The Geometry of the Articulatory Region That Produces a Speech Sound

Chao Qin and Miguel Á. Carreira-Perpiñán

EECS, School of Engineering, University of California, Merced, USA
{cqin,mcarreira-perpinan}@ucmerced.edu

Abstract

It is known that some speech sounds are produced by more than a single vocal tract shape. Here, we study to what extent individual articulators (e.g. the tongue tip) are constrained by a given acoustic feature vector. We use parametric and nonparametric methods for articulatory inversion and quantify the error incurred by inversion methods, and the dimensionality and multimodality of the inverse region in articulatory space that corresponds to a speech sound.

Index Terms: articulatory inversion, nonuniqueness, neural networks, mode-finding.

1. Introduction

Articulatory inversion is the problem of recovering the sequence of shapes of the vocal tract (from the glottis to the lips) that produces a given acoustic utterance [1]. It is a difficult problem because the forward mapping from articulators to acoustics is many-to-one, i.e., different vocal tracts can produce the same acoustics. This makes its inverse not only highly nonlinear but also one-to-many. Many methods have been proposed to perform articulatory inversion. However, in this paper we focus not on specific inversion methods, but on characterizing the degree and nature of the nonuniqueness of the inverse mapping itself. Although specific instances of nonuniqueness are known (e.g. /ɹ/), little is known about the geometry (size, dimensionality, multimodality) of the region in articulatory space (of the whole vocal tract, or of individual articulators) that corresponds to sounds occurring in normal speech.

One question we address in this paper is whether the problem of articulatory inversion is simpler when trying to recover only a portion of the vocal tract rather than all of it. It is known that a certain amount of nonuniqueness exists during normal speech in the vocal tract [2]. However, while (1) nonuniqueness of one articulator implies (2) nonuniqueness of the entire vocal tract, (2) does not necessarily imply (1) for all articulators. For example, two different vocal tract shapes that produce the same acoustics might place the lips in the same position. In fact, it is conceivable that certain articulators are uniquely determined by the acoustics for every phoneme. Thus, recovering certain articulators only may be an easier problem, and articulatory inversion methods could benefit from this. A less fundamental but practically important argument is that by considering a portion of the vocal tract, the dimension of the space to model decreases, so the efficiency and robustness of the methods increases (in particular of probabilistic methods such as [3]).

Recovering only a portion of the vocal tract is of interest in several applications. For example, recovering the shape of the lips and anterior tongue is useful for facial animation [5]. Recovering the geometry of the velum could be useful as an aid in the diagnosis of dysarthria (which is characterized by hypernasalization, caused by an impairment of the velopharyngeal function). Also, it has been suggested [6] that linguistic information is coded in the geometry of the frontal cavity of the vocal tract, whereas speaker-dependent aspects are controlled by the geometry of the back cavity.

Several studies of the inverse mapping exist. Some of these are based on vocal tract models, that is, articulatory synthesizers based on a tube-like geometric model of the vocal tract, controlled by a few parameters [7, 8]. For example, Boë et al [8], using Maeda's model, argued that the lip area and the location and dimension of the oral constriction used in French vowel production could be derived from the first 3 formants, even though the complete shape of the vocal tract could not be recovered. However, as argued by Hogden et al [9] and others, these studies contain significant uncertainties. For example, vocal tract models have the problem of ensuring not only that vocal tract shapes are physically feasible, but also that they are actually used in normal speech. Some of these problems can be avoided by using measured articulatory data. Several such studies exist (e.g. [10, 9]), although they are often limited to small datasets (often just vowels, represented only by their first 3 formants). We use two large articulatory databases that cover most sounds in American (XRMB) and British (MOCHA) English. However, these databases include information only up to the velum, with no information about the pharyngeal region of the vocal tract. No other public database that we know of includes data about the entire vocal tract during large enough amounts of conversational speech. Thus, our work will be limited to the lips–velum portion of the vocal tract.

In the following two sections, we quantify how difficult it is to recover portions of the vocal tract and individual articulators from the instantaneous acoustics. Section 2 uses model-based inversion methods, in particular neural nets and radial basis functions. These methods cannot model multivalued mappings, but get good results if there is little nonuniqueness, and are useful as a baseline. Section 3 uses nonparametric methods based on searching the articulatory data for frames matching the given acoustics within a certain tolerance. These methods can deal with multivalued mappings and rely on fewer assumptions about the data. This extends previous work [2] where we studied the nonuniqueness of the whole vocal tract.

