I-vector representation based on GMM and DNN for audio classification

Najim Dehak

Center for Language and Speech Processing
Electrical Computer Engineering Department
Johns Hopkins University

Thanks to Patrick Cardinal, Lukas Burget, Fred Richardson, Douglas Reynolds, Pedro Torres-Carrasquillo, Hasan Bahari and Hugo Van hamme

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Outline

- Introduction
- Gaussian Mixture Model (GMM) for sequence modeling
  - GMM means adaptation (I-vector)
    - Speaker recognition tasks (data visualization)
    - Language recognition tasks (data visualization)
  - GMM weights adaptation
- Deep Neural Network (DNN) for sequence modeling
  - DNN layer activation subspaces
  - DNN layer activation path with subspace approaches
  - Experiments and results
- Conclusions

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Introduction

- The i-vector approach has been largely used in several speech classification tasks (speaker, language, Dialect recognition, speaker diarization, Speech recognition, Clustering…)
- The I-vector is a compact representation that summarizes what is happening in a given speech recording
- Classical i-vector approach is based on Gaussian Mixture Model (GMM)
  - GMM means
  - Applying subspace-based approaches for GMM weights
- We applying the subspace approaches to model neurons activation
  - Building an i-vector approach on the top of Deep Neural Network (DNN)
  - Modeling the neurons activation on the DNN using subspace techniques
    - Similar to the GMM weight adaptation approaches

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Outline

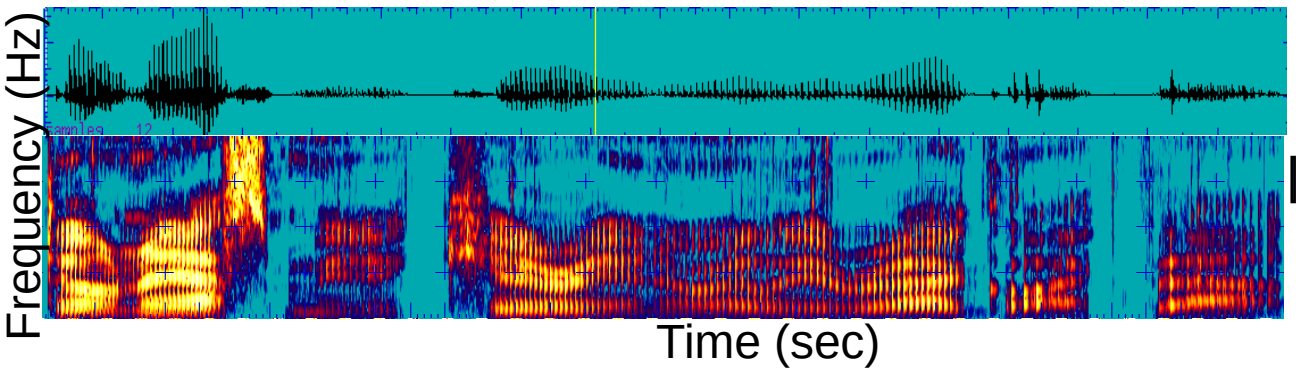
- Introduction
- Gaussian Mixture Model for sequence modeling
  - GMM means adaptation (I-vector)
    - Speaker recognition tasks (data visualization)
    - Language recognition tasks (data visualization)
  - GMM weights adaptation
- Deep Neural Network for sequence modeling
  - DNN layer activation subspaces
  - DNN layer activation path with subspace approaches
  - Experiments and results
- Conclusions

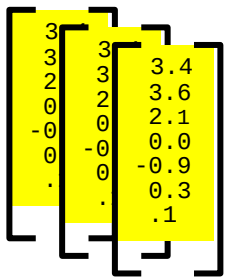
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Modeling Sequence of Features

Gaussian Mixture Models

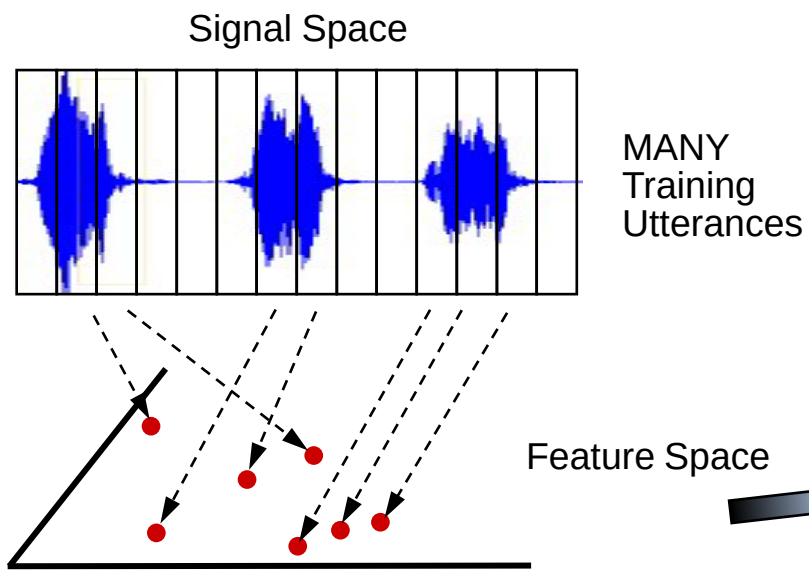
- For most recognition tasks, we need to model the distribution of feature vector sequences

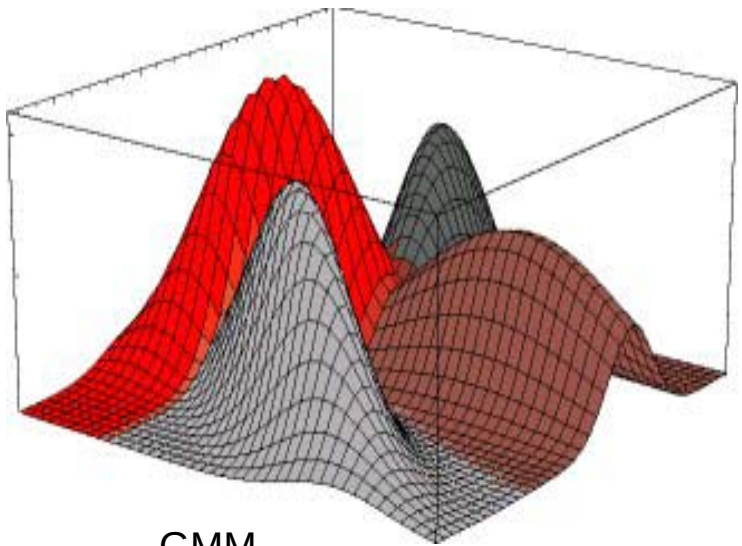




100 vec/sec

- In practice, we often use Gaussian Mixture Models (GMMs)





GMM

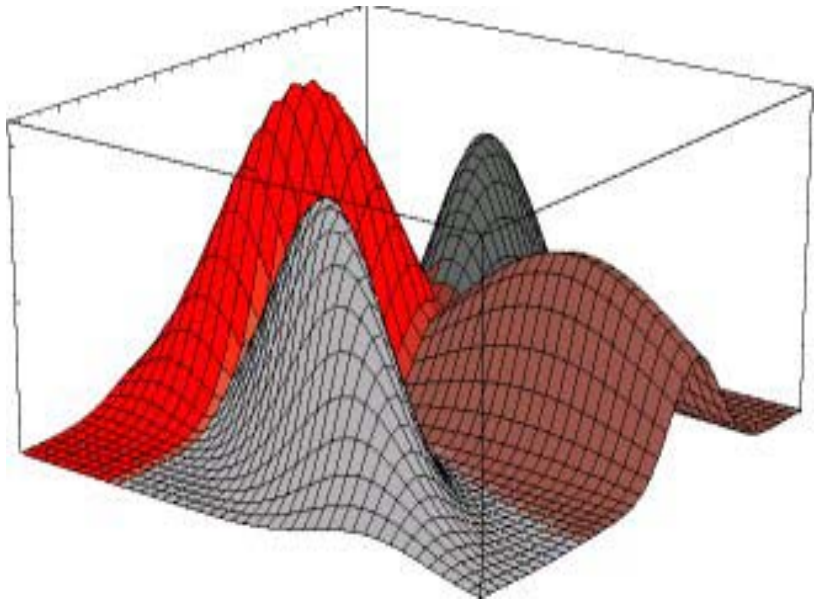
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Gaussian Mixture Models

- A GMM is a weighted sum of Gaussian distributions

$$p(\vec{x} \mid \lambda_s) = \sum_{i=1}^{M} p_i b_i(\vec{x})$$

$$\lambda_s = (p_i, \vec{\mu}_i, \Sigma_i)$$



p_i = mixture weight (Gaussian prior proability)

$\vec{\mu}_i$ = mixture mean vector

Σ_i = mixture covariance matrix

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} \left|\Sigma_i\right|^{1/2}} \exp\left(-\tfrac{1}{2}(\vec{x} - \vec{\mu}_i)' \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i)\right)$$

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

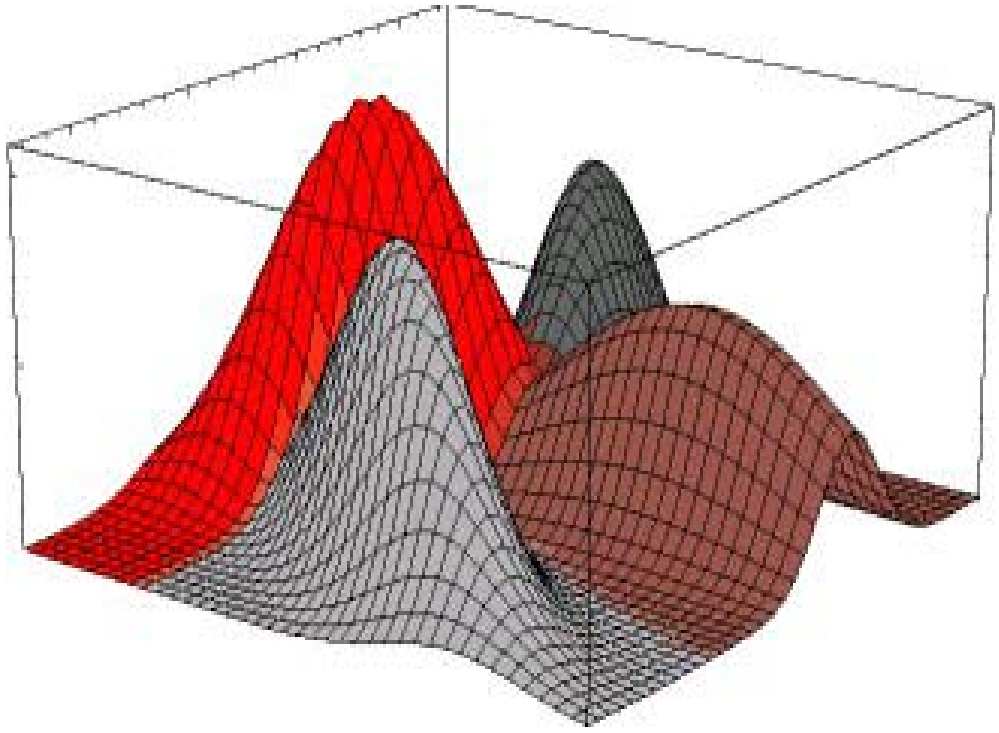
MAP Adaptation

- Target model is often trained by adapting from an Universal Background Model (UBM) [Douglas Reynolds 2000]
  - Couples models together and helps with limited target training data
- Maximum A Posteriori (MAP) Adaptation (similar to EM)
  - Align target training vectors to UBM
  - Accumulate sufficient statistics
  - Update target model parameters with smoothing to UBM parameters
- Adaptation only updates parameters representing acoustic events seen in target training data
  - Sparse regions of feature space filled in by UBM parameters
- Usually we only update the means of the Gaussians

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

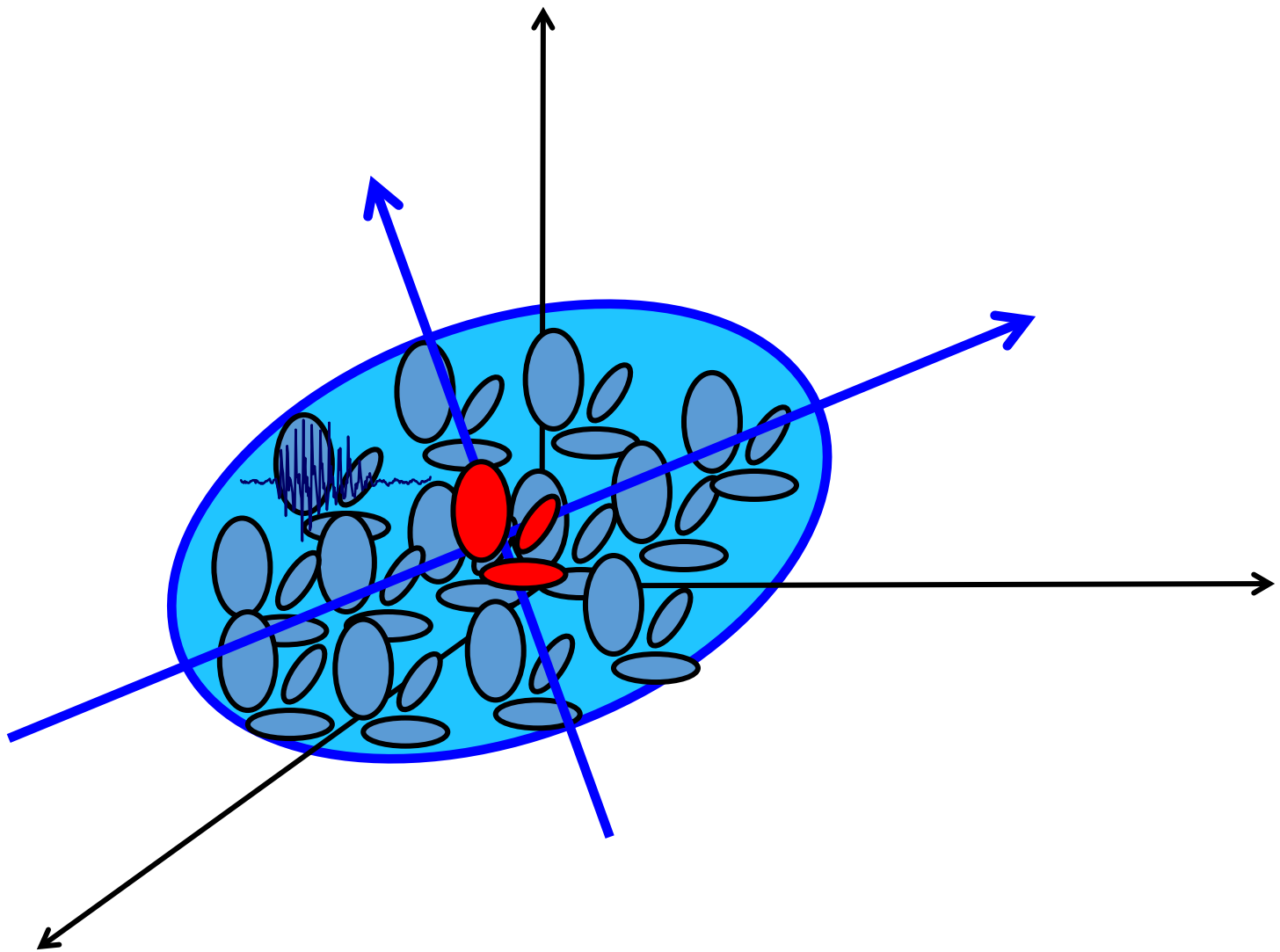
GMM means adaptation: Intuition

- The way the UBM adapts to a given speaker ought to be somewhat constrained
  - There should exist some relationship in the way the mean parameters move relative to speaker to another
  - The Joint Factor Analysis [Kenny 2008] explored this relationship
    - Jointly model between- and within-speaker variabilities
  - Support Vector Machine GMM supervector [Campbell 2006]



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

I-vector : Total variability space



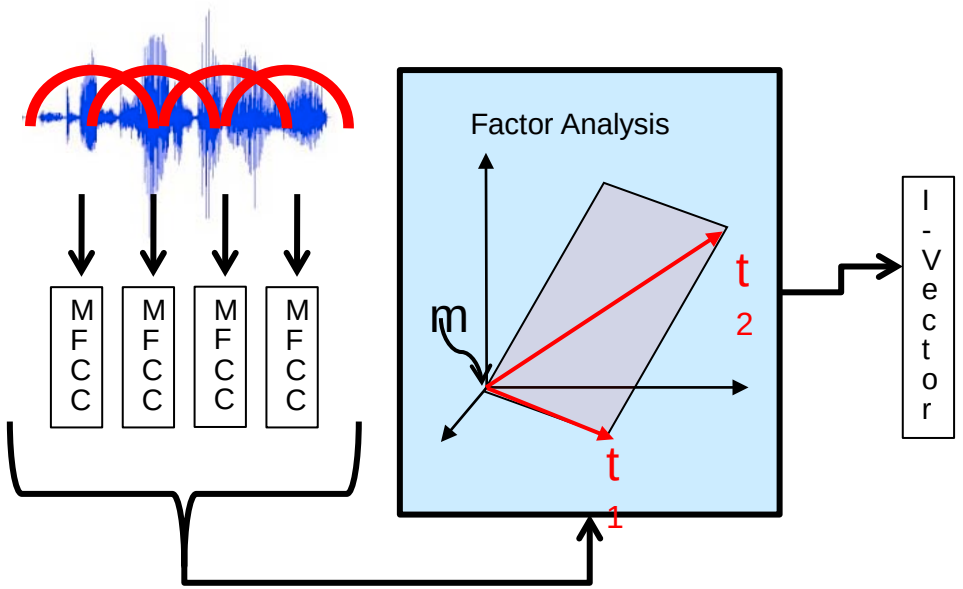
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

I-Vector

- Factor analysis as feature extractor
- Speaker and channel dependent supervector

$$\mathbf{M} = \boldsymbol{m} + \boldsymbol{Tw}$$

  - T is rectangular, low rank (total variability matrix)
  - w standard Normal random (total factors – intermediate vector or i-vector)



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Outline

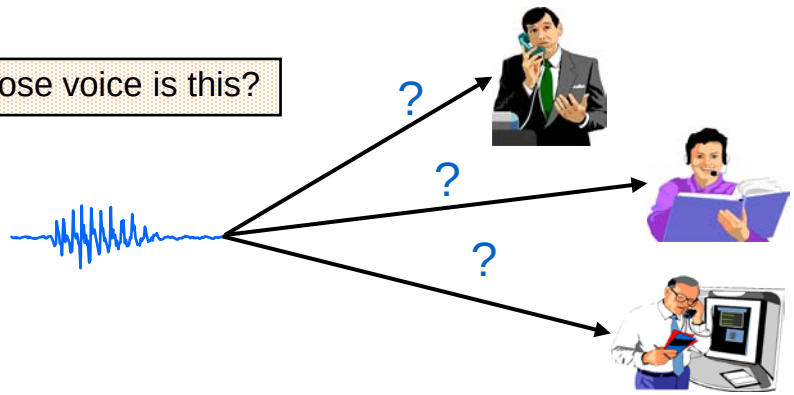
- Introduction
- Gaussian Mixture Model for sequence modeling
  - GMM means adaptation (I-vector)
    - Speaker recognition tasks (data visualization)
    - Language recognition tasks (data visualization)
  - GMM weights adaptation
- Deep Neural Network for sequence modeling
  - DNN layer activation subspaces
  - DNN layer activation path with subspace approaches
  - Experiments and results
- Conclusions

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Speaker Recognition tasks

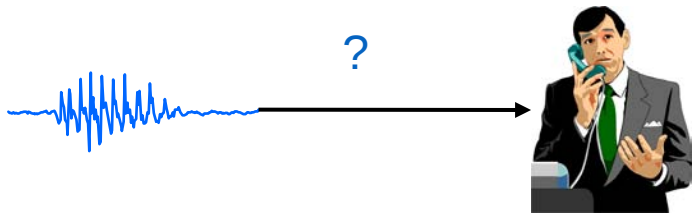
Speaker Identification

Whose voice is this?



