An Empirical Investigation of the Nonuniqueness in the Acoustic-to-Articulatory Mapping

Chao Qin and Miguel Á. Carreira-Perpiñán

Dept. of Computer Science and Electrical Engineering
OGI School of Science and Engineering, Oregon Health and Science University
20000 NW Walker Road, Beaverton, OR 97006, USA
{cqin,miguel}@csee.ogi.edu

Abstract

Articulatory inversion is the problem of recovering the sequence of vocal tract shapes that produce a given acoustic speech signal. Traditionally, its difficulty has been attributed to nonuniqueness of the inverse mapping, where different vocal tract shapes can produce the same acoustics. However, evidence for the nonuniqueness has been restricted to theoretical studies, or to data from atypical speech or very specific sounds. We present a systematic large-scale study using articulatory data for normal speech from the Wisconsin XRDB. We find that nonuniqueness does exist for some sounds, but that the majority of normal speech is produced with a unique vocal tract shape.

Index Terms: acoustic-to-articulatory mapping, articulatory inversion, X-ray microbeam database

1. Introduction

The acoustic-to-articulatory mapping problem, or articulatory inversion, consists of recovering the sequence of vocal tract shapes that produce a given acoustic speech signal [1]. A solution (even partial) of this problem would have important applications in speech recognition, synthesis and coding, and language learning and speech therapy. The problem is hard because the inverse mapping is strongly nonlinear and multivalued, i.e., multiple vocal tract (VT) shapes can produce the same acoustics—unlike the forward, or articulatory-to-acoustic, mapping, which though nonlinear is univalued. Many techniques have been proposed over the years to solve it (see [1, 2, 3] for reviews), some using mapping approximators (e.g. neural nets), while others try to address the multivalued nature of the mapping directly (e.g. codebooks or density models).

In this paper, we perform a systematic study of the nonuniqueness of the inverse mapping using real human speech production data (for a single speaker). The traditional evidence for the nonuniqueness of the inverse mapping can be summarised as follows.

Articulatory compensation   One example are bite-block experiments, where a speaker is capable of producing acoustic signals perceptually close to the intended sounds even when the jaw is fixed in an unnatural postion by a bite-block [4]. Another example are ventriloquists, who can produce intelligible speech without moving their lips.

Theoretical or modelling studies   Webster's horn equation, a second-order linear differential equation, describes wave propagation in the vocal tract. It can be shown that for lossless vocal tracts with fixed boundary conditions, the area functions $A(x)$ and $1/A(L-x)$ (where L is the length of the vocal tract) produce the same acoustics. Computational studies based on articulatory models also indicate nonuniqueness, e.g. Atal et al. [5, 1] manipulated an articulatory model to demonstrate how very different VT shapes could produce acoustics with nearly identical values of the first three formants.

The American English /ɹ/   Speakers of rhotic dialects of American English use many different articulatory configurations for the approximant consonant /ɹ/ (the 'r' as in 'perk' or 'rod'), which are all acoustically characterised by an extremely low frequency of the third formant (often close to that of the second formant). These configurations differ most in the palatal constriction and 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 [6, 7]. These different configurations occur both within and across speakers: some speakers may use one type of configuration exclusively while others switch between two or three different types in different phonetic contexts and according to prosodic variations.

However, the question of to what extent does nonuniqueness occur in normal human speech remains unclear, because (1) articulatory models are only crude approximations to the geometry of the human VT, (2) the vocal acrobatics of ventriloquists and bite-block experiments are not representative of the typical behaviour of the human VT, and (3) the /ɹ/ remains an isolated, exceptional case. The only way to answer the question is by studying large amounts of human articulatory data, which have become available in recent years (Wisconsin XRDB [8], MOCHA database [9]— though none of these two databases provide information about the lower VT). We approach the problem by analysing simultaneously recorded articulatory and acoustic data using statistical machine learning techniques. The goal of this paper is to report a preliminary such study. We describe our methodological setup, give experimental results and discuss them.

2. Methodological setup

The XRDB [8] provides pairs $(\mathbf{x}, \mathbf{y})$ for articulatory vectors $\mathbf{x}$ and acoustic vectors $\mathbf{y}$. Our basic idea is to fix one acoustic vector $\mathbf{y}_n$ and search the database for articulatory vectors $\{\mathbf{x}_m\}$ that approximately map to $\mathbf{y}_n$ (*inversion*). Then, a *clustering* algorithm determines whether the point cloud $\{\mathbf{x}_m\}$ is unimodal

or not. Repeating this for every acoustic vector $\mathbf{y}$ allows exploration of the nonuniqueness of the inverse mapping for a full range of sounds. Let us consider each step in detail.

