Unsupervised Clustering Approaches for Domain Adaptation in Speaker Recognition Systems

Stephen H. Shum

CSAIL
MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

Douglas A. Reynolds

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Daniel Garcia-Romero

Alan McCree

human language technology
center of excellence

Unsupervised Clustering Approaches for Domain Adaptation in Speaker Recognition Systems

Domain Adaptation & Transfer Learning

- **Most current statistical learning techniques assume (incorrectly) that the training and test data come from the same underlying distribution.**
- **Labeled data may exist in one domain, but we want a model that can also perform well on a related, but not identical, domain.**
- **Hand-labeling data in a new domain is difficult and expensive.**
- **What can we do to leverage the original, labeled, "out-of-domain" data when building a model to work on new, unlabeled, "in-domain" data?**

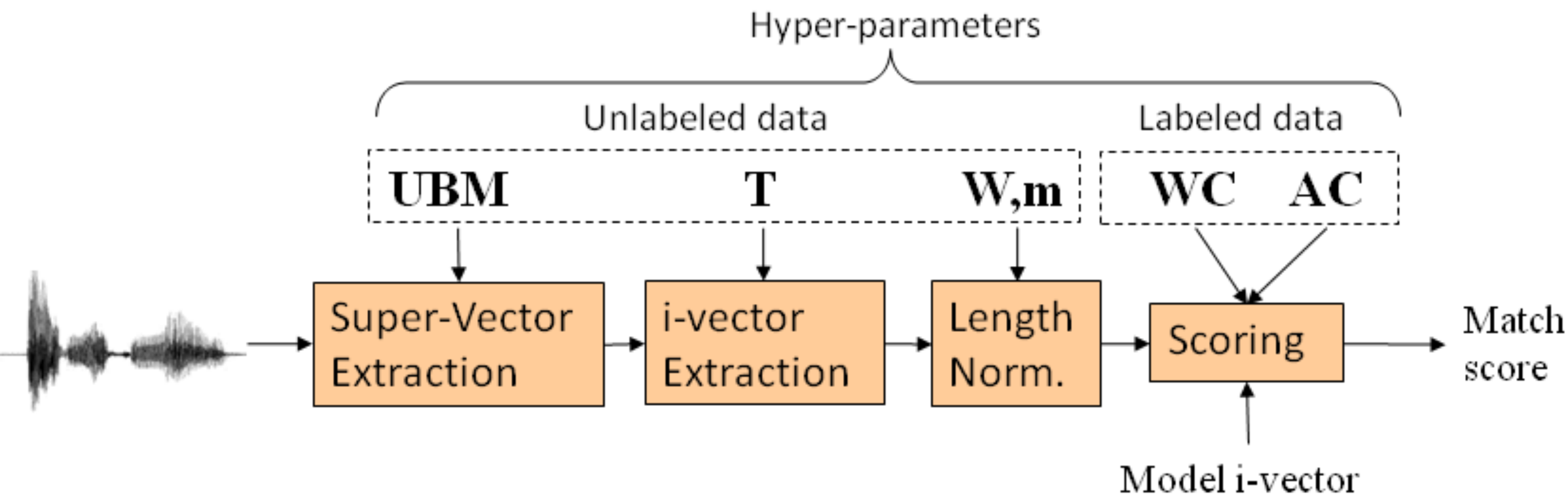
[2] Hal Daume III and Daniel Marcu, "Domain adaptation for statistical classifiers," Journal of Artificial Intelligence Research, 2006.

Unsupervised Clustering Approaches for <u>Domain Adaptation</u> in Speaker Recognition Systems

The i-vector approach

- Segment-length independent, low-dimensional, vector-based summary representation of audio
- Allows the use of large amounts of previously collected and labeled audio to characterize and exploit speaker and channel (i.e., all non-speaker) variabilities.
  – 1000's of speakers making 10's of calls
- Unrealistic to expect that most applications will have access to such a large set of labeled data from matched conditions.

