Articulatory inversion of American English /ɹ/ by conditional density modes

Chao Qin and Miguel Á. Carreira-Perpiñán

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

Abstract

Although many algorithms have been proposed for articulatory inversion, they are often tested on synthetic models, or on real data that shows very small proportions of nonuniqueness. We focus on data from the Wisconsin X-ray microbeam database for the American English /ɹ/ displaying multiple, very different articulations (retroflex and bunched). We propose a method based on recovering the set of all possible vocal tract shapes as the modes of a conditional density of articulators given acoustics, and then selecting feasible trajectories from this set. This method accurately recovers the correct /ɹ/ shape, while a neural network has errors twice as large.

Index Terms: acoustic-to-articulatory mapping, density models, neural networks, dynamic programming.

1. Introduction

Articulatory inversion, the problem of recovering the sequence of vocal tract shapes that produce a given utterance, is difficult because the mapping from articulators to acoustics is very nonlinear, and because some acoustics can be produced by multiple, different vocal tract shapes (see reviews in [1, 2, 3]). Although many different approaches have been proposed for articulatory inversion, the problem is still not generally solved. In this paper, we focus on two aspects of the problem: we propose an algorithm that directly addresses the multivalued nature of the inverse mapping, and we demonstrate it with real articulatory data for a well-known case that clearly display multiple articulations, the American English /ɹ/ (see [4] and references therein).

Much of the early evidence on multiple articulations comes from synthetic models (e.g. [5]) or unnatural conditions (e.g. bite-block experiments), with less actual evidence from normal, real speech. Our recent work [6, 7], based on large-scale articulatory databases (Wisconsin XRMB [8], MOCHA [9]), shows that only a small portion (around 15%) of all acoustic frames correspond to more than one vocal tract shape. One such case is the American English /ɹ/. The performance of inversion algorithms has been often evaluated with synthetic datasets [5], or with real datasets for which nonuniqueness is of lesser or no importance (for example, with vowels). Other work (e.g. [10]) has used a large, real articulatory dataset (MOCHA) but the performance with the small proportion of it that shows nonuniqueness was not quantified. In section 3 we focus exclusively on utterance subsequences containing /ɹ/, thus with a large proportion of nonuniqueness.

Many of the existing inversion algorithms estimate univalued mappings that provide a single vocal tract shape by applying a nonlinear mapping (such as a neural network) to a given acoustic frame [1, 10]. Asymptotically this is equivalent to estimating the distribution of vocal tract shapes conditioned on the acoustic frame and taking its mean. These methods essentially ignore the existence of multiple articulations; consequently, when multiple, different shapes exist (as for /ɹ/), they return an average shape that is incorrect (see our experiments). However, when applied to a dataset containing little nonuniqueness, this problem is masked, as they do perform well with uniquely determined shapes. In section 2, we propose a method (based on [11, 2]) that directly addresses the nonuniqueness by explicitly estimating all the modes (rather than the mean) of this conditional distribution. If the density model is estimated using acoustic and articulatory data from normal speech, these candidate shapes are (in principle) both *feasible and typical*: they satisfy physical limitations (such as the tongue not penetrating the palate) and correspond to the patterns and idiosyncrasies of normal speech (unlike, say, bite-block speech). The reason is that the modes of a distribution are located on high-density areas of the space—unlike the mean, which can sit in infeasible or atypical areas in between modes. The sequence of vocal tract shapes is finally determined among all the candidates at each frame by minimising a smoothness constraint (inspired by the economy of skilled movements) by dynamic programming.

Related work. There exist several methods that address the nonuniqueness more directly by considering some type of temporal constraints [1, 12, 13]. The closest approach to ours are codebook methods [1, 12]. These create a very large codebook (10^5+ entries) of articulatory and acoustic shapes by finely sampling the articulatory input of a synthetic vocal tract model and computing its acoustic output. This codebook is then searched using each acoustic frame as index, and dynamic programming is used to return a smooth articulatory trajectory. The fundamental problem of this approach is the difficulty of constructing the codebook and the slowness of its search. It is difficult to generate a comprehensive set of feasible, realistic shapes typical of normal speech by sampling a synthetic model—by its very nature it is an approximation. In order to represent well the very nonlinear articulatory-acoustic manifold, many codebook vectors are required. Quantising this (e.g. with k-means) often produces infeasible shapes that require a complex postprocessing, and the final result suffers from quantisation error. One approach to reduce the time and space complexity is to cluster the codebook into regions where the articulatory-acoustic mapping is a bijection and fit a neural net in each such region [12], but doing this reliably this is very difficult. Our approach of learning a continuous density model from real articulatory-acoustic vector pairs eliminates some of these problems: high density is naturally assigned to feasible, typical areas of the articulatory-acoustic manifold (as defined by the dataset); and the density model can be characterised with a relatively small number of parameters, yet quantisation error is eliminated. In fact, the codebook can be seen as a coarse, nonparametric density estimate (a multidimensional histogram), and the codebook search corresponds to finding the modes of our conditional density.