Inversion   This requires a distance between acoustic vectors $\mathbf{y}$. We use LPC coefficients because they are closely related to the vocal tract spectral envelope, which allows direct visualisation of spectral differences and formant structures; and because they have been shown to perform well with articulatory inversion techniques [10]. As acoustic distance $d(\mathbf{y}, \mathbf{y}')$ we use the Itakura distance [11], which emphasises 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. Next, we fix a reference distance r for which we consider two acoustic vectors to be roughly the same sound. Since we observed that feature vectors for consecutive frames in an utterance are a distance 0.06–0.1 apart, we chose $r = 0.4$. While the specific set $\{\mathbf{x}_m\}$ of articulatory vectors returned for an acoustic vector $\mathbf{y}$ depends on r, the output of the clustering algorithm does not as long as r is neither too small (too few $\mathbf{x}$ are returned, erasing clusters) nor too large (too many $\mathbf{x}$ are returned, obscuring clusters). An approximate inversion of this type is unavoidable given the discrete nature of the data. Roweis [12] proposed a similar search strategy using a Mahalanobis distance but returning a fixed number K of neighbours, which is much harder to estimate since K depends on the particular acoustic vector $\mathbf{y}$. In summary, the inversion for an acoustic vector $\mathbf{y}$ returns a set $\{\mathbf{x}_m : d(\mathbf{y}_m, \mathbf{y}) \le r\}$.

Clustering   The large size of the database requires an automatic procedure to determine whether the resulting point cloud $\{\mathbf{x}_m\}$ is multimodal. The complex shape of the cloud and the fact that we do not know how many clusters it contains prevents us from using parametric models such as a Gaussian mixture. Instead, we use a nonparametric kernel density estimate with Gaussian kernel and bandwidth σ, i.e., we define a density $p(\mathbf{x}) \propto \sum_m G\left(\frac{\mathbf{x}-\mathbf{x}_m}{\sigma}\right)$. Inspection of many point clouds suggested using $\sigma = 6$ mm (again, the results were not sensitive to small variations of σ). Finally, we find the modes of $p(\mathbf{x})$ using a mean-shift algorithm [13], which iterates a hill-climbing algorithm initialised at every $\mathbf{x}_m$ and collects all the resulting, distinct modes. Strongly unimodal clouds $\{\mathbf{x}_m\}$ yield a single mode while clustered or elongated clouds yield several modes.

3. Experimental results

We use the Wisconsin X-ray microbeam database (XRDB [8]), which records, simultaneously with the acoustic wave, the positions of 8 pellets in the midsaggital plane of the VT (see fig. 1), sampled at 147 Hz, for various types of speech (isolated words, prose, etc.). The XRDB measurement error for the pellets is 0.7 mm. 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 well; 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 45 000+ vectors $(\mathbf{x}, \mathbf{y})$ with $\mathbf{x} \in \mathbb{R}^{16}$ and $\mathbf{y} \in \mathbb{R}^{20}$.

Fig. 2 shows the result of inverting an acoustic vector $\mathbf{y}$ corresponding to the sound /θ/. The cloud $\{\mathbf{y}_m\}$ of acoustic vectors at distance $\le r = 0.4$ is strongly unimodal but the cor-
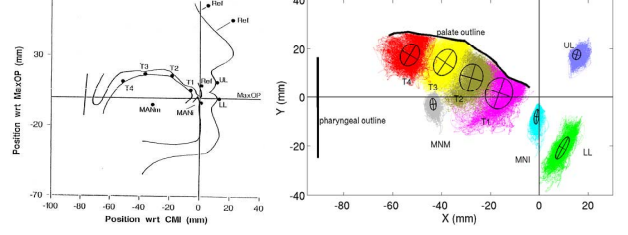

Figure 1: *Left*: pellet locations in the XRDB. *Right*: plot of the entire dataset for speaker jw11; each pellet's data uses a different colour and shows a contour line of one standard deviation centred at its mean.

responding cloud of articulatory vectors $\{\mathbf{x}_m\}$ is clearly multimodal, and our mode-finding algorithm detects this (modes in green). The LPC spectral envelope confirms the acoustic similarity of the cloud and the VT representation shows the corresponding VT shapes to be significantly different. We also verified that adding dynamic features (Δ, Δ^2) to the LPC vector did not separate the acoustic cloud (if they did, this might have disambiguated the VT shape). Thus, widely different VT shapes produce approximately the same acoustic sound /θ/, indicating nonuniqueness of the instantaneous inverse mapping in this particular case.

