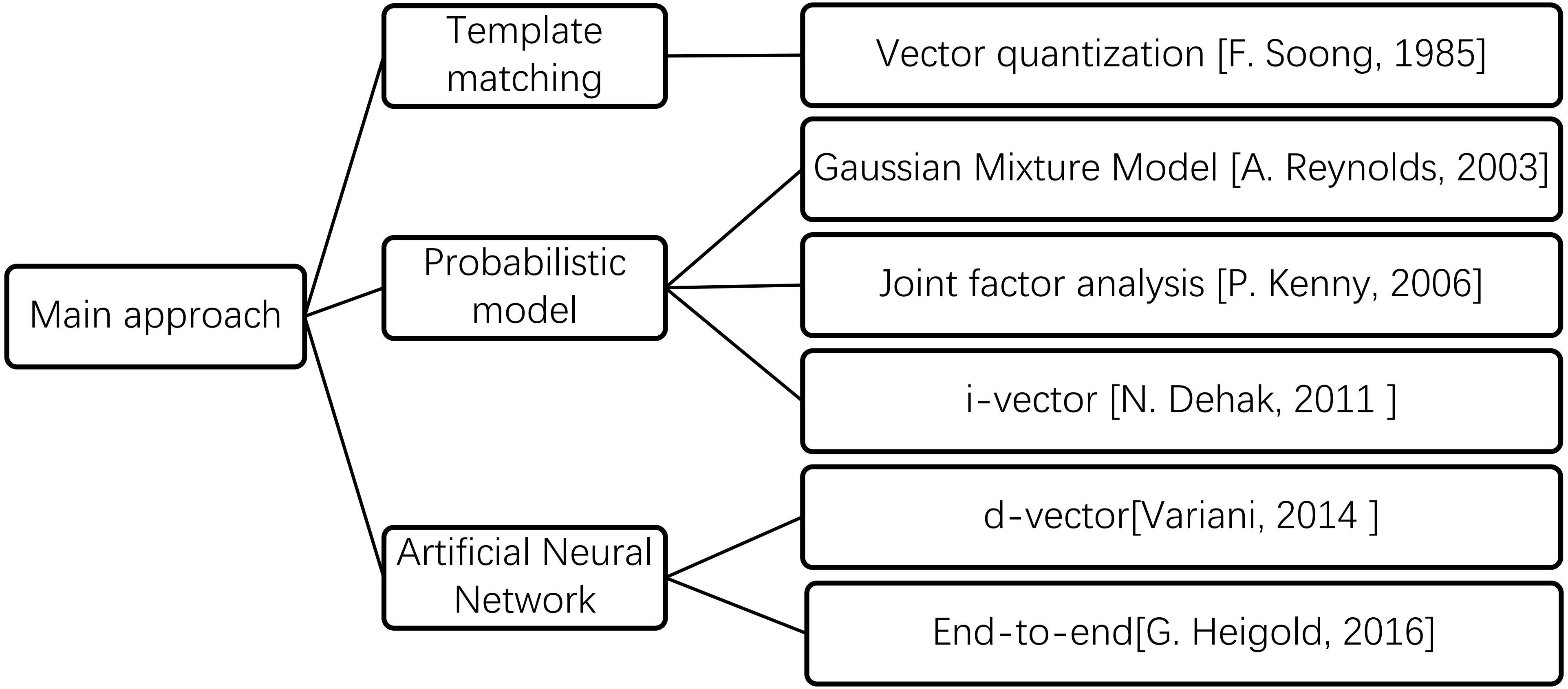
Introduction: Feature Extraction

- Converting the raw speech signal into a sequence of acoustic feature vectors carrying characteristic information about the signal



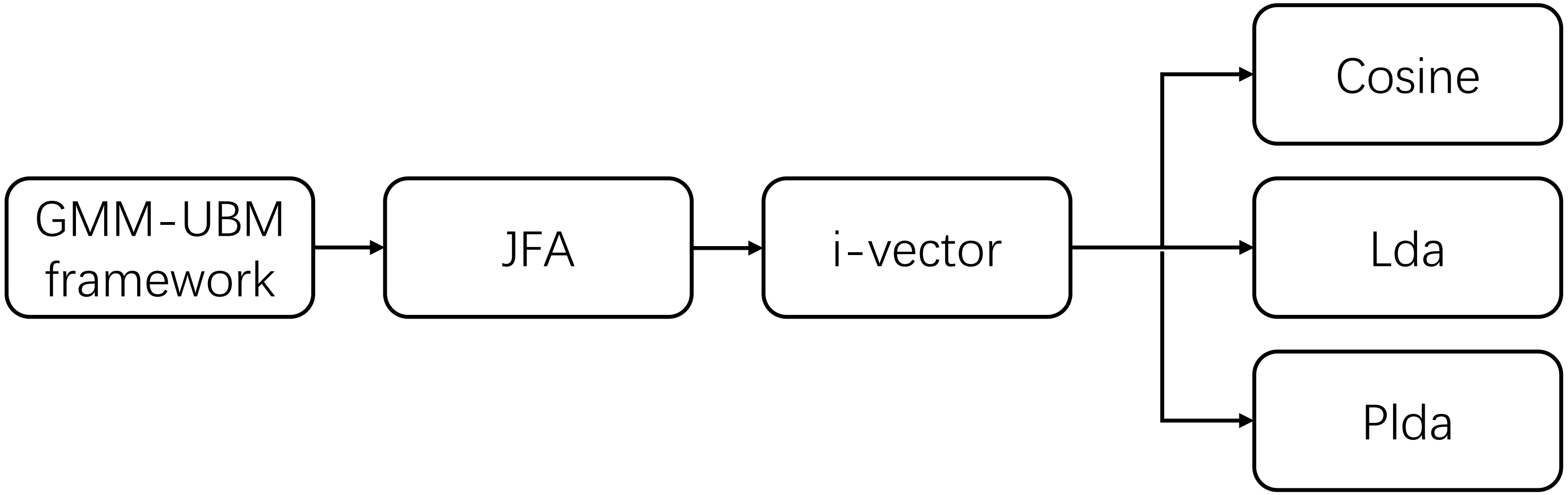
Introduction: Pattern matching

- Main approaches in pattern matching for speaker recognition main



The i-vector methodology of speaker recognition

- Over recent years, ivector has demonstrated state-of-the-art performance for speaker recognition.