2. Inversion by conditional density modes

The method of [11, 2] offers a general framework for missing data problems. We specialise this inverse problems as follows. In a first stage performed offline, given a training set of articulatory-acoustic vector pairs $\{(\mathbf{x}_n, \mathbf{y}_n)\}$ (obtained from the Wisconsin XRMB), we learn a Gaussian-mixture density model for this data. This can be done with the EM algorithm, or we can use a kernel density estimate if the dataset is not large (our case for the American English /ɹ/). We will use this joint density model to derive conditional densities $p(\mathbf{x}|\mathbf{y}_n)$ given an acoustic frame $\mathbf{y}_n$. For Gaussian mixtures with diagonal covariances, this is a trivial computation. It is also possible to learn directly a conditional density model $p(\mathbf{x}|\mathbf{y})$, for example with a mixture density model [10].

To invert a given acoustic sequence $\mathbf{y}_1, \dots, \mathbf{y}_T$, we first find for each frame all the modes of the conditional density $p(\mathbf{x}|\mathbf{y}_t)$. We do this with the mean-shift algorithm of [14], where a fixed-point iteration is initialised from each centroid of the conditional mixture. Each mode represents a candidate vocal tract shape for the acoustic frame $\mathbf{y}_t$. Then, we minimise the following objective function over the set of all modes at each frame:

$$\mathcal{C}(\mathbf{x}_1, \dots, \mathbf{x}_T) = \textstyle\sum_{t=1}^{T-1} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|. \qquad (1)$$

This is a shortest-path problem that can be solved in $\mathcal{O}(T\nu^2)$ with dynamic programming (where ν is the average number of modes at each frame). The articulatory trajectory returned represents a minimum-energy sequence of motions and represents the fact that physical articulators move slowly. If a forward, acoustic-to-articulatory mapping $\mathbf{f}$ is available (this could be learned from the training set $\{(\mathbf{x}_n, \mathbf{y}_n)\}$), a second term $\lambda \sum_{t=1}^{T} \|\mathbf{f}(\mathbf{x}_t) - \mathbf{y}_t\|$ (with a weight $\lambda > 0$) can be added to the objective (1) to eliminate spurious density modes $\mathbf{x}_t$ that do not map near the target acoustic vector.

The computational bottleneck of the method is the mode-finding step (depending on the number of components in the mixture). This can be drastically accelerated (at a small approximation error) by thresholding out mixture components far from the acoustic vector $\mathbf{y}_t$, and by accelerated variations of mean-shift [15].

By the nature of Gaussian mixtures, the number of modes returned at each frame is finite. With redundant systems, in which excess degrees of freedom result in a continuous set of inverses, the modes will be a quantised version of this.

In the experiments, we will refer to this algorithm as `dpmode`, and compare it to: (1) the conditional mean (`cmean`) of the density (which reconstructs $\mathbf{x}_t$ with $\mathrm{E}\{\mathbf{x}|\mathbf{y}_t\}$ independently for each frame); (2) a neural network (`rbf`), which is asymptotically equivalent to `cmean`; and (3) to an oracle method `cmode` where the dynamic programming picks the optimal mode at each frame (i.e., the one closest to the true $\mathbf{x}_t$); this provides a lower bound in the error achievable by `dpmode`.