Fig. 3 shows several other sounds for which we find multimodality (/ɪ/, /l/, /w/), again clearly showing similar spectral envelopes (with Itakura distances < 0.1) but qualitatively different VT shapes, in particular the tongue. In contrast, fig. 4 shows several sounds for which the mapping is unimodal (/æ/, /uː/, /y/): the VT shapes are essentially the same for a given acoustic vector (with Itakura distances up to 0.5). In each sound, the acoustic vectors shown come from different utterances and different contexts, i.e., they are not just consecutive frames in the same utterance.

While this demonstrates the use of multiple VT shapes to produce the same acoustic sound for certain sounds, how frequently does nonuniqueness occur overall? We found that around 5% of the acoustic vectors yield a multimodal cloud in articulatory space. This suggests that, while nonuniqueness does happen, by and large it is an infrequent situation.

Fig. 5 gives further indirect evidence for nonuniqueness. For each acoustic vector we computed the standard deviation per dimension of articulatory space (specifically, $\frac{1}{16}\sqrt{\lambda}$ where λ is the top eigenvalue of the 16×16 covariance matrix of the articulatory cloud, i.e., the principal variance) and plotted their distribution. If the articulatory clouds were strongly unimodal, we would expect a symmetric distribution, but instead the distribution is skewed with a long right tail. This indicates that, while most articulatory clouds have a standard deviation ≈ 7 mm, a small proportion has a larger std deviation (up to 25 mm).

4. Conclusion

Our experiments give direct, quantitative evidence of the presence of nonuniqueness of the inverse mapping in normal human speech; but also suggest that, while some sounds are indeed produced in multiple ways, most times a unique VT shape is used. This is also consistent with the fact that most techniques for articulatory inversion yield similar reconstruction errors (around 2 mm), whether they use methods for univalued
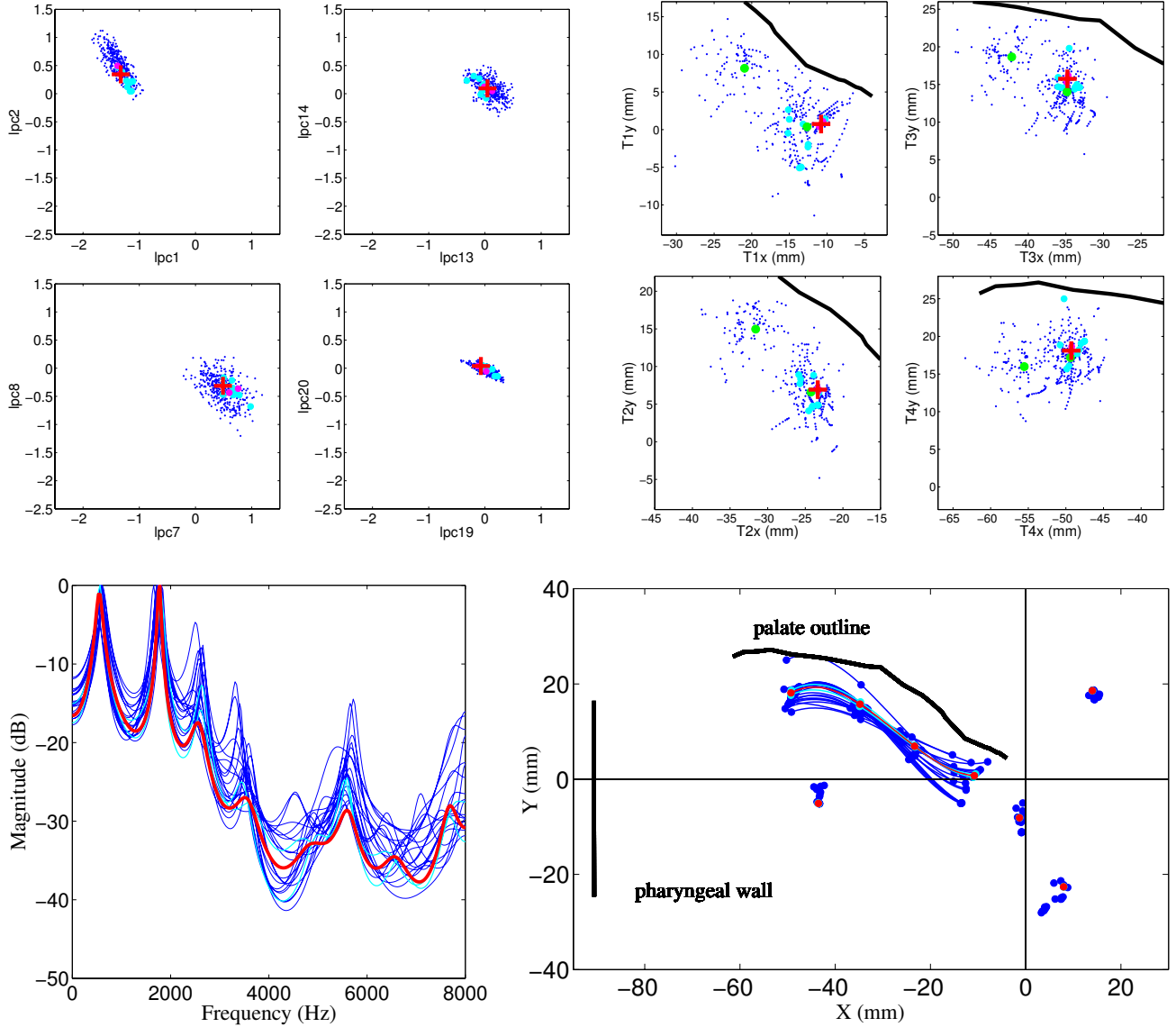
Figure 2: An example of nonuniqueness in articulatory space for a reference acoustic vector $\mathbf{y} = /\theta/$ (red, utterance tp004 "<u>th</u>ings"). *Top left*: point cloud $\{\mathbf{y}_m\}$ in acoustic space with $d(\mathbf{y}_m, \mathbf{y}) \leq r$ (containing ~ 400 points), clearly unimodal. For further resolution, we show points at distances r, $r/2$, $r/4$ in different colours (blue, cyan, magenta, respectively). *Top right*: corresponding point cloud $\{\mathbf{x}_m\}$ in articulatory space, clearly multimodal (modes in green). *Bottom left*: $\{\mathbf{y}_m\}$ as spectral envelopes. *Bottom right*: $\{\mathbf{x}_m\}$ as VT shapes (we fit a spline to the 4 tongue pellets for visualisation). The bottom plots show a subset of the curves to avoid clutter.