Joint factor analysis

- Factor analysis is a statistical method which is used to describe the variability among the observed variables in terms of potentially lower number of unobserved variables called factors
- Factor analysis is a latent variable model for feature extraction
- Joint factor analysis (JFA) was the initial paradigm for speaker recognition

$$\underbrace{\mathbf{u}}_{\text{Speaker Supervector}} = \underbrace{\mathbf{m}}_{\text{UBM}} + \mathbf{V}\overbrace{\mathbf{y}}^{\text{Speaker factors}} + \mathbf{U}\underbrace{\mathbf{x}}_{\text{Channel factors}} + \mathbf{D}\overbrace{\mathbf{z}}^{\text{Residual factors}}$$

Joint factor analysis

- A supervector for a speaker should be decomposable into speaker independent, speaker dependent, channel dependent, and residual components
- Each component is represented by low-dimensional factors, which operate along the principal dimensions of the corresponding component
- Speaker dependent component, known as the eigenvoice, and the corresponding factors

Eigenvoice matrix

$$\mathbf{V} \cdot \mathbf{y} = \begin{bmatrix} | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \ldots & \mathbf{v}_N \\ | & | & & | \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Each speaker factor controls an eigendimension of the eigenvoice matrix

Low dimensional eigenvoice factors

- GMM supervector u for a speaker can be decomposed as

Speaker-dependent component

Speaker-dependent resuidual component

$$\mathbf{u} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} + \mathbf{D}\mathbf{z}$$

Speaker supervector

Speaker-independent component

Channel-dependent component

Where:

m is a speaker-independent supervector from UBM

V is the eigenvoice matrix

y ~ N(0, I) is the speaker factor vector

U is the eigenchannel matrix

x ~ N(0, I) is the channel factor vector

D is the residual matrix, and is diagonal

z ~ N(0, I) is the speaker-specific residual factor vector

Training procedure

- We train the JFA matricies in the following order [Kenny et al., 2007a]
  - 1. Train the eigenvoice matrix V, assuming that U and D are zero
  - 2. Train the eigenchannel matrix U given the estimate of V, assuming that D is zero
  - 3. Train the residual matrix D given the estimates of V and U
- Using these matrices, we compute y for speaker, x for channel, and z for residual factors
- We compute the final score by using these matrices and factors

Total variability

- Subspaces U and V are not completely independent
- A combined total variability space was used [Dehak et al., 2011]

Speaker factors

Residual factors

$$\mathbf{u} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} + \mathbf{D}\mathbf{z}$$

Speaker Supervector

UBM

Channel factors

$$\Downarrow$$

$$\mathbf{u} = \mathbf{m} + \mathbf{T}\mathbf{w}$$

Speaker supervector

UBM

Intermediate/identity vector (i-vector)

Total variability matrix

i-vector

- An i-vector system uses a set of low-dimensional total variability factors (w) to represent each conversation side. Each factor controls an eigen-dimension of the total variability matrix (T), and are known as the i-vectors.
- Unlike JFA or other FA methods, the i-vector approach does not make a distinction between speaker and channel
- define a total variability space, contains speaker and channel variabilities simultaneously

Training total variability space

- Rank of T is set prior to training
- T and w are latent variables
- EM algorithm is used
- Training total variability matrix T is similar to training V except that training T is performed by using all utterances from a given speaker but as produced by different speakers
- Random initialization for T
- Each ot has dimension D. Number of Gaussian components is M. Dimension of supervector is $M \cdot D$
- UBM diagonal covariance matrix Σ ($MD \times MD$) is introduced to model the residual variability not captured by T

i-vector extraction

0^{th} order statistics $N_c(u) = \sum_t \gamma_c(\mathbf{o}_t)$ of an utterance u

1^{th} order statistics $F_c(u) = \sum_t \gamma_c(\mathbf{o}_t)\mathbf{o}_t$

2^{nd} order statistics $S_c(u) = \mathrm{diag}\left(\sum_t \gamma_c(\mathbf{o}_t)\mathbf{o}_t\mathbf{o}_t^\top\right)$ where

$$\gamma_c(\mathbf{o}_t) = p(c|\mathbf{o}_t, \boldsymbol{\theta}_{\mathrm{ubm}}) = \frac{\pi_c p(\mathbf{o}_t|\mathbf{m}_c, \boldsymbol{\Sigma}_c)}{\sum_{j=1}^{M} \pi_i p(\mathbf{o}_t|\mathbf{m}_j, \boldsymbol{\Sigma}_j)}$$

Centralized 1^{th} and 2^{nd} order statistics

$$\tilde{F}_c(u) = \sum_{t=1}^{T} \gamma_c(\mathbf{o}_t)(\mathbf{o}_t - \mathbf{m}_c)$$

$$\tilde{S}_c(u) = \mathrm{diag}\left(\sum_{t=1}^{T} \gamma_c(\mathbf{o}_t)(\mathbf{o}_t - \mathbf{m}_c)(\mathbf{o}_t - \mathbf{m}_c)^\top\right)$$

where $\mathbf{m}_c$ is the subvector corresponding to mixture component c

i-vector extraction

Sufficient statistics

$$N(u) = \begin{bmatrix} N_1(u)\cdot\mathbf{I}_{D\times D} & 0 & \cdots & 0 \\ 0 & N_2(u)\cdot\mathbf{I}_{D\times D} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & N_M(u)\cdot\mathbf{I}_{D\times D} \end{bmatrix} \quad \tilde{F}(u) = \begin{bmatrix} \tilde{F}_1(u) \\ \tilde{F}_2(u) \\ \vdots \\ \tilde{F}_M(u) \end{bmatrix}$$

EM algorithm [Kenny et al., 2005]

- Initialize $\mathbf{m}$, $\mathbf{\Sigma}$ and $\mathbf{T}$
- E-step: for each utterance u, calculate the parameters of the posterior distribution of $\mathbf{w}(u)$ using the current estimates of $\mathbf{m}, \mathbf{\Sigma}, \mathbf{T}$
- M-step: update $\mathbf{T}$ and $\mathbf{\Sigma}$ by solving a set of linear equations in which $\mathbf{w}(u)$'s play the role of explanatory variables
- Iterate until data likelihood given the estimated parameters converges

E-step: posterior distribution of w(u)

- For each utterance u, we calculate the matrix L(u) T and w are latent variables

$$\mathbf{L}(u) = \mathbf{I} + \mathbf{T}^\top \mathbf{\Sigma}^{-1} N(u) \mathbf{T}$$