Speaker Verification

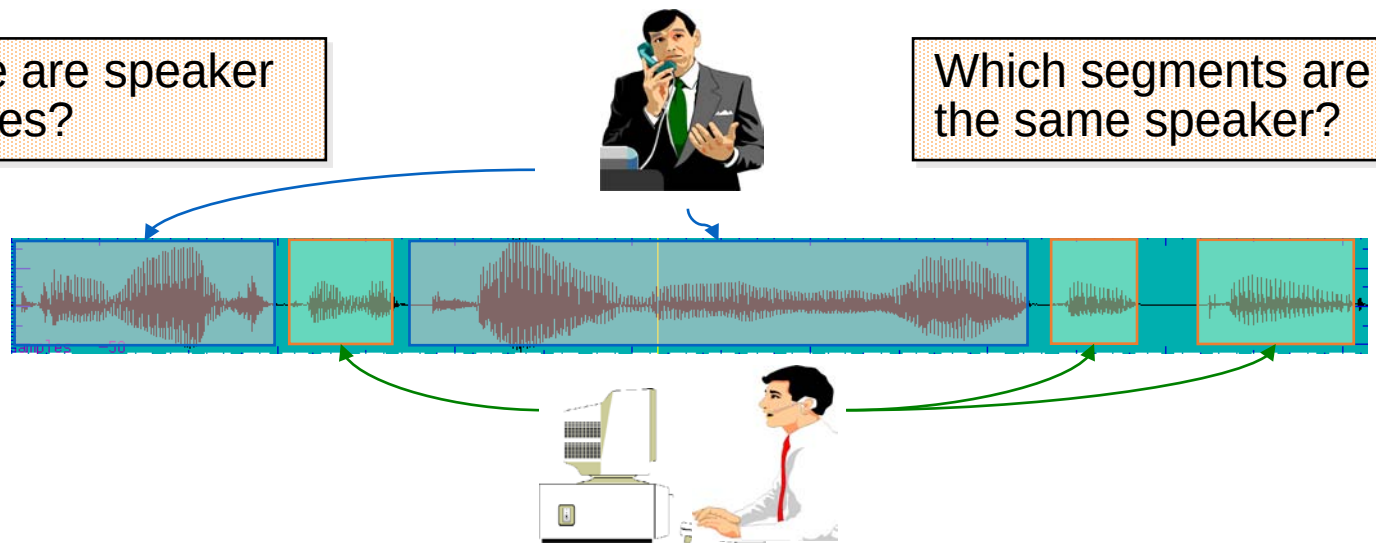
Is this Bob's voice?



Speaker Diarization : Segmentation and clustering

Where are speaker changes?

Which segments are from the same speaker?

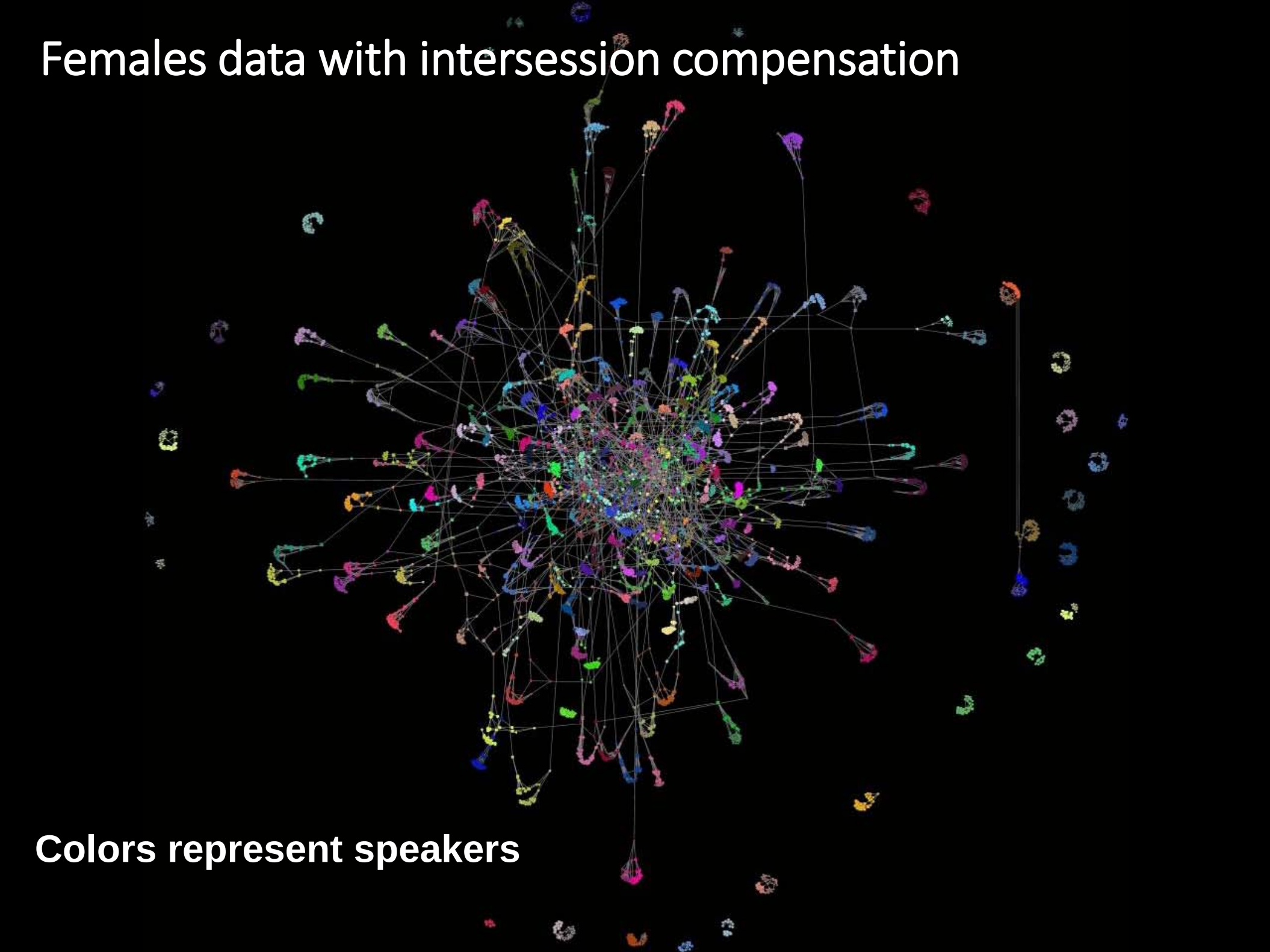


JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Data Visualization based on Graph

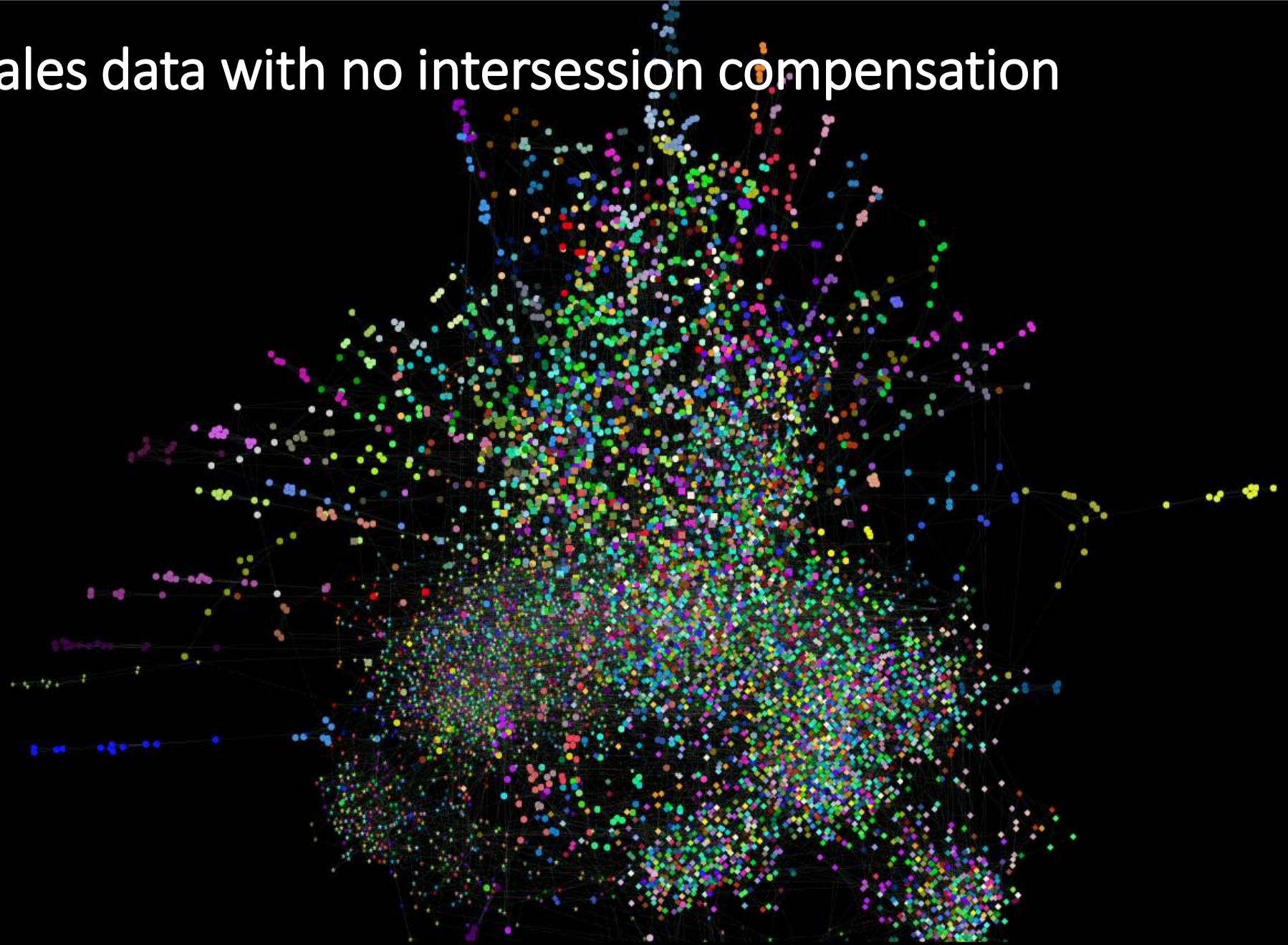
- Nice performance of the cosine similarity for speaker recognition
- **Data visualization using the Graph Exploration System (GUESS) [Eytan 06] (Zahi Karam)**
- Represent segment as a node with connections (edges) to nearest neighbors (3 NN used)
  - NN computed using blind TV system (with and without channel normalization)
- Applied to 5438 utterances from the NIST SRE10 core
  - Multiple telephone and microphone channels
- Absolute locations of nodes not important
- Relative locations of nodes to one another is important:
  - The visualization clusters nodes that are highly connected together
- Meta data (speaker ID, channel info) not used in layout
- Colors and shapes of nodes used to highlight interesting phenomena

