An Utterance Comparison Model for Speaker Clustering Using Factor Analysis

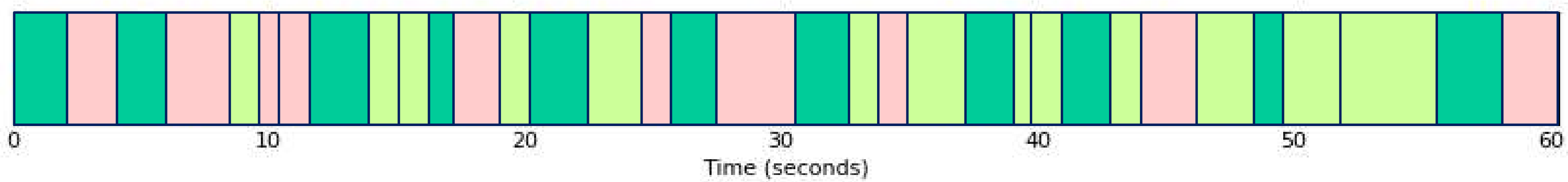
Woojay Jeon
Changxue Ma
Dusan Macho

Speaker Clustering

- **Definition**
  - Cluster a set of *speaker-homogeneous* speech utterances such that each cluster corresponds to a unique speaker
  - Each utterance contains speech from only one speaker
  - The source speakers and the number of speakers are unknown
- **Applications**
  - Speech recognition : Use a predefined set of speaker clusters to do robust speaker adaptation when test data is very limited
  - Speaker diarization : “Who spoke when”

Speaker Diarization

- **Given an unlabelled, random recording of an unknown number of unknown speakers talking, determine the parts spoken by each person.**



Cluster 1

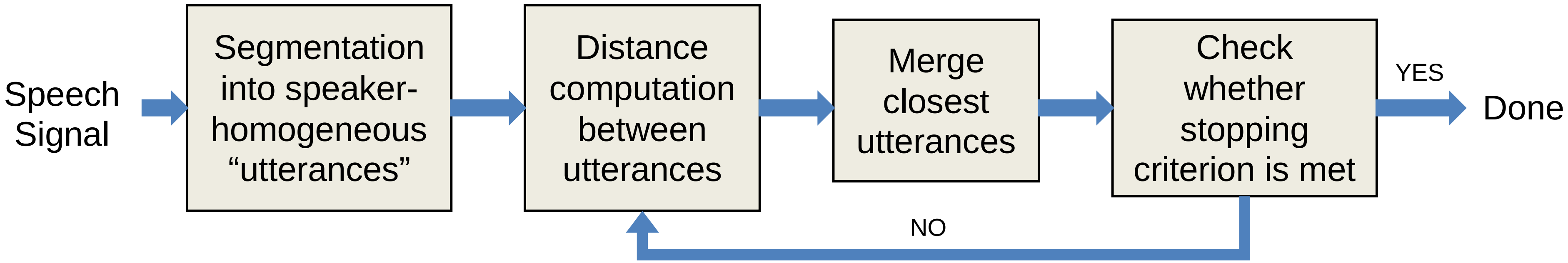
Clustering error example 3: there are less speakers than clusters

Cluster 2

Cluster 3

- **If the number of clusters == the number of speakers, and each cluster actually contains speech by only one person, we have perfect speaker diarization.**

“Classic” Speaker Diarization Method



Popular Distance Measures

- Given two arbitrary speech utterances X_A and X_B, what is the "distance" between them?
- Generalized Likelihood Ratio (GLR)

$$\mathrm{GLR}_{A,B} = \log \frac{P\left(X_A, X_B \middle| \boldsymbol{\lambda}_{A,B}\right)}{P\left(X_A \middle| \boldsymbol{\lambda}_A\right) P\left(X_B \middle| \boldsymbol{\lambda}_B\right)}$$