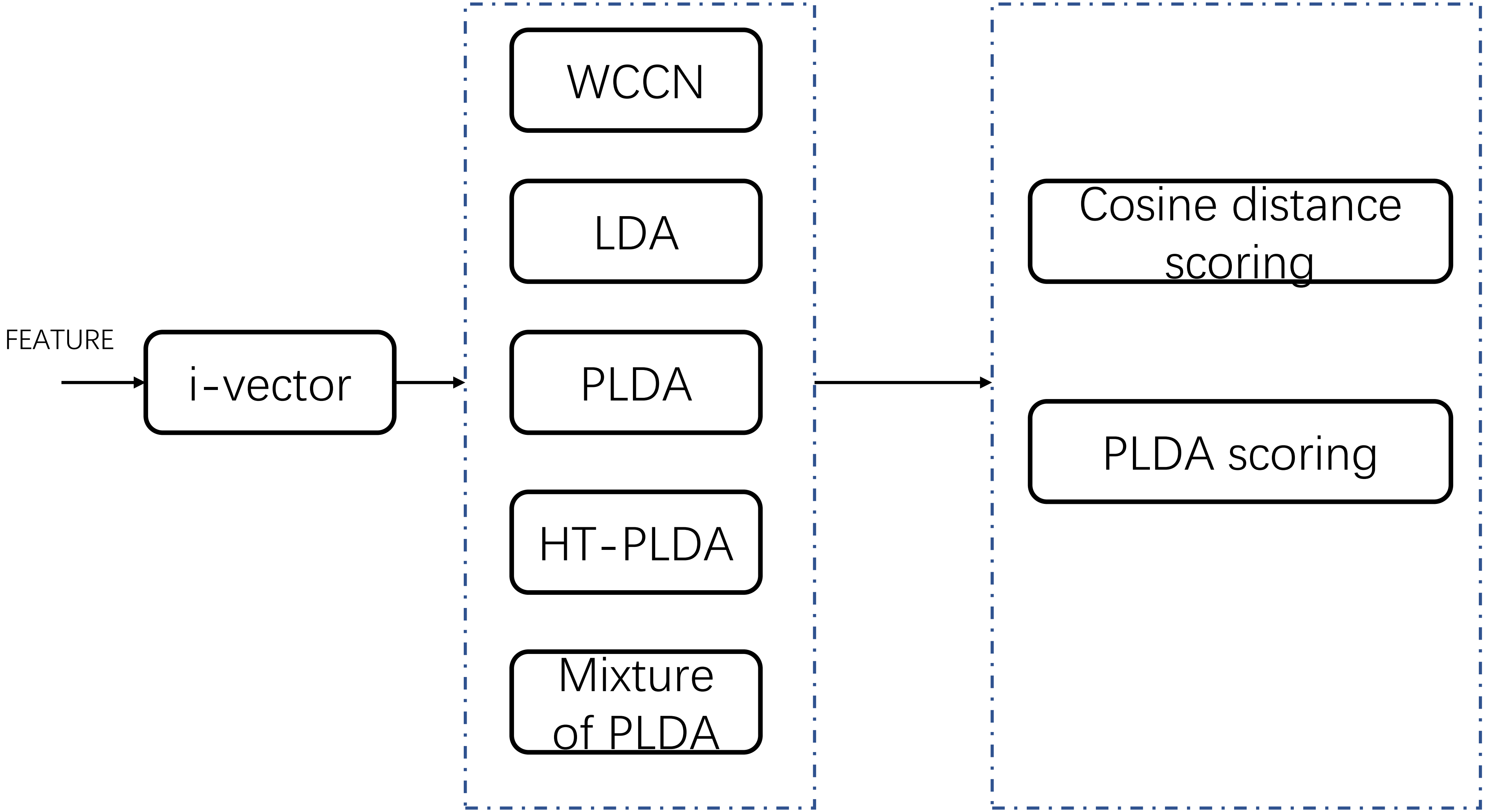
- Posterior distribution of w(u) conditioned on the acoustic observations of an utterance u is Gaussian with mean

$$\mathbb{E}[\mathbf{w}(u)] = \mathbf{L}^{-1}(u) \mathbf{T}^\top \mathbf{\Sigma}^{-1} \tilde{F}(u)$$

- and covariance matrix

$$\mathrm{Cov}(\mathbf{w}(u), \mathbf{w}(u)) = \mathbf{L}^{-1}(u)$$

Intersession compensation and scoring method for ivector



Linear discriminant analysis

- I-vectors from JFA model are used in linear discriminant analysis (LDA)

$$\mathbf{u} = \mathbf{m} + \mathbf{T}\mathbf{w} \longleftarrow \text{Factor analysis}$$

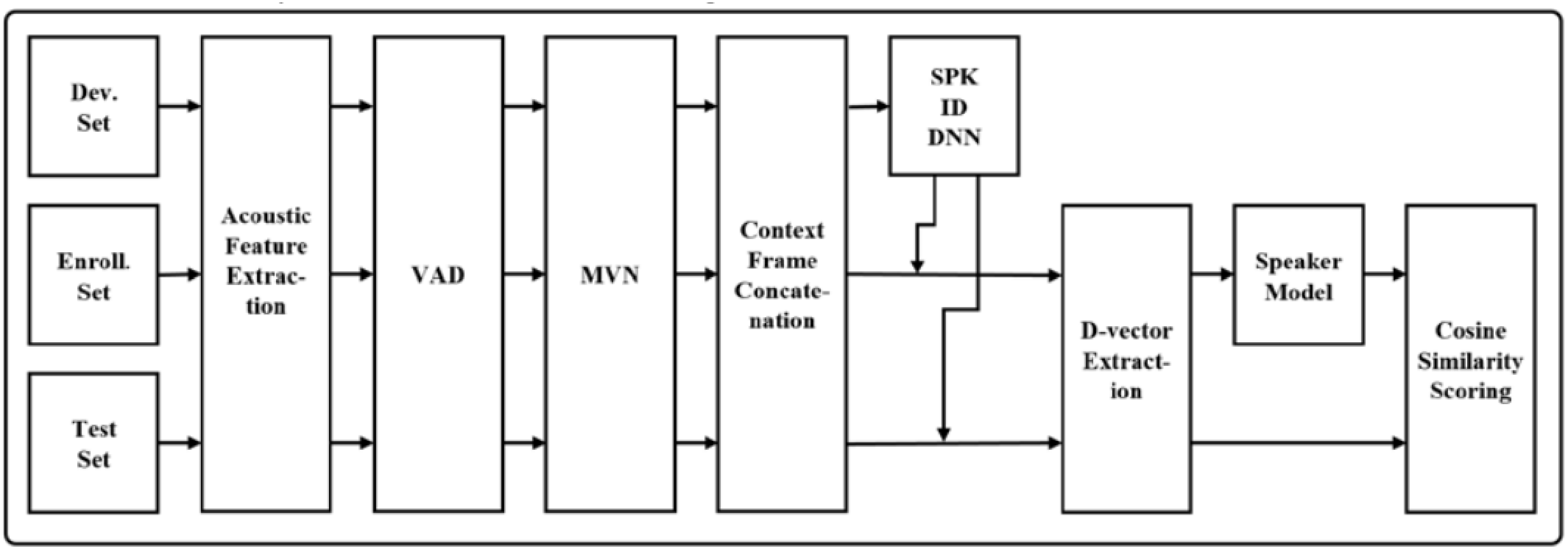
$$\mathcal{W} = \mathbf{A}\mathbf{w} \longleftarrow \text{Linear discriminant analysis}$$