2. Prediction Error of Individual Articulators in Inverse Models

Dataset. We used the MOCHA-TIMIT database ([11], see fig. 1) which records, simultaneously with the acoustic wave, positions of 7 receiver coils in the midsaggital plane of the vocal tract (VT) shape, sampled at 500 Hz. We used the dataset from speaker `fsew0`, which is divided into a training, validation and test set of $10\,000$ frames, $4\,000$ frames, and 15 unseen

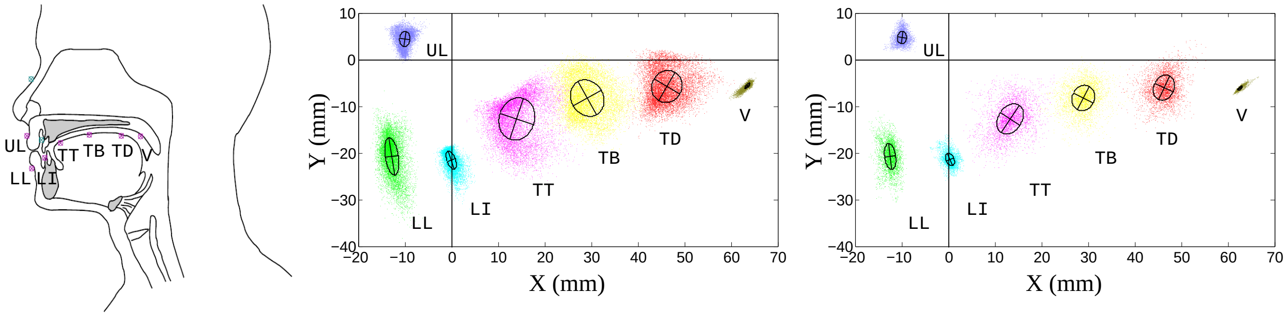

Fig. 1. *Left*: pellet locations in the MOCHA database. *Middle*: plot of the entire dataset for speaker `fsew0` (this is the corrected dataset by mean-filtering in [13]); each pellet's data uses a different color and shows a 1–stdev contour of its covariance $\mathbf{\Sigma}_r$ centered at its mean. *Right*: distribution of estimation errors at each pellet, centered at each pellet's mean, and contour for its covariance $\mathbf{\Sigma}_e$.

	`UL`		`UL...LL`		`UL...LI`		`UL...TT`		`UL...TB`		`UL...TD`		`UL...V`		RBF		Individual	
	RMSE	corr	RMSE	corr	RMSE	corr	RMSE	corr	RMSE	corr	RMSE	corr	RMSE	corr	RMSE	corr	RMSE	corr
`ULx`	1.02	0.48	1.02	0.49	1.01	0.49	1.01	0.49	1.01	0.49	1.01	0.49	1.00	0.51	0.99	0.51	1.02	0.48
`ULy`	1.36	0.58	1.36	0.58	1.35	0.58	1.36	0.58	1.35	0.59	1.34	0.59	1.36	0.57	1.33	0.60	1.36	0.58
`LLx`			1.34	0.48	1.31	0.49	1.32	0.49	1.34	0.48	1.32	0.49	1.32	0.49	1.28	0.51	1.35	0.47
`LLy`			2.99	0.70	2.95	0.70	2.95	0.70	2.96	0.70	2.96	0.70	2.96	0.70	2.93	0.71	2.95	0.71
`LIx`					0.94	0.48	0.95	0.48	0.96	0.47	0.94	0.49	0.94	0.48	0.92	0.51	0.95	0.47
`LIy`					1.35	0.74	1.33	0.75	1.34	0.74	1.33	0.75	1.33	0.75	1.32	0.75	1.35	0.74
`TTx`							2.78	0.71	2.74	0.72	2.76	0.72	2.74	0.72	2.71	0.73	2.79	0.71
`TTy`							3.06	0.77	2.99	0.78	3.04	0.77	3.06	0.77	3.01	0.78	3.05	0.77
`TBx`									2.36	0.77	2.38	0.76	2.37	0.77	2.36	0.77	2.44	0.75
`TBy`									2.65	0.73	2.63	0.74	2.63	0.74	2.60	0.74	2.65	0.74
`TDx`											2.22	0.74	2.21	0.74	2.19	0.75	2.26	0.72
`TDy`											2.75	0.59	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

