The Geometry of the Articulatory Region That Produces a Speech Sound

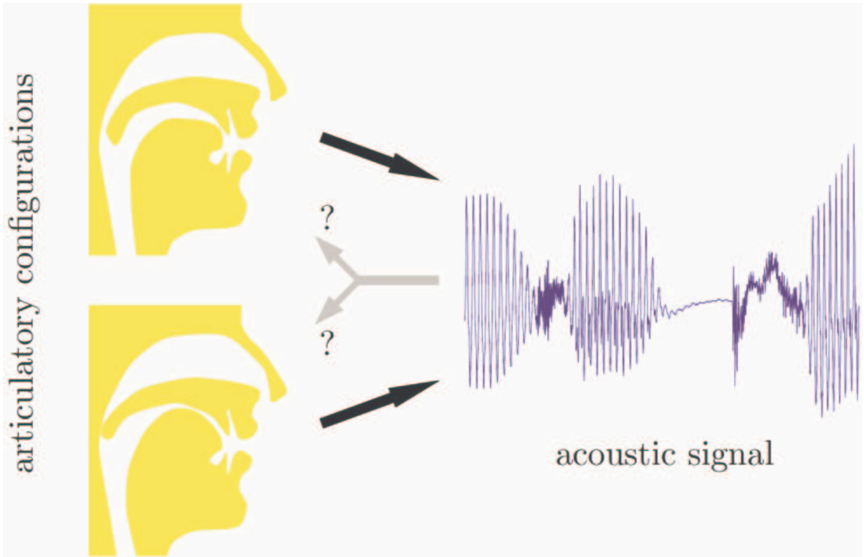
Chao Qin
EECS, School of Engineering, UC Merced, USA
November 2009

Outline

- Introduction and motivation
- Nonuniqueness of the inverse mapping
- Prediction error of individual articulators
- Nonuniqueness of individual articulators
- Conclusions

Introduction

- Articulatory inversion
  - Recovering vocal tract shapes from acoustics
  - Still an open research problem!



- Nonuniqueness of the inverse mapping
  - Model-based approaches: Atal et al'78, Boe et al'92
  - Data-driven approaches: Qin&Carreira-Perpiñán'07

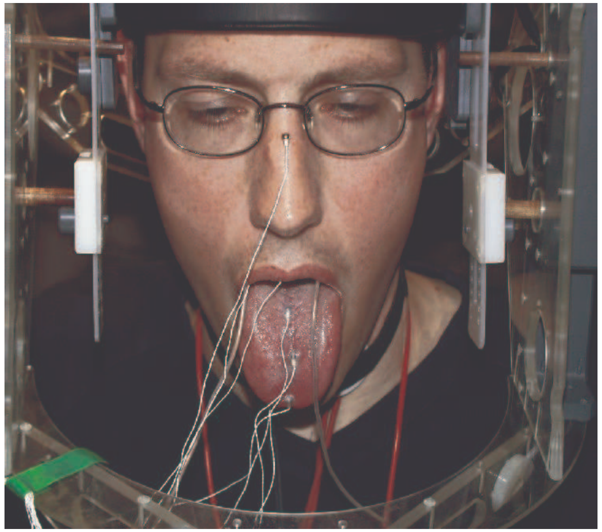
Introduction

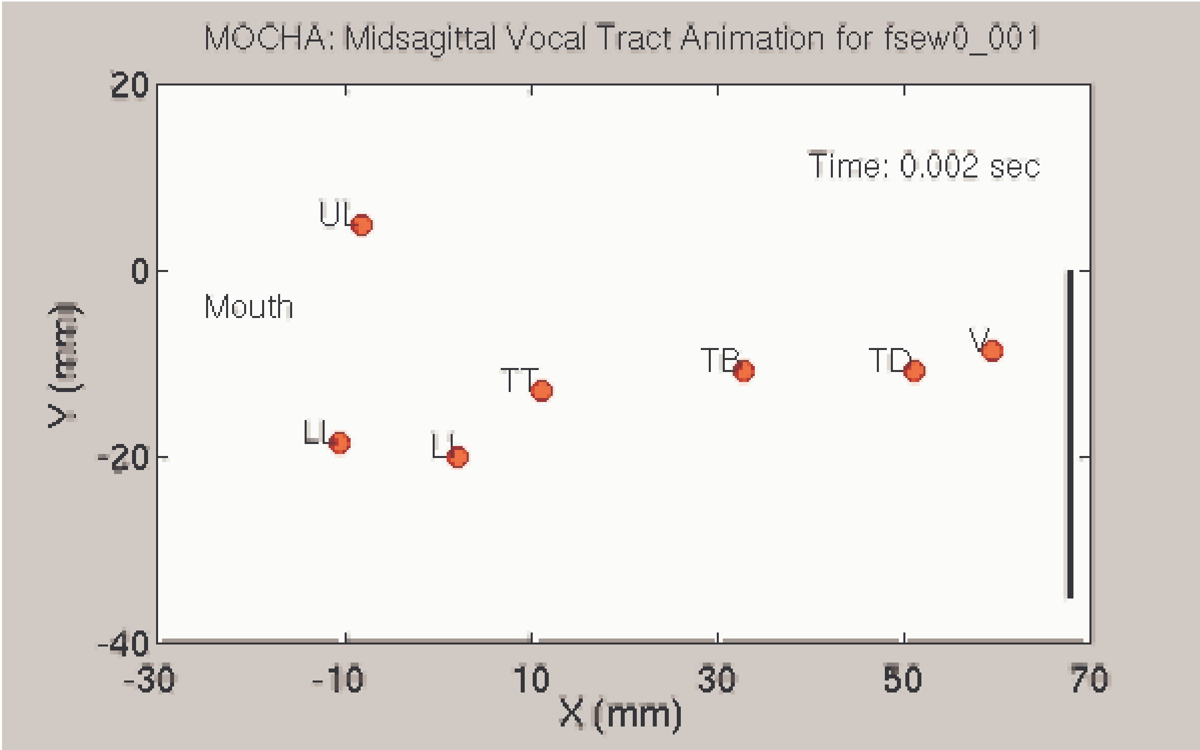
Nonuniqueness of any articulator	$\Rightarrow$	Nonuniqueness of the entire VT
Nonuniqueness of the entire VT	$\not\Rightarrow$	Nonuniqueness of every articulator