Data usage (labeled & unlabeled) in an i-vector system



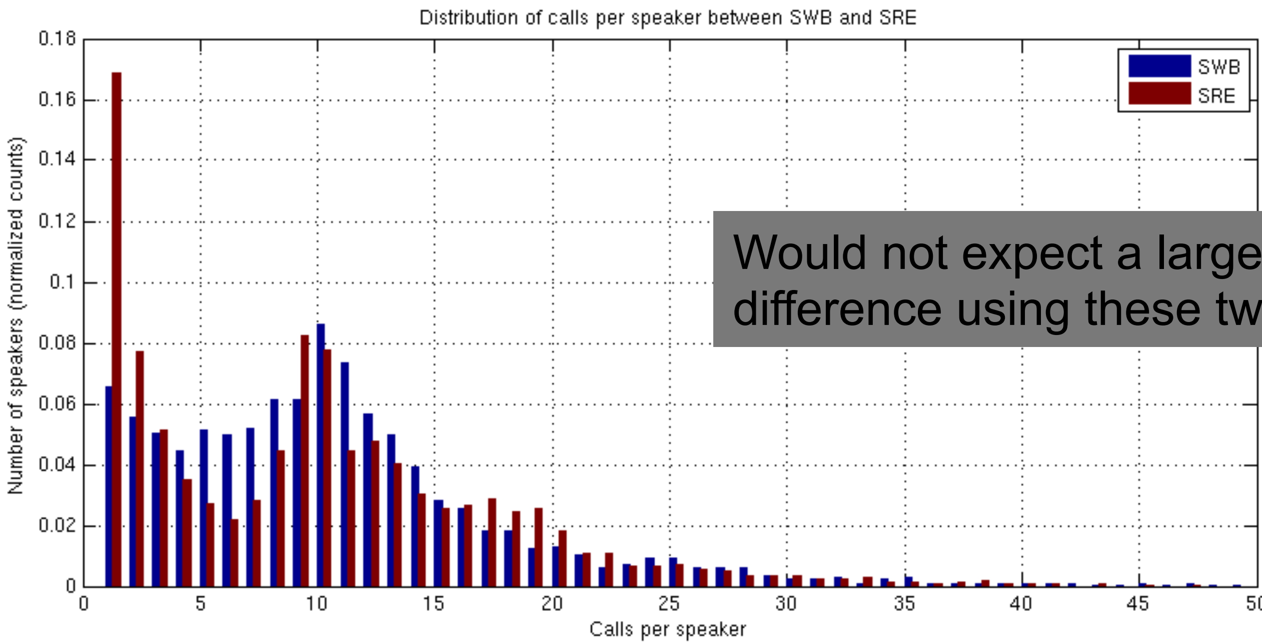
Demonstrating Mismatch

- **Enroll and score**
  - SRE10 telephone speech
- **Matched, "in-domain" SRE data**
  - All telephone calls from all speakers from SRE 04, 05, 06, and 08 collections
- **Mismatched "out-of-domain" SWB data**
  - All calls from all speakers from Switchboard-I and Switchboard-II collections

Demonstrating Mismatch

- **Summary statistics for SRE & SWB lists**

Hyper list	# Spkrs	# Males	# Females	# Calls	Avg # calls/spkr	Avg # phone_num/spkr
SWB	3114	1461	1653	33039	10.6	3.8
SRE	3790	1115	2675	36470	9.6	2.8



Would not expect a large performance difference using these two sets of data.

Demonstrating Mismatch

- **Baseline / Benchmark Results (Equal Error Rate – EER)**

UBM & T	Whitening	WC & AC	JHU	MIT
SWB	SWB	SWB	6.92%	7.57%
SWB	SRE	SWB	5.54%	5.52%
SWB	SRE	SRE	2.30%	2.09%
SRE	SRE	SRE	2.43%	2.48%

- **Focus on the performance gap caused by using SRE instead of SWB labels (SWB/SRE) for WC & AC**
  - **Continue using SWB for UBM&T and SRE for Whitening**

Challenge Task Rules

- **Allowed to use SWB data *and* their labels**
- **Allowed to use SRE data but <u>not</u> their labels**
- **Evaluate on SRE10.**

Exploring the Domain Mismatch

- ~~Speaker ages?~~
- ~~Languages spoken?~~
  - ~~SWB contains only English~~
  - ~~SRE contains 20+ different languages~~

[11] Carlos Vaquero, "Dataset Shift in PLDA-based Speaker Verification," in *Proceedings of Odyssey*, 2012.