Females data with intersession compensation



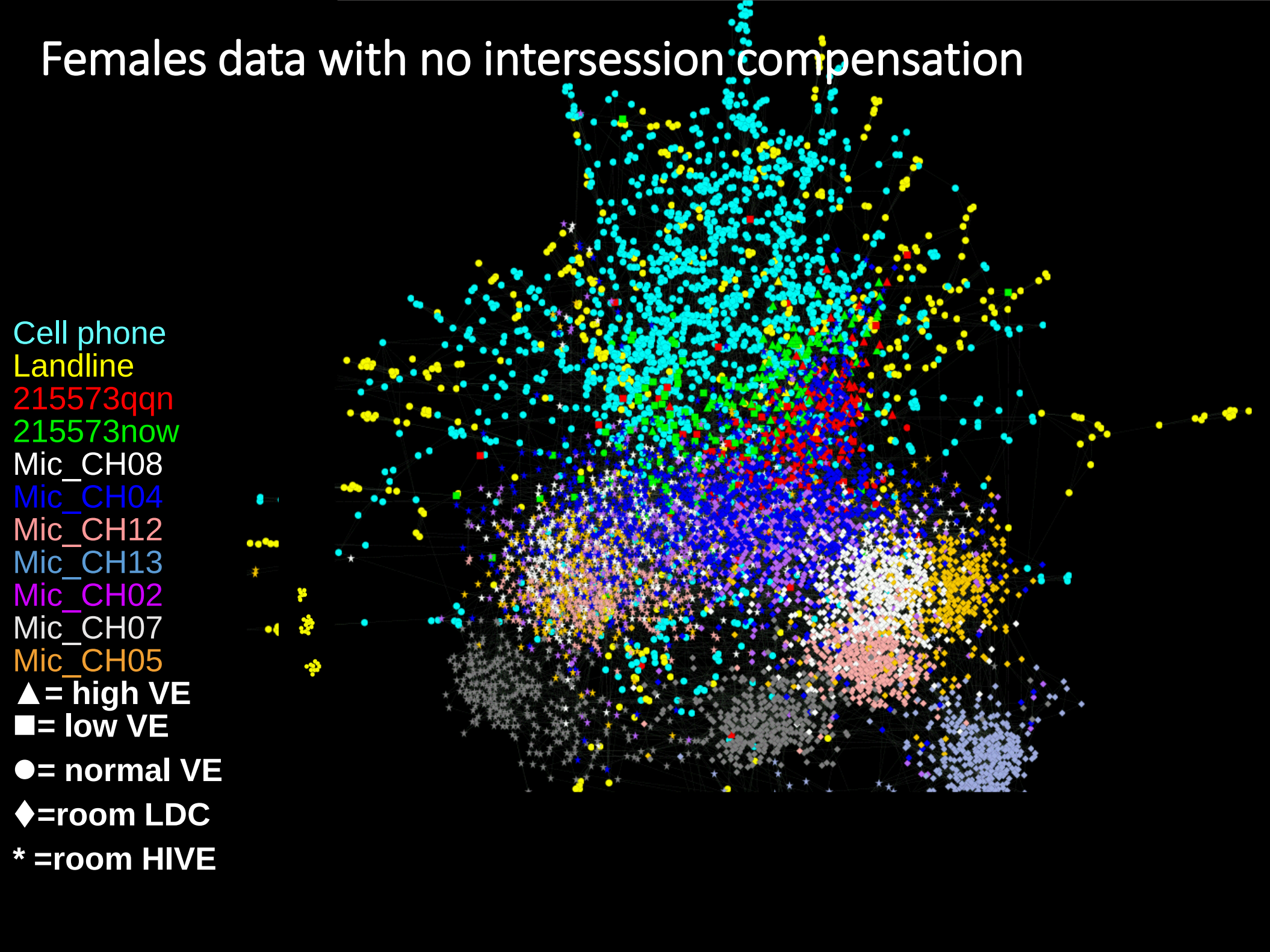
Colors represent speakers

Females data with no intersession compensation

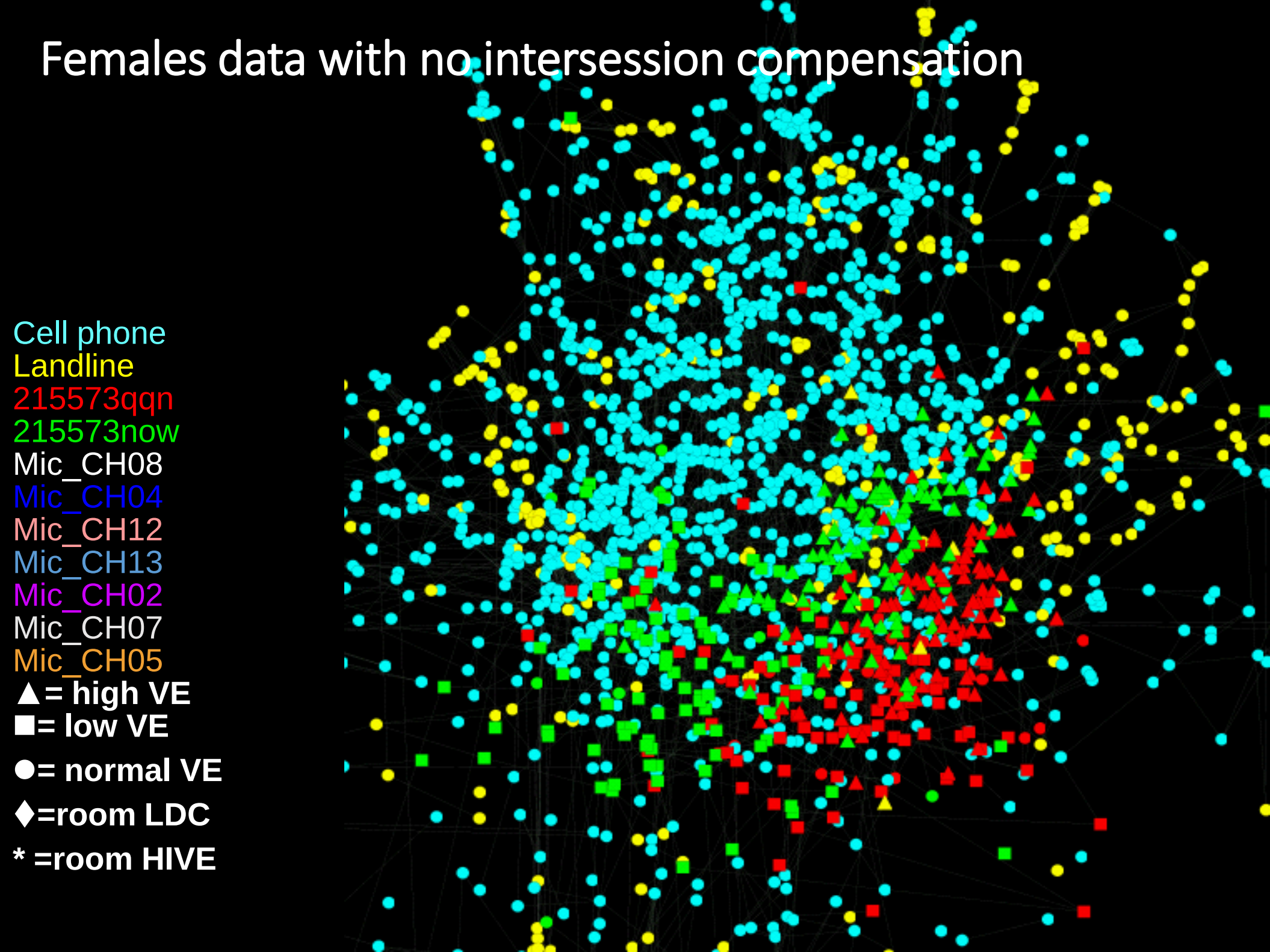


Colors represent speakers

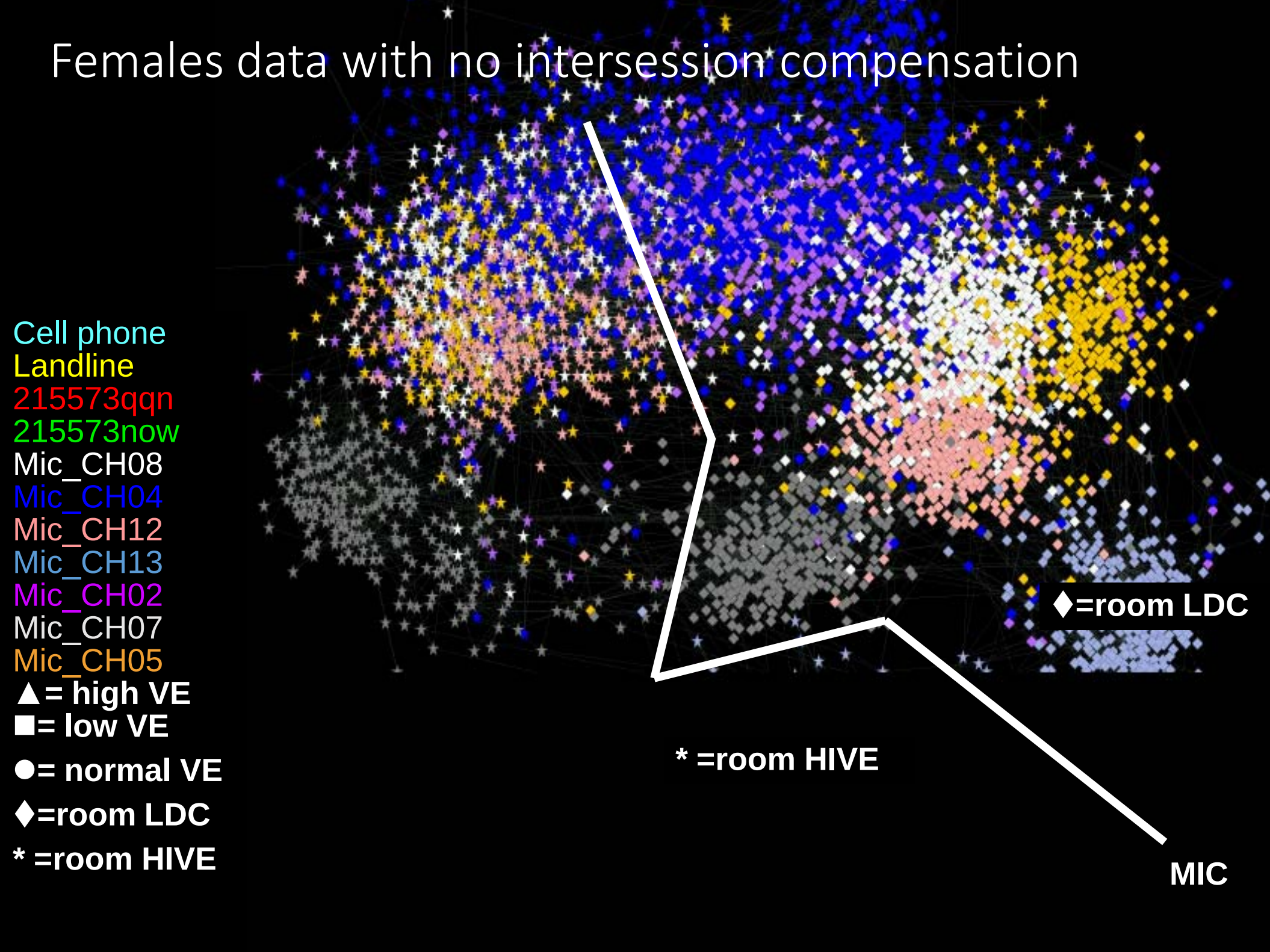
Females data with no intersession compensation



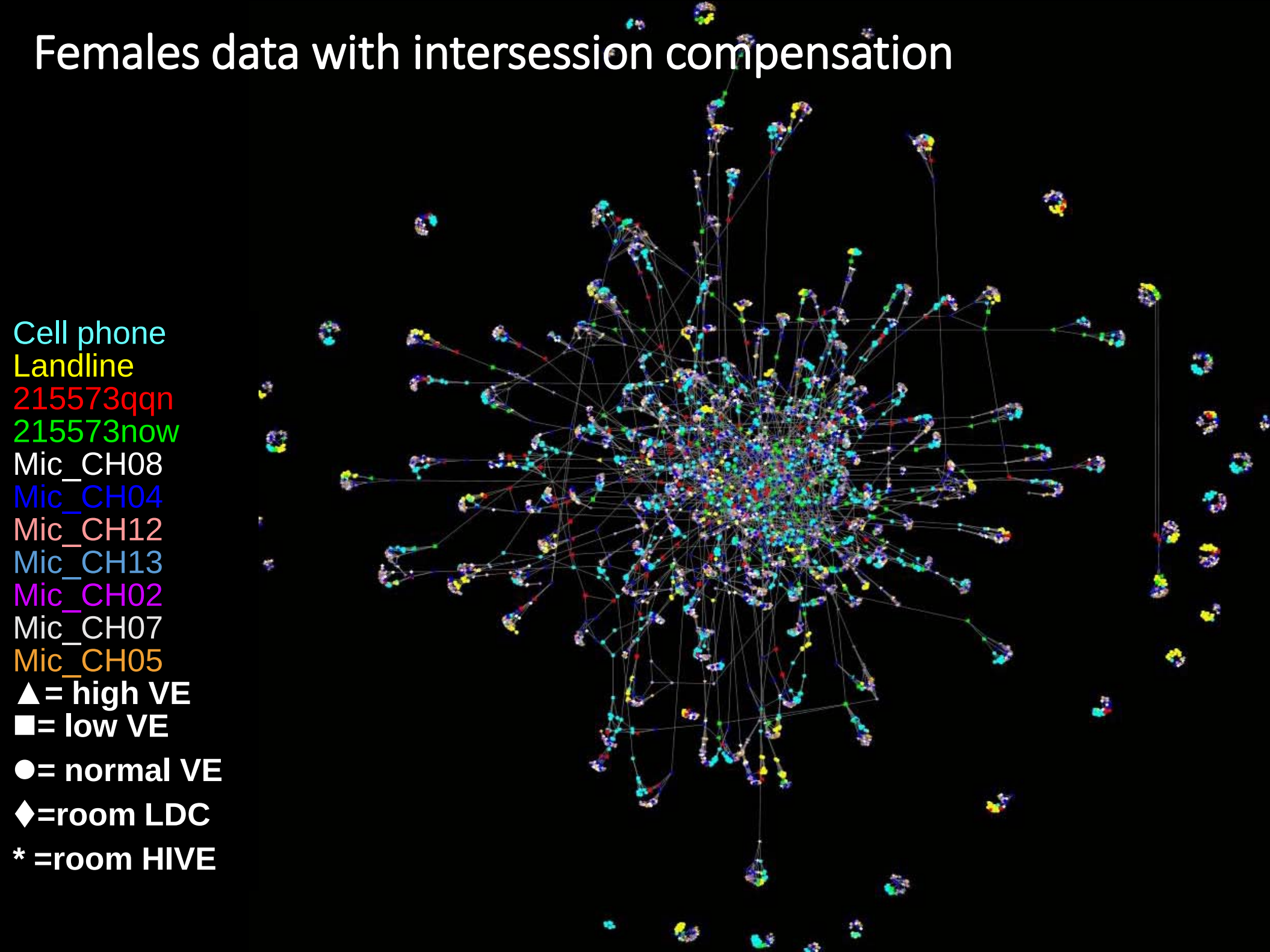
Females data with no intersession compensation



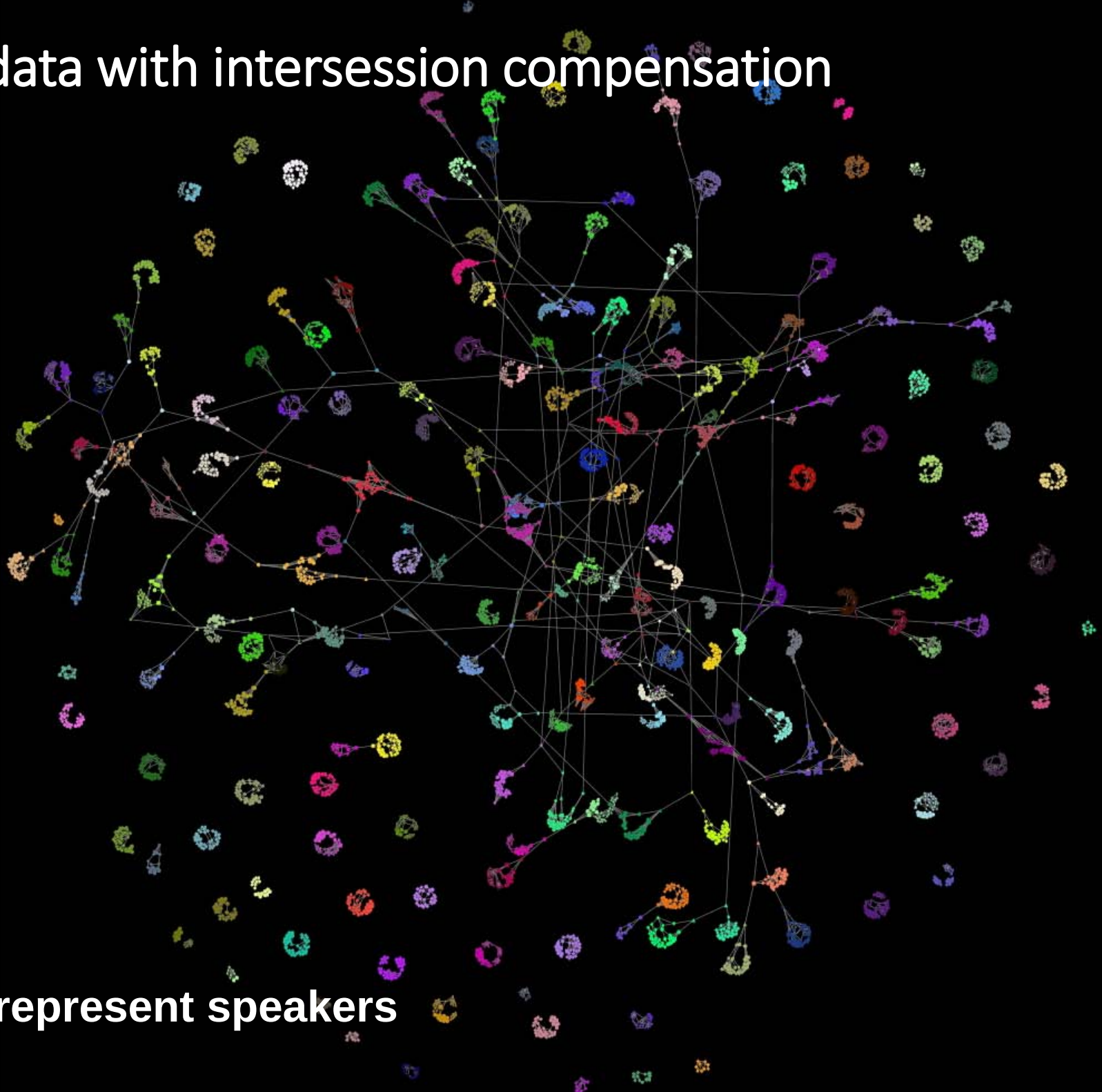
Females data with no intersession compensation