- Questions
  - Is recovering a portion of the vocal tract simpler than recovering the entire VT?
  - How to quantify the difficulty?
- Why recovering portions of the vocal tract?
  - Useful for facial animation (lips and anterior tongue) and diagnosis of speech disorders (velum height) in dysarthria
  - Useful for separating linguistic information from speakers' idiosyncrasy
- Approaches
  - Parametric methods: model-based inversion
  - Nonparametric methods: fewer assumptions

PART I:
Prediction Error of Individual Articulators in Inverse Models

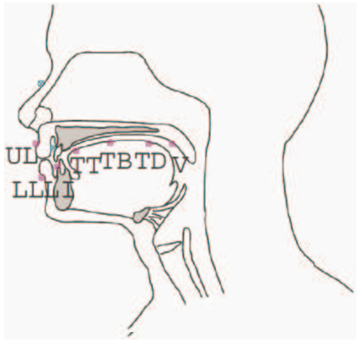
Articulatory databases

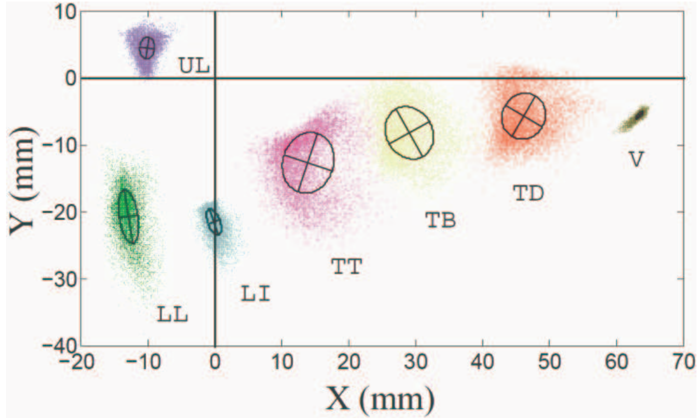




Prediction error of individual articulators

- Dataset
  - MOCHA-TIMIT
    - Train: 10000 frames
    - Valid: 4000 frames
    - Test: 15 utterances
  - EMA after "mean-filtering"
  - 12-order line spectral frequency (LSF)





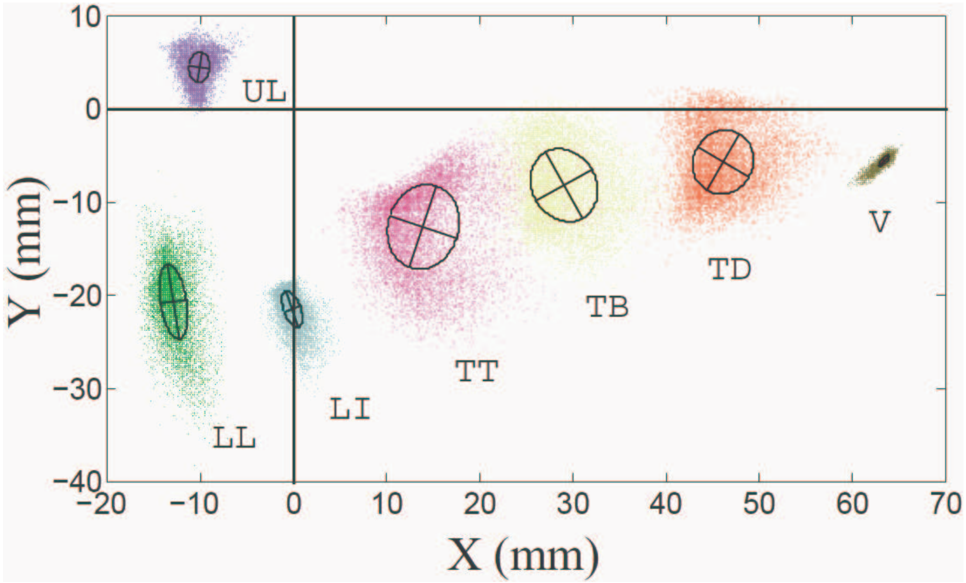
- Inversion by neural networks
  - 7 MLPs for different portions of the front VT
  - 6 MLPs for individual articulators
  - 1 RBF for entire vocal tract:
- Model parameters
  - MLPs: single layer with 100 hidden units
  - RBF: regularization $\lambda = 0.1$, $M = 600$ basis functions, bandwidth $\sigma = 0.1$
- Per-articulator error and correlation are similar whether to recover portions of VT or the whole VT