- Both methods used to reduce the dimensionality of speaker model
- A is chosen such that within-speaker variability Sw is minimized and between-speaker variability Sb is maximized within the space
- A is found by eigenvalue method via maximizing

$$\mathcal{J}(\mathbf{A}) = \mathrm{Tr}\{S_w^{-1} S_b\}$$

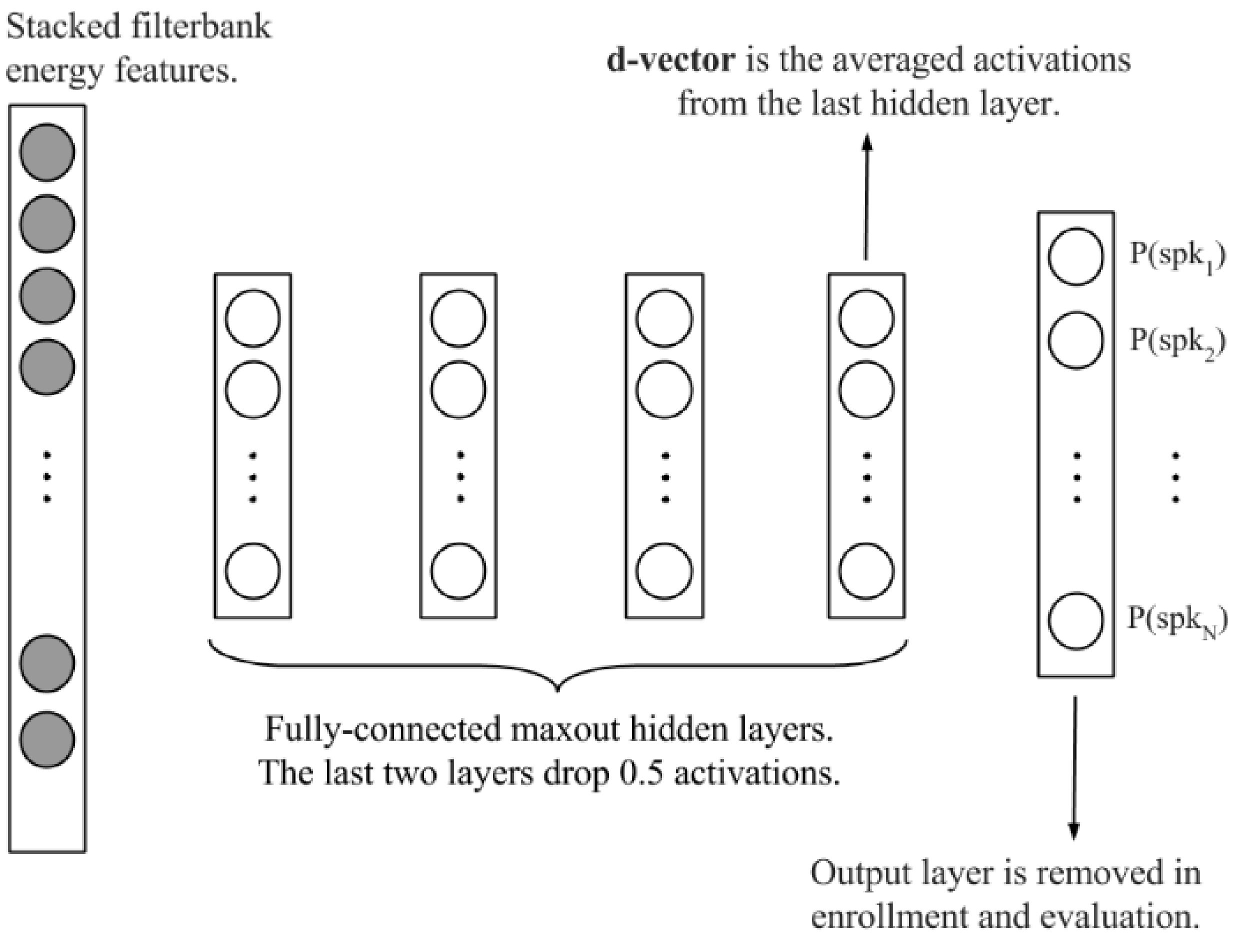
The d-vector methodology of speaker recognition

- Pipeline process employed in conventional d-vector based speaker verification system.



The d-vector methodology of speaker recognition

- What is d-vector?



What is d-vector?

- Using a DNN architecture as a speaker feature extractor

- For every frame of a given utterance belonging to a new speaker, we compute the output activations of the last hidden layer using standard feedforward propagation in the trained DNN
- Then accumulate those activations to form a new compact representation of that speaker, the d-vector.

The d-vector methodology of speaker recognition

- The reason of use the output from the last hidden layer instead of the softmax output layer:
    - First, we can reduce the DNN model size for runtime by pruning away the output layer, and this also enables us to use a large number of development speakers without increasing DNN size at runtime.
    - Second, we have observed better generalization to unseen speakers from the last hidden layer output.

Enrollment and evaluation

- Given a set of utterances X_s from a speaker s,

$$X_s = \{O_{s_1}, O_{s_2}, \ldots, O_{s_n}\}$$

- With observations O_{s_i}

$$O_{s_i} = \{o_1, o_2, \ldots, o_m\}$$

- the process of enrollment can be described as follows:
  - First, we use every observation o_j in utterance O_{s_i}, together with its context, to feed the supervised trained DNN. The output of the last hidden layer is then obtained, L2 normalized, and accumulated for all the observations o_j in O_{s_i}.
  - Then we refer to the resulting accumulated vector as the d-vector associated with the utterance O_{s_i}.
  - The final representation of the speaker s is derived by averaging all d-vectors corresponding for utterances in X_s.