3. Experiments

Dataset. We use data from the American English /ɹ/ from speaker `jw11` in XRMB. /ɹ/ is a sound with well-documented nonuniqueness both within and across speakers, where the tongue shapes have traditionally been divided into contrasting categories of retroflex (tongue tip raised, tongue dorsum lowered) and bunched (tongue dorsum raised, tongue tip lowered), though there really seems to exist a continuum between them [4, 17] (see fig. 1). Since there are no phonetic labels available in the XRMB, we manually choose frames corresponding to /ɹ/ from the entire database containing $45\,760$ paired articulatory-acoustic frames (as in [6], we use 20-order LPC as acoustic features). We validate them by listening to the acoustics. Most frames in this dataset correspond to initial /ɹ/, e.g. "right" and "row"; some correspond to middle /ɹ/, e.g. "program", which last shortly. The final dataset of /ɹ/ (fig. 1 shows the various tongue shapes it can adopt) contains a training set of 402 frames and 6 test trajectories from different utterances. Among the latter, 3 are retroflex (e.g. "right" in `tp099`, "roll" in `tp096`) and the other 3 bunched (e.g. "rag" in `tp017`, "row" in `tp050`). Each trajectory contains a small stretch of /ɹ/ and possibly its neighboring sounds. Given the low relative frequency of /ɹ/ among all sounds, our dataset must necessarily be small. The dataset is available from the authors.

Methods. We compare `dpmode` with `cmean`, `rbf`, and `cmode`. We use a Gaussian kernel density estimate for the joint density, from which we derive the conditional density $p(\mathbf{x}|\mathbf{y})$. The bandwidth was set to $\sigma = 11$ in pilot experiments.

Modes of the conditional density. Figs. 4–5 (upper panel) show the conditional modes over time for all testing trajectories. The middle frames in each test trajectory clearly show a multimodal conditional density, and the modes reliably identify both retroflex and bunched shapes, though occasionally additional, spurious modes exist. In these middle frames, `cmean` is the average of the two canonical shapes, which is generally an invalid shape. The start and end frames tend to show a unimodal distribution, which implies the neighbouring sounds have little nonuniqueness. In these cases, the conditional modes still identify the correct shape, but `cmean` performs well. This demonstrates that the density method is effective at identifying the multiple articulations that correspond to a given acoustics.

Inversion results. Figs. 4–5 (lower panel) show the reconstructions at each frame. `dpmode` significantly outperforms `cmean` and `rbf` and picks the correct shapes (either retroflex or bunched but not in between). Fig. 3 plots the aggregated inversion errors on all test trajectories with each methods. On average, `dpmode` achieves an RMSE of less than 1.3 mm while `cmean` or `rbf` have an RMSE of over 1.9 mm. The advantage of `dpmode` is strongest in reconstructing the tongue coils. The RMSE values by `dpmode` in the presence of pervasive nonuniqueness are comparable with other methods that achieved RMSE of 1.6 to 1.9 mm in tasks with little nonuniqueness [16, 10]. The correlation values for some articulators (`UL`, `MNI`, `MNM`) are remarkably low for all methods; the reason is that in our short utterances those articulators barely move. The average RMSE (correlation) for the tongue and all coils are as follows:

	`rbf`	`cmean`	`cmode`	`dpmode`
Tongue	2.66 (0.67)	3.08 (0.55)	1.19 (0.94)	1.20 (0.94)
All	2.07 (0.52)	2.26 (0.48)	1.27 (0.72)	1.30 (0.71)

The `dpmode` value is very close to the `cmode` one, indicating the dynamic programming is effective at selecting the right modes from the ones provided by the density model.

4. Conclusions

We have proposed an articulatory inversion algorithm that uses a density model to predict (possibly multiple) feasible, typical vocal tract shapes for a given acoustics, and disambiguates a sequence by choosing the smoothest path among these shapes. The algorithm correctly recovers either a retroflex or a bunched