Experimental results: vocal tract inversion

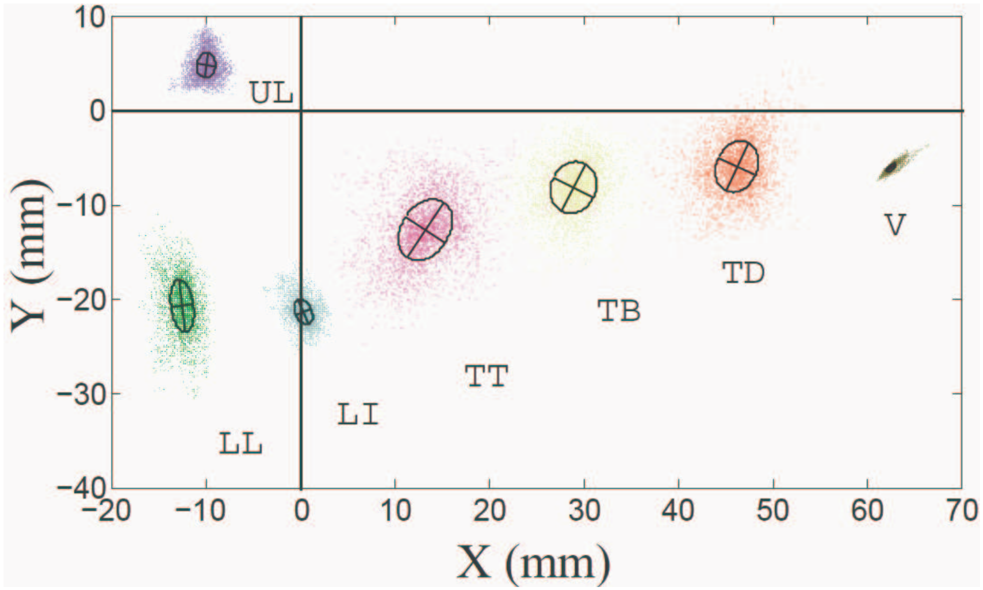
	Portions of the VT by MLPs		Whole VT by RBF		Individual articulator by MLPs	
	RMSE	Correlation	RMSE	Correlation	RMSE	Correlation
ULx	1.00	0.51	0.99	0.51	1.02	0.48
ULy	1.36	0.57	1.33	0.60	1.36	0.58
LLx	1.32	0.49	1.28	0.51	1.35	0.47
LLy	2.96	0.70	2.93	0.71	2.95	0.71
LIx	0.94	0.48	0.92	0.51	0.95	0.47
LIy	1.33	0.75	1.32	0.75	1.35	0.74
TTx	2.74	0.72	2.71	0.73	2.79	0.71
TTy	3.06	0.77	3.01	0.78	3.05	0.77
TBx	2.37	0.77	2.36	0.77	2.44	0.75
TBy	2.63	0.74	2.60	0.74	2.65	0.74
TDx	2.21	0.74	2.19	0.75	2.26	0.72
TDy	2.75	0.59	2.72	0.59	2.78	0.59
Vx	0.51	0.69	0.52	0.68	0.52	0.68
Vy	0.46	0.70	0.46	0.70	0.46	0.70

Normalized estimation error

The entire dataset for speaker fsew0



Estimation errors: $e_j^i = a_j^i - \hat{a}_j^i$



Relative estimation error for each articulator

	UL	LL	LI	TT	TB	TD	V
RBF	0.84	0.77	0.83	0.72	0.70	0.74	0.82
MLP	0.85	0.79	0.84	0.72	0.71	0.75	0.82

$$\left(\frac{1}{2}\mathrm{tr}(\Sigma_r^{-1/2}\Sigma_e\Sigma_r^{-1/2})\right)^{1/2} \stackrel{\sim}{\Rightarrow} \left(\frac{1}{2}\sum_{i=1}^{2}\lambda_e^i / \lambda_r^i\right)^{1/2}$$

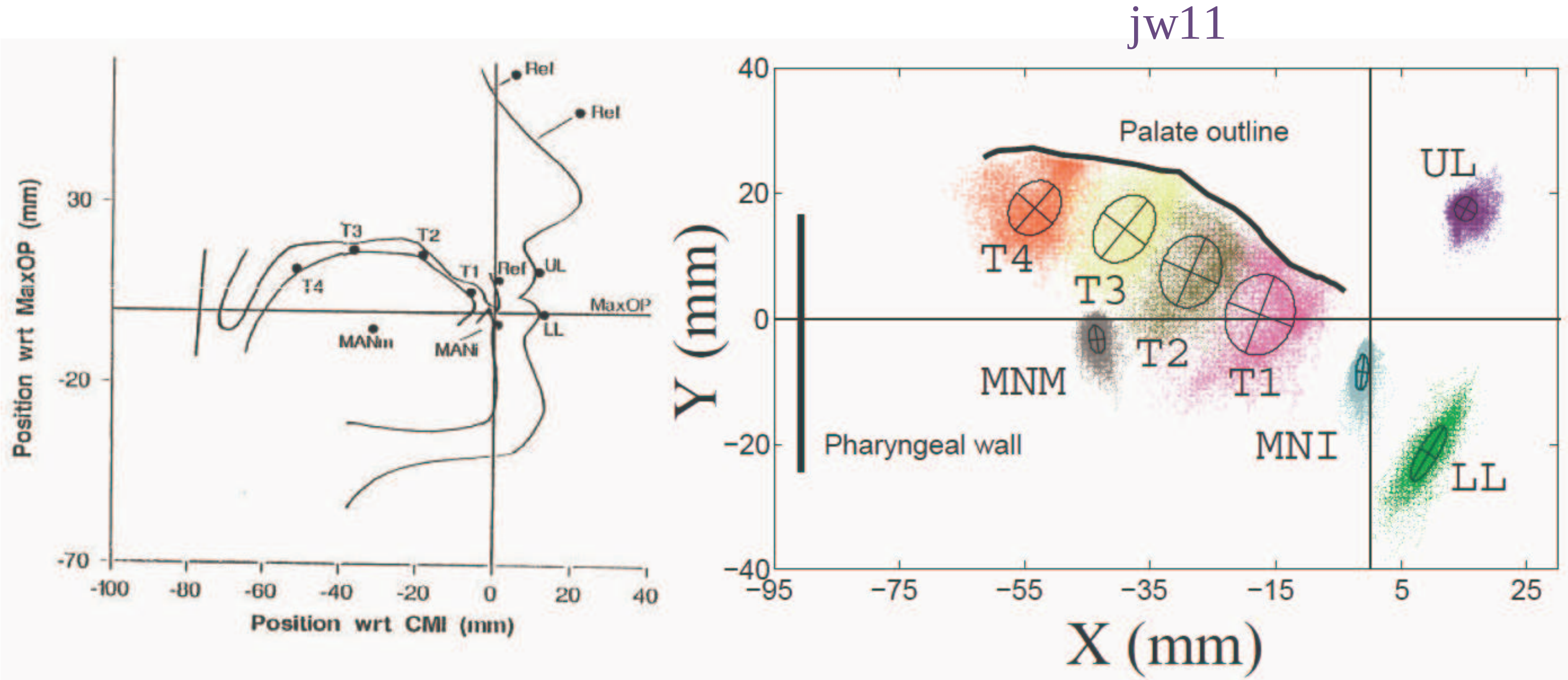
Σ_r : covariance of each articulator's position

