

Blind Inversion of Multidimensional Functions for Speech Enhancement

John Hogden¹, Patrick Valdez¹, Shigeru Katagiri², Erik McDermott².

1) Modeling, Algorithms, and Informatics Group
Los Alamos National Laboratory, Los Alamos, U.S.A.
hogden@lanl.gov, pfvaldez@lanl.gov

2) NTT Communications Science Laboratories
NTT Corporation, Kyoto, Japan
katagiri@cslab.kecl.ntt.co.jp, mcd@cslab.kecl.ntt.co.jp

Abstract

We discuss speech production in terms of a mapping from a low-dimensional articulator space to low-dimensional manifold embedded in a high-dimensional acoustic space. Our discussion highlights the advantages of using an articulatory representation of speech. We then summarize mathematical results showing that, because articulator motions are bandlimited, a large class of mappings from articulation to acoustics can be blindly inverted. Simulation results showing the power of the inversion technique are also presented. One of the most interesting simulation results is that some many-to-one mappings can also be inverted. These results explain earlier experimental results that the studied technique can recover articulator positions. We conclude that our technique has many advantages for speech processing, including invariance with respect to various nonlinearities and the ability to exploit context more easily.

1. Introduction

Many researchers have proposed learning the mapping between acoustics and the speech articulator positions for use in speech recognition, e.g. [1-4]. We discuss a method called MALCOM [5], which is *blind* in the sense that it appears to learn the mapping between acoustics and articulation using only acoustic data and the knowledge that articulator trajectories are bandlimited. Interestingly, for a set of vowel-to-vowel transitions, correlations between the positions of pellets placed on the tongue body and MALCOM estimates of those positions were between 93% and 97%. However, it was not clearly understood why the correlations were so high. In this paper, we give the first proof that bandlimited articulatory trajectories make blind inversion possible.

MALCOM actually refers to a class of algorithms, not all of which can do the blind inversion. To prevent confusion, the algorithm that accomplishes the blind inversion (called simplified MALCOM in [6]) is hereby given its own name based on the tasks it can perform – Mapping Inversion, Manifold Inference, and Contextual Recovery of Information (MIMICRI). These tasks are illustrated below.

Many of the difficulties we face in speech processing can be succinctly described in terms of a warped manifold

of articulator positions embedded in high-dimensional acoustic space [7]. Figure 1 lays out the fundamental idea. Let the square region in the “articulation” plot represent the (hypothetical) set of all articulator positions that can be used by a particular speaker. In this example, we only show two dimensions of the articulator space, maybe the horizontal and vertical positions of the tongue body, but the actual articulator space would need more dimensions. The curve through the articulator space represents a single articulator trajectory and is shown as a smooth curve to convey that articulator trajectories are band-limited [4].

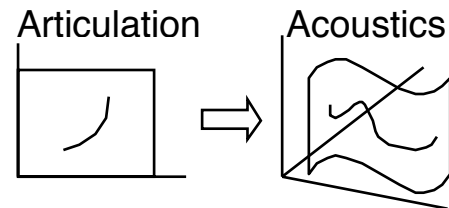


Figure 1: Geometric view of speech production

The “acoustics” plot in Figure 1 shows the (hypothetical) acoustic space that results from mapping the articulator positions to acoustic feature vectors, e.g., melcepstra. Necessarily, the set of acoustic feature vectors has the same intrinsic dimensionality as the articulator space, but is warped and embedded in the high-dimensional acoustic space.

Additive acoustic noise will make some acoustic signals lie off the manifold. This is a problem for recognition algorithms because such algorithms typically involve estimating probability density functions (PDFs) over the acoustic space. Consider that, if we estimate a 20 dimensional PDF over a 10 dimensional manifold, nearly all of the PDF is off the manifold. This implies that many of the PDF parameters we estimate are really estimating the distribution of noise. Thus, recognition performance is not robust to a change in the noise environment [8].

If we can project the manifold onto a space with the same dimensionality as articulation, we should be able to increase noise robustness while also decreasing the number of parameters that need to be estimated by our recognition algorithms. Thus, a dimensionality reduction technique, e.g. [9, 10] may help recognition.

However, note that the warping applied to the manifold will change depending on the speaker and the microphone used to collect the speech data. Dimensionality reduction will not eliminate differences between the nonlinear warpings, but will merely give us low-dimensional representations that are, at best, nonlinearly related to the articulator positions. Being able to invert back to the articulator positions would provide us with the additional advantages of having a signal that is invariant to microphone effects, and, possibly, to speaker differences.

Furthermore, nonlinear transformations typically increase the bandwidth of signals. Suppose that we make a second-order Taylor series approximation of a nonlinear function applied to a signal. Since squaring in the time domain is equivalent to convolving in the frequency domain, we can expect nonlinear transformations to double (or more) the bandwidth of a signal. This also has implications for signal processing. For example, noise in articulator trajectories can be decreased by simply using a low-pass filter. Using a low-pass filter amounts to using context (a convolution window) to get better estimates of the signal. However, in the acoustic domain, using context is much more difficult because increased bandwidth prohibits using simple filtering. In fact, state-of-the-art speech recognition tools use only first-order hidden Markov models.

In the next section, we give a mathematical argument that, since articulator trajectories are bandlimited, any memoryless, one-to-one function of the trajectories can be inverted, even if we are not allowed to directly measure the function, and even if we are not allowed to observe the articulator trajectories.

2. Mapping Inversion -- Mathematics

Assume that an observable signal (e.g. acoustics), $\mathbf{y} = [y_1 \ y_2 \ \dots \ y_T]'$, where the prime denotes transpose, is a memoryless, one-to-one function of an unobservable but constrained signal