Table 1. Results (RMSE in mm and correlations, for the `x` and `y` coordinates of each articulator) of recovering the whole vocal tract (VT) or portions of it using neural networks and an RBF net. The first 7 columns are for MLPs trained to recover different portions of the VT; the last column is for MLPs trained to recover individual articulators.

utterances, respectively. We used a mean-filtering procedure as in [13] to normalize the raw EMA data, and further downsampled the EMA data from 500 to 100 Hz to match the acoustic frame rate of 10 ms. As acoustic features, we used line spectral frequency (LSF, order 12, 20 ms window) with dynamic features, which was found in [14] to be best suited for inversion.

Inversion methods. We use neural networks for the inverse mapping. We train: 1) 7 multilayer perceptrons (MLP), each to recover a portion of the front VT from the acoustics, as follows: `UL+LL+LI+TT+TB+TD+V`, `UL+LL+LI+TT+TB+TD`, `UL+LL+LI+TT+TB`, `UL+LL+LI+TT`, `UL+LL+LI`, `UL+LL`, and `UL`. 2) 6 MLPs, each to recover an individual articulator (`LL`, `LI`, `TT`, `TB`, `TD`, `V`). 3) One radial-basis function (RBF) network to recover the entire front VT (`UL+LL+LI+TT+TB+TD+V`). Note that the result for a given articulator (e.g. `UL`) may be different among different MLPs (e.g. `UL+LL` or `UL+LL+LI`) because the first-layer weights are shared among all outputs. However, for RBFs where the first layer (the means and widths of the RBFs) is trained based only on the inputs (as commonly done), the outputs are independent from each other. All MLPs have a single hidden layer with 100 hidden units. We trained them (using Netlab `http://www.ncrg.aston.ac.uk/netlab`) with conjugate gradient descent and early stopping. For each MLP, we picked the one with lowest test error over 10 random initializations. The RBF is trained with weight regularization ($\lambda = 0.1$ and has $M = 600$ basis functions with width $\sigma = 0.1$.

Table 1 shows the root-mean-square error (RMSE) in mm and the correlation in $[-1, 1]$ for the neural nets and RBF net. As can be seen, one can achieve approximately the same error whether recovering part of the VT with dedicated networks, or the whole of it with a single network. Thus, although some previous work has focused on inverting specific articulators (e.g. the velum height in [12]), our results suggest that one may just as well fit a single mapping to recover the entire vocal tract. The RMSE and correlations we obtain are comparable to others [13] (noting that we do not use an acoustic context window and our dataset is much smaller than in [13]). We obtained similar results with different number of hidden units and different initialization strategies.

Fig. 1 shows the distribution of the errors $\mathbf{e}^i_j = \mathbf{a}^i_j - \hat{\mathbf{a}}^i_j$ between the true ($\mathbf{a}^i_j$) and estimated ($\hat{\mathbf{a}}^i_j$) articulator in frame j, for each articulator i. We plot these errors as vectors centered at each articulator's mean. We see that the covariance of each articulator's error $\mathbf{\Sigma}_e$ is roughly aligned and proportional to the covariance of that articulator's position $\mathbf{\Sigma}_r$ (compared to fig. 1), except for `TB`, which is roughly spherical. To evaluated quantitatively the errors over different articulators we used the a relative error $(\frac{1}{2}\,\mathrm{tr}\,(\mathbf{\Sigma}_r^{-1/2}\mathbf{\Sigma}_e\mathbf{\Sigma}_r^{-1/2}))^{1/2}$. If $\mathbf{\Sigma}_r$ and $\mathbf{\Sigma}_e$ have the same eigenvectors sorted by decreasing eigenvalue (and thus are aligned), then this relative error becomes $(\frac{1}{2}\sum_{i=1}^{2}\lambda^i_e/\lambda^i_r)^{1/2}$ in terms of the eigenvalues. Table 2 lists these relative estimation errors for each articulator. Overall, they are similar for all articulators, although somewhat larger values occur on `UL`, `LI` and `V`, which have smaller ranges than the tongue.