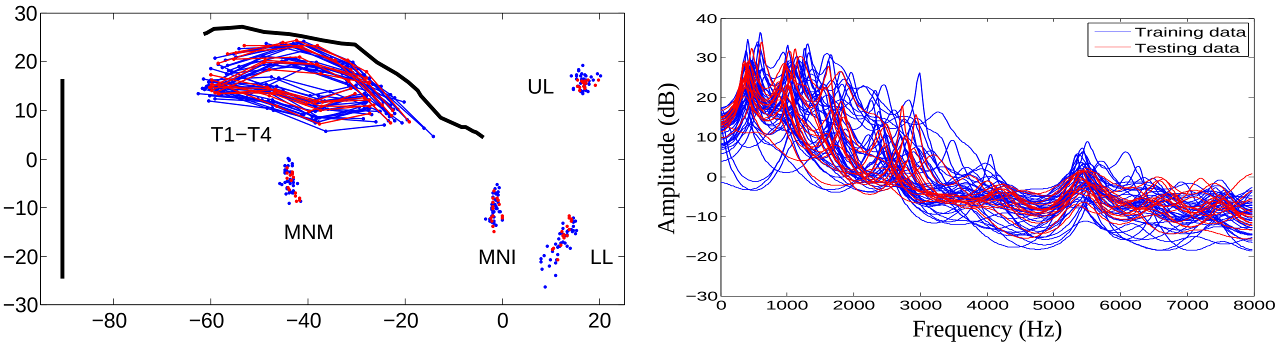

Fig. 1. *Left*: datasets in the articulatory space (2D position in mm of each coil, with tongue coils `T1` to `T4` connected by line segments). *Right*: datasets in the acoustic space (spectral envelopes). Only a subset of frames shown to avoid clutter.

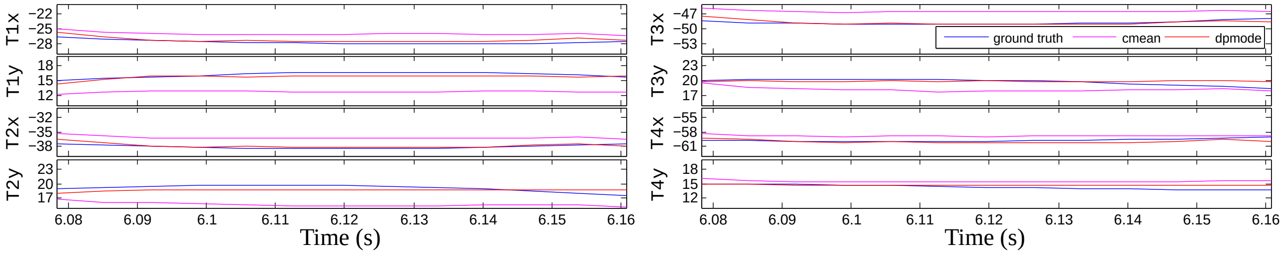

Fig. 2. Temporal true trajectories (blue) and their reconstructions with `dpmode` (red) and `cmean` (magenta) for the bunched /ɹ/ in utterance `tp050` “r̲ow” (for coils `T1` to `T4`). `rbf` (not shown to avoid clutter) performs similarly to `cmean`.

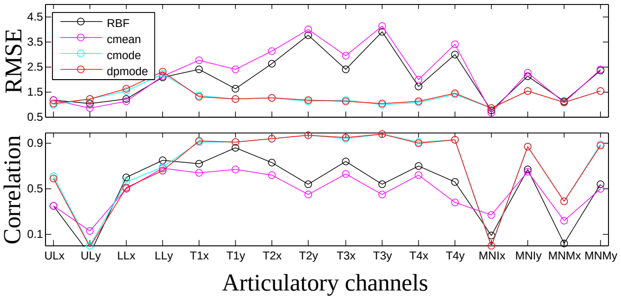

Fig. 3. Inversion error (RMSE in mm) and correlation (in $[-1, 1]$) per articulator channel for each method.

shape for the American English /ɹ/, while a neural network recovers an incorrect average of both. Since the algorithm is computationally more costly than a neural network, but nonunique articulations are overall infrequent in speech, a practical strategy may be to apply the algorithm selectively by detecting first the presence of nonuniqueness from the acoustics.

Acknowledgments. Work funded by NSF award IIS–0711186.

5. References

[1] J. Schroeter and M. M. Sondhi, “Techniques for estimating vocal-tract shapes from the speech signal,” *IEEE TSAP*, vol. 2, 1994.

[2] M. Á. Carreira-Perpiñán, “Continuous latent variable models for dimensionality reduction and sequential data reconstruction,” Ph.D. dissertation, University of Sheffield, UK, 2001.

[3] S. Maeda, M.-O. Berger, O. Engwall, Y. Laprie, P. Maragos, B. Potard, and J. Schoentgen, “Acoustic-to-articulatory inversion: Methods and acquisition of articulatory data,” Nov. 15 2006, EU Project no. 2005–021324.