mappings (e.g. neural nets) or for multivalued mappings (e.g. codebooks or density models). An open direction of future research is to quantify nonuniqueness over different types of sounds, e.g. vowels (which appear to have a one-to-one mapping), consonants, liquids, etc.

While we think our work is the first systematic, large-scale study of nonuniqueness, it is important to note its limitations, which future studies may improve upon: (1) We considered a single speaker from the XRDB, and did not study the relevance of the acoustic context for the cases where nonuniqueness occurred. (2) The VT representation provided by the XRDB is incomplete, lacking data about the lower VT. It is thus possible that sounds that are produced with the same upper VT

shape do differ in the lower VT, thus increasing the frequency of nonuniqueness. Techniques such as dynamic MRI may offer a full representation of the VT in the future.

5. Acknowledgements

6. References

[1] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal tract shapes from the speech signal," *IEEE*
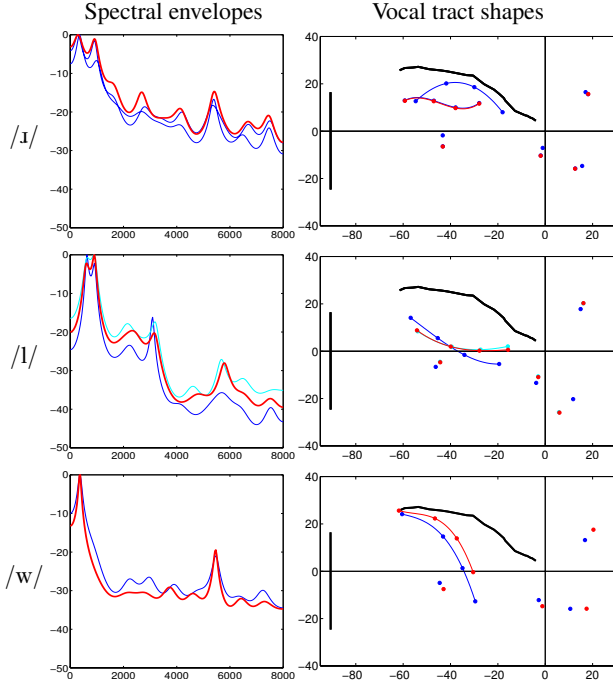
Figure 3: Examples of sounds showing multimodality. Spectral envelopes: magnitude (dB) vs frequency (Hz); VT shapes: X (mm) vs Y (mm). Utterances: /ɹ/, `tp009` "<u>r</u>ow"; /l/, `tp037` "<u>l</u>ong"; /w/, `tp044` "<u>w</u>ork". For /ɹ/, the well-known retroflex and bunched tongue shapes [6, 7] are evident. We show only a very small subset of the curves and points in the cloud to avoid clutter; colour scheme as in fig. 2.
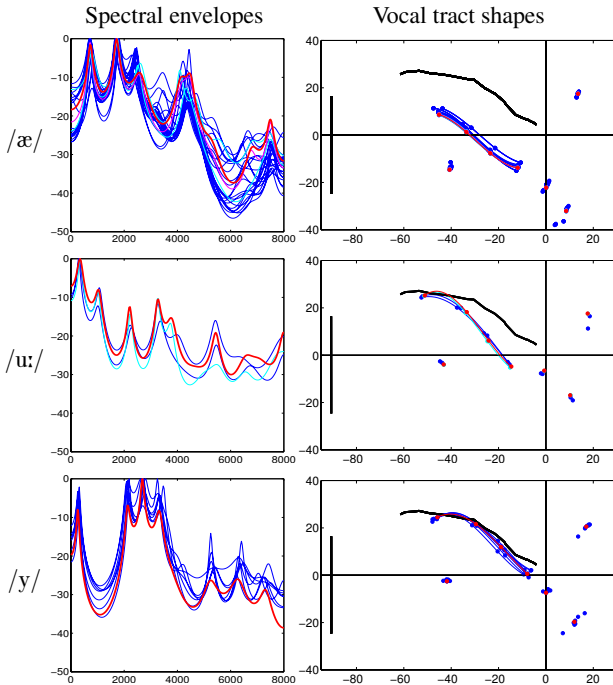

Figure 4: Examples of sounds with unimodality. Utterances: /æ/, `tp001` "h<u>a</u>s"; /uː/, `tp001` "sch<u>oo</u>l"; /y/, `tp040` "y<u>o</u>u".