	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

Table 2. Relative estimation error for each MOCHA articulator.

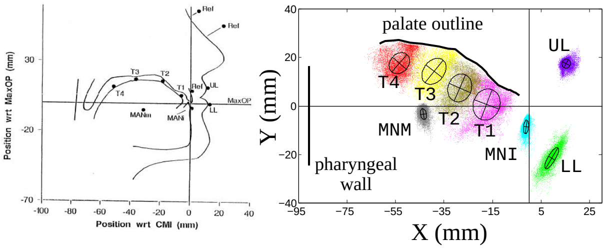

Fig. 2. *Left*: pellet locations in the XRMB database. *Right*: plot of the entire dataset for speaker `jw11`; each pellet's data uses a different color and shows a 1–stdev contour of its covariance centered at its mean.

3. Nonuniqueness of Individual Articulators

In [2], we studied nonuniqueness of the entire vocal tract. Our approach was to estimate nonparametrically the conditional density in articulatory space given an acoustic feature vector, and then search for modes in this density. Here, we use this approach to study the nonuniqueness of individual articulators. As noted earlier, nonuniqueness of the entire VT shape does not imply nonuniqueness of each articulator. Likewise, a Gaussian mixture in XY space with modes at $(\pm 1, 0)$ only has one mode in Y space. We already noted from our previous work [2] that the same sound could be produced by very different tongue shapes but with almost the same upper lip position. The basic idea in our approach is to fix one acoustic vector $\mathbf{y}_n$ and search the database for its *inverse set*, i.e., all articulatory vectors $\{\mathbf{x}_m\}$ that approximately map to $\mathbf{y}_n$ (*inversion*). Then, we apply a *clustering* algorithm to determine whether the inverse set $\{\mathbf{x}_m\}$ in each articulator's 2D space (e.g. `ULx` and `ULy`) is unimodal or not, and compute statistics from the inverse set for each articulator. Repeating this for every acoustic vector $\mathbf{y}_n$ in the database allows an exploration of the nonuniqueness of the inverse mapping for a wide range of sounds, and a characterization of the geometry of the inverse set. Let us consider each step in detail.

Dataset. In this study, we use the Wisconsin X-ray microbeam database (XRMB [15]), which records, simultaneously with the acoustic wave, the positions of 8 pellets in the midsaggital plane of the VT (see fig. 2), sampled at 147 Hz, for various types of speech (isolated words, prose, etc.). The XRMB measurement error for the pellets is 0.7 mm. As acoustic features, we use linear predictive coding (LPC) coefficients because they are closely related to the vocal tract spectral envelope, which allows direct visualization of spectral differences and formant structures. We use LPC of order 20 to obtain an accurate formant structure (for order 12, F3 is smoothed out in e.g. /ɪ/). The acoustic feature vectors use a window and step size to yield 147 Hz as with the articulatory features; we removed silent frames using energy-based endpoint detection. We use a single speaker (`jw11`, male, 90 utterances including isolated words, prose passages, etc.), resulting in a dataset of $45\,760$ vectors $(\mathbf{x}, \mathbf{y})$ with $\mathbf{x} \in \mathbb{R}^{16}$ and $\mathbf{y} \in \mathbb{R}^{20}$. Due to the fact that LPC is not effective at modeling unvoiced sounds, e.g. fricative and plosive, we eliminated those unvoiced frames (roughly 5%) from the original dataset of $45\,760$ vectors, making the final dataset $43\,260$ vectors.