Σ_e : covariance of each articulator's error

PART II:

Nonuniqueness of Individual Articulators

Wisconsin X-ray microbeam database



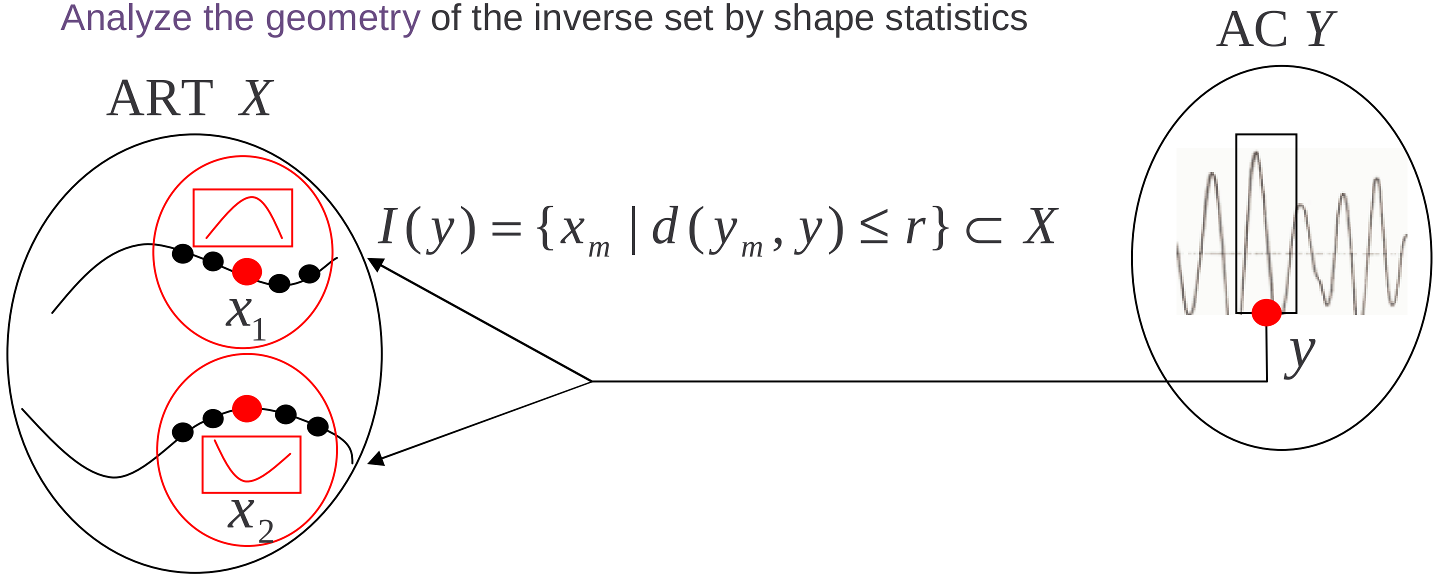
$$\{x_n, y_n\}_{n=1}^{43260}$$

$$x_n \in \Re^{16D} : \text{articulatory positions}$$

$$y_n \in \Re^{20D} : 20\text{-order LPC}$$

Multimodality of the inverse set

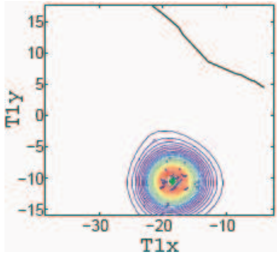
- Nonparametric algorithm
  - Search multimodality in individual 2D articulatory space (like Qin&Carreira-Perpiñán'07)
  - Analyze the geometry of the inverse set by shape statistics



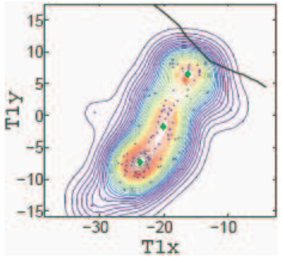
1. Given an acoustic vector y
2. Find its inverse set $I(y)$
3. Count number of modes of kernel density estimate of bandwidth $\sigma = 6\ \mathrm{mm}$
4. Compute shape statistics: eigenvalues of the covariance matrix
5. Repeat for all acoustic vectors in the dataset