- Cross-Likelihood Ratio (CLR)

$$\mathrm{CLR}_{A,B} = \log \frac{P\left(X_A \middle| \boldsymbol{\lambda}_B\right)}{P\left(X_A \middle| \boldsymbol{\lambda}_A\right)} + \log \frac{P\left(X_B \middle| \boldsymbol{\lambda}_A\right)}{P\left(X_B \middle| \boldsymbol{\lambda}_B\right)}$$

- Bayesian Information Criterion (BIC) Distance

$$\mathrm{BICD}_{A,B} = BIC\left(X_A, X_B \,\mathrm{separate}\right) - BIC(X_A, X_B \ \mathrm{merged})$$

where $\mathrm{BIC} = \left(\mathrm{Log\ Likelihood\ of\ Observations}\right) - \frac{1}{2} \cdot \alpha \cdot \left(\mathrm{num\ params}\right) \cdot \log\left(\mathrm{num\ frames}\right)$

A Better Distance Measure?

- The previous distance measures are purely mathematical constructs
  - Lack of a rigorous justification on how they can compare utterances based on ***physical*** speaker similarity
  - The only physical element is the feature set (MFCCs)
- No statistical training is involved
- Somewhat ad-hoc
- “Trainable” distance metrics [Aronowitz 07], Eigenvoice-based methods [Falthauser 01], [Castaldo 08] have been proposed to address these problems
- Eigenvoice, Eigenchannels, and Factor Analysis [Kenney 08] provide an elegant, analytic framework for modeling interspeaker and intraspeaker variability

An “Utterance Comparison Model”

- Given two arbitrary speech utterances X_A and X_B, define the distance as **the probability that the two utterances were spoken by the same person**

$$X_A = \{\mathbf{x}_{a,1}, \mathbf{x}_{a,2}, \cdots, \mathbf{x}_{a,A}\},\ X_B = \{\mathbf{x}_{b,1}, \mathbf{x}_{b,2}, \cdots, \mathbf{x}_{b,B}\}$$

- Assuming each speaker is w_i, and the posterior probability $P(w_i \mid X)$ is known, we have

$$P(H_1 | X_A, X_B) = \sum_{i=1}^{W} P(w_i | X_A) P(w_i | X_B)$$

where W is the population of the world

- We also have

$$P(H_0 | X_a, X_b) = \sum_{i=1}^{W} \sum_{j=1, j \neq i}^{W} P(w_i | X_a) P(w_j | X_b)$$

- Using $\sum_{i=1}^{W} P(w_i | X) = 1$, it is easy to show that

$$P(H_0 | X_a, X_b) + P(H_1 | X_a, X_b) = 1$$

Factor Analysis

- Factor analysis says that

$$\mathbf{s} = \mathbf{m} + V\mathbf{y} + U\mathbf{z}$$

$$\mathbf{y} \sim N[0, I],\ \mathbf{z} \sim N[0, I]$$

  - $\mathbf{s}$: speaker-dependent GMM's mean supervector
  - $\mathbf{m}$: Universal Background Model(UBM)'s mean supervector
  - $\mathbf{y}$: speaker factor vector
  - $\mathbf{z}$: channel factor vector
  - $\mathbf{V}$: Eigenvoice matrix models *inter*-speaker variabilities
  - $\mathbf{U}$: Eigenchannel matrix models *intra*-speaker variabilities

- Assuming each unique speaker w_i is mapped to a unique speaker factor vector $\mathbf{y}_i$, the utterance comparison model becomes

$$P(H_1 | X_A, X_B) = \sum_{i=1}^{W} P(\mathbf{y}_i | X_A) P(\mathbf{y}_i | X_B)$$

$\rightarrow$ The equation still has no practical value

Mold into an Analytical Form

- First instinct:

$$P\left(H_1 \middle| X_A, X_B\right) = \sum_{i=1}^{W} P\left(\mathbf{y}_i \middle| X_A\right) P\left(\mathbf{y}_i \middle| X_B\right) \approx \int_{-\infty}^{\infty} p\left(\mathbf{y} \middle| X_a\right) p\left(\mathbf{y} \middle| X_b\right) d\mathbf{y}$$

$\rightarrow$ WRONG!