Exploring the Domain Mismatch

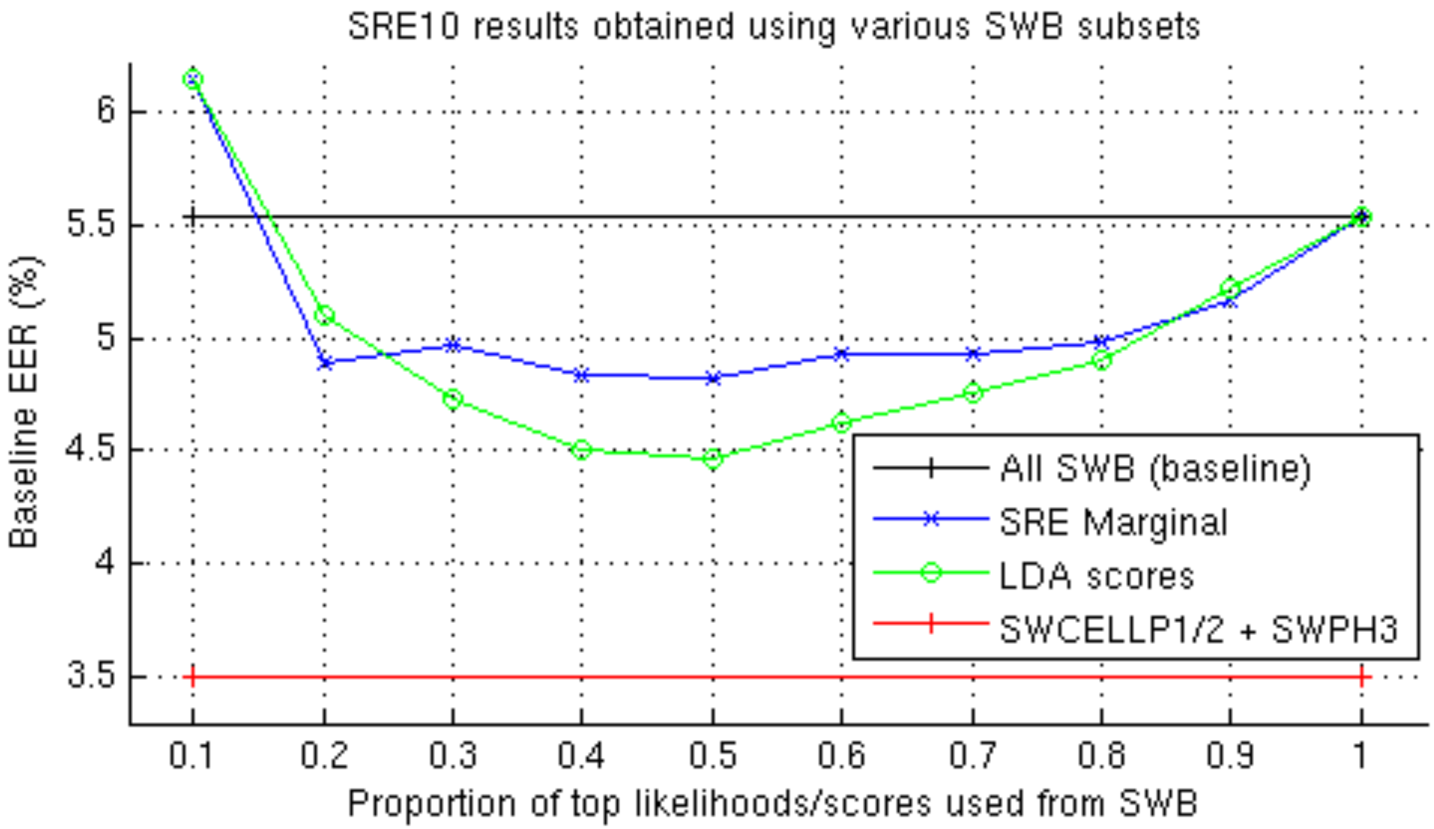
- **SWB subsets**
  - **SWPH0 (1992)**
  - **SWPH1 (1996)**
  - **SWPH2 (1997)**
  - **SWPH3 (1997-1998)**
  - **SWCELLP1 (1999)**
  - **SWCELLP2 (2000)**

WC & AC	EER (%)
SWCELLP1/2	4.67%
+ SWPH3	**3.51%**
+ SWPH1/2	4.85%
+SWPH0	5.54%

[13] Hagai Aronowitz, “Inter-Dataset Variability Compensation for Speaker Recognition,” in *Proceedings of ICASSP*, 2014.

Exploring the Domain Mismatch

- **Naïve "adaptation" via automatic subset selection**



Unsupervised Clustering Approaches for Domain Adaptation in Speaker Recognition Systems

Proposed (Bootstrap) Framework

- Begin with Σ_{SWB} (WC) and Φ_{SWB} (AC).
- Use PLDA and Σ_{SWB} , Φ_{SWB} to compute pairwise affinity matrix, $\mathbf{A}$, on SRE data.
- Cluster $\mathbf{A}$ to obtain hypothesized speaker labels.
- Use labels to obtain Σ_{SRE} and Φ_{SRE}
- Linearly interpolate (via α_{WC} and α_{AC}) between prior (SWB) and new (SRE) covariance matrices to obtain final hyper-parameters:

$$\Sigma_{\mathrm{F}} = \alpha_{\mathrm{WC}} \cdot \Sigma_{\mathrm{SRE}} + (1 - \alpha_{\mathrm{WC}}) \cdot \Sigma_{\mathrm{SWB}}$$

$$\Phi_{\mathrm{F}} = \alpha_{\mathrm{AC}} \cdot \Phi_{\mathrm{SRE}} + (1 - \alpha_{\mathrm{AC}}) \cdot \Phi_{\mathrm{SWB}}$$

- Iterate?