Females data with intersession compensation

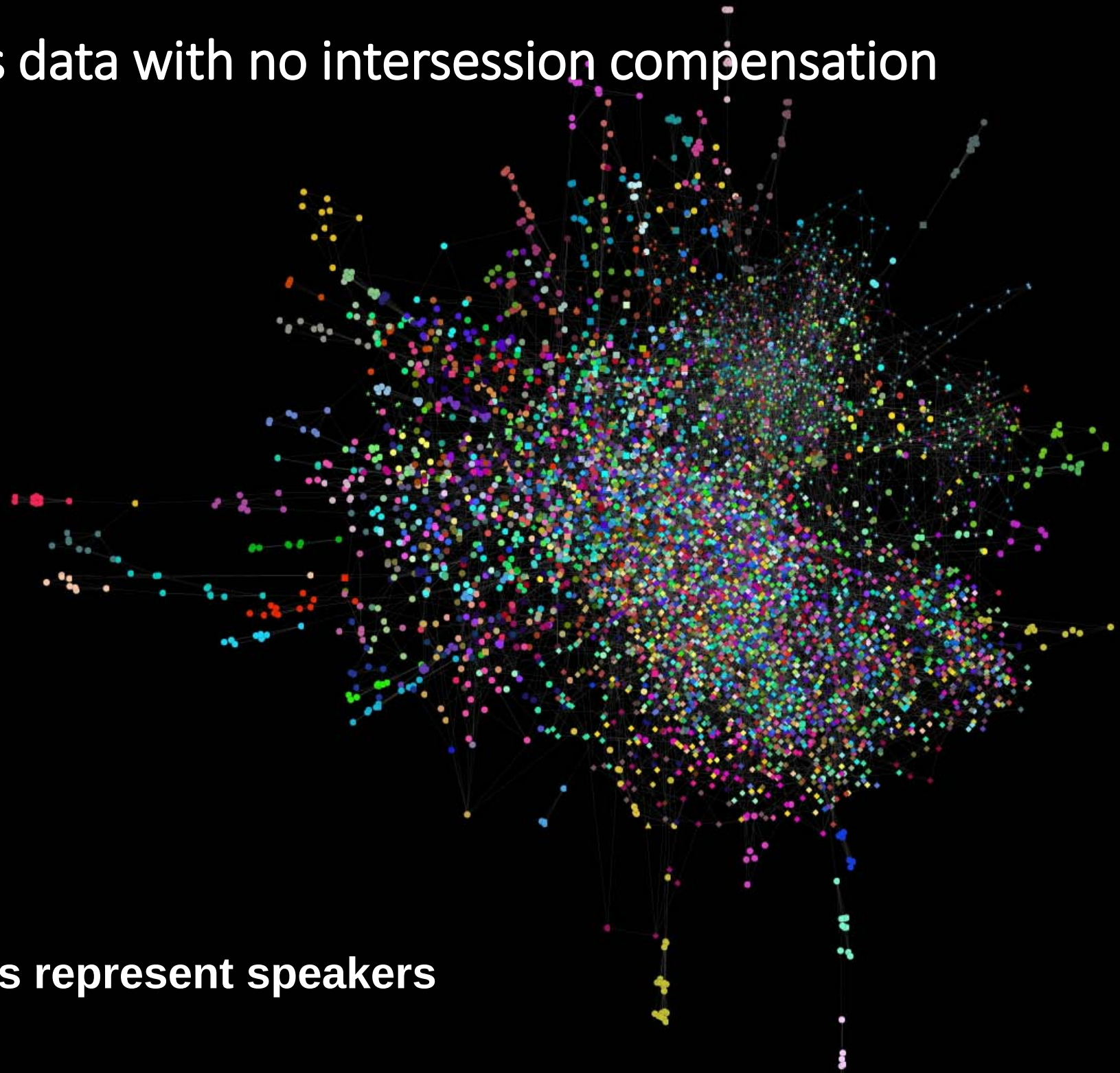


Males data with intersession compensation



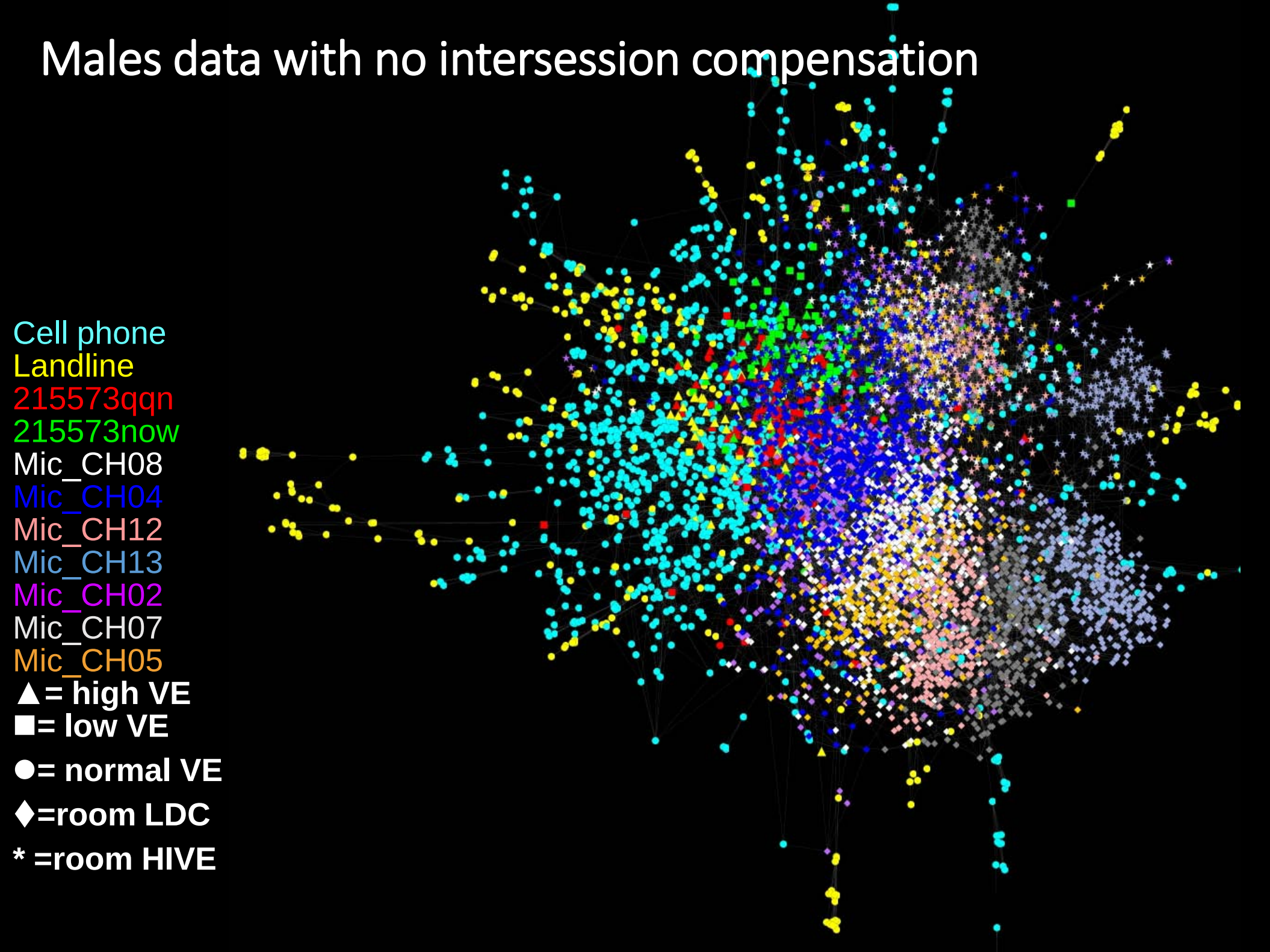
Colors represent speakers

Males data with no intersession compensation

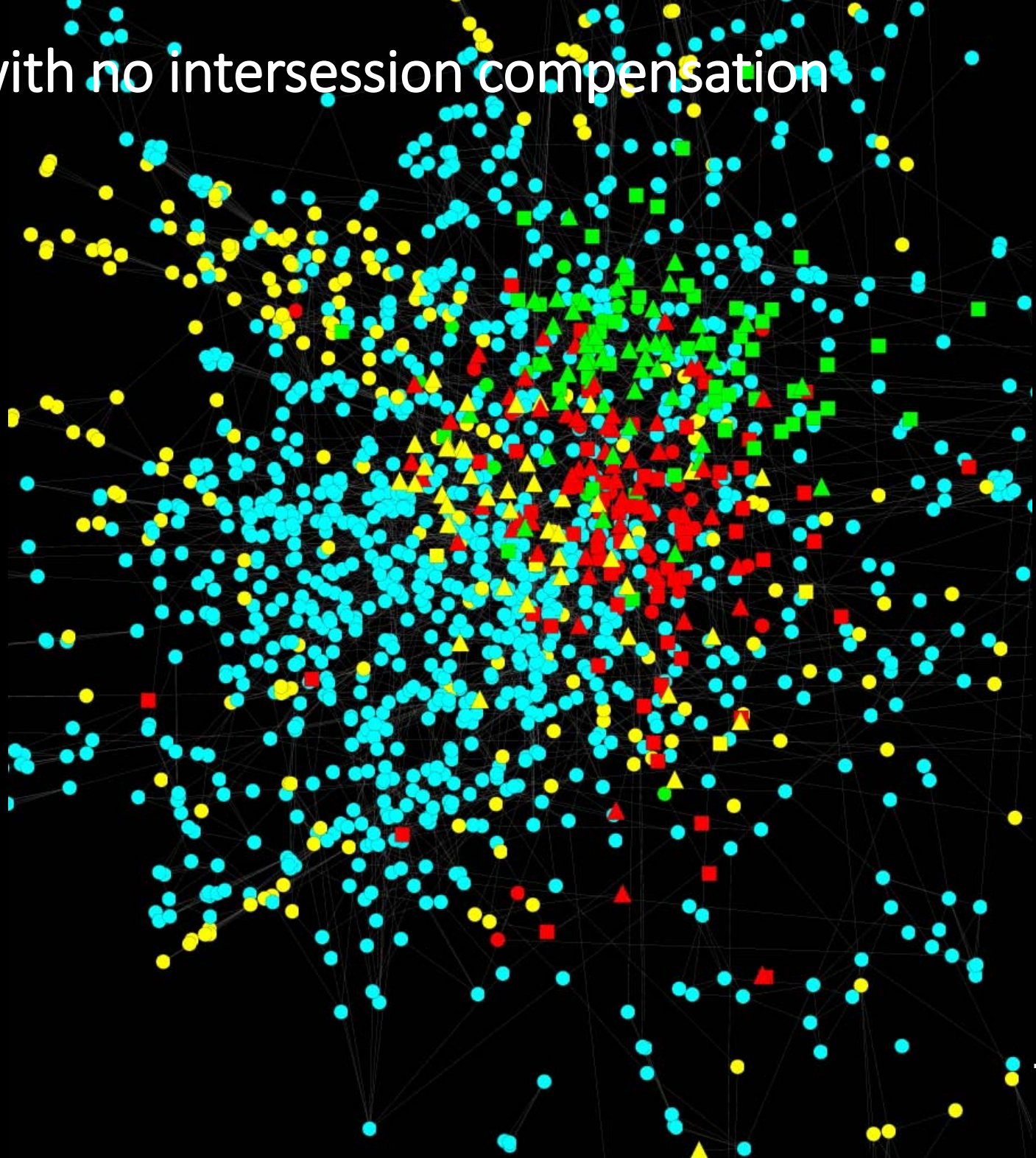


Colors represent speakers

Males data with no intersession compensation



Males data with no intersession compensation

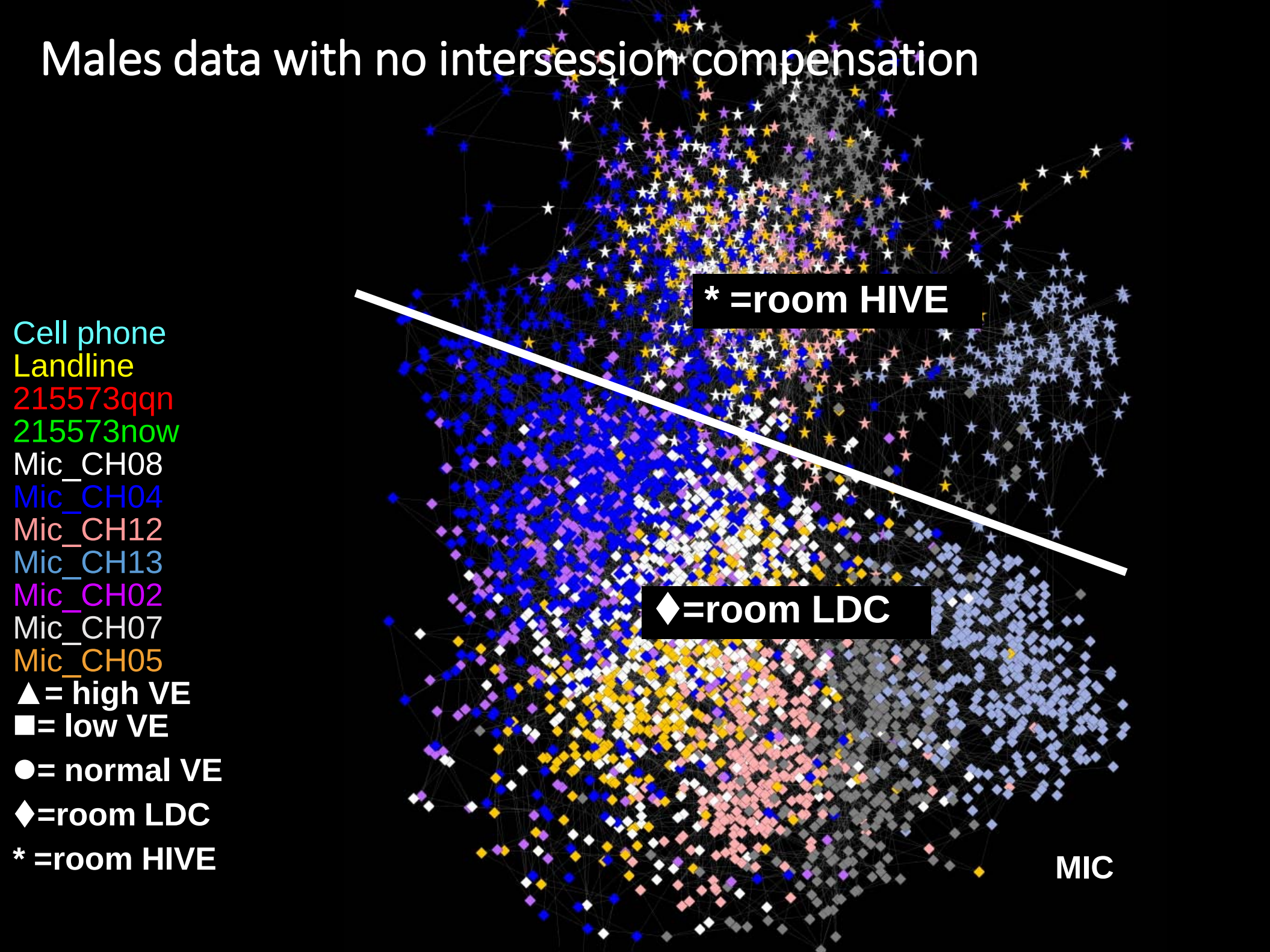


Cell phone
Landline
215573qqn
215573now
Mic_CH08
Mic_CH04
Mic_CH12
Mic_CH13
Mic_CH02
Mic_CH07
Mic_CH05

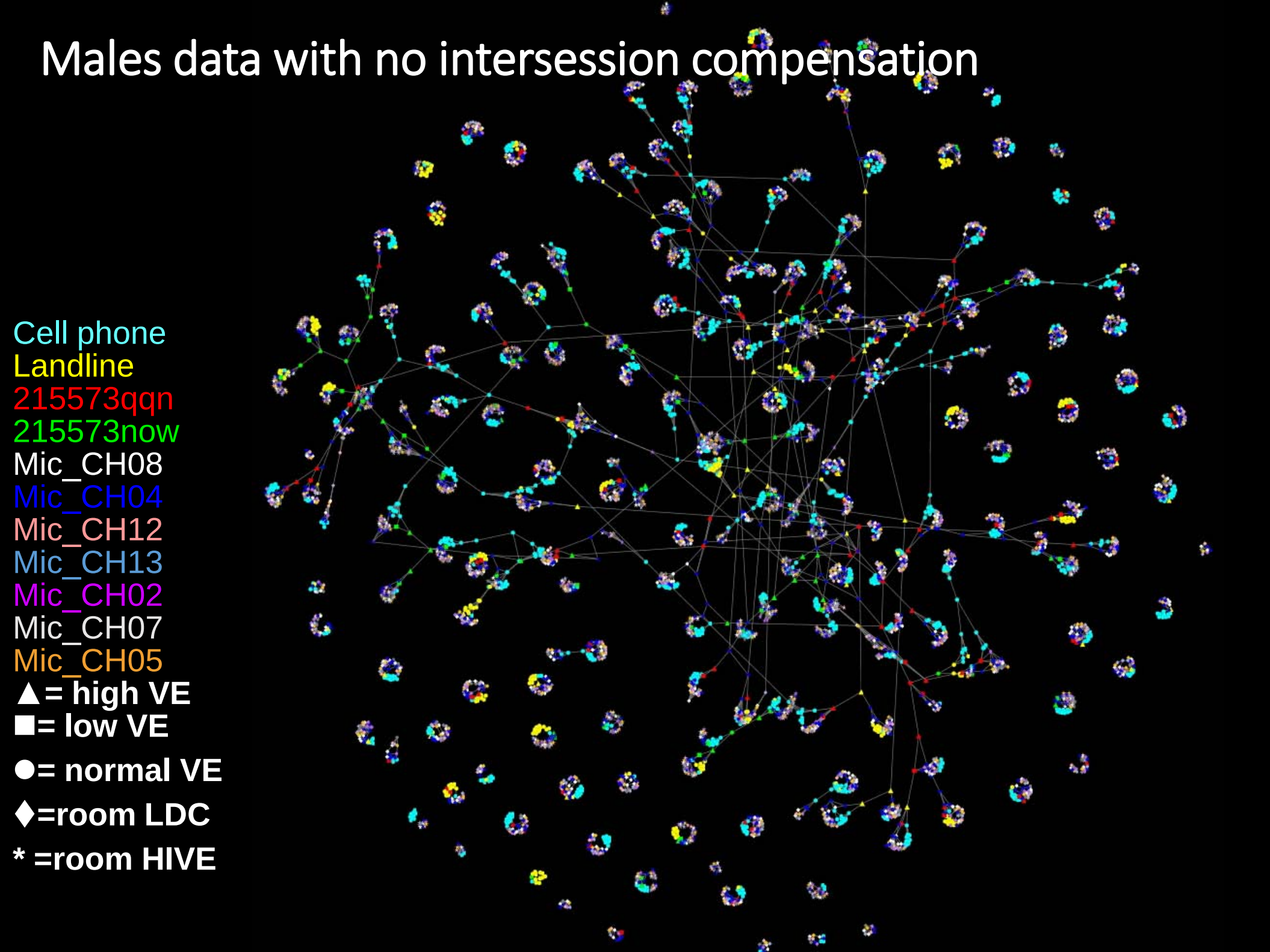
▲= high VE
■= low VE
●= normal VE
♦=room LDC
*** =room HIVE**

TEL

Males data with no intersession compensation



Males data with no intersession compensation



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

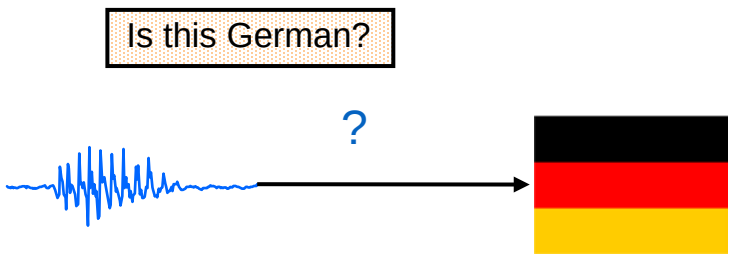
Outline

- Introduction
- Gaussian Mixture Model for sequence modeling
  - GMM means adaptation (I-vector)
    - Speaker recognition tasks (data visualization)
    - Language recognition tasks (data visualization)
  - GMM weights adaptation
- Deep Neural Network for sequence modeling
  - DNN layer activation subspaces
  - DNN layer activation path with subspace approaches
  - Experiments and results
- Conclusions

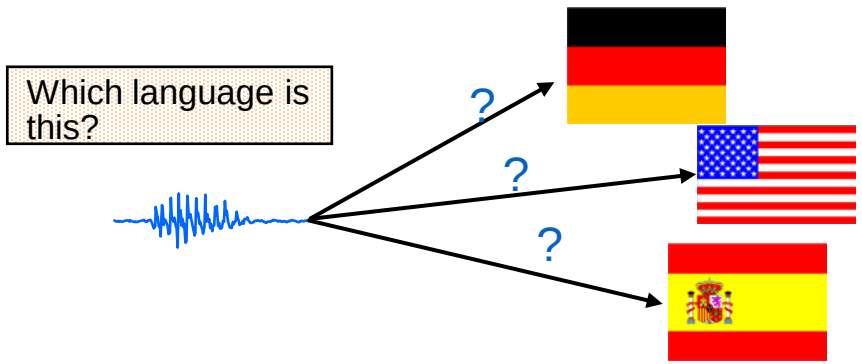
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Language Recognition Tasks

Language Verification



Language Identification



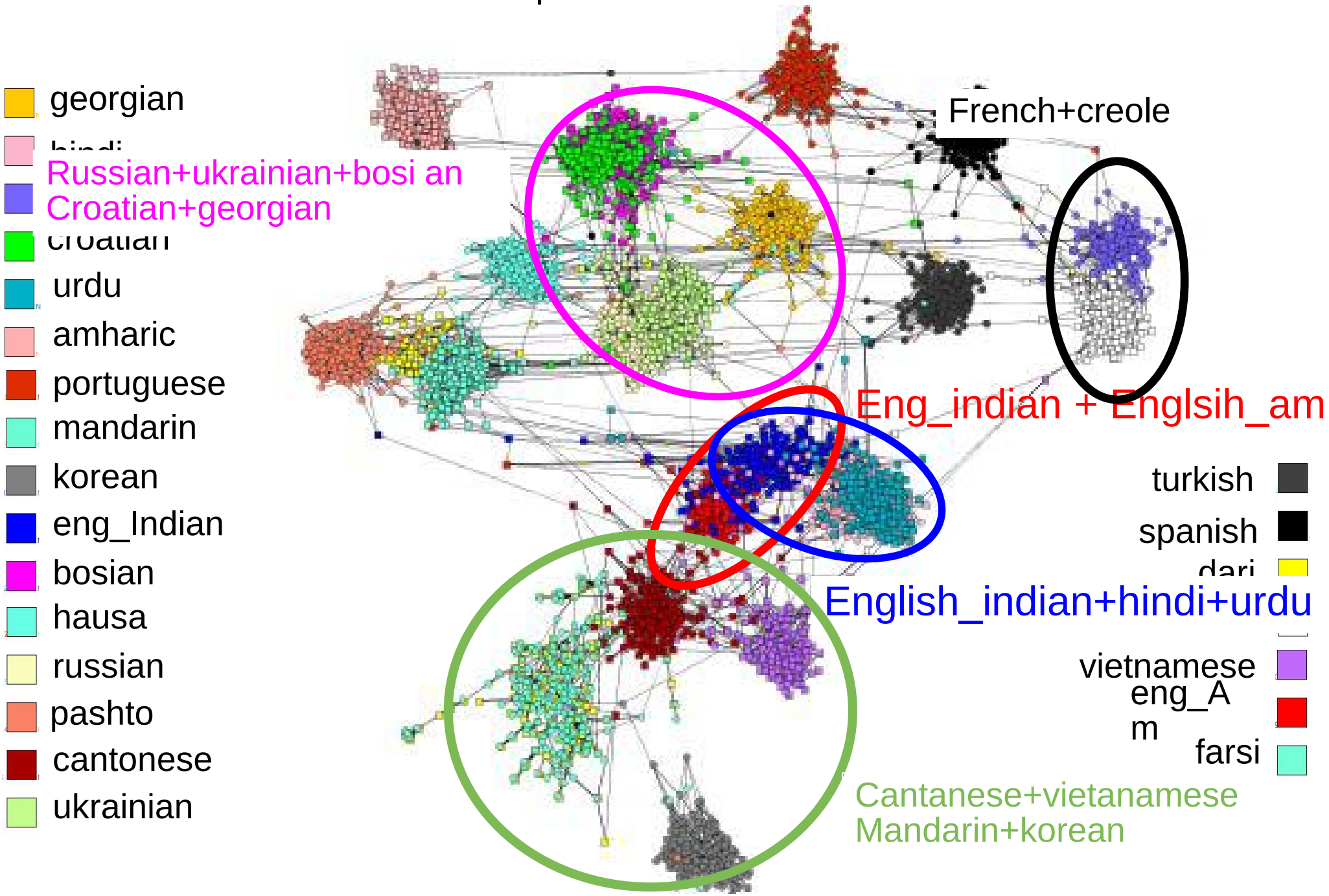
CSAIL

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Data Visualization based on Graph

- Work at Exploring the variability between different languages.
  - Visualization using the Graph Exploration System (GUESS) [Eytan 06]
- Represent segment as a node with connections (edges) to nearest neighbors (3 NN used)
  - Euclidean distance after i-vectors length normalization.
  - NN computed using TV system (with and without intersession compensation normalization)
  - Intersession compensation :
    - Linear Discriminant Analysis + Within Class Covariance Normalization
- Applied to 4600 utterances from 30s condition of the NIST LRE09
  - 200 utterances for Language class
- Absolute locations of nodes not important
- Relative locations of nodes to one another is important:
  - The visualization clusters nodes that are highly connected together
- Colors represent Language Classes

With intersession compensation



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Outline

- Introduction
- Gaussian Mixture Model for sequence modeling
  - GMM means adaptation (I-vector)
    - Speaker recognition tasks (data visualization)
    - Language recognition tasks (data visualization)
  - GMM weights adaptation
- Deep Neural Network for sequence modeling
  - DNN layer activation subspaces
  - DNN layer activation path with subspace approaches
  - Experiments and results
- Conclusions

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

GMM weights adaptation approaches

- GMM weights adaptation
  - Maximum Likelihood estimation
  - Non-negative matrix factorization
  - Subspace Multinomial Model
  - Non-negative Factor Analysis

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Weight Adaptation: ML

Given a speech signal $\chi = \{x_1, \cdots, x_t, \cdots, x_\tau\}$

Components posterior probability for each frame is

$$\gamma_{c,t} = P(c \mid x_t, \theta_{\text{UBM}}) = \frac{b_c P_c(x_t \mid \mu_c, \Sigma_c)}{\sum_{i=1}^{C} b_i P_i(x_t \mid \mu_i, \Sigma_i)}$$

Objective: maximizing auxiliary function

$$\Omega(\lambda, \hat{\lambda}) = \sum_{t=1}^{\tau} \sum_{c=1}^{C} \gamma_{c,t} \log \omega_c \, p(x_t \mid \mu_c, \Sigma_c)$$

Adapted weight is obtained by maximizing

$$\omega_c = \frac{\sum_{t}^{\tau} \gamma_{c,t}}{\tau}$$

CSAIL

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Weight Adaptation: NMF

Main assumption: $\omega_c = A_c h$

both row-vector A_c and column-vector h are non-negative

Calculating A_c and h: maximizing auxiliary function

$$\Omega(\lambda, \hat{\lambda}) = \sum_{t=1}^{\tau} \sum_{c=1}^{C} \gamma_{c,t} \log \omega_c$$

$\gamma_{c,t}$ Occupation count for a class *c* and segment *t*

A_c and h can be obtained iteratively by (E-M) algorithm like.

Maximizing Likelihood is equivalent to minimizing Kullback–Leibler divergence, hence A_c and h can be obtained using the iterative updating rules of standard NMF.