- By using calculus and probability theory, the correct form can be derived as

$$\boxed{\begin{aligned} P\left(H_1 \middle| X_a, X_b\right) &\approx \frac{1}{W} \int_{-\infty}^{\infty} \frac{p\left(\mathbf{y} \middle| X_a\right) p\left(\mathbf{y} \middle| X_b\right)}{p\left(\mathbf{y}\right)} d\mathbf{y} \\ &= \frac{1}{W} \frac{1}{p\left(X_a\right)} \frac{1}{p\left(X_b\right)} \int_{-\infty}^{\infty} p\left(X_a \middle| \mathbf{y}\right) p\left(X_b \middle| \mathbf{y}\right) p\left(\mathbf{y}\right) d\mathbf{y} \end{aligned}}$$

- Want **closed form expression** of this. Need to resolve $p(X)$ and $p(X \mid \mathbf{y})$.

Use Eigenvoices

- Simplify the problem by ignoring the intraspeaker variability, i.e.,

$$\mathbf{s} = \mathbf{m} + V\mathbf{y} + \cancelto{0}{U}\mathbf{z}$$

$$\mathbf{y} \sim N\left[0, I\right]$$

$$\begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_M \end{bmatrix} = \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \vdots \\ \mathbf{m}_M \end{bmatrix} + \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_M \end{bmatrix} \mathbf{y}$$

- For utterance X_A with A feature vectors, we have

$$p\left(X_A \middle| \mathbf{y}\right) = \prod_{t=1}^{A} p\left(\mathbf{x}_t \middle| \mathbf{y}\right) = \prod_{t=1}^{A} \sum_{k=1}^{M} c_k N\left(\mathbf{x}_t ; \mathbf{m}_k + V_k \mathbf{y}, C_k\right)$$

Two Identities

- Let $N(\mathbf{x};\mathbf{m},C)$ denote the d-dimensional Gaussian pdf.
- Identity I Any Gaussian can be written as a Gaussian with respect to the mean

$$N\left(\mathbf{x};\mathbf{m}_1+\mathbf{m}_2,C\right)=N\left(\mathbf{m}_1;\mathbf{x}-\mathbf{m}_2,C\right)$$

- Identity II The product of two Gaussians is also a (unnormalized) Gaussian

$$N\left(A_{a\times d}\mathbf{x}_{d\times 1};\mathbf{m}_1,C_1\right)N\left(B_{b\times d}\mathbf{x};\mathbf{m}_2,C_2\right)$$

$$=(2\pi)^{-(a+b-d)/2}\left(\frac{|C_1||C_2|}{|D|}\right)^{-1/2}\cdot N\left(\mathbf{x};D\mathbf{d},D\right)$$

$$\cdot\exp\left[-\frac{1}{2}\left\{-d^T D\mathbf{d}+\mathbf{m}_1{}^T C_1{}^{-1}\mathbf{m}_1+\mathbf{m}_2{}^T C_2{}^{-1}\mathbf{m}_2\right\}\right]$$

$$D_{d\times d}{}^{-1}=A^T C_1{}^{-1}A+B^T C_2{}^{-1}B,\quad D=D^T$$

$$\mathbf{d}_{d\times 1}=A^T C_1{}^{-1}\mathbf{m}_1+B^T C_2{}^{-1}\mathbf{m}_2$$

“One Gaussian” Assumption

- Assume that each vector in X_A was “generated” by only one Gaussian in the GMM

$$p\left(X_A \middle| \mathbf{y}\right) = \prod_{t=1}^{A} p\left(\mathbf{x}_t \middle| \mathbf{y}\right) = \prod_{t=1}^{A} \sum_{k=1}^{M} c_k N\left(\mathbf{x}_t; \mathbf{m}_k + V_k \mathbf{y}, C_k\right)$$

$$\rightarrow \prod_{t=1}^{A} N\left(\mathbf{x}_t; \mathbf{m}_t + V_t \mathbf{y}, C_t\right)$$

$$= \prod_{t=1}^{A} N\left(V_t \mathbf{y}; \mathbf{x}_t - \mathbf{m}_t, C_t\right)$$