Shape statistics of the inverse set

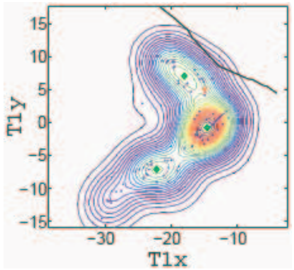
- Characterizing the geometry by the shape statistics
  - Eigenvalues of the covariance matrix from the inverse set
  - $\lambda_1 \geq \lambda_2$ measure the spread of the inverse set along its principal axes



1. λ_2 and λ_1 are small $\Rightarrow$ tightly concentrated and 0D manifold

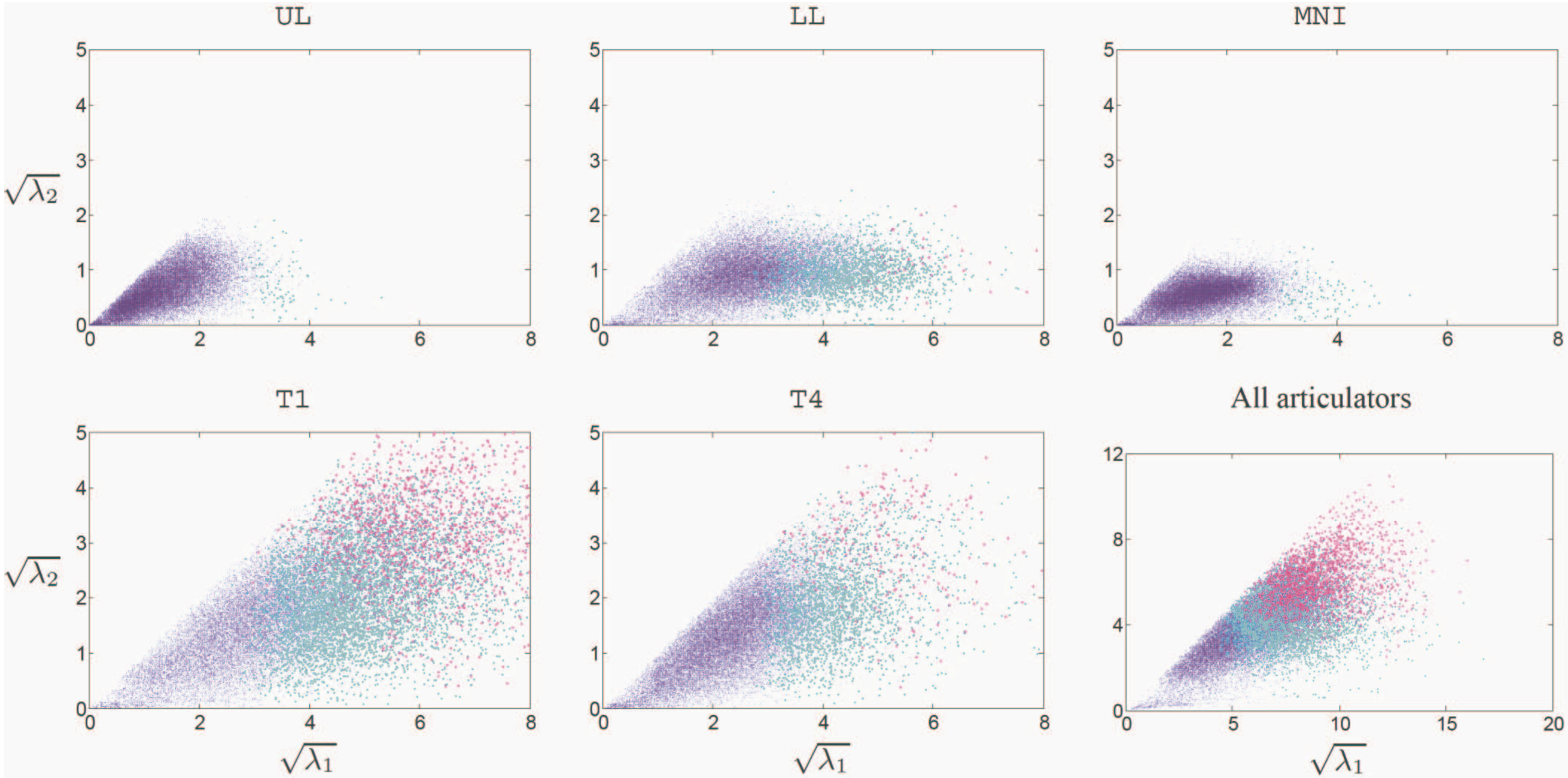


2. $\lambda_2 << \lambda_1 \Rightarrow$ elongated shape and 1D manifold
3. Otherwise $\Rightarrow$ complex shape