Xueru Zhang, Kris Demuynck and Hugo Van hamme. Rapid speaker adaptation in latent speaker space with non-negative matrix factorization. *Speech Communication*, volume 55, No. 9, pages 893--908, 2013.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Weight Adaptation: Subspace Multinomial Model

- Log likelihood of multinomial model with C discrete classes

$$\log(P(\chi)) = \sum_t \sum_c \gamma_{c,t} \log \phi_c$$

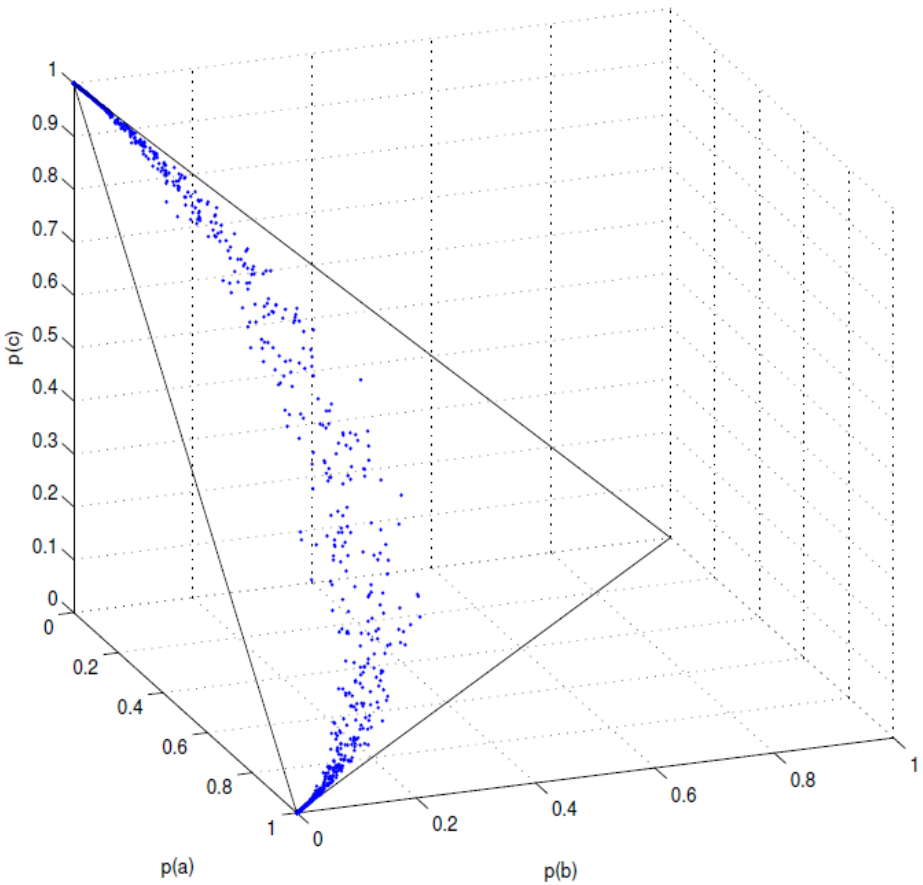
- $\gamma_{c,t}$ Occupation count for a class *c* and segment *t*
- ϕ_c are probabilities of multinomial distribution

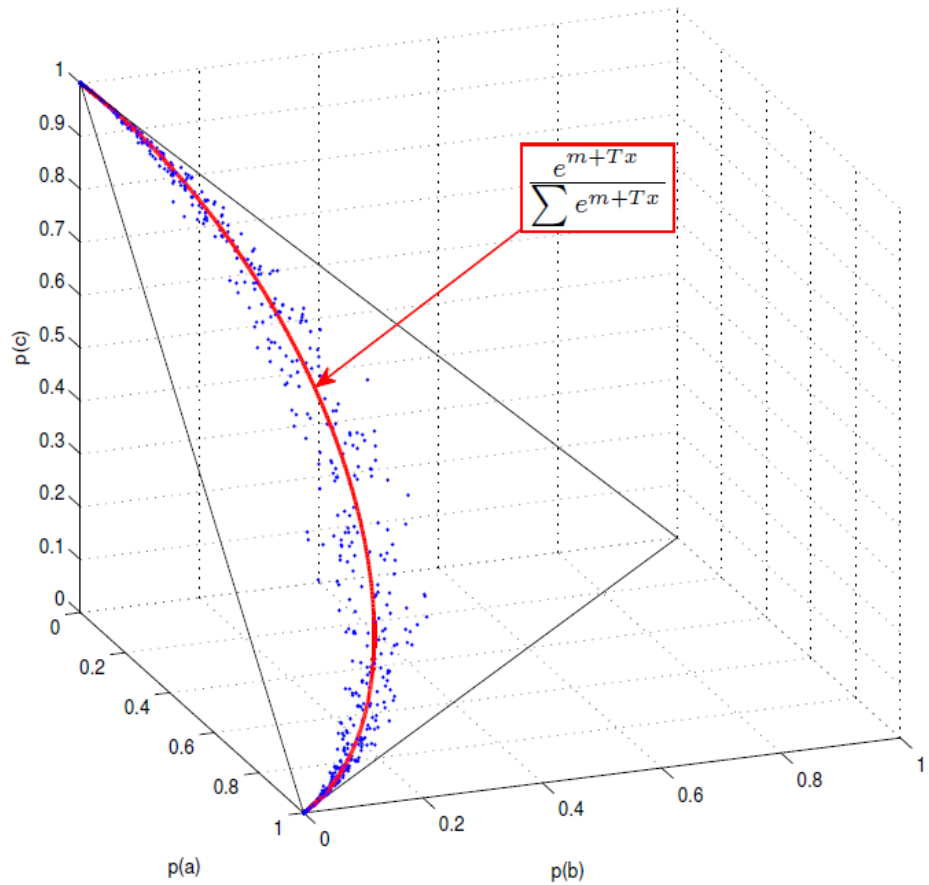
$$\phi_c = \frac{\exp(b_c + L_c r)}{\sum_{j=1}^{C} \exp(b_c + L_c r)}$$

M. Kockmann, L. Burget, O. Glembek, L. Ferrer, and J. Cernocky, “Prosodic speaker verification using subspace multinomial models with intersession compensation,” in *Eleventh Annual Conference of the International Speech Communication Associa- tion*, 2010
Mehdi Soufifar,Lukas Burget, Oldrich Plchot, Sandro Cumani, Jan "Honza" Cernocky, REGULARIZED SUBSPACE MULTINOMIAL MODEL FOR PHONOTACTIC IVECTOREXTRACTION, In Proc. Interspeech 2013, Lyon, France, 2013

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Weight adaptation subspace SMM





Ondrej Glembek and Pavel Matejka and Lukas Burget and Tomas Mikolov ”Advances in Phonotactic Language Recognition” *Proc. Interspeech 2008.*

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Weight Adaptation: Non-negative Factor Anslys

- Main assumption of NFA

$$\omega = b + Lr$$

  - L and r consist of positive and negative values

- E-M algorithm like to update L and r.
  - Updating r : is calculated assuming that L is known
  - Updating L :is calculated assuming that r is known for all s utterances in the training dataset

H. Bahari, N. Dehak, H. Van Hamme, L. Burget, A. M. Ali and J. Glass ”Non-negative Factor Analysis of Gaussian Mixture Model Weight Adaptation for Language and Dialect Recognition” to appear in IEEE Transactions on Audio, Speech and Language Processing

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

NFA training

$$\max \gamma'(\chi)\log(b+Lr)$$

$$subject\ to$$

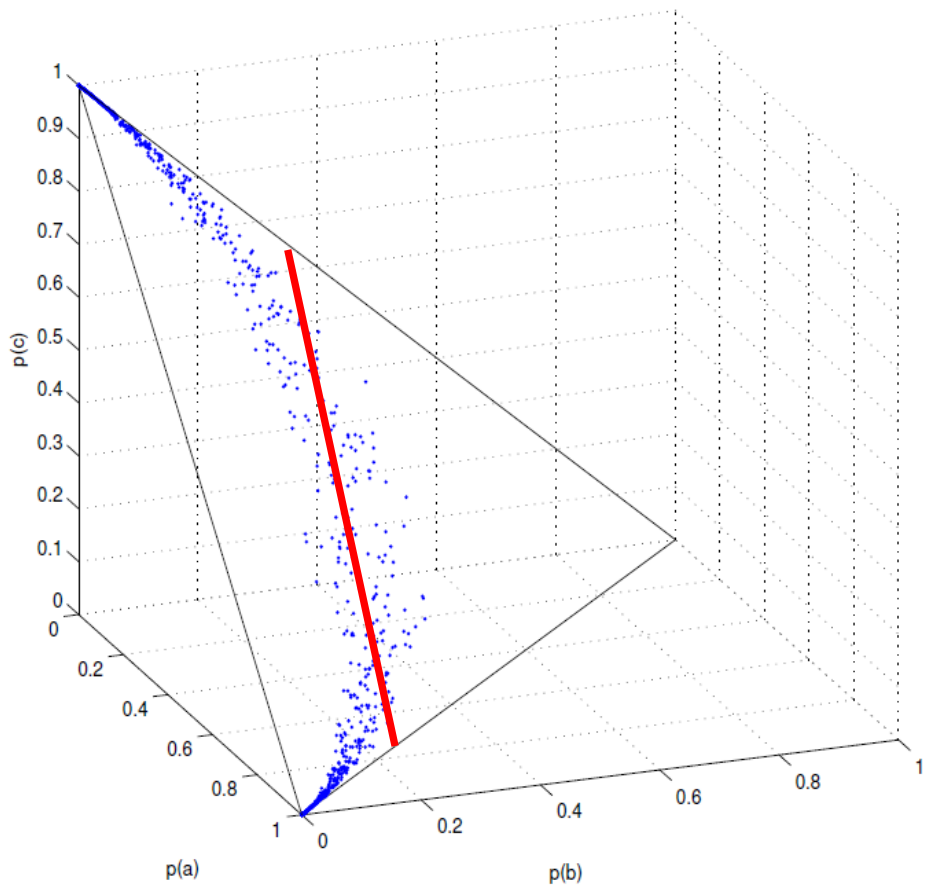
$$g(b+Lr)=1 \quad Equality\ constraint$$

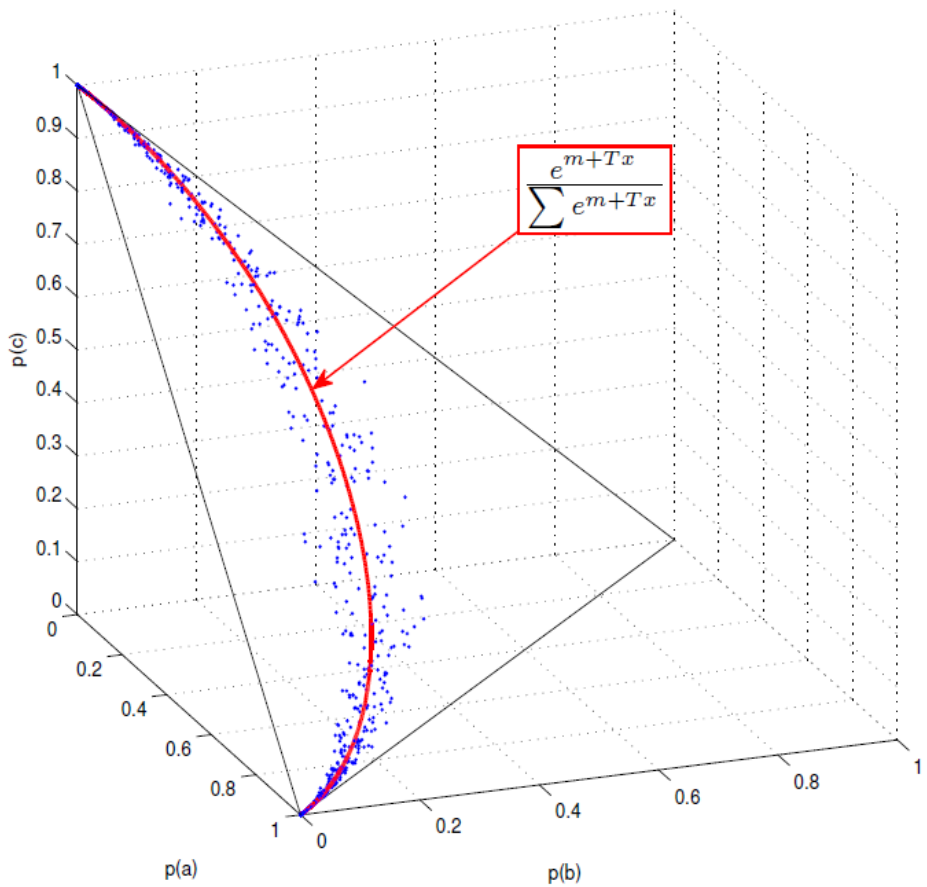
$$b+Lr>0 \quad Inequality\ constraint$$

- $\gamma(\chi)$ is the zero sufficient statistics
- g is a vector of 1

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

NFA vs SMM Weight adaptation subspace





Ondrej Glembek and Pavel Matejka and Lukas Burget and Tomas Mikolov ”Advances in Phonotactic Language Recognition” *Proc. Interspeech 2008.*

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Discrete I-vector applications

- Prosodic
  - M. Kockmann, L. Burget, O. Glembek, L. Ferrer, and J. Cernocky, “Prosodic speaker verification using subspace multinomial models with intersession compensation,” in *Eleventh Annual Conference of the International Speech Communication Association*, 2010
- Phonotactic
  - Ondrej Glembek and Pavel Matejka and Lukas Burget and Tomas Mikolov ”Advances in Phonotactic Language Recognition” *Proc. Interspeech 2008.*
  - Mehdi Soufifar,Lukas Burget, Oldrich Plchot, Sandro Cumani, Jan "Honza" Cernocky, REGULARIZED SUBSPACE MULTINOMIAL MODEL FOR PHONOTACTIC IVECTOREXTRACTION, In Proc. Interspeech 2013, Lyon, France, 2013
- GMM Weights adaptation
  - H. Bahari, N. Dehak, H. Van Hamme, L. Burget, A. M. Ali and J. Glass ”Non-negative Factor Analysis of Gaussian Mixture Model Weight Adaptation for Language and Dialect Recognition” to appear in IEEE Transactions on Audio, Speech and Language Processing

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Outline

- Introduction
- Gaussian Mixture Model for sequence modeling
  - GMM means adaptation (I-vector)
    - Speaker recognition tasks (data visualization)
    - Language recognition tasks (data visualization)
  - GMM weights adaptation
- Deep Neural Network for sequence modeling
  - DNN layer activation subspaces
  - DNN layer activation path with subspace approaches
  - Experiments and results
- Conclusions

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Modeling Sequence of Features DNN (motivation)

- Purely Unsupervised Feature Learning from Images
  - Deep Sparse Auto-encoders
  - on 10 million unlabeled YouTube thumbnails

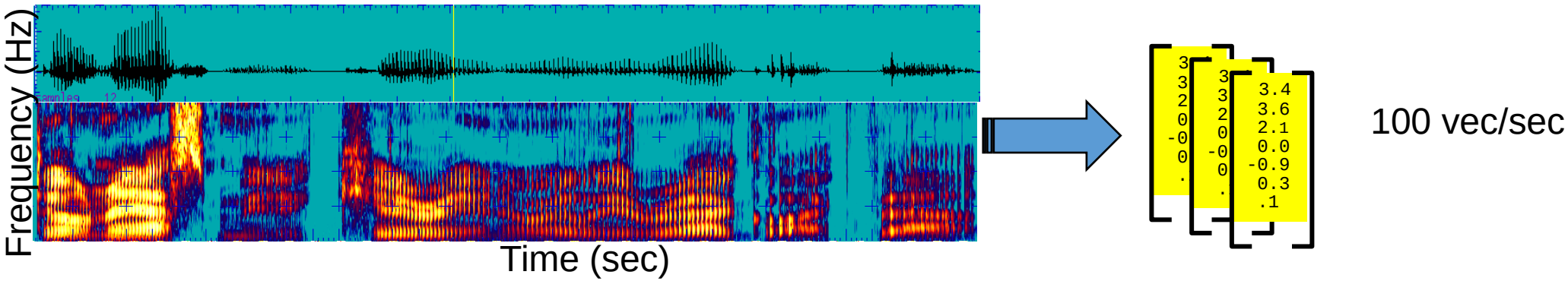
  - Optimal stimulus for cat neuron

Quoc V. Le “DistBlief - Large Scale distributed Deep Neuronal Network” Slides

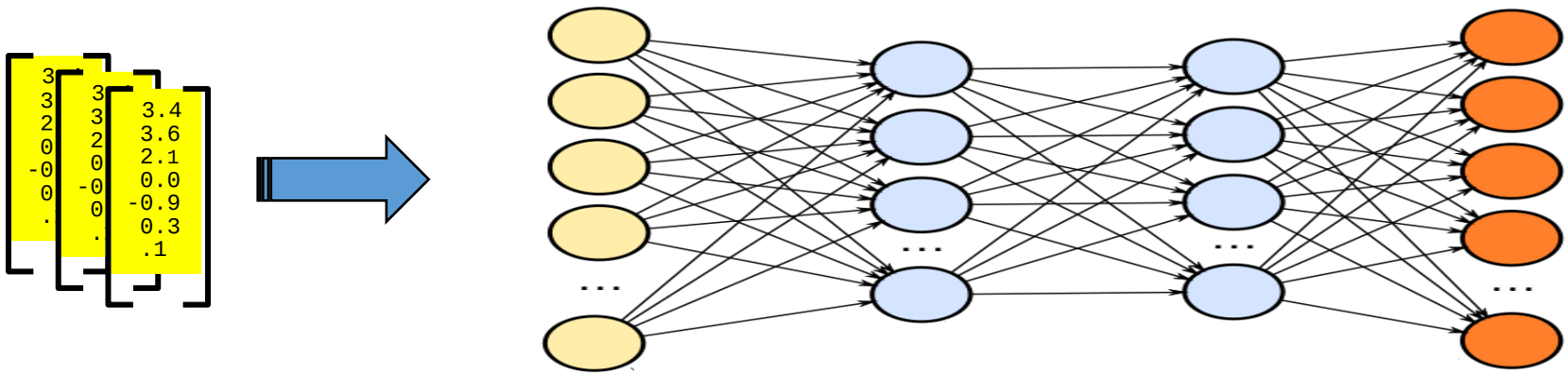
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Modeling Sequence of Features DNN

- For most recognition tasks, we need to model the distribution of feature vector sequences



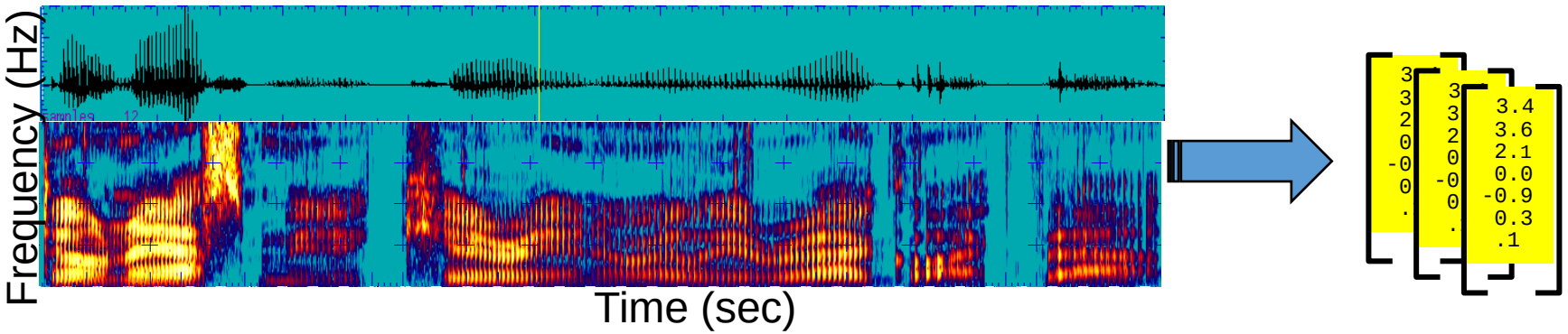
- Use these features with a DNN modeling