- How to decide which mixture?
  - One way is to obtain $\mathbf{y}_{ML}$ via maximum likelihood estimation, which then fully describes all parameters in the GMM, then for each $\mathbf{x}_t$ find the Gaussian with the maximum “occupation” probability

Iteratively Apply Identity II

- Apply Identity II to the first two pairs:

$$p\left(X_A \middle| \mathbf{y}\right) = \prod_{t=1}^{A} N\left(V_t\mathbf{y}; \mathbf{x}_t - \mathbf{m}_t, C_t\right)$$

$$= \underbrace{N\left(V_1\mathbf{y}; \mathbf{x}_1 - \mathbf{m}_1, C_1\right) \cdot N\left(V_2\mathbf{y}; \mathbf{x}_2 - \mathbf{m}_2, C_2\right)}_{\text{Apply Identity II}} \cdot N\left(V_3\mathbf{y}; \mathbf{x}_3 - \mathbf{m}_3, C_3\right) \cdots N\left(V_A\mathbf{y}; \mathbf{x}_A - \mathbf{m}_A, C_A\right)$$

- The result is

$$p\left(X \middle| \mathbf{y}\right) = \left(2\pi\right)^{-(2d-v)/2} \left(\frac{\left|C_1\right|\left|C_2\right|}{\left|D_2\right|}\right)^{-1/2}$$
$$\cdot \exp\left\{-\frac{1}{2}\left[\left(-\mathbf{d}_2{}^T D_2 \mathbf{d}_2\right) + f_2\right]\right\} N\left(\mathbf{y}; D_2\mathbf{d}_2, D_2\right)$$
$$\cdot N\left(V_3\mathbf{y}; \mathbf{x}_3 - \mathbf{m}_3, C_3\right) \cdots N\left(V_A\mathbf{y}; \mathbf{x}_A - \mathbf{m}_A, C_A\right)$$

where

$$D_2{}^{-1} = V_1^T C_1^{-1} V_1 + V_2{}^T C_2{}^{-1} V_2$$

$$\mathbf{d}_2 = V_1^T C_1^{-1}\left(\mathbf{x}_1 - \mathbf{m}_1\right) + V_2{}^T C_2{}^{-1}\left(\mathbf{x}_2 - \mathbf{m}_2\right)$$

$$f_2 = \left(\mathbf{x}_1 - \mathbf{m}_1\right)^T C_1^{-1}\left(\mathbf{x}_1 - \mathbf{m}_1\right) + \left(\mathbf{x}_2 - \mathbf{m}_2\right)^T C_2{}^{-1}\left(\mathbf{x}_2 - \mathbf{m}_2\right)$$

Iteratively Apply Identity II (cont'd)

- Keep going, and notice a pattern, resulting in

$$p\left(X_A \mid \mathbf{y}\right) = \alpha\left(X_A\right) N\left(y; D_A \mathbf{d}_A, D_A\right)$$

where

$$\alpha\left(X_A\right) = \left(2\pi\right)^{-(Ad-v)/2} \left(\frac{1}{\left|D_A\right|} \prod_{t=1}^{A} \left|C_t\right|\right)^{-1/2}$$

$$\cdot \exp\left\{\frac{1}{2}\left[\mathbf{d}_A{}^T D_A \mathbf{d}_A - \sum_{t=1}^{A} \left(\mathbf{x}_t - \mathbf{m}_t\right)^T C_t^{-1} \left(\mathbf{x}_t - \mathbf{m}_t\right)\right]\right\}$$

$$D_A^{-1} = \sum_{t=1}^{A} V_t^T C_t^{-1} V_t$$

$$\mathbf{d}_A = \sum_{t=1}^{A} V_t^T C_t^{-1} \left(\mathbf{x}_t - \mathbf{m}_t\right)$$

$$f_A = \sum_{t=1}^{A} \left(\mathbf{x}_t - \mathbf{m}_t\right)^T C_t^{-1} \left(\mathbf{x}_t - \mathbf{m}_t\right)$$