- These shape statistics only depend on the acoustic distance $r = 0.2$

Eigenvalue plots for some articulators



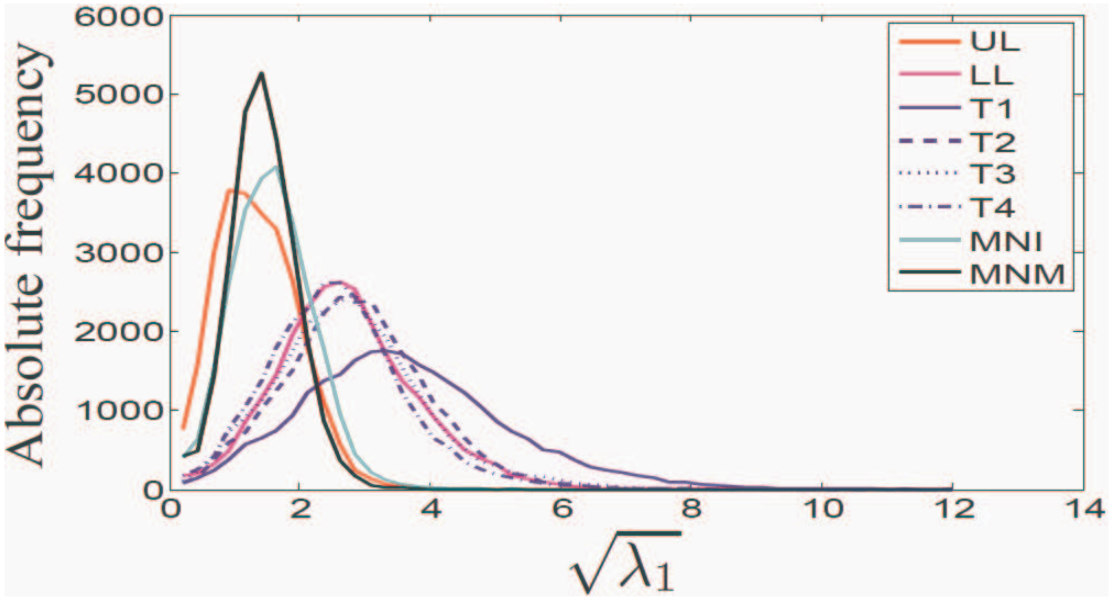
Percentage of nonuniqueness in the dataset

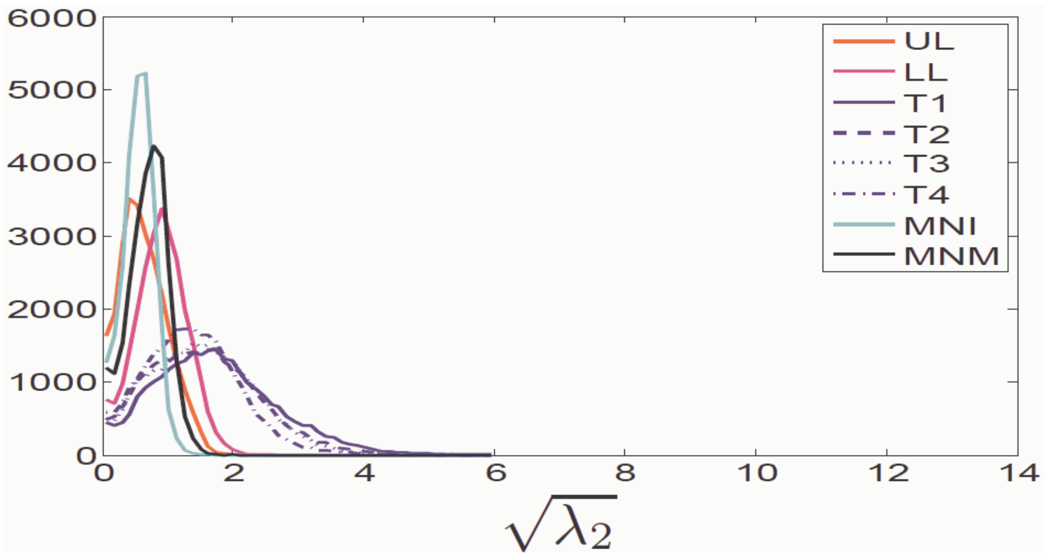
Extremely infrequent

Quite infrequent

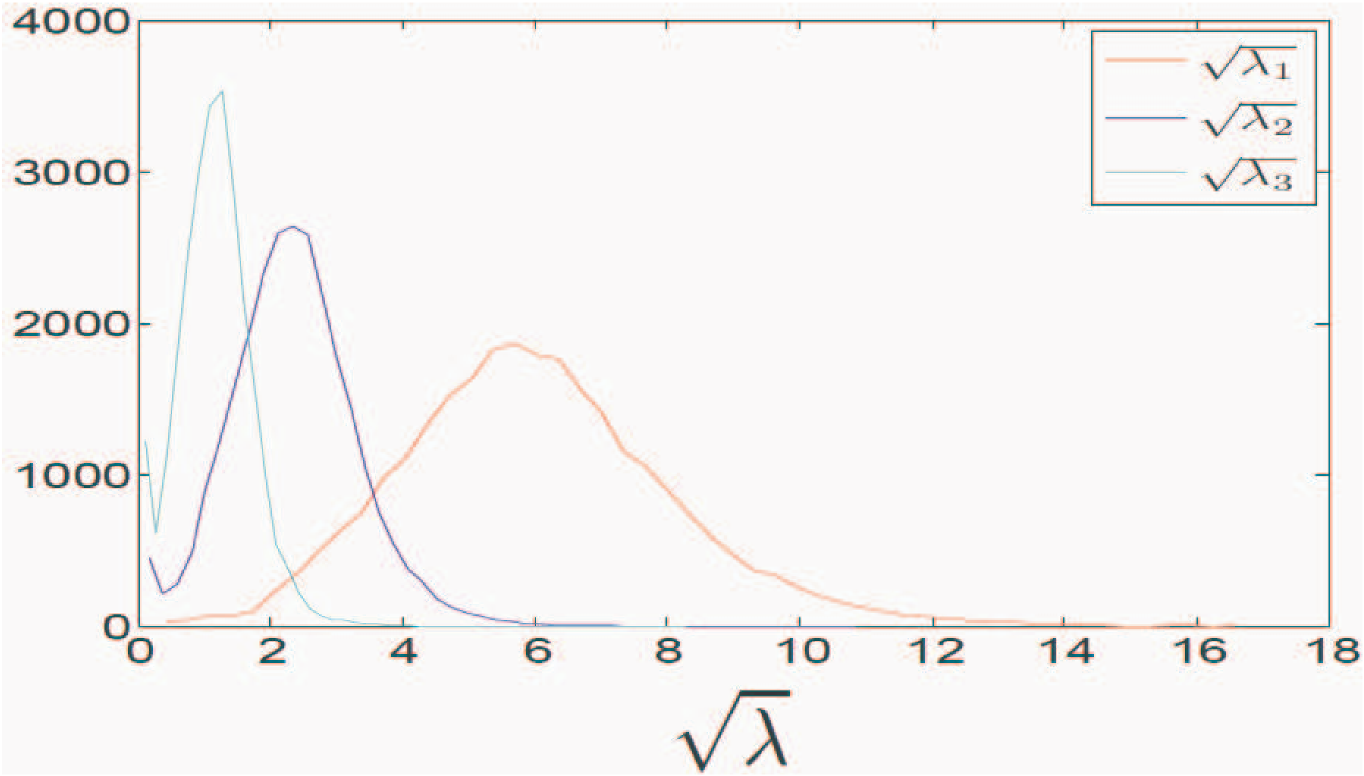
modes	UL	LL	MNI	MNM	T1	T2	T3	T4	all
1	99.6	93.1	99.5	99.8	78.3	88.3	89.1	91.4	78.1
2	0.4	6.7	0.5	0.2	17.6	10.2	9.6	7.9	16.7
3+	0	0.2	0	0	4.1	1.5	1.3	0.7	5.2