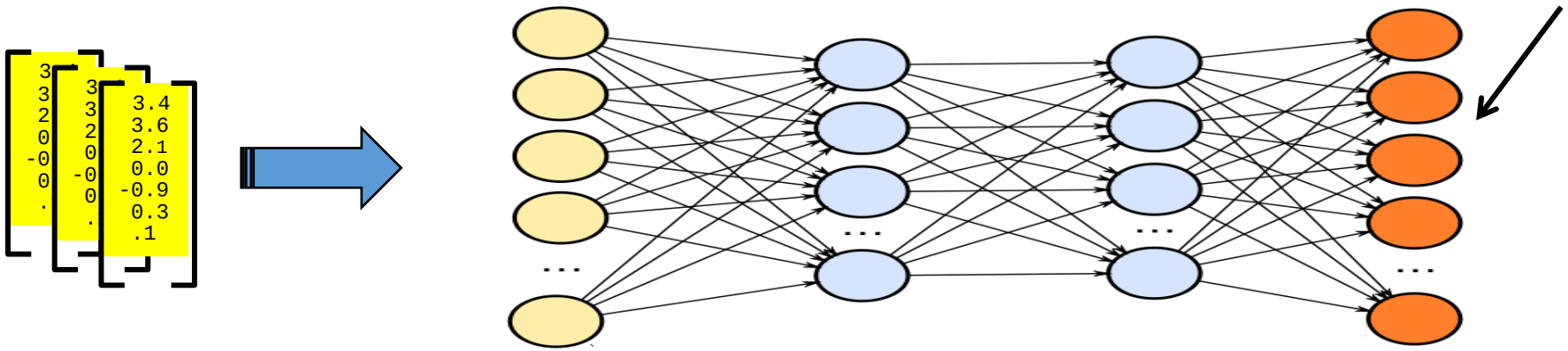
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Modeling Sequence of Features DNN

- For most recognition tasks, we need to model the distribution of feature vector sequences



- Use these features with a DNN modeling

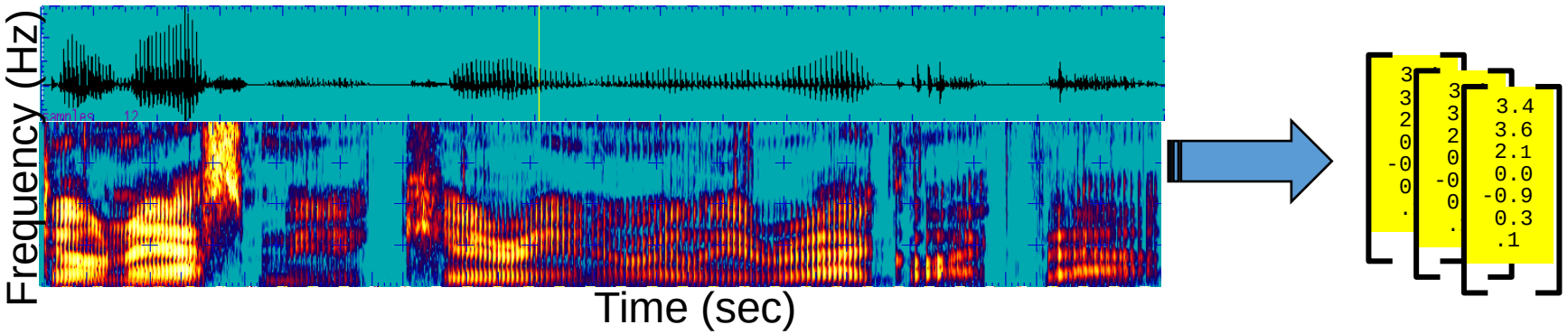


[1] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, “Automatic language iden- tification using deep neural networks,” in *Proc. ICASSP*, 2014, pp. 5374–5378.

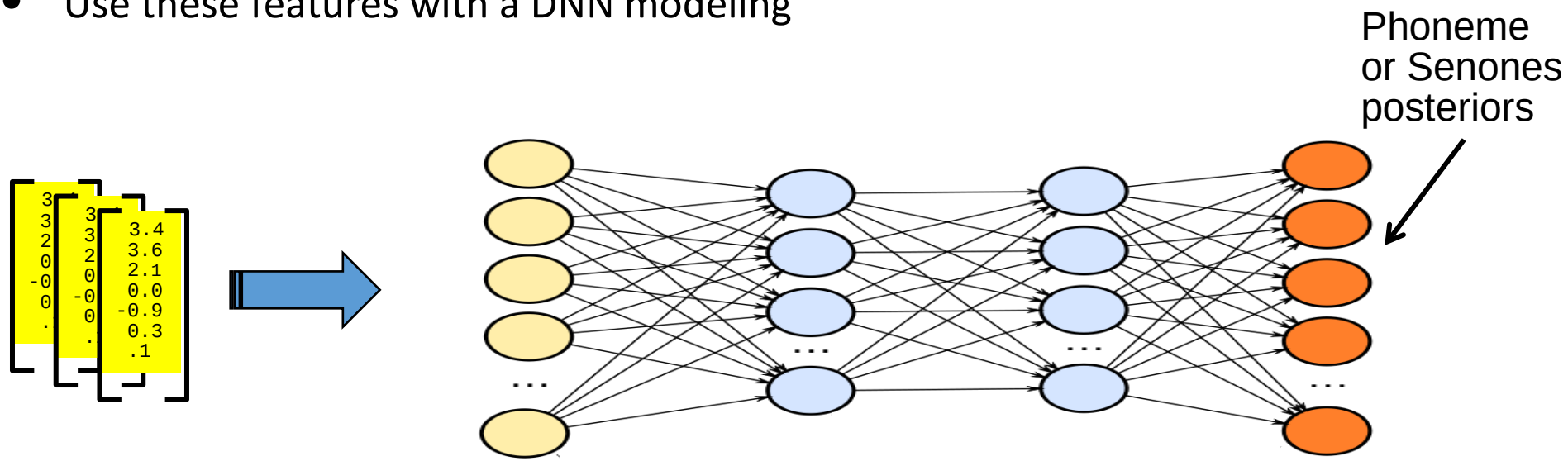
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Modeling Sequence of Features DNN

- For most recognition tasks, we need to model the distribution of feature vector sequences



- Use these features with a DNN modeling

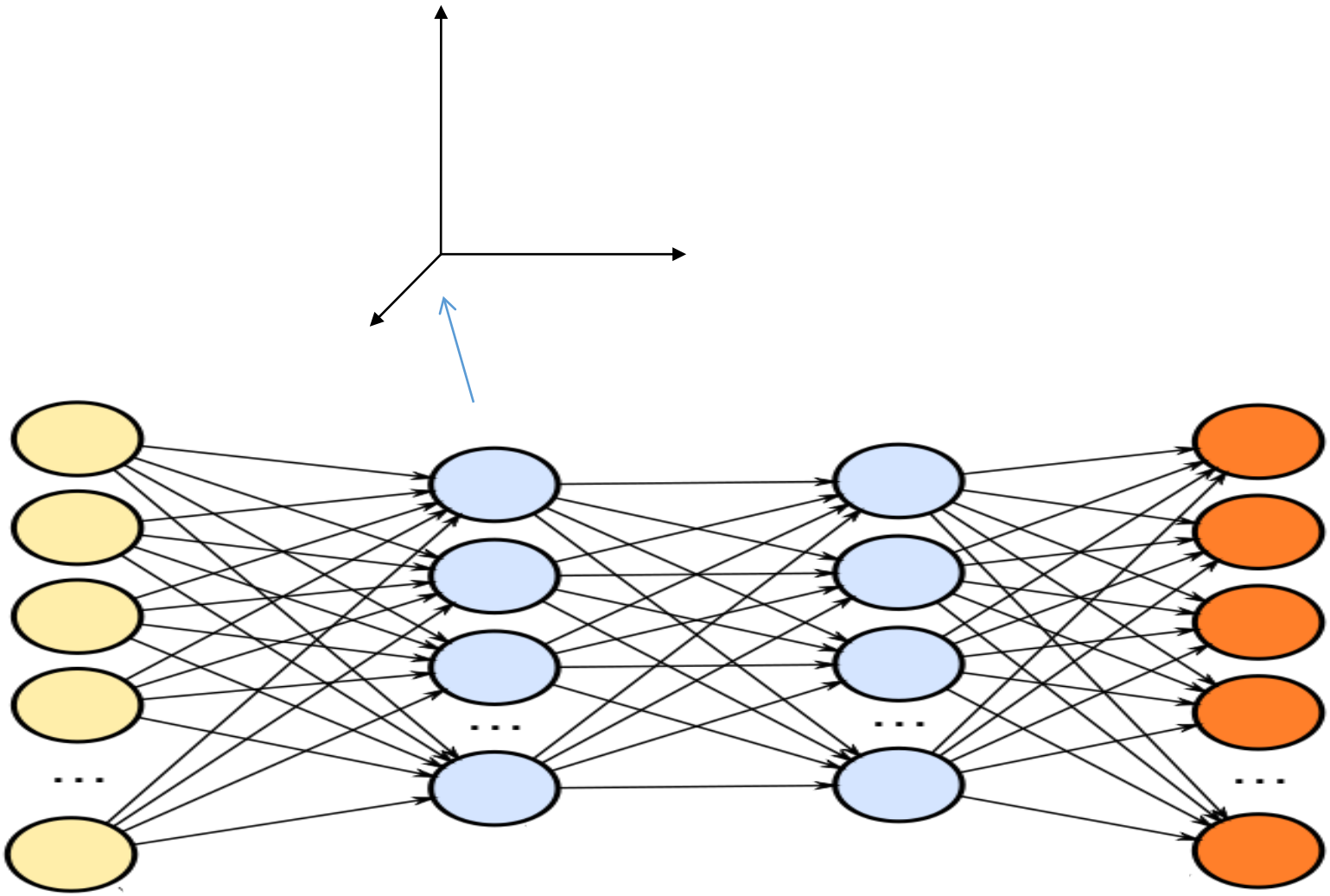


JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

The use of DNNs in general

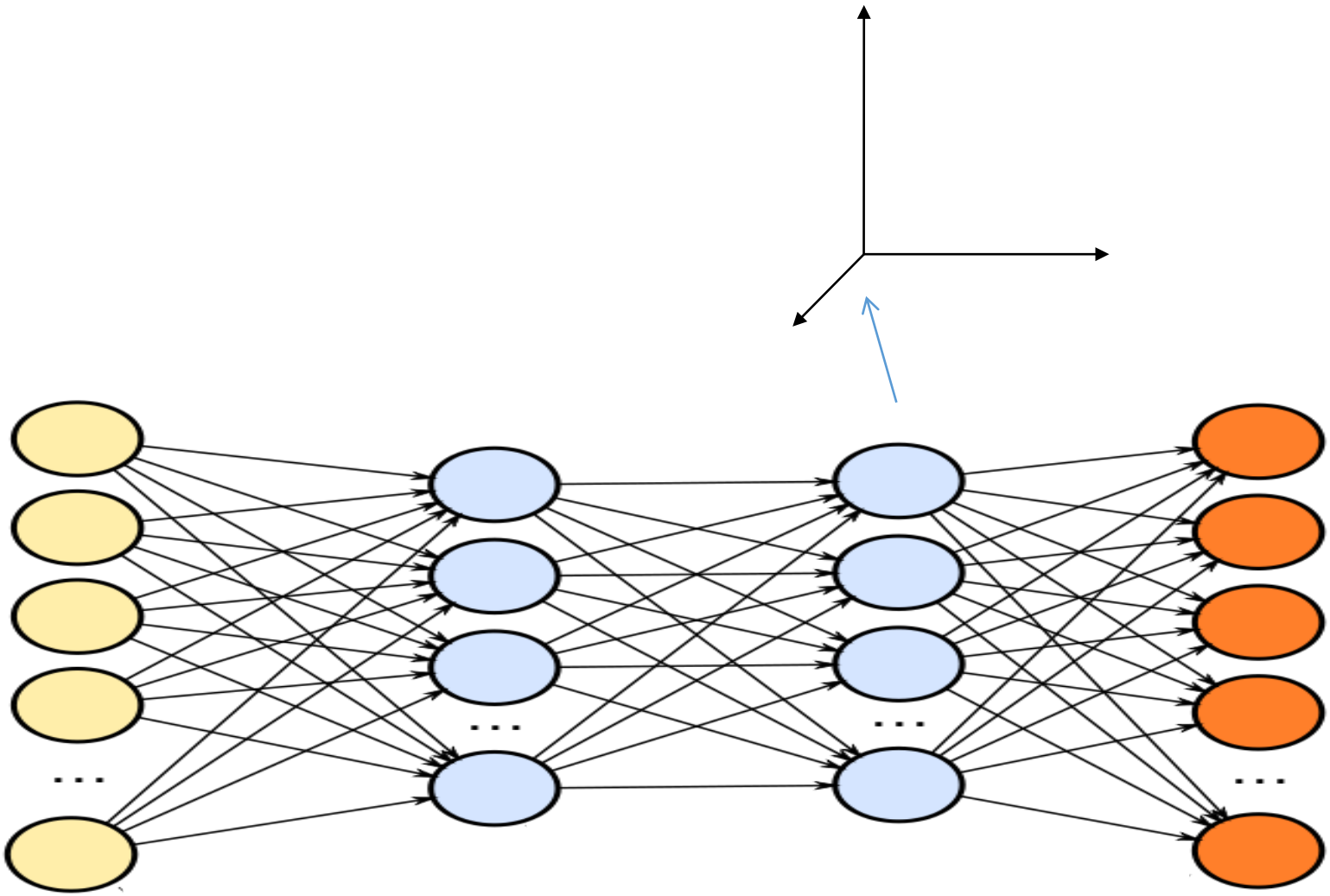
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

DNN layer activation subspaces



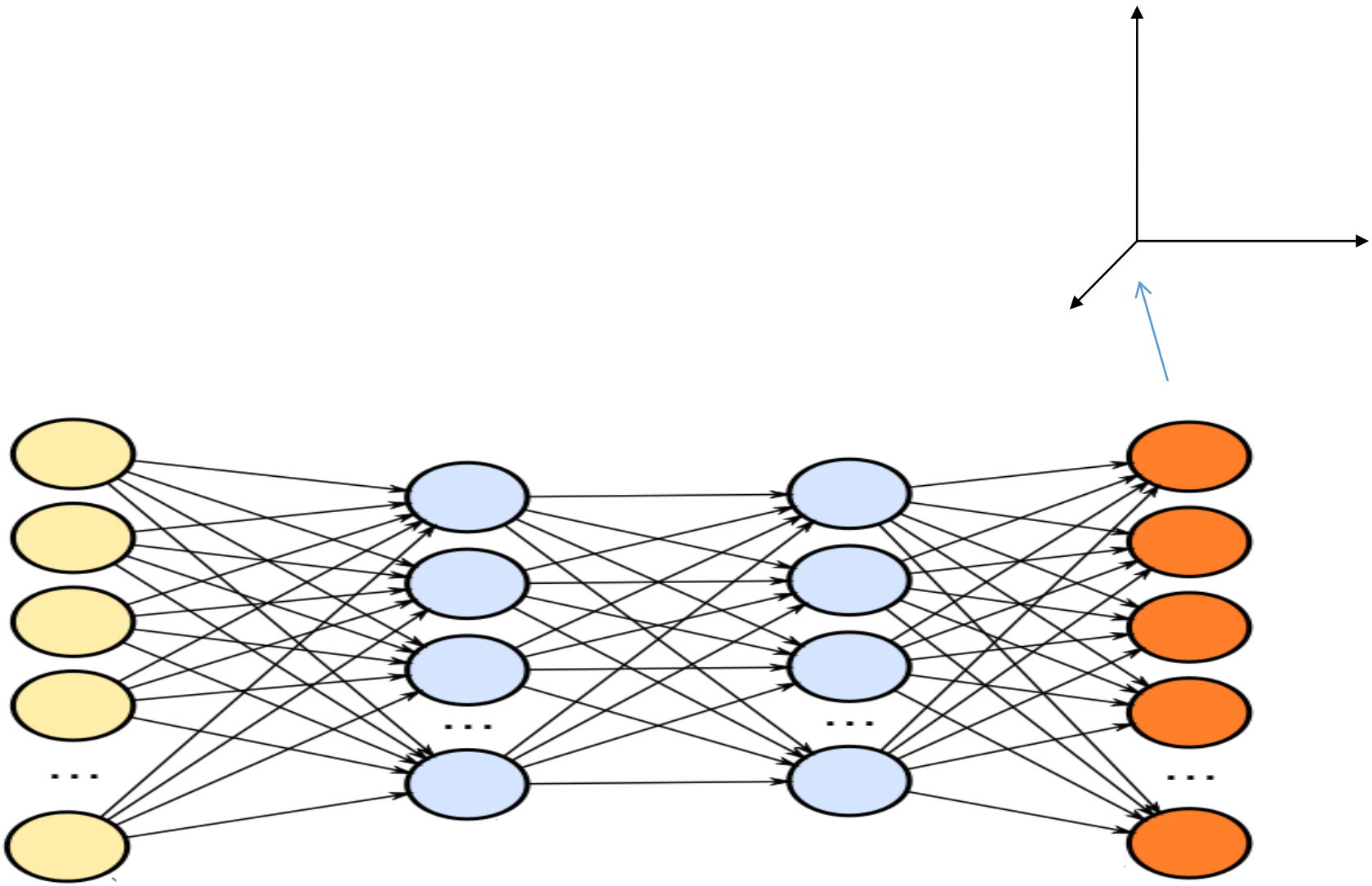
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

DNN layer activation subspaces



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

DNN layer activation subspaces

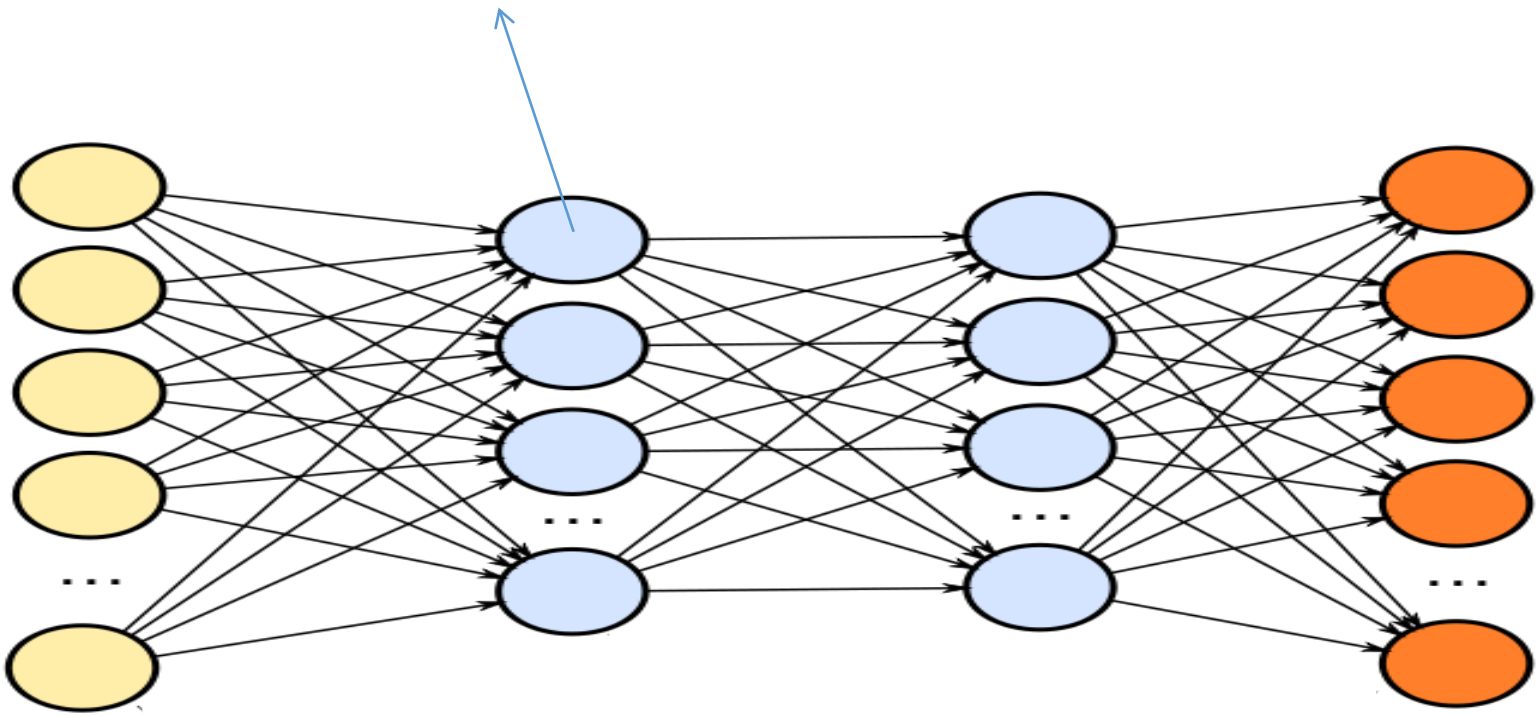


JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

DNN layer activation subspaces Counts accumulation (1/3)