(Unsupervised) Clustering

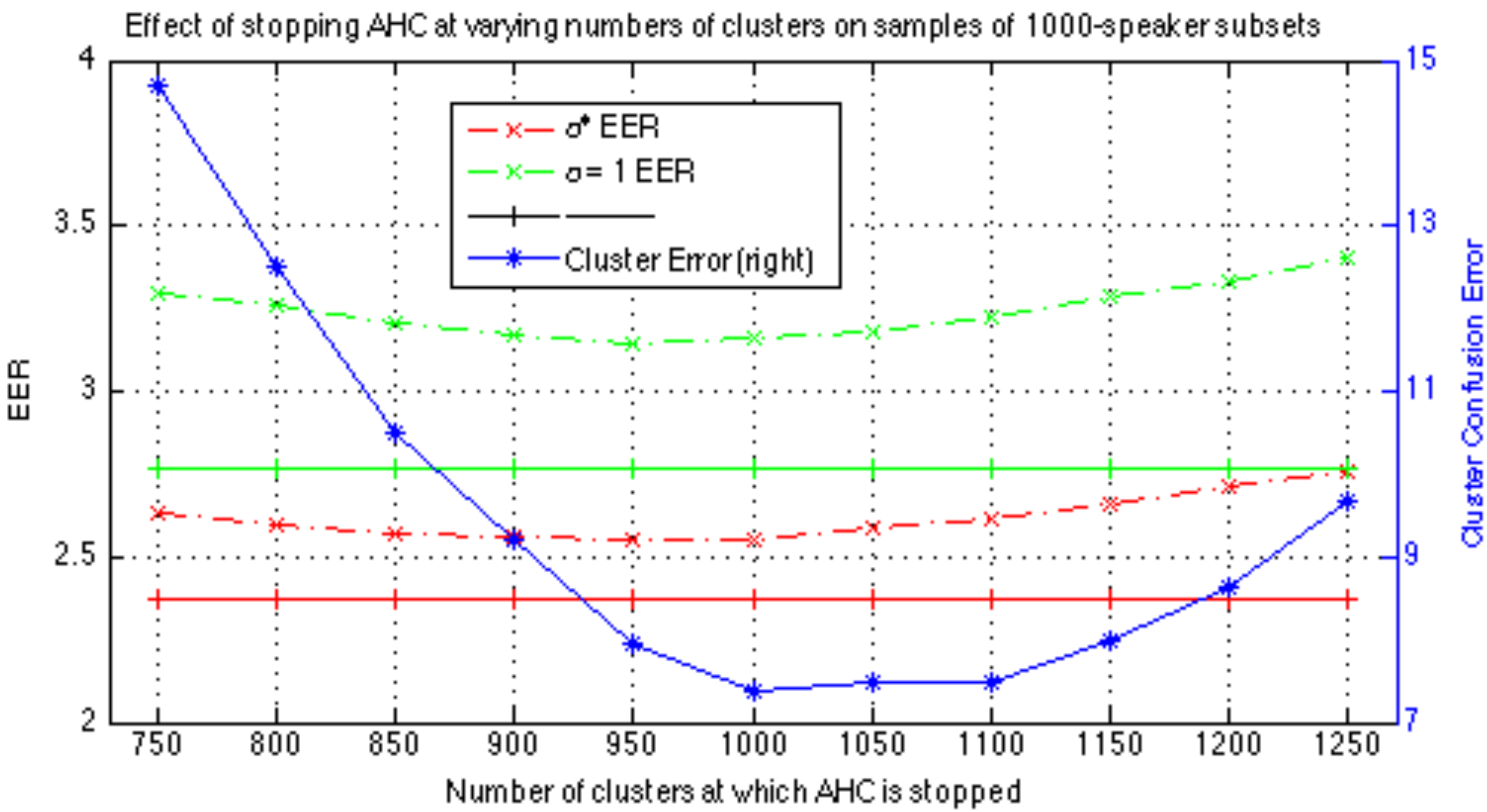
- **Agglomerative hierarchical clustering (AHC)**
  - **Requires as input the number of clusters at which to stop**

- **Graph-based random walk algorithms**
  - **Infomap [24]**
  - **Markov Clustering (MCL) [25]**

[24] Martin Rosvall and Carl T. Bergstrom, “Maps of Random Walks on Complex Networks Reveal Community Structure”, in *Proceedings of the National Academy of Sciences*, 2008.
[25] Stijn van Dongen, Graph Clustering by Flow Simulation, Ph.D. Thesis, University of Utrecht, May 2000.