Expression for the Prior

- This also allows us to obtain a closed form solution for the pdf

$$
\begin{aligned}
p(X_A) &= \int_{-\infty}^{+\infty} p(X_A|\mathbf{y})\, p(\mathbf{y})\, d\mathbf{y} \\
&= \int_{-\infty}^{+\infty} \alpha(X_A)\, N(\mathbf{y}; D_A\mathbf{d}_A, D_A)\, N(\mathbf{y}; \mathbf{0}, I)\, d\mathbf{y} \\
&= \alpha(X_A)\, \beta(X_A) \int_{-\infty}^{+\infty} N(\mathbf{y}; J_A\mathbf{d}_A, J_A)\, d\mathbf{y} \\
&= \alpha(X_A)\, \beta(X_A)
\end{aligned}
$$

where

$$
\beta(X) = (2\pi)^{-v/2} \left(\frac{|D_A|}{|J_A|} \right)^{-1/2} \exp\left\{ \frac{1}{2} \left[\mathbf{d}_A{}^T J_A \mathbf{d}_A - \mathbf{d}_A{}^T D_A \mathbf{d}_A \right] \right\}
$$

$$
J_A{}^{-1} = D_A{}^{-1} + I
$$

The Closed-Form Utterance Comparison Model

- We have

$$P\left(H_1 \middle| X_a, X_b\right) = \frac{1}{W} \frac{1}{p\left(X_a\right)} \frac{1}{p\left(X_b\right)} \int_{-\infty}^{\infty} p\left(X_a \middle| \mathbf{y}\right) p\left(X_b \middle| \mathbf{y}\right) p\left(\mathbf{y}\right) d\mathbf{y}$$

$$= \frac{1}{W} \frac{1}{\alpha\left(X_a\right)\beta\left(X_a\right)} \frac{1}{\alpha\left(X_b\right)\beta\left(X_b\right)} \int_{-\infty}^{\infty} \alpha\left(X_a\right)\alpha\left(X_b\right)$$

$$N\left(\mathbf{y}; D_A \mathbf{d}_A, D_A\right) N\left(\mathbf{y}; D_B \mathbf{d}_B, D_B\right) N\left(\mathbf{y}; \mathbf{0}, I\right) d\mathbf{y}$$

The Closed-Form Utterance Comparison Model

- Use Identity II again, simplify, and the final form is

$$P\left(H_1 \middle| X_A, X_B\right) = \frac{1}{W}\left(\frac{|J_A||J_B|}{|D|}\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left\{-\mathbf{d}^T D\mathbf{d} + \mathbf{d}_A{}^T J_A \mathbf{d}_A + \mathbf{d}_B{}^T J_B \mathbf{d}_B\right\}\right]$$

where

$$D_A{}^{-1} = \sum V_{A,t}{}^T C_{A,t}{}^{-1} V_{A,t}$$

$$J_A{}^{-1} = I + D_A{}^{-1}$$

$$\mathbf{d}_A = \sum_{t=1}^{A} V_{A,t}{}^T C_{A,t}{}^{-1}\left(\mathbf{x}_{A,t} - \mathbf{m}_{A,t}\right)$$

$$D^{-1} = J_A{}^{-1} + J_B{}^{-1} - I$$

$$\mathbf{d} = \mathbf{d}_A + \mathbf{d}_B$$

- Hence, for each utterance, we need only $\{\mathbf{d}_A, J_A\}$ to compute the utterance comparison function.