- Compute the posteriors for each frame
- Normalize them to sum to one
- Accumulate over time (γ_j)

$$P_j \leftarrow \frac{P_j}{\sum_i^{i=K} P_i}$$

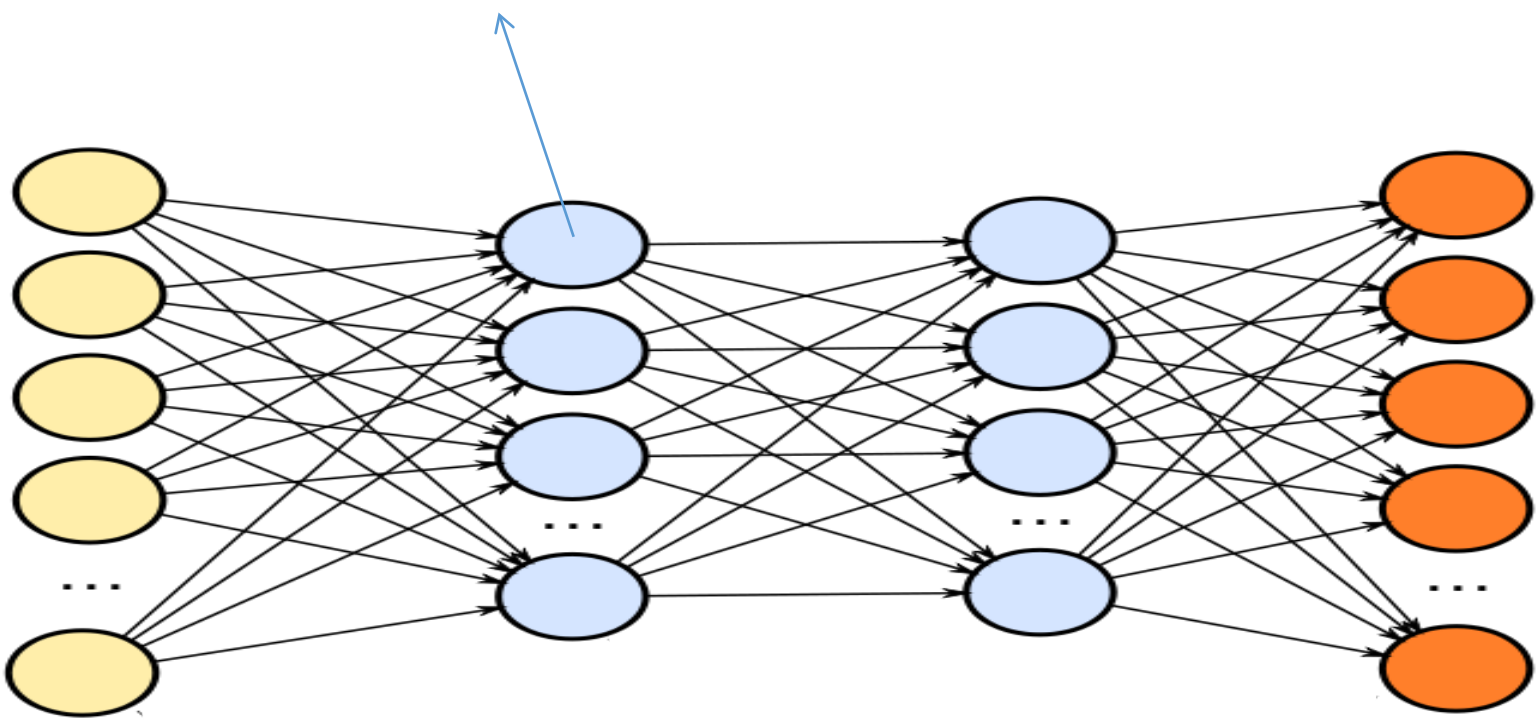


JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

DNN layer activation subspaces Counts accumulation (2/3)

- Compute the softmax for each frame
- Accumulate over time (γ_j)

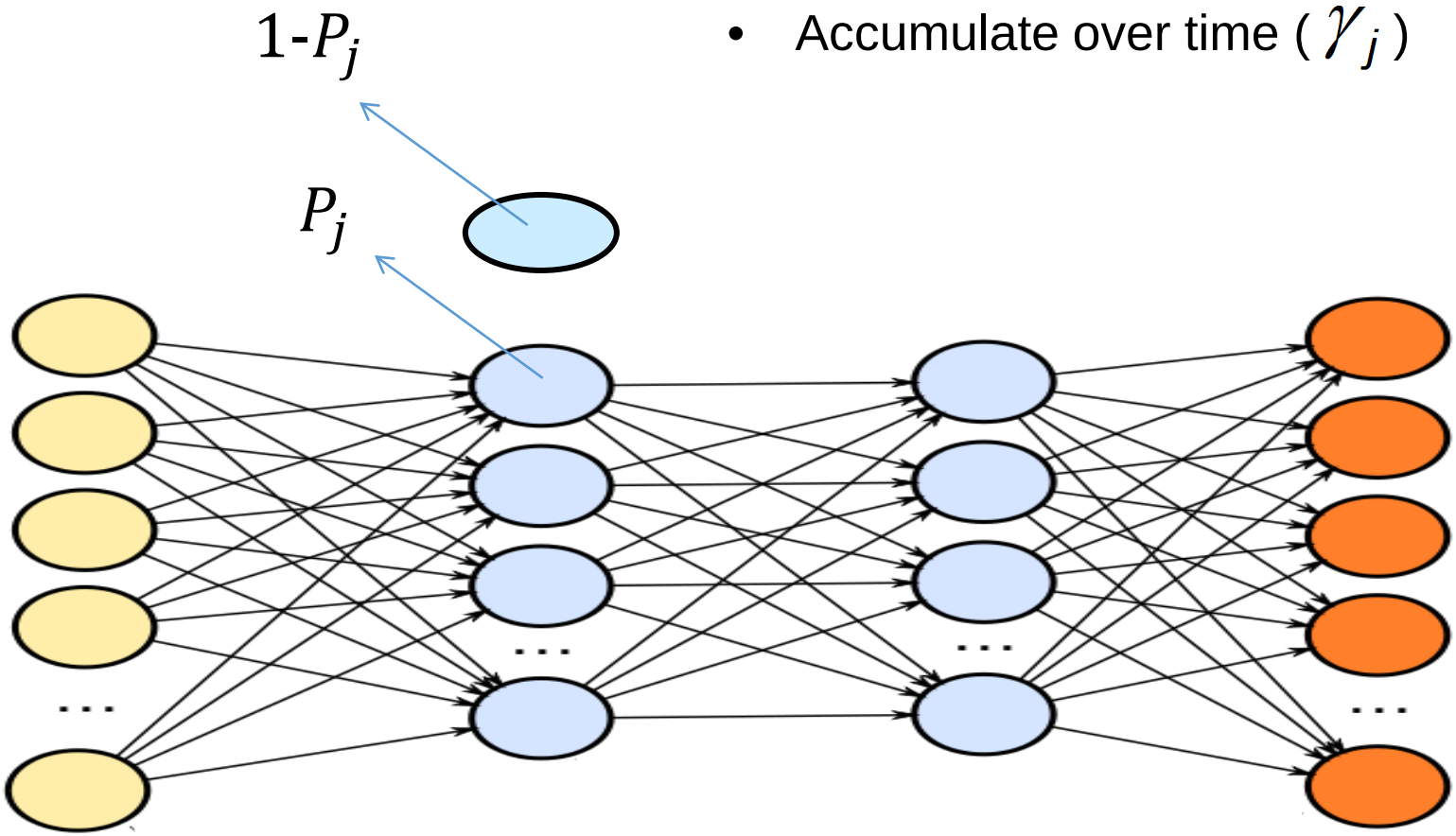
$$P_j = \frac{e^{x_j}}{\sum_i^{i=K} e^{x_i}}$$



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

DNN layer activation subspaces Counts accumulation (3/3)

- Compute the posteriors and the complement for each frame
- Accumulate over time (γ_j)



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Weight Adaptation: Non-negative Factor Analysis

- Main assumption of NFA

$$\omega = b + Lr$$

  - L and r consist of positive and negative values

- E-M algorithm like to update L and r
  - Update r is calculated assuming that L is known
  - Update L is calculated assuming that r is known for all s utterances in the training dataset

H. Bahari, N. Dehak, H. Van Hamme, L. Burget, A. M. Ali and J. Glass ”Non-negative Factor Analysis of Gaussian Mixture Model Weight Adaptation for Language and Dialect Recognition” IEEE Transactions on Audio, Speech and Language Processing 2014

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

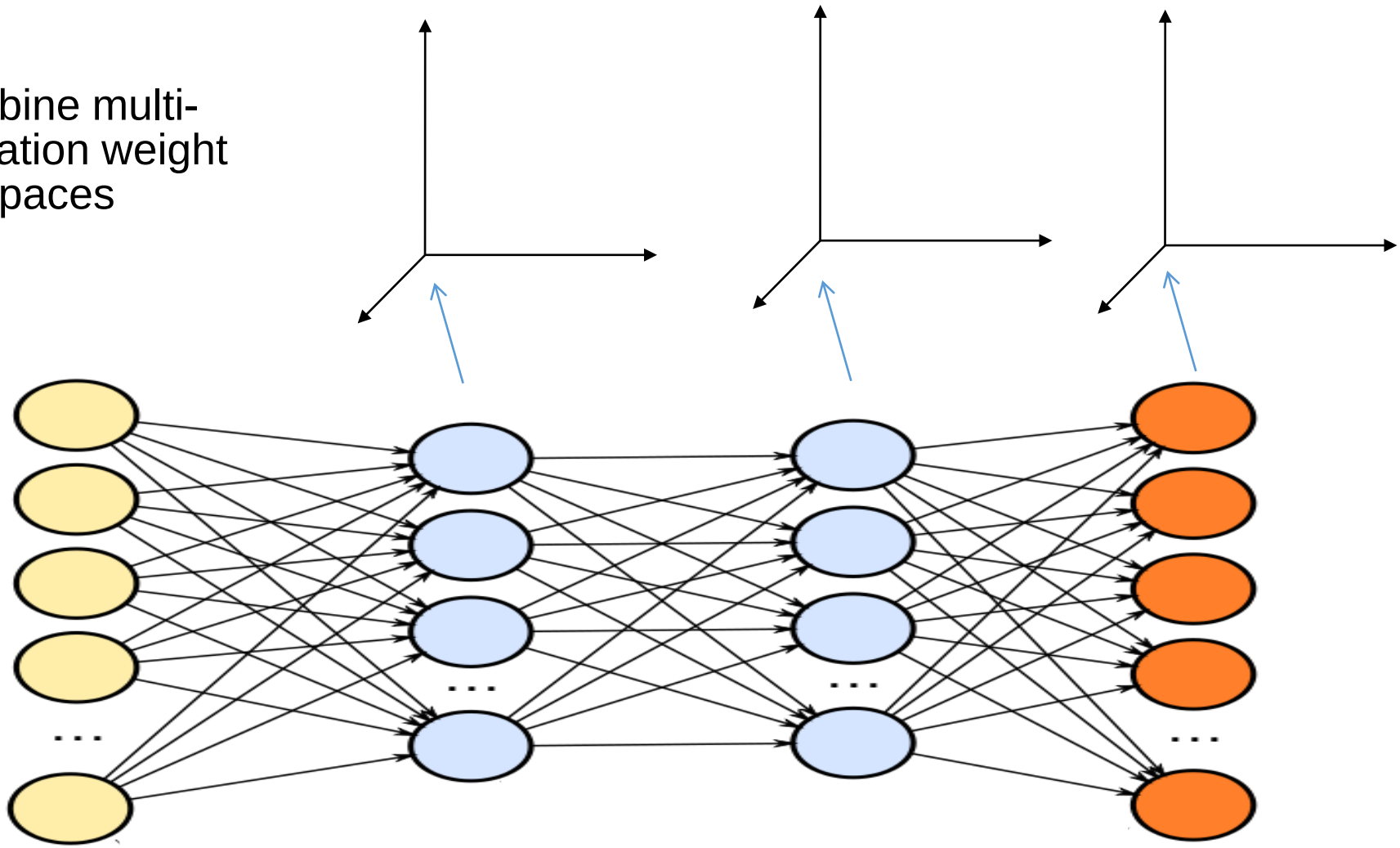
Outline

- Introduction
- Gaussian Mixture Model for sequence modeling
  - GMM means adaptation (I-vector)
    - Speaker recognition tasks (data visualization)
    - Language recognition tasks (data visualization)
  - GMM weights adaptation
- Deep Neural Network for sequence modeling
  - DNN layer activation subspaces
  - DNN layer activation path with subspace approaches
  - Experiments and results
- Conclusions

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

DNN layer activation subspaces modeling the activation paths

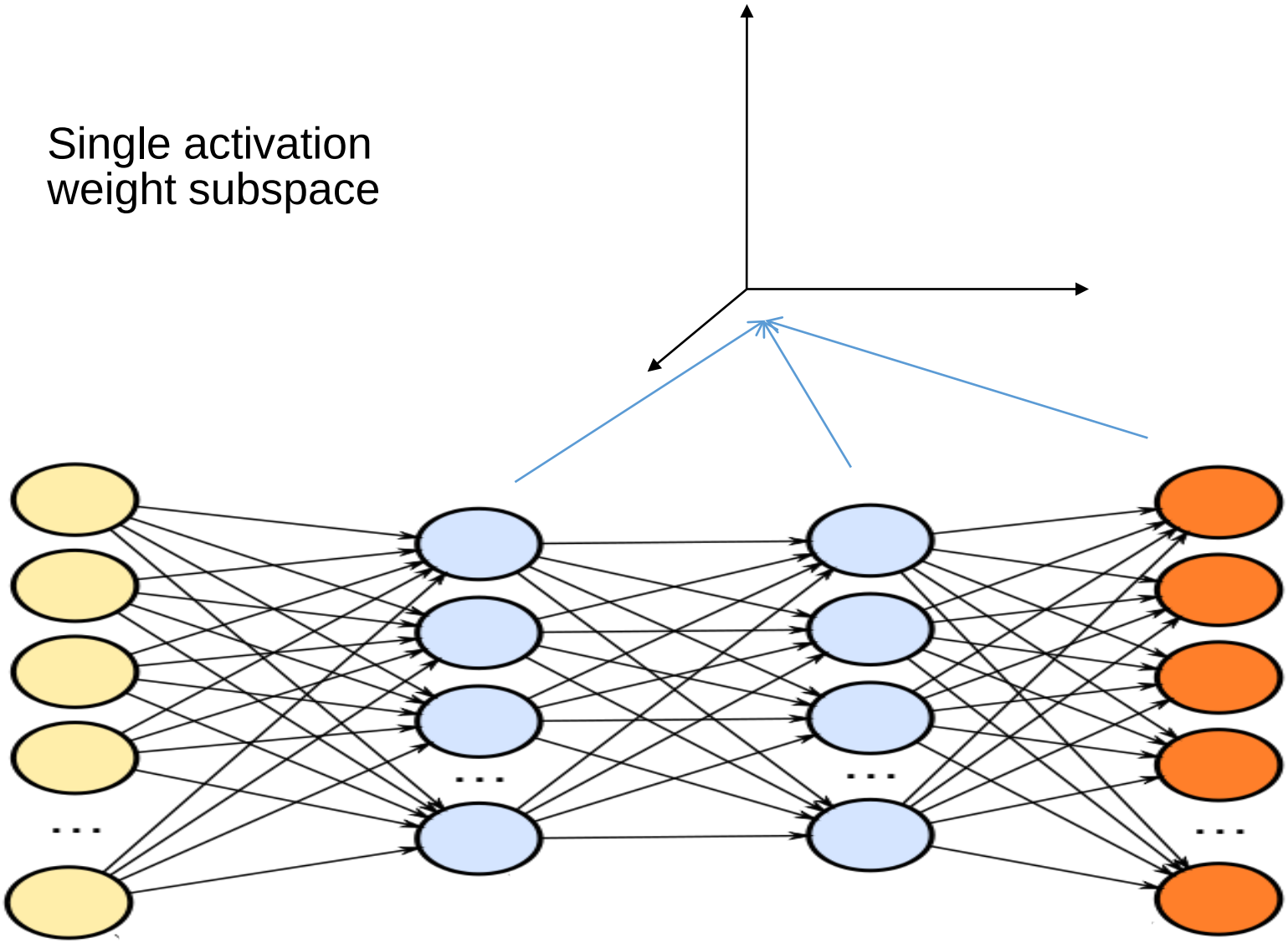
Combine multi-activation weight subspaces



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

DNN layers activation subspace modeling the activation paths

Single activation
weight subspace



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Multi-Weight Adaptation: NFA

- Main assumption of NFA

$$\omega_l = b_l + Lr$$

  - L and r is common for all the layers with the constraint that each layer posterior should sum to one
- E-M algorithm like to update L and r
  - Updating r : is calculated assuming that L is known
  - Updating L : is calculated assuming that r is known for all s utterances in the training dataset (with the constraint that each layer posterior should some to one)

An Extension of Hasan's works : H. Bahari, **N. Dehak**, H. Van Hamme, L. Burget, A. M. Ali and J. Glass "Non-negative Factor Analysis of Gaussian Mixture Model Weight Adaptation for Language and Dialect Recognition" to appear in IEEE Transactions on Audio, Speech and Language Processing

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Outline

- Introduction
- Gaussian Mixture Model for sequence modeling
  - GMM means adaptation (I-vector)
    - Speaker recognition tasks (data visualization)
    - Language recognition tasks (data visualization)
  - GMM weights adaptation
- Deep Neural Network for sequence modeling
  - DNN layer activation subspaces
  - DNN layer activation path with subspace approaches
  - Experiments and results
- Conclusions

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

QCRI dataset for Arabic Dialect identification

- Broadcast news and show recordings

Dialects	Train	Development	Evaluation
MSA	1480	254	207
Egyptian	1116	463	139
Gulf	1181	221	218
Levantine	1074	186	132
Maghbrebi	1506	234	193

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

DNN subspace results

- Five Dialect classes
- DNN with five hidden layers (1024,512,512,512,512)
  - Speech frames as input and Dialect classes as output
  - Cosine scoring with LDA
  - The best subspace dimension (1000,500,500,500,500)
- Performance Metric: Error Rate

System	Dev	Eval
i-vector	19.97%	21.66%
DNN output (scores average)	16.50%	16.48%

[1] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, “Automatic language iden- tification using deep neural networks,” in *Proc. ICASSP*, 2014, pp. 5374–5378.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