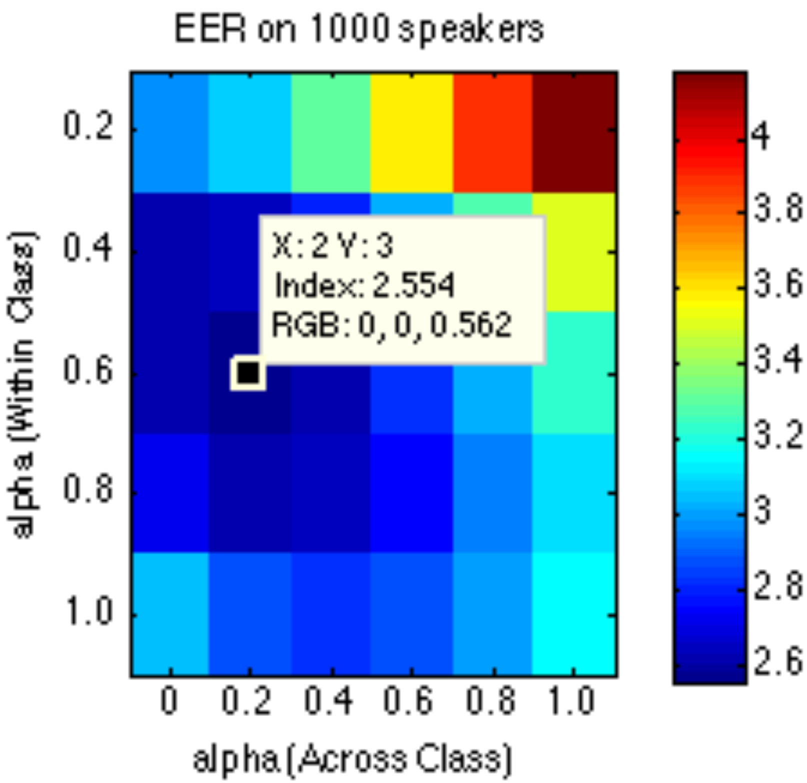
Initial Findings

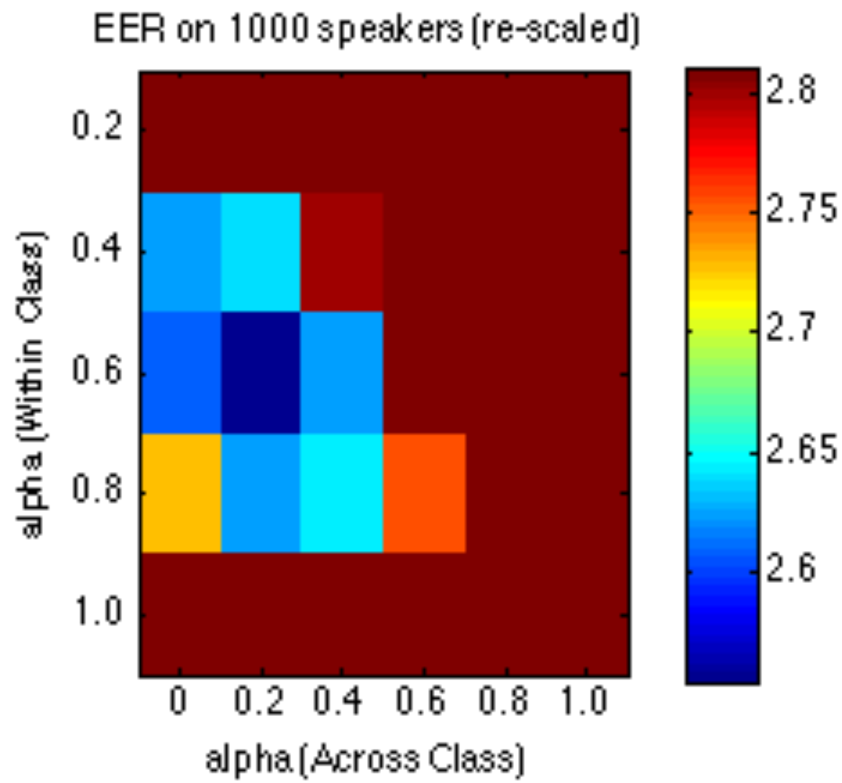
- **In the presence of interpolation (0 < α < 1), an imperfect clustering is forgivable.**