[4] C. Y. Espy-Wilson, S. E. Boyce, M. Jackson, S. Narayanan, and A. Alwan, “Acoustic modeling of American English /ɹ/,” *JASA*, vol. 108, no. 1, pp. 343–356, 2000.

[5] 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,” *JASA*, pp. 1535–1555, 1978.

[6] C. Qin and M. Á. Carreira-Perpiñán, “An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping,” in *Proc. Interspeech*, 2007, pp. 74–77.

[7] ——, “The geometry of the articulatory region that produces a speech sound,” in *43rd Annual Asilomar Conference on Signals, Systems, and Computers*, 2009, pp. 1742–1746.

[8] J. R. Westbury, *X-Ray Microbeam Speech Production Database User’s Handbook V1.0*, University of Wisconsin, Jun. 1994.

[9] A. A. Wrench, “A multi-channel/multi-speaker articulatory database for continuous speech recognition research,” in *Phonus 5*, Institute of Phonetics, University of Saarland, 2000, pp. 1–13.

[10] K. Richmond, S. King, and P. Taylor, “Modelling the uncertainty in recovering articulation from acoustics,” *Computer Speech and Language*, vol. 17, no. 2–3, pp. 153–172, Apr.–Jul. 2003.

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

[12] M. G. Rahim, C. C. Goodyear, W. B. Kleijn, J. Schroeter, and M. M. Sondhi, “On the use of neural networks in articulatory speech synthesis,” *JASA*, vol. 93, no. 2, pp. 1109–1121, 1993.

[13] T. Toda, A. W. Black, and K. Tokuda, “Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model,” *Speech Communication*, 2008.

[14] M. Á. Carreira-Perpiñán, “Mode-finding for mixtures of Gaussian distributions,” *IEEE TPAMI*, vol. 22, no. 11, pp. 1318–1323, 2000.

[15] ——, “Acceleration strategies for Gaussian mean-shift image segmentation,” in *CVPR*, 2006, pp. 1160–1167.

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

[17] J. R. Westbury, M. Hashi, and M. J. Lindstrom, “Differences among speakers in lingual articulation for American English /ɹ/,” *Speech Communication*, vol. 26, no. 3, pp. 203–226, Nov. 1998.