Evaluation

- During the evaluation phase, we first extract the normalized d-vector from the test utterance.
- Then we compute the cosine distance between the test d-vector and the claimed speaker's d-vector.
- Finally a verification decision is made by comparing the distance to a threshold.

Kaldi d-vector Baseline System： Toolkits and database

- Kaldi toolkits [D. Povey, 2011]
- Database： THCHS-30

Table 2: Statistics of THCHS-30 database

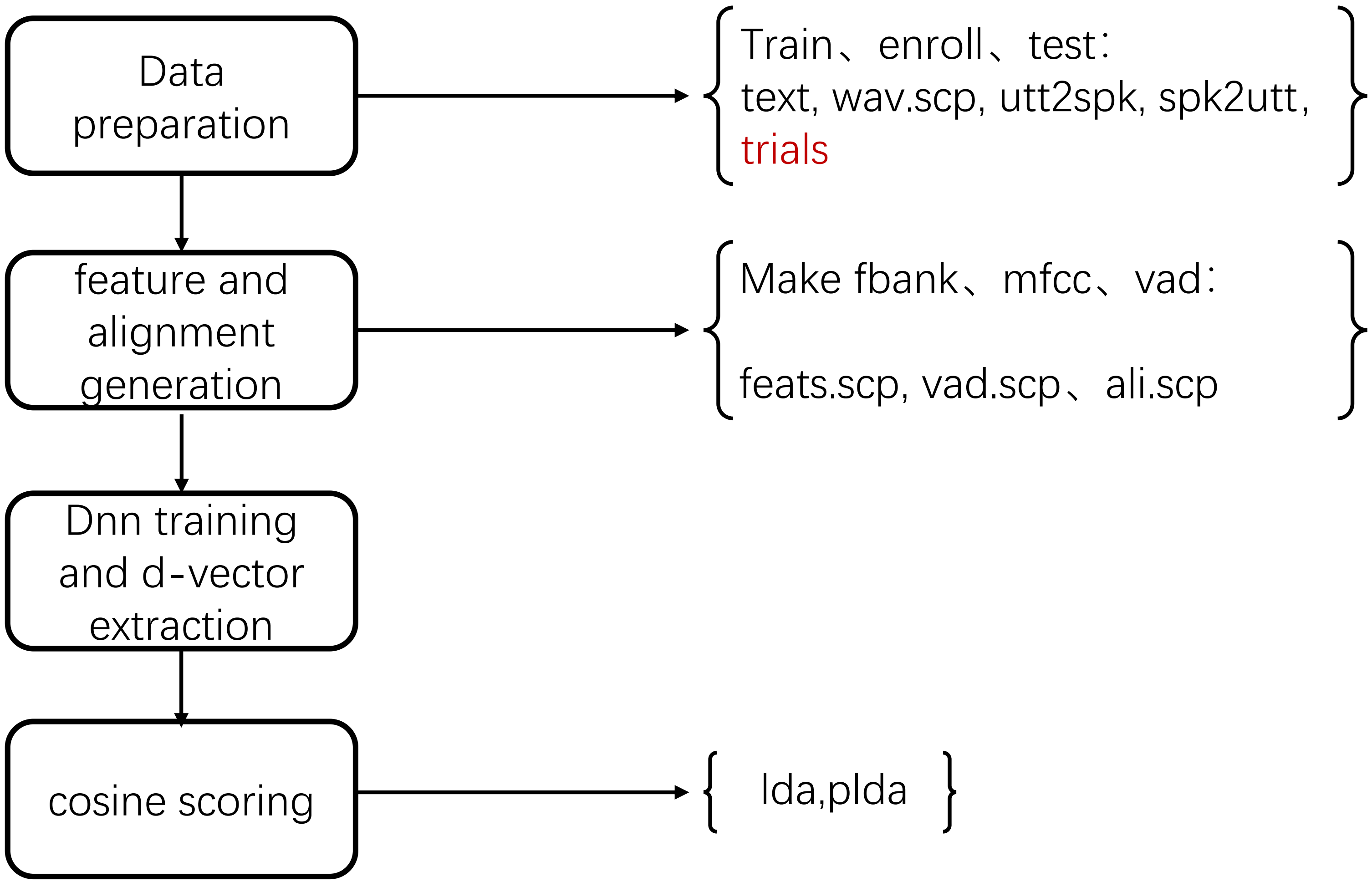
Data Set	Speaker	Male	Female	Age	Utterance	Duration (hour)
Training	30	8	22	20-55	10893	27.23h
Test	10	1	9	19-50	2496	6.24h

Data Set	Speaker	Utterance
Training	50	10000
Enroll	10	1000
Test	10	1495

setup

- i-vector
  - 2048 Gaussian Mixtures
  - 400-dimensional ivector.
  - 150-dimensional lda/plda.
- d-vector
  - 400-dimensional d-vector.
  - 150-dimensional lda/plda.

Kaldi d-vector Recipe



results

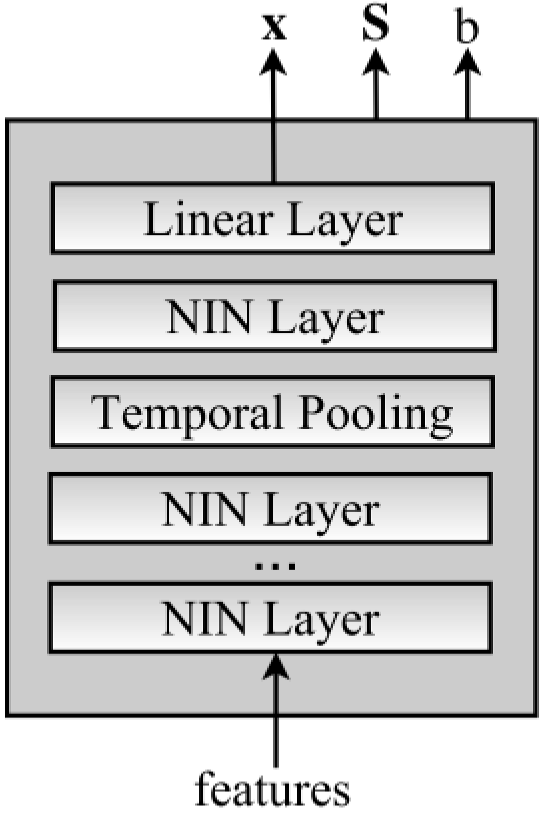
%EER	cosine	lda	plda
i-vector	0.64	0.07	0.07
d-vector	3.08	1.07	2.21

The end-to-end methodology of speaker recognition

- Overview:
  - The architecture is a feed-forward DNN that extracts statistics over a sequence of stacked MFCCs and maps it to a speaker embedding.
  - The objective function operates on pairs of embeddings, and maximizes a same-speaker probability for embeddings from the same speaker
  - minimizes the same probability for pairs of embeddings from different speakers.