Searching for multimodality in the inverse set in each articulator's space. This requires a distance $d(\mathbf{y}, \mathbf{y}')$ between acoustic vectors; we use the Itakura distance [16], which emphasizes the role of the formants and is a reasonable approximation to a perceptual distance. The VT shape representation is simpler: each component of the articulatory vector $\mathbf{x}$ is the horizontal or vertical coordinate (in mm) of a pellet, and we can use the Euclidean distance. Next, we fix an acoustic reference distance $r = 0.2$ for which we consider two acoustic vectors to be roughly the same sound. In [2], we used $r = 0.4$, but further analyses indicate that this can be too large and include some frames that have different phonetic identities in the inverse set, hence affecting the mode search. We find that the size of the inverse set varies considerably depending on the acoustic vector, so that searching for the K nearest vectors as in [17], instead of for those vectors with $d(\mathbf{y}, \mathbf{y}') \leq r$, results in missing true inverses or including false inverses depending on the value of K. We also discard those acoustic vectors whose inverse set contains less than 10 vectors so as to obtain meaningful statistics. An approximate inversion of this type is unavoidable given the discrete nature of the data. In summary, the inversion for an acoustic vector $\mathbf{y}$ returns a set $\{\mathbf{x}_m : \; d(\mathbf{y}_m, \mathbf{y}) \leq r\}$. To search for modes in this inverse set, we first fit to it a nonparametric kernel density estimate $p(\mathbf{x}) \propto \sum_m G\left(\frac{\mathbf{x}-\mathbf{x}_m}{\sigma}\right)$ with Gaussian kernel and bandwidth σ. Then, we find the modes of $p(\mathbf{x})$ using a mean-shift algorithm [4], which iterates a hill-climbing algorithm initialized at every $\mathbf{x}_m$ and collects all the resulting, distinct modes. We use $\sigma = 6$ mm; we found this to be a reasonable value based on visual inspection of the 2D inverse sets.

Shape statistics of the inverse set. The number of modes gives only partial information about the geometry of the inverse set. If the latter is a manifold of dimension zero, it can consist of one or more tight clusters (and so one or more modes). If it has dimension one and is thus elongated, it may also consist of one or more modes along it. We report additional shape statistics for each inverse set $\{\mathbf{x}_m\}$ (for a given acoustic vector) based on its covariance matrix. Its eigenvalues $\lambda_1 \geq \lambda_2$ measure the spread of the inverse set along its principal axes. If $\lambda_2 \approx \lambda_1$, the inverse set is usually distributed as a round cloud. If both λ_1 and λ_2 are quite small, the inverse set is tightly concentrated and may be considered a zero-dimensional manifold. If $\lambda_2 \ll \lambda_1$, the inverse set has an elongated shape, perhaps corresponding to a 1D manifold. These shape statistics only depend on the acoustic reference distance r, but on no other parameters (e.g. σ, since they are not obtained from the kernel density estimate). We also explore visually the inverse sets for many acoustic vectors to try to characterize their shape.

Exploratory analysis of the geometry and dimensionality of the inverse set. Fig. 3 shows the distributions of the square-root of λ_2 vs λ_1 (in mm) for selected articulators and for the entire VT. Each point corresponds to one acoustic vector and is colored according to the number of modes of its inverse set. Points can lie roughly on the diagonal or below and to the right of it, corresponding to circular and elongated shapes, respectively. The following table lists the percentage of frames with 1, 2 or more modes in each articulator space and in the entire VT space:

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

This shows that multimodality occurs in all articulators, i.e., for each articulator there are acoustic vectors for which multiple VT shapes exist that differ in that articulator (and possibly others). As noted in [2], the percentage of multimodal frames in the entire VT shape is small (here, 21.9%). However (and unlike in table 2), there are marked differences among articulators. Multimodality is very infrequent for UL, MNI, and MNM (upper lip, teeth), which mostly show circular, tight inverse sets, that may be considered as 0D manifolds. Multimodality is more frequent for the tongue (T1–T4, in particular the tip, T1) and the lower lip (LL).

Fig. 4 shows the histogram of each square-root eigenvalue for individual articulators and for the entire VT. T1 to T4 and LL have higher variability than other articulators (UL, MNI, MNM). Many frames satisfy $\sqrt{\lambda_3} \le \sqrt{\lambda_2} \le \sqrt{\lambda_1} \le 4$ mm and can be considered as tight inverse sets. The full-VT histogram shows that λ_1 is typically quite larger than λ_2 and λ_3, and that the latter are more comparable. Thus, many inverse sets in the 14D space are somewhat or considerably elongated in 1D; this can also be seen in the 2D projections in fig. 5. They are particularly common with the lips and teeth but also with the tongue. We suspect this may be the result of rigid 1D motion (for example, the jaw can mostly rotate around its axis, so the lower teeth track a circle) of an articulator that has little effect on the acoustics, or more generally a coordinated motion of several articulators. Finally, as reported in [2], we also find clearly multimodal sets with two or more tight clusters (0D manifolds).