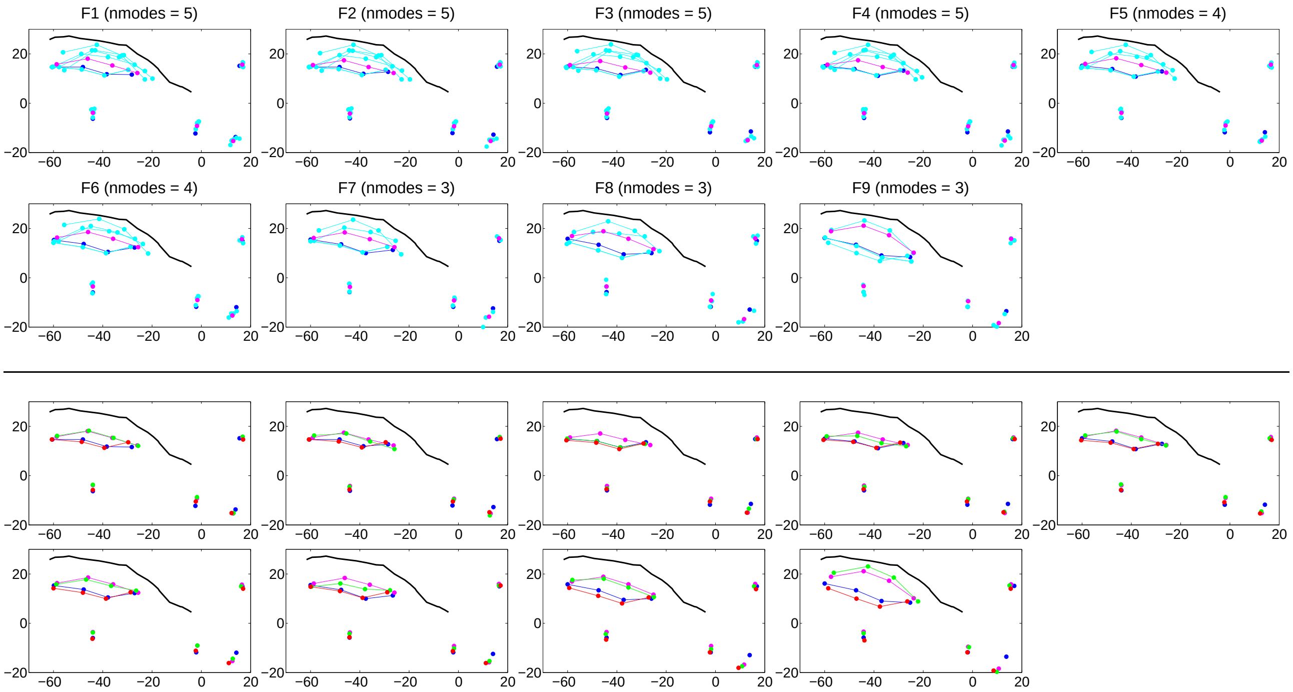

Fig. 4. Sequence of reconstructions for the retroflex /ɹ/ in utterance `tp096` "roll". *Top panels*: plots of the conditional modes (cyan; number of modes above each plot), `cmean` (magenta) and true value (blue) of all articulators. Tongue coils `T1`–`T4` are connected by segments. *Bottom panels*: plots of the reconstructions by all methods at each time frame: `dpmode` (red), `cmean` (magenta), `rbf` (green), and true (blue).

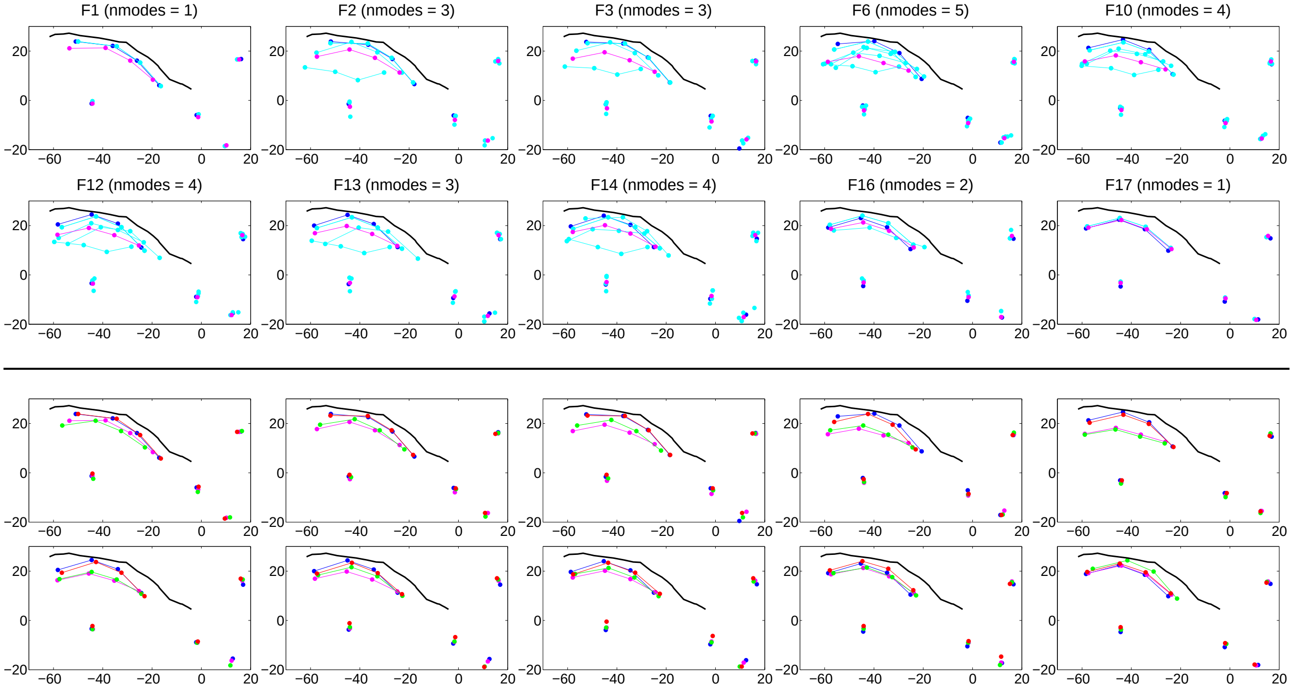

Fig. 5. Like fig. 4 but for the bunched /ɹ/ in the utterance `tp040` "rag".