Neural Network Architecture

- **x**:speaker embedding.
- The symmetric matrix **S** and offset **b** are constant outputs (independent of the input)
- The network activations are a type of network-in-network (NIN) nonlinearity



(a) DNN Architecture

Training

- We model the probability of embeddings x and y belonging to the same speaker by the logistic function in Equation 1.

$$Pr(\mathbf{x},\mathbf{y}) = \frac{1}{1+e^{-L(\mathbf{x},\mathbf{y})}} \qquad (1)$$

- Equation 2 is a PLDA-like quantity defines the distance between two embeddings.

$$L(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{y} - \mathbf{x}^T\mathbf{S}\mathbf{x} - \mathbf{y}^T\mathbf{S}\mathbf{y} + b \qquad (2)$$

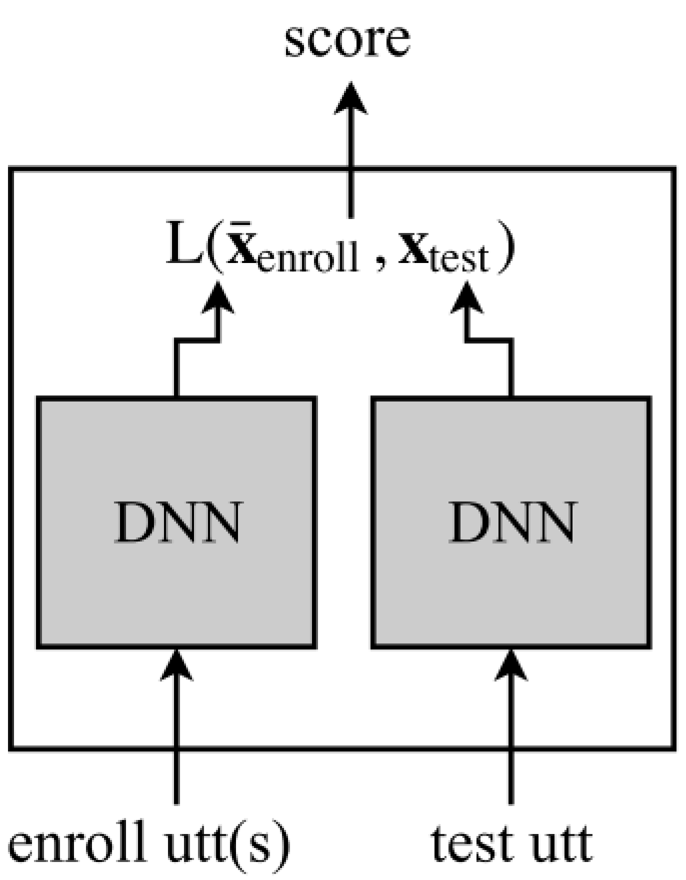
- P_{diff} and $Psame$ be the set of different-speaker and same-speaker pairs, respectively.
- The objective function (Equation 3) is the log probability of the correct choice for each pair.

$$E = -\sum_{\mathbf{x},\mathbf{y}\in P_{\text{same}}} ln\left(Pr(\mathbf{x},\mathbf{y})\right) - K\sum_{\mathbf{x},\mathbf{y}\in P_{\text{diff}}} ln\left(1-Pr(\mathbf{x},\mathbf{y})\right) \quad (3)$$

Scoring

- enroll and test utterances are scored by the distance metric used in the objective function (Equation 2)

$$L(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y} - \mathbf{x}^T\mathbf{S}\mathbf{x} - \mathbf{y}^T\mathbf{S}\mathbf{y} + b$$