Fig. 5 shows inverse sets (in the tongue 2D spaces) representative of the variety of shapes we find: compact unimodal (e.g. vowels), compact multimodal (e.g. “the” or /ɹ/ in “row” or “real”), or elongated 1D shapes (e.g. glides /l/, /w/). Other sounds (e.g. /m/) seem to show very complex tongue shapes.

In summary, we find most inverse sets are compact unimodal, but among the remaining ones, we find many that are elongated in a 1D shape (possibly indicating rigid motion of a non-critical articulator) or that consist of two compact but separated clusters (distinct 0D manifolds). Beyond this, we find sets with more complex shapes too.

Relation with critical articulators. The issue of nonuniqueness of the vocal tract shape is related but not identical to that of critical articulators [10]. The latter refers to the sensitivity of the acoustics as a function of small changes in different articulators. For a given phoneme, a critical articulator is one such that motions of it can strongly alter the sound, while motions of a non-critical articulator have a small effect on the sound. For example, the lower lip is critical for producing /b/ (since slightly opening the lips alters the acoustics strongly), but the tongue dorsum is not; this is reflected in a low variance of the lower lip's position over different realizations of /b/ sounds. In contrast, nonuniqueness (strictly defined) means entirely different vocal tract shapes produce exactly the same acoustics. Depending on how loosely we define nonuniqueness (i.e., how much acoustic variation we tolerate), a non-critical articulator may or may not result in nonuniqueness. More importantly, a critical articulator need not be uniquely determined. For example, the tongue dorsum in /ɹ/ has a bimodal distribution of two tight clusters; thus, while small variations of the tongue can change the acoustics significantly, entirely different tongue shapes result in almost the same acoustics.

4. Conclusion

Our results, based on parametric and nonparametric inversion techniques, suggest that nonuniqueness affects all the vocal tract articulators that we considered (in particular the tongue). However, for any given acoustic sound some or even all articulators may be strongly constrained. The set of articulatory shapes that correspond to a given sound (within a small Itakura distance in acoustic space) is usually tightly concentrated around a roughly spherical region in articulator space (dimension 0). However, many sounds do show more complex shapes: multimodality (dimension 0), very elongated in a straight or curved path (dimension 1), or even more complex. When averaged over a large dataset containing most English sounds, the inversion error using a neural net of each articulator normalized by its range of variation is approximately the same over all articulators; yet, the tongue and lower lip are much less constrained by the sound than the teeth and upper lip. How these results depend on specific classes of sounds or different speakers is a topic of future research. **Acknowledgments.** Work funded by NSF awards IIS–0754089 (CAREER) and IIS–0711186. XRMB funded (in part) by NIDCD grant R01 DC 00820.

5. References

[1] J. Schroeter and M. M. Sondhi, “Techniques for estimating vocal tract shapes from the speech signal,” *IEEE Trans. ASSP*, 1994.

[2] C. Qin and M. Á. Carreira-Perpiñán, “An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping,” in *Proc. Interspeech*, 2007.

[3] M. Á. Carreira-Perpiñán, “Reconstruction of sequential data with probablistic models and continuity constraints,” in *NIPS*, 2000.

[4] M. Á. Carreira-Perpiñán, “Mode-finding for mixtures of Gaussian distributions,” *IEEE Trans. PAMI*, 2000.

[5] M. M. Cohen, J. Beskow, and D. W. Massaro, “Recent developments in facial animation: An inside view,” in *Proc. Third Auditory-Visual Speech Processing Conference (AVSP)*, 1998.

[6] H. Hermansky and D. J. Broad, “The effective second formant F2' and the vocal tract front-cavity,” in *ICASSP*, 1989.

[7] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, “Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique,” *J. Acoustic Soc. Amer.*, 1978.

[8] L. Boë, P. Perrier, and G. Bailly, “The geometric vocal tract variables controlled for vowel production,” *J. Phonetics*, 1992.

[9] J. Hogden, A. Löfqvist, V. Gracco, I. Zlokarnik, P. E. Rubin, and E. Saltzman, “Accurate recovery of articulator positions from acoustics: New conclusions based on human data,” *J. Acoustic Soc. Amer.*, 1996.