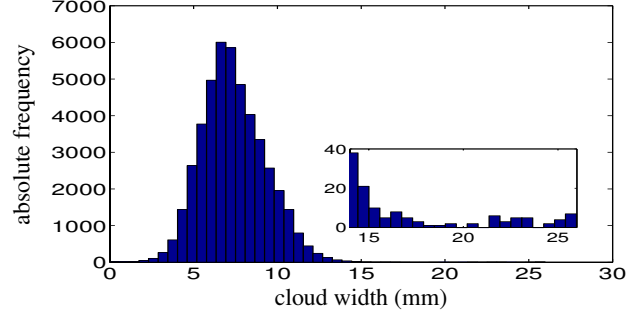

Figure 5: Histogram of the cloud width (standard deviation per dimension) for the inverse mapping. The inset blows up the right tail.

Trans. ASSP, vol. 2, no. 1, pp. 133–150, Jan. 1994.

[2] Miguel Á. Carreira-Perpiñán, *Continuous Latent Variable Models for Dimensionality Reduction and Sequential Data Reconstruction*, Ph.D. thesis, Dept. of Computer Science, University of Sheffield, UK, 2001.

[3] K. Richmond, *Estimating Articulatory Parameters from the Acoustic Speech Signal*, Ph.D. thesis, University of Edinburgh, 2001.

[4] B. Lindblom, J. Lubker, and T. Gay, "Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation," *J. of Phonetics*, vol. 7, no. 2, pp. 147–161, 1979.

[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," *J. Acous. Soc. Amer.*, vol. 63, no. 5, pp. 1535–1555, May 1978.

[6] 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.

[7] C. Y. Espy-Wilson, S. E. Boyce, M. Jackson, S. Narayanan, and A. Alwan, "Acoustic modeling of American English /ɹ/," *J. Acous. Soc. Amer.*, vol. 108, 2000.

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

[9] A. A. Wrench, "A multi-channel/multi-speaker articulatory database for continuous speech recognition research," in *Phonus*. Institute of Phonetics, 2000.

[10] C. Qin and M. Á. Carreira-Perpiñán, "A comparison of acoustic features for articulatory inversion," in *EUROSPEECH 2007)*, to appear.

[11] Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[12] S. Roweis, *Data Driven Production Models for Speech Processing*, Ph.D. thesis, Caltech, 1999.

[13] M. Á. Carreira-Perpiñán, "Mode-finding for mixtures of gaussian distributions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1318–1323, Aug. 2000.