Initial Findings

- **Automatic estimation of α^***
  - **Open and unsolved, but not a huge problem**





Results So Far

- **Via clustering and optimal adaptation**

	$\hat{K}$	Perfect	Hypothesized	Gap (%)
AHC	3790*	2.23	2.58	16%
Infomap+AHC	3196	—	**2.53**	**13%**
MCL+AHC	3971	—	2.61	17%

- **Initial baseline and benchmark**

UBM & T	Whitening	WC & AC	JHU
SWB	SRE	SWB	5.54%
SWB	SRE	SRE	2.30%

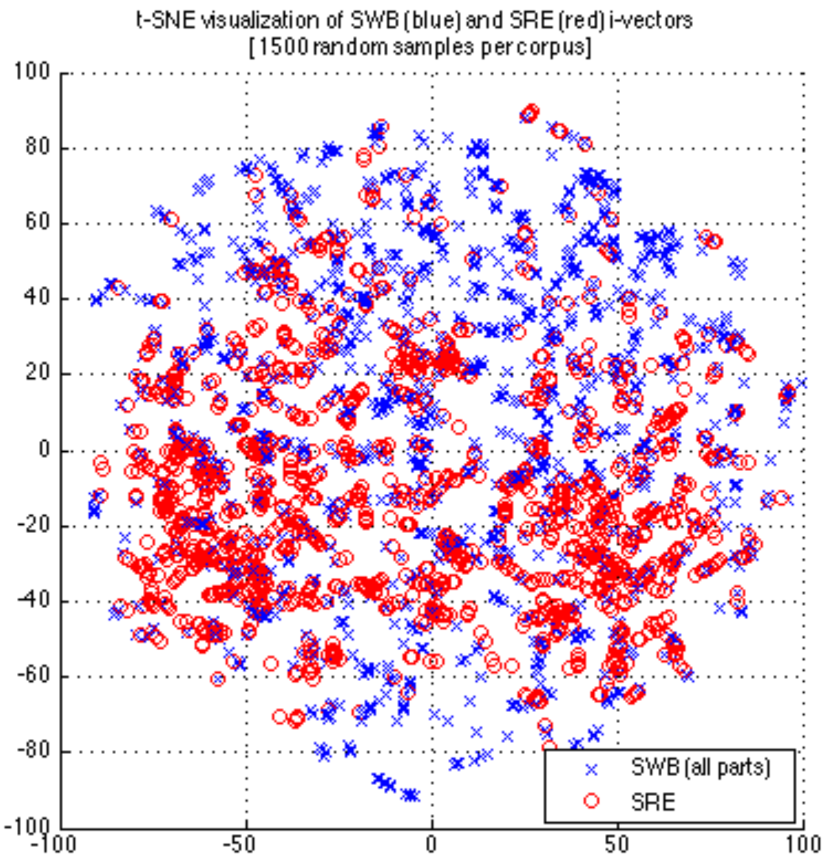
Take-home Ideas

- In the presence of interpolation, α, an imprecise estimate of the number of clusters is forgivable.
- Range of adaptation parameters yield decent results.
  - The selection of optimal values is still an open question.
- Best automatic system so far obtains SRE10 performance that is within 15% of a system that has access to all speaker labels.

What's Next?