Histogram plots for each articulator

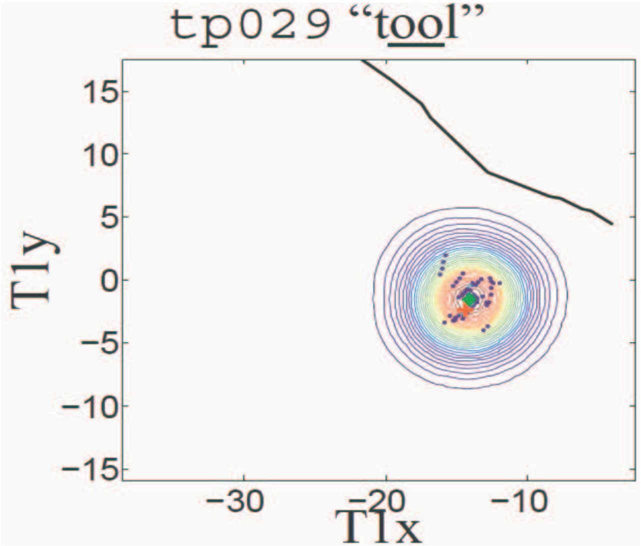


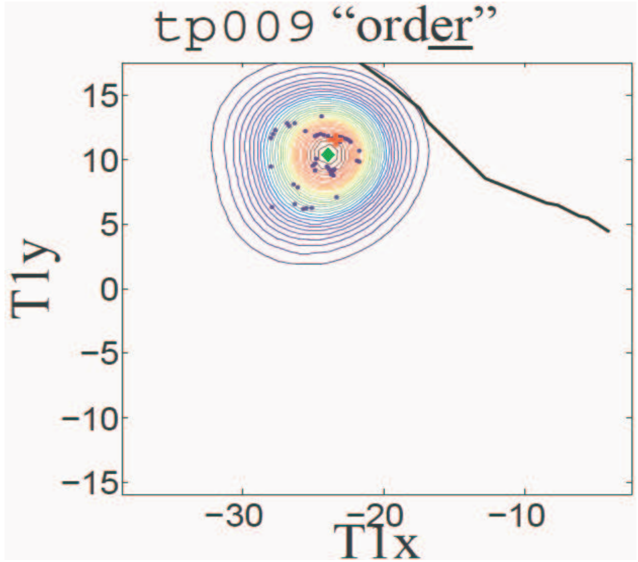


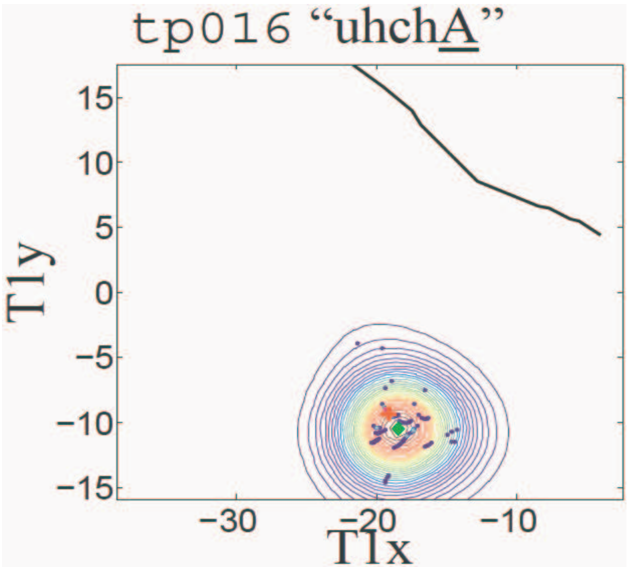
Histogram plot for the entire vocal tract