[10] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zacks, and S. Levy, “Inferring articulation and recognizing gestures from acoustics with a neural network trained on X-ray microbeam data,” *J. Acoustic Soc. Amer.*, 1992.

[11] A. A. Wrench and W. J. Hardcastle, “A multichannel articulatory speech database and its application for automatic speech recognition,” in *Proc. 5th Speech Production Workshop*, 2000.

[12] K. Richmond, “Estimating velum height from acoustics during continuous speech,” in *Proc. Eurospeech*, 1999.

[13] K. Richmond, “Estimating articulatory parameters from the acoustic speech signal,” PhD thesis, Univ. of Edinburgh, 2002.

[14] C. Qin and M. Á. Carreira-Perpiñán, “A comparison of acoustic features for articulatory inversion,” in *Proc. Interspeech*, 2007.

[15] J. R. Westbury, *X-Ray Microbeam Speech Production Database User's Handbook Version 1.0*, 1994.

[16] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[17] S. Roweis, “Data driven production models for speech processing,” Ph.D. dissertation, California Institute of Technology, 1999.

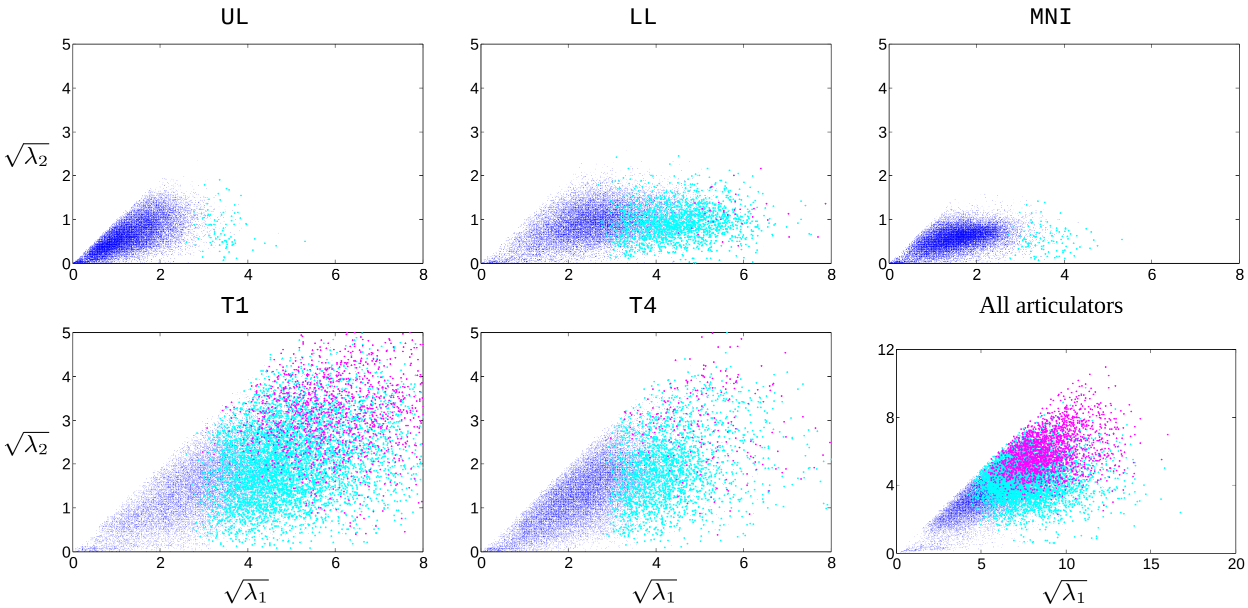

Fig. 3. Scatterplots of the eigenvalues ($\sqrt{\lambda_2}$ vs $\sqrt{\lambda_1}$) of the covariance matrix of the inverse set for some articulators. The colors (blue, cyan, magenta) indicate the number of modes (1, 2, 3+, respectively) for each inverse set.