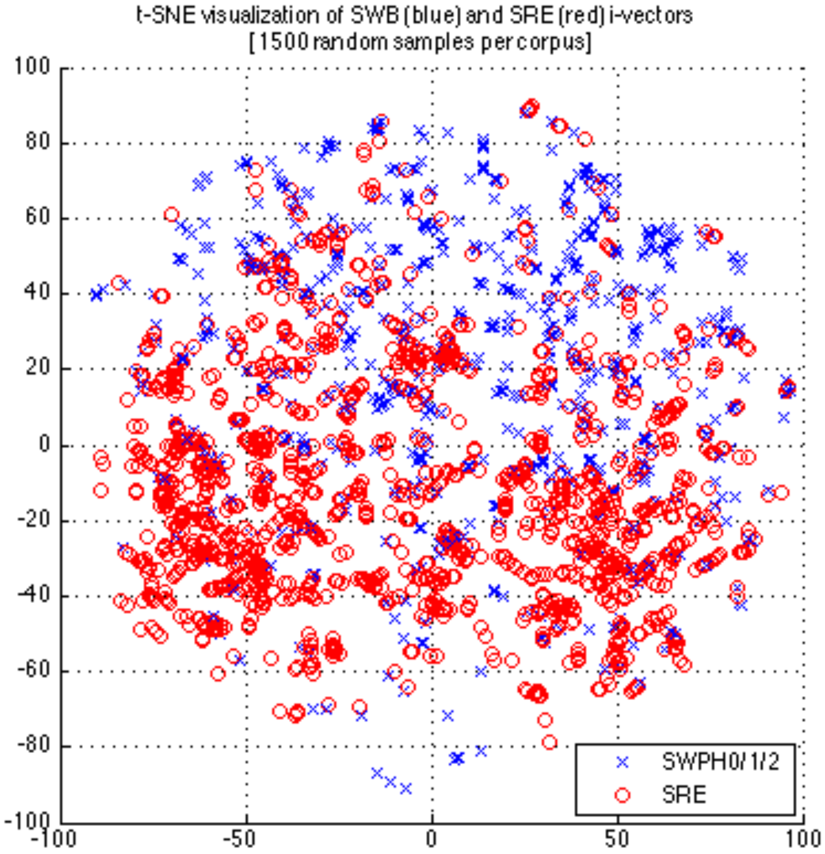
- Telephone – Telephone domain mismatch
  - Simple solutions work well already.
  - Explicitly identifying the source of the performance degradation via metadata analysis, etc.
- Telephone – Microphone domain mismatch
  - Expected to be a more difficult problem
- Out-of-domain detection
  - Not unlike outlier/novelty detection