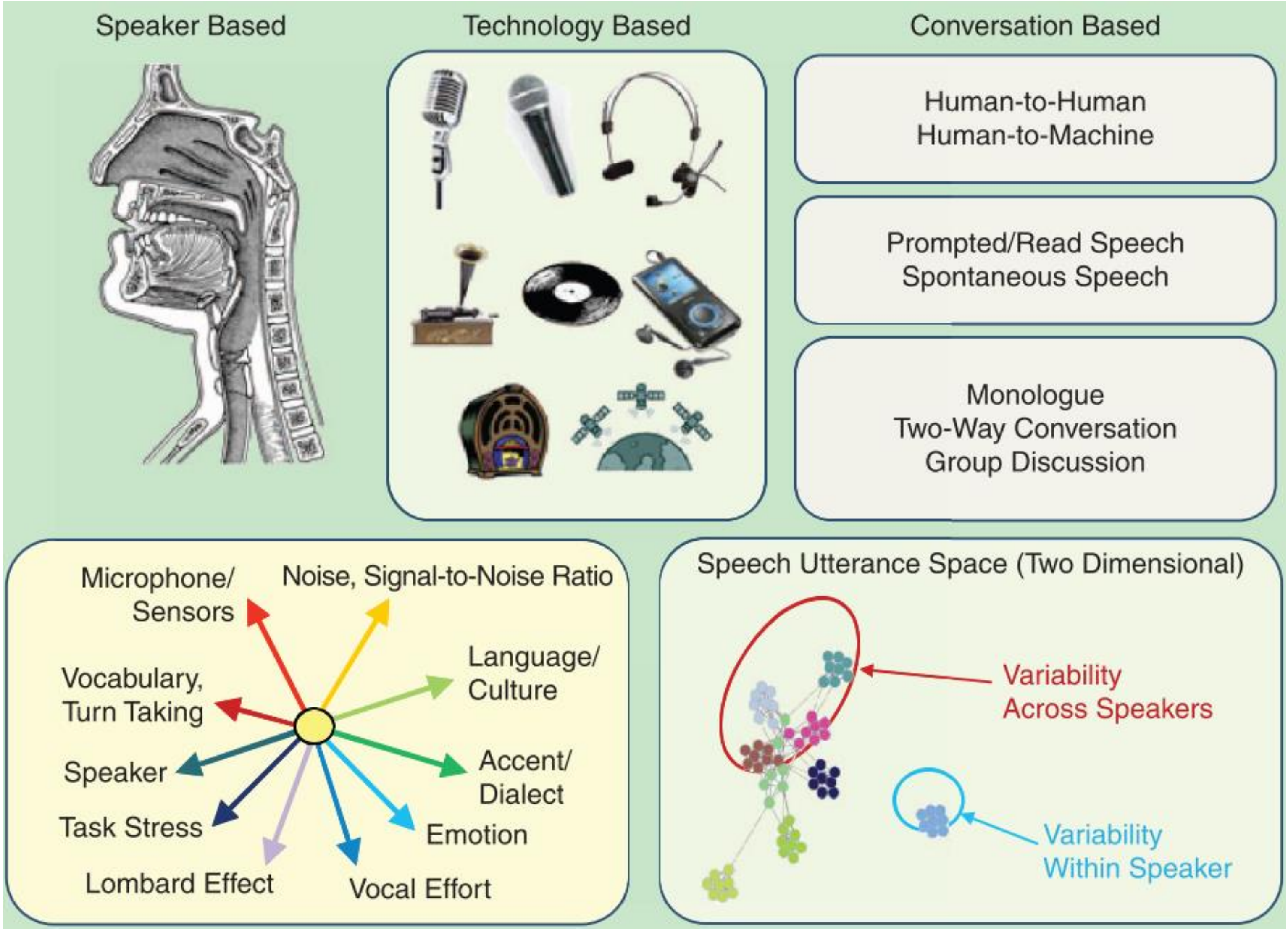


(b) Scoring Schema

Variations in speaker recognition

- **speaker** based
- **conversation** based
- **technology** based

Variations in speaker recognition



Speaker-based variability sources

- these reflect a range of changes in how a speaker produces speech and will affect system performance for speaker recognition.
- These can be thought of as intrinsic or within-speaker variability and include the following factors:
  - Situational task stress
  - Vocal effort/style
  - Emotion
  - Physiological
  - Disguise

Conversation-based variability sources

- these reflect different scenarios with respect to the voice interaction with either another person or technology system, or differences with respect to the specific language or dialect spoken, and can include :
  - human-to-human
    - language or dialect spoken
    - if speech is read/prompted (through visual display or through headphones), spontaneous, conversational, or disguised speech
    - monologue, two-way conversation, public speech in front of an audience or for TV or radio, group discussion
  - human-to-machine
    - prompted speech: voice input to a computer
    - voice input for telephone/dialog system/computer
    - input: interacting with a voice-based system

Technology-based variability sources

- these include how and where the audio is captured and the following issues:
  - electromechanical—transmission channel, handset (cell, cordless, and landline) microphone
  - environmental—background noise (stationary, impulsive, time-varying, etc.), room acoustics , reverberation , and distant microphone
  - data quality—duration, sampling rate, recording quality, and audio codec/compression.

Variations in speaker recognition

- These multifaceted sources of variation pose the greatest challenge in accurately modeling and recognizing a speaker
- Additive noise and transmission channel variability have received much attention recently.
- Higher-level knowledge may become important in these cases.
  - eg： a person’s voice (spectral characteristics) may change due to his or her current health (e.g., a cold) or aging, the person’s accent or style of speech remains generally the same

age

- Eigenageing Compensation approach[Finnian Kelly,2013]
- Analogous to eigenchannel compensation, the proposed eigenageing compensation method operates by adapting a speaker model to a test sample based on a predetermined ageing subspace.
- The aim of eigenageing compensation is to model the ageing change in speakers, and then use this to adapt a speaker model at verification time to a sample of unknown age

language

- training speaker models using both enrollment and test languages
- incorporated Language Identification (LID) as a first layer in speaker verification to detect the test language and then use an appropriate model trained on that language for scoring.
    - Both of these studies, however, require a Gaussian Mixture Model-Universal background Model (GMM-UBM) speaker verification system, that is no longer state-of-the-art in speaker verification
- A language dependent subspace is estimated using a Joint Factor Analysis (JFA) framework and then suppressed as a nuisance attribute
    - This approach is close to the state-of-the art system, but requires significant multi- lingual seed data to train the system
- adding small amounts of multi-lingual data to a Probabilistic Linear Discriminant Analysis (PLDA) development set and achieve a significant improvement.