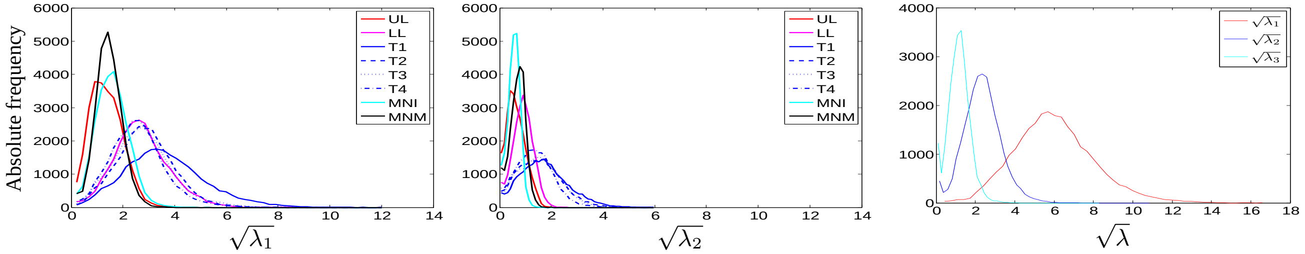

Fig. 4. *Left, middle*: histograms of $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$ for each articulator. *Right*: histogram of $\sqrt{\lambda_1}$, $\sqrt{\lambda_2}$, $\sqrt{\lambda_3}$ for the entire vocal tract.

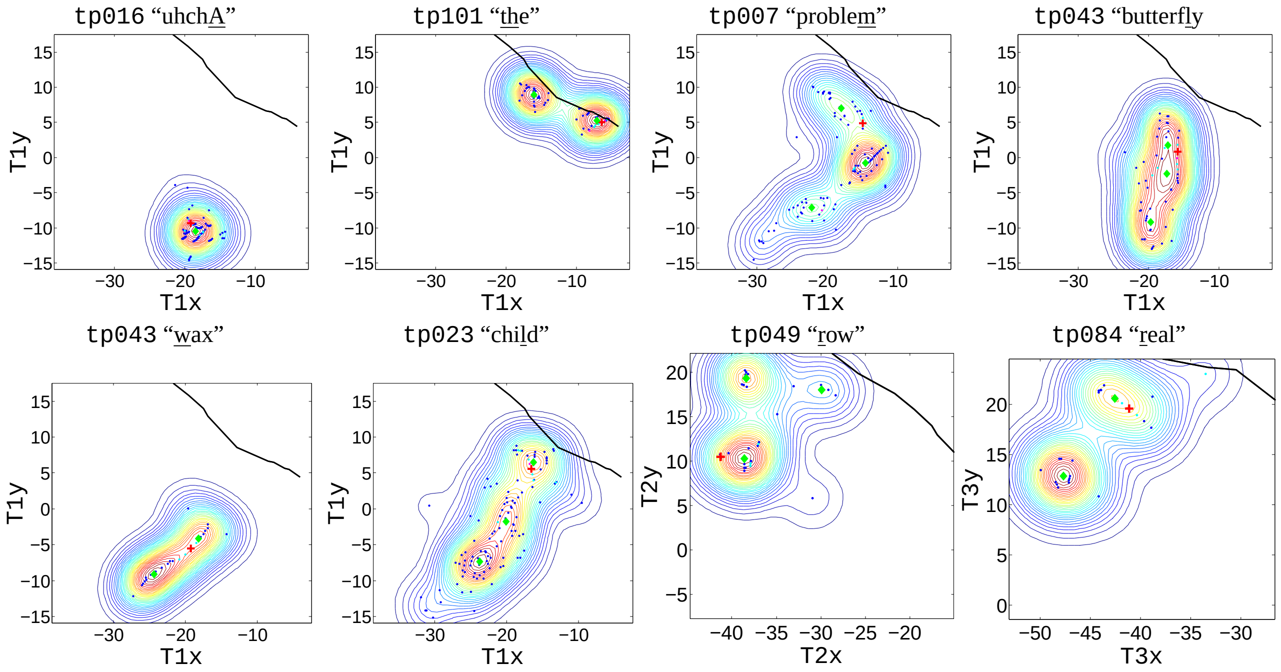

Fig. 5. Sample plots of the inverse sets (blue dots) for a given sound $\mathbf{y}_n$ in the XRMB database (speaker jw11) in the space of T1 to T3, density contours, modes (green dot) and palate (black line). The red mark is the articulatory vector $\mathbf{x}_n$ corresponding to the sound.