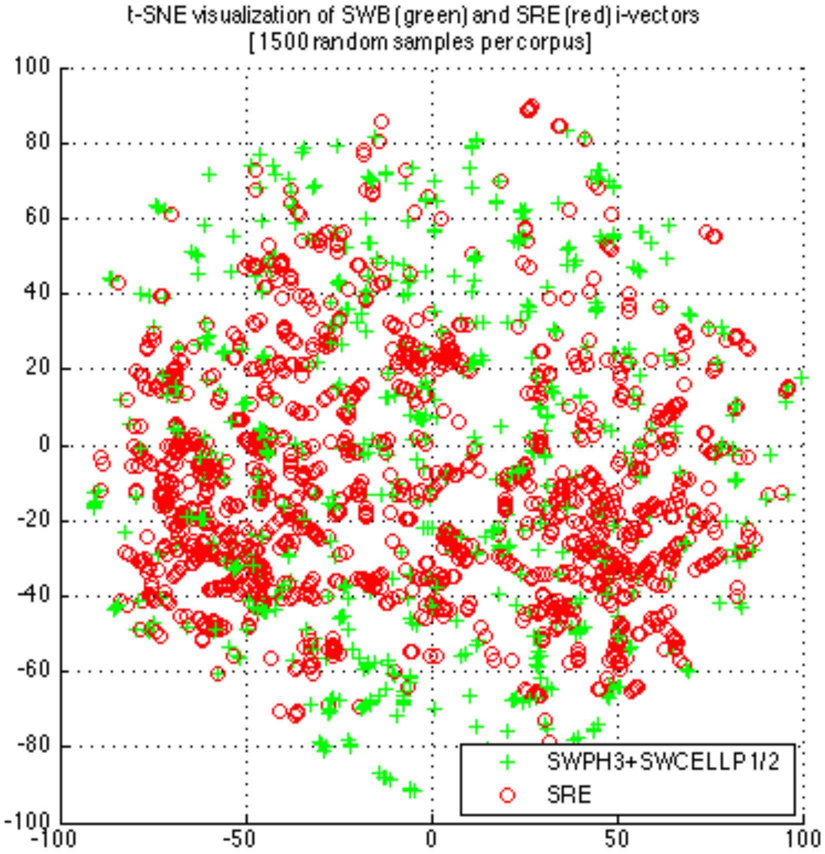
Telephone vs. Telephone



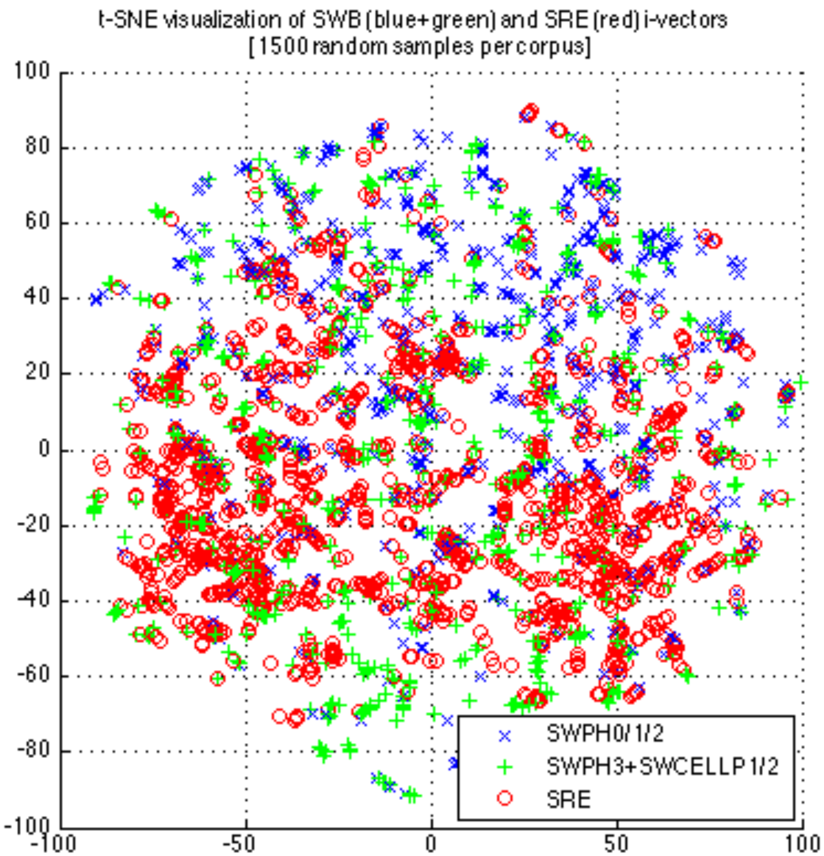
[--] Laurens van der Maaten and Geoffrey Hinton, “Visualizing data using t-SNE,” Journal of Machine Learning Research, 2008.

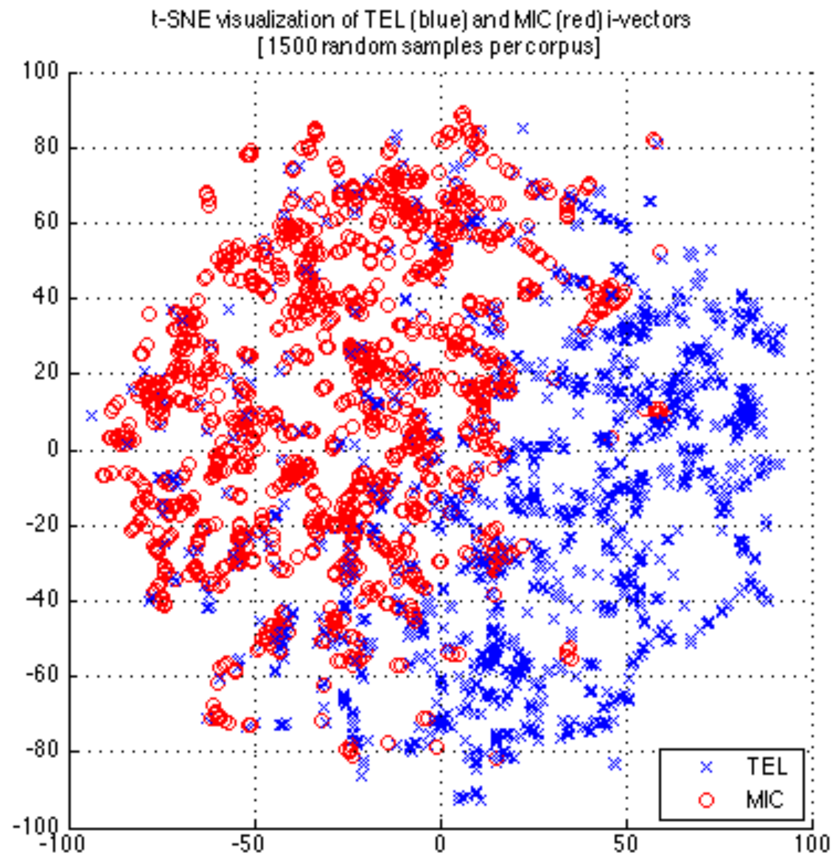
Telephone vs. Telephone





Telephone vs. Microphone





TEL = {SWB, SRE};
MIC = {SRE 05, 06, 08 microphone}

Microphone vs. Microphone