Experiment Using CALLHOME Corpus

Table 1. Clustering accuracy for CALLHOME utterances

	Proposed	CLR	GLR	EV
I_{PUR}	86.82	85.50	83.70	48.60
I_{SPK}	81.97	74.21	64.06	64.24

- 680 phone conversations with the number of speakers ranging from 2 to 7 (more than half have 2)
- The features were 12 MFCC coefficients+E/D
- A harmonicity-based Voice Activity Detector was used to drop out non-speech frames
- The eigenvoices were trained using PCA on MAP-adapted speaker-dependent GMMs.
- Each GMM had 256 Gaussians, and the number of eigenvoices was set to 20.
- The cluster purity and speaker number accuracy were measured to evaluate performance

Extension of Model Including Eigenchannels

- Use both Eigenvoices and Eigenchannels

$$\mathbf{s} = \mathbf{m} + V\mathbf{y} + U\mathbf{z}$$

$$\mathbf{y} \sim N[0, I],\ \mathbf{z} \sim N[0, I]$$

$$\begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_M \end{bmatrix} = \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \vdots \\ \mathbf{m}_M \end{bmatrix} + \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_M \end{bmatrix} \mathbf{y} + \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_M \end{bmatrix} \mathbf{z}$$

- We have

$$p(X|\mathbf{y}) = \int_{-\infty}^{+\infty} p(X, \mathbf{z}|\mathbf{y})\, d\mathbf{z} = \int_{-\infty}^{+\infty} p(X|\mathbf{y}, \mathbf{z})\, p(\mathbf{z}|\mathbf{y})\, d\mathbf{z} = \int_{-\infty}^{+\infty} p(X|\mathbf{y}, \mathbf{z})\, p(\mathbf{z})\, d\mathbf{z}$$

$$p(X|\mathbf{y}, \mathbf{z}) = \prod_{t=1}^{A} p(\mathbf{x}_t|\mathbf{y}, \mathbf{z}) = \prod_{t=1}^{A} N(\mathbf{x}_t; \mathbf{m}_t + V_t\mathbf{y} + U_t\mathbf{z}, C_t)$$

$$= \prod_{t=1}^{A} N(U_t\mathbf{z}; \mathbf{x}_t - (\mathbf{m}_t + V_t\mathbf{y}), C_t)$$

The Extended Closed-Form Utterance Comparison Model

- After a bold investment of masochistic man-hours, we can obtain

$$P\left(H_1 \middle| X_a, X_b\right) = \frac{1}{W}\left(\frac{\left|H_A\right|\left|H_B\right|}{\left|G\right|}\right)^{-1/2} \exp\left[-\frac{1}{2}\left\{-\mathbf{g}^T G \mathbf{g} + \mathbf{g}_A{}^T H_A \mathbf{g}_A + \mathbf{g}_B{}^T H_B \mathbf{g}_B\right\}\right]$$

where

$$D_A{}^{-1} = \sum_{t=1}^{A} V_t{}^T C_t{}^{-1} V_t, D_A{}^T = D_A$$

$$E_A{}^{-1} \equiv \sum_{t=1}^{A} U_t{}^T C_t{}^{-1} U_t,\ E_A{}^T = E_A$$

$$F_A \equiv \sum_{t=1}^{A} U_t{}^T C_t{}^{-1} V_t$$

$$K_A{}^{-1} = E_A{}^{-1} + I,\ K_A{}^T = K_A$$

$$H_A{}^{-1} = D_A{}^{-1} - F_A{}^T K_A F_A + I,\ H_A{}^T = H_A$$

$$G^{-1} = H_A{}^{-1} + H_B{}^{-1} - I$$

$$\mathbf{d}_{A(v\times 1)} = \sum_{t=1}^{A} V_t{}^T C_t{}^{-1}\left(\mathbf{x}_t - \mathbf{m}_t\right)$$

$$\mathbf{e}_{A(u\times 1)} \equiv \sum_{t=1}^{A} U_t{}^T C_t{}^{-1}\left(\mathbf{x}_t - \mathbf{m}_t\right)$$

$$\mathbf{g}_A = \mathbf{d}_A - F_A{}^T K_A \mathbf{e}_A$$

$$\mathbf{g} = \mathbf{g}_A + \mathbf{g}_B$$

- Using this form with Eigenchannels improved the accuracy of the CALLHOME task by one or two percent points

The End

- Questions?