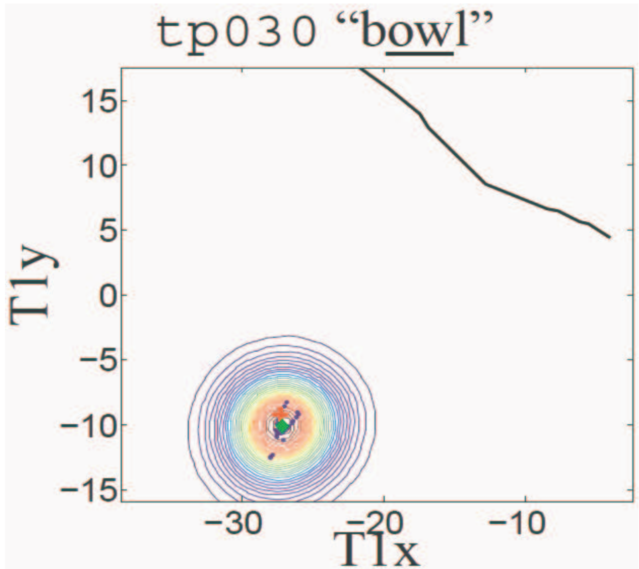


Unique frames in T1 space

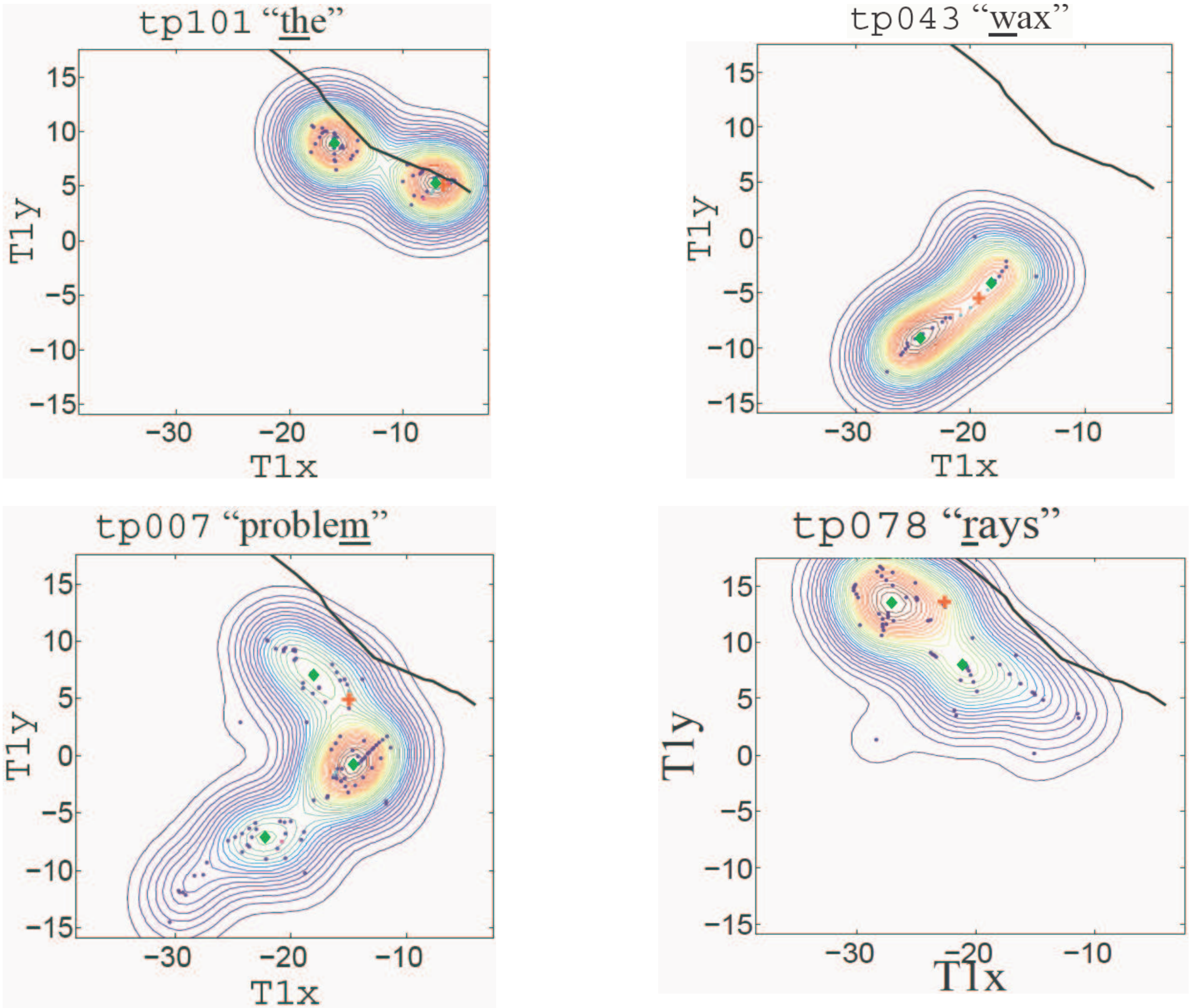








Nonunique frames in T1 space



Conclusion

- Nonuniqueness affects all the articulators of the vocal tract
- Some or even all articulators may be strongly constrained
- The normalized inversion error by neural nets is approximately the same over all articulators
- Generally, the set of articulatory shapes that correspond to a given sound is relatively constrained around a roughly spherical region in articulatory space (0D manifold, eg. vowels)
- Many frames do show more complex shapes: very elongated in a straight or curved path (1D manifold, eg. glides /l/ and /w/) or multimodality (>=2D manifold, eg. /r/) or even more complex (eg. /m/)

Acknowledge

- Work funded by NSF award IIS-0754089 and IIS-0711186