DNN subspace results

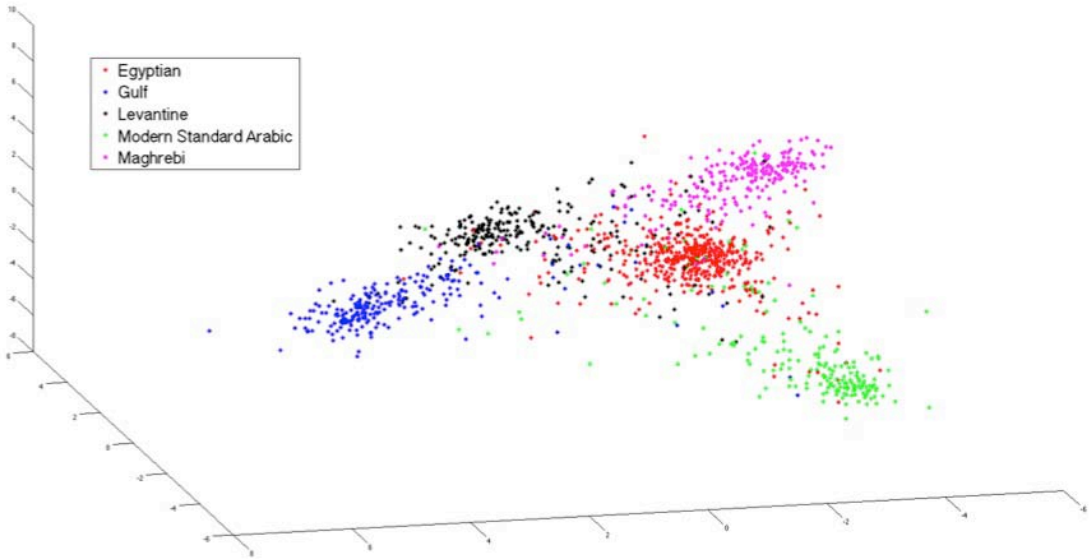
- Five Dialect classes
- DNN with five hidden layers (1024,512,512,512,512)
  - Speech frames as input and Dialect classes as output
  - The best subspace dimension (1000,500,500,500,500)
  - Cosine scoring with LDA

- Performance Metric: Error Rate

System	Dev	Eval
i-vector	19.97%	21.66%
DNN output (scores average)	16.50%	16.48%
Hidden layer 1	14.40%	14.80%
Hidden layer 2	13.14%	13.97%
Hidden layer 3	12.80%	13.40%
Hidden layer 4	12.48%	12.26%
Hidden layer 5	12.79%	12.58%

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

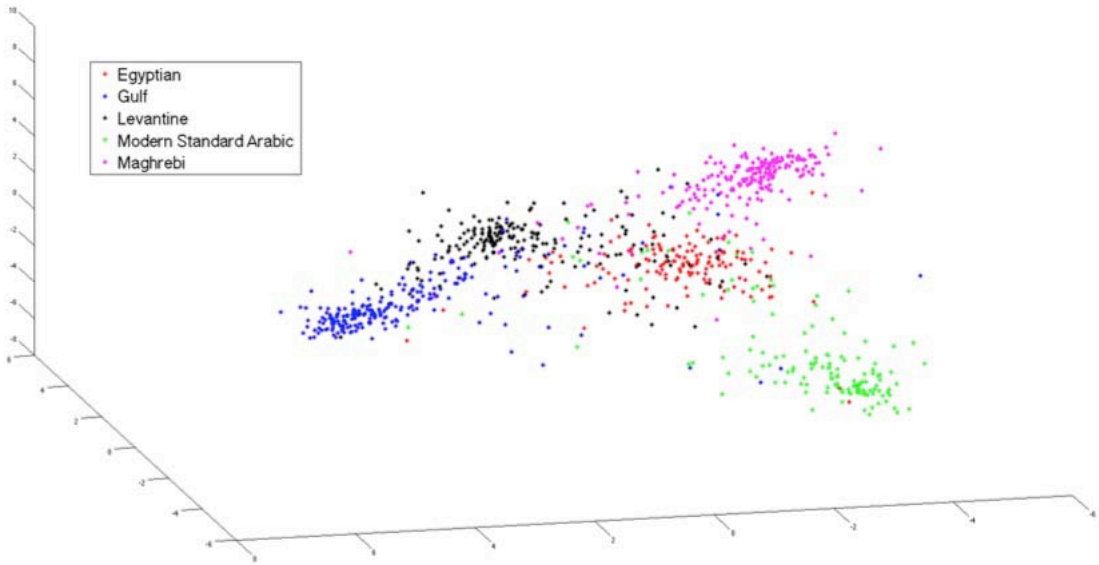
Data visualization on DNN subspace



Layer 4

Development data

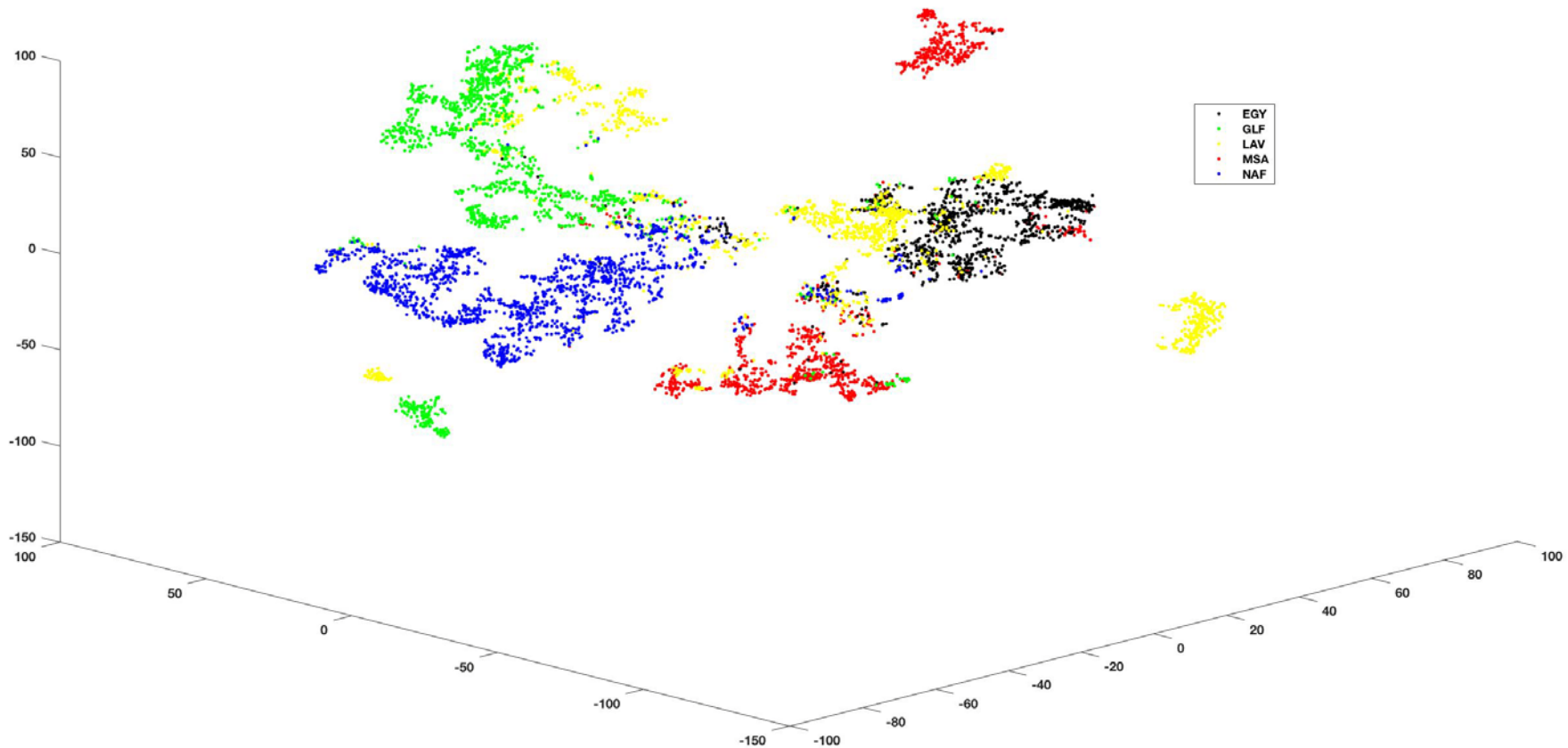
Evaluation data



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Data visualization on DNN subspace

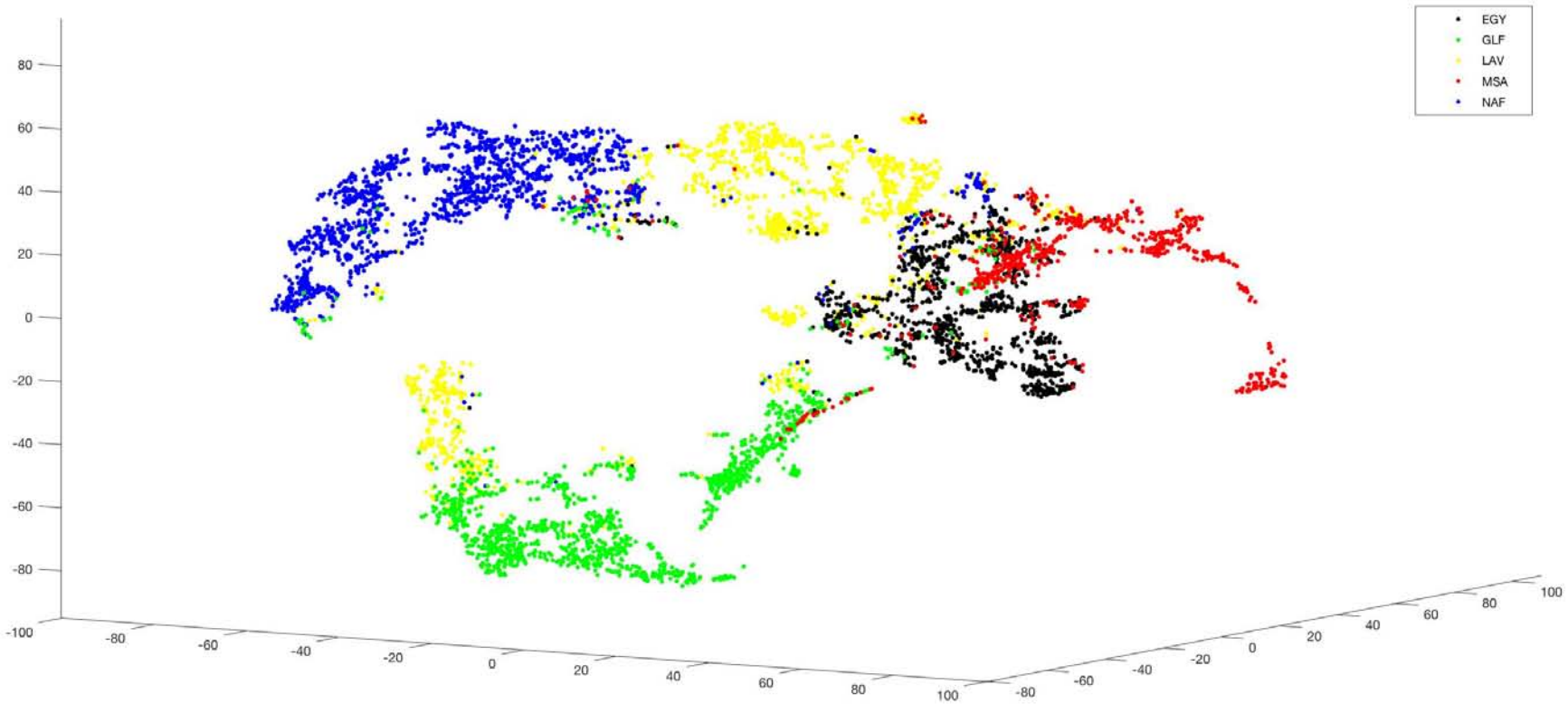
- Layer 5 (raw I-vectors)
- t-sne software



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Data visualization on DNN subspace

- Layer 5 (length normalized I-vectors)
- t-sne software



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

DNN subspace results

- Combining the layers in one subspace

System	Dev	Eval
i-vector	19.97%	21.66%
DNN	16.50%	16.48%
Last hidden DNN layer - NFA	12.79%	12.58%

CSAIL

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

DNN subspace results

- Combining the layers in one subspace

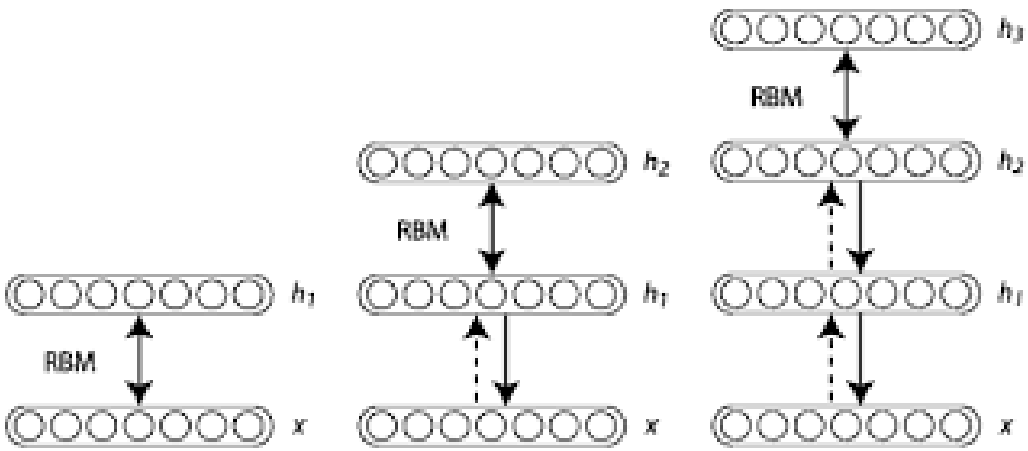
System	Dev	Eval
i-vector	19.97%	21.66%
DNN	16.50%	16.48%
Last hidden DNN layer - NFA	12.79%	12.58%
All hidden DNN layers - NFA	10.10%	10.48%

CSAIL

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Stacked RBMs auto-encoders

- Five Dialect classes
- Stacked RBMs auto-encoders
- RBMs with five hidden layers (1024,512,512,512,512)
  - Speech frames as input
  - The best subspace dimension (1000,500,500,500,500)
  - Cosine scoring with LDA



CSAIL

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

DNN subspace results

- Five Dialect classes
- RBMs with five hidden layers (1024,512,512,512,512)
  - Speech frames as input
  - The best subspace dimension (1000,500,500,500,500)
- Performance Metric: Error Rate

System	Dev	Eval
i-vector	19.97%	21.66%
DNN output (scores average)	16.50%	16.48%
Hidden layer 1	21.46%	19.55%
Hidden layer 2	23.67%	24.19%
Hidden layer 3	20.67%	21.49%
Hidden layer 4	23.18%	22.77%
Hidden layer 5	24.14%	23.26%

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Results on subset of NIST 2009 LRE

- Similar to Google works which selected 6 languages
  - We use both collects CTS and VOA (different to [1])

		TRAIN(hrs)	TEST(cuts)
farsi	cts	132.56	46
farsi	voa	22.62	338
hindi	cts	141.75	240
hindi	voa	9.25	397
korean	cts	130.80	133
korean	voa	9.85	318
mandarin	cts	444.48	580
mandarin	voa	11.63	390
russian	cts	71.17	229
russian	voa	20.68	254
vietnamese	cts	115.41	251
vietnamese	voa	13.77	27

[1] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, “Automatic language identification using deep neural networks,” in *Proc. ICASSP*, 2014, pp. 5374–5378.

CSAIL

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Results on subset of NIST 2009 LRE

- 6 language classes
- Subspace Multinomial Model
- DNN with five hidden layers (1024, 1024, 1024, 1024, 1024)
  - Speech frames as inputs
  - Language Id classes as outputs
  - Linear Gaussian Classifier
- Performance Metric: Cavg

System	30s	10s	3s
i-vector	0.0068	0.0287	0.1140
DNN output (scores average)	0.0231	0.0344	0.0930

[1] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," in *Proc. ICASSP*, 2014, pp. 5374–5378.

CSAIL

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Results on subset of NIST 2009 LRE

- 6 language classes
- Subspace Multinomial Model
- DNN with five hidden layers (1024, 1024, 1024, 1024, 1024)
  - Speech frames as inputs
  - Language Id classes as outputs
  - Linear Gaussian Classifier
- Performance Metric: Cavg

System	30s	10s	3s
i-vector	0.0068	0.0287	0.1140
DNN output (scores average)	0.0231	0.0344	0.0930
Hidden layer 1	0.0411	0.0971	0.2415
Hidden layer 2	0.0257	0.0672	0.2161
Hidden layer 3	0.0146	0.0423	0.1650
Hidden layer 4	0.0103	0.0336	0.1534
Hidden layer 5	0.0046	0.0162	0.0837

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Results on subset of NIST 2009 LRE

- 6 language classes
- Subspace Multinomial Model
- DNN with five hidden layers (1024, 1024, 1024, 1024, 1024)
  - Speech frames as inputs
  - Language Id classes as outputs
  - Linear Gaussian Classifier
- Performance Metric: Cavg

System	30s	10s	3s
i-vector	0.0068	0.0287	0.1140
DNN output (scores average)	0.0231	0.0344	0.0930
Hidden layer 5	0.0046	0.0162	0.0837

CSAIL

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Results on subset of NIST 2009 LRE

- 6 language classes
- Subspace Multinomial Model
- DNN with five hidden layers (1024, 1024, 1024, 1024, 1024)
  - Speech frames as inputs
  - Language Id classes as outputs
  - Linear Gaussian Classifier
- Performance Metric: Cavg

System	30s	10s	3s
i-vector	0.0068	0.0287	0.1140
DNN output (scores average)	0.0231	0.0344	0.0930
Hidden layer 5	0.0046	0.0162	0.0837
All Hidden layers	0.0050	0.0154	0.0742

CSAIL

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Outline

- Introduction
- Gaussian Mixture Model for sequence modeling
  - GMM means adaptation (I-vector)
    - Speaker recognition tasks (data visualization)
    - Language recognition tasks (data visualization)
  - GMM weights adaptation
- Deep Neural Network for sequence modeling
  - DNN layer activation subspaces
  - DNN layer activation path with subspace approaches
  - Experiments and results
- Conclusions

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Conclusions

- I-vector is an elegant way to represent a speech utterance
- GMM weights adaptation approaches can be used to model the DNN layer activations

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Conclusions

- I-vector is an elegant way to represent a speech utterance
- GMM weights adaptation approaches can be used to model the DNN layer activations
- Take Home:
  - The majority of the DNN information is contained in the hidden layers (not the output layer)
  - Looking to the DNN information in each single layer separately may be not the best way to exploit all information modeled in the DNN
  - Subspace approaches seem to be a great tool to aggregate frames DNN propagation over time
    - Even if the DNN is frame based trained

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

future works and directions

- Modeling long term speech information
  - TDNNs and RNNs (LSTM) instead of DNNs
- Explore more DNN architectures that will be more suitable for SID
- Explore better Auto-encoder DNNs
  - Variational Bayes Auto-encoder (Joint estimation of all hidden layer instead of greedy learning of RBMs)
  - Modeling speech (speech Coding)
- Explore sparse activations (dropout)
- Alternate activation functions

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Questions

- SLT 2016 at Puerto Rico

Welcome

IEEE Spoken Language Technology (SLT) Workshop series try to promote the progress made on the different field of the spoken language technologies and their impact in the human life. The SLT 2016 will be held at San Juan in Puerto Rico (US territory) on December 13-16. San Juan is an island in the Atlantic Ocean and presents an amazing location in terms of access and tourism.

The theme for SLT 2016 will be “Machine learning from signal to concepts”. The SLT 2016 workshop will be partnering with National Institute of Standards Technology workshops on speaker and language recognition that will happen just before the SLT 2016 workshop.

See you all there