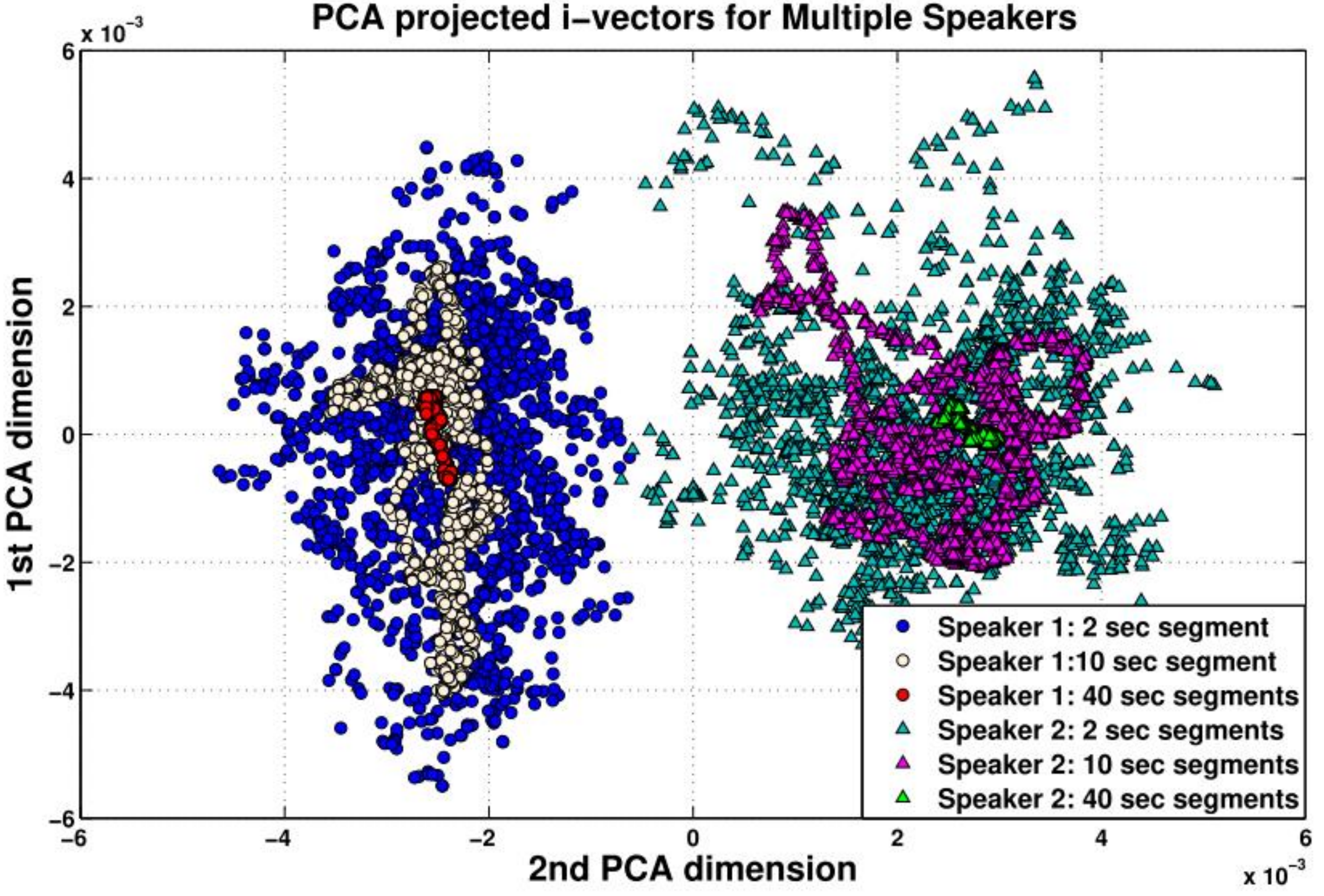
noise

- The effect of environmental noise on the recording is at least twofold:
  - the noise is added to the speech signal at the transducer, leading to a lower SNR at the receiver's end.
  - the Lombard reflex in human speakers will cause the speaker to change the vocal effort and simultaneously changing their voice's spectral characteristics.
- priori knowledge method：
  - filtering techniques：spectral subtraction、Kalman filtering……
  - noise compensation :PMC、Jacobian environmental adaptation
- missing-feature approaches
  - base the recognition only on the data with little or no contamination

  ……

duration

- The main challenge in achieving high performance with short duration is the increase in intra-speaker variability of estimated parameters.



duration

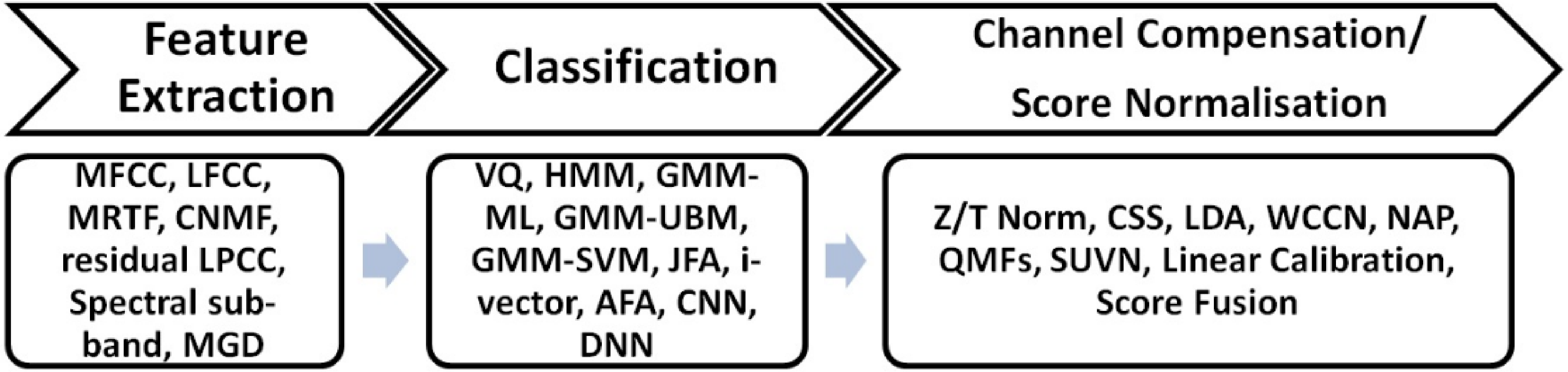


Fig. 6: Diagrammatic representation of different methods used in three sub-system levels of ASV to mitigate the problem of short utterance.

References

[1] J. P. Campbell, “Speaker recognition: A tutorial,” Proceedings of the IEEE, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] D. A. Reynolds, “An overview of automatic speaker recognition technology,” in Acoustics, speech, and signal processing (ICASSP), 2002 IEEE international conference on, vol. 4. IEEE, 2002, pp. IV–4072.

[3] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” Speech communication, vol. 52, no. 1, pp. 12–40, 2010.

[4] J. H. Hansen and T. Hasan, “Speaker recognition by machines and humans: A tutorial review,” IEEE Signal processing magazine, vol. 32, no. 6, pp. 74–99, 2015.

[5]P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, pp. 1435–1447, 2007.

[6] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front- end factor analysis for speaker verification,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, 2011.

[7]S. Ioffe, “Probabilistic linear discriminant analysis,” Computer Vision ECCV 2006, Springer Berlin Heidelberg, pp. 531–542, 2006.

[8] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, “Deep neural networks for extracting baum-welch statistics for speaker recognition,” Odyssey, 2014.

References

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

[10] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, “End-to-end text- dependent speaker verification,” in Acoustics, Speech and Signal Pro- cessing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016, pp. 5115–5119.

[11]D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, “Deep neural network-based speaker embeddings for end-to-end speaker verification,” in SLT’2016, 2016.

[12]L. Li, Y. Chen, Y. Shi, Z. Tang, and D. Wang, “Deep speaker feature learning for text-independent speaker verification,” arXiv preprint arXiv:1705.03670, 2017.

[13]Tomi Kinnunen, Haizhou Li. An Overview of Text-Independent Speaker Recognition: from Features to Supervectors. Speech Communication, Elsevier : North-Holland, 2009, 52 (1), pp.12. <10.1016/j.specom.2009.08.009>. <hal-00587602>

[14] J. Ming, T.J. Hazen, J.R. Glass, and D.A. Reynolds, “Robust speaker recognition in noisy conditions,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 15, no. 5, pp. 1711–1723, 2007.

References

[15] Arnab Poddar1 Md Sahidullah2 Goutam Saha1,” Verification with Short Utterances: A Review of Challenges, Trends and Opportunities”.

[16] Perrachione, T.K. (in press). "Speaker recognition across languages" in S. Frühholz & P. Belin (Eds.), The Oxford Handbook of Voice Perception, Oxford: Oxford University Press.

[17]Andreas Lanitis, “A survey of the effects of aging on bio- metric identity verification,” International Journal ofBio- metrics, vol. 2, no. 1, pp. 34–52, 2009.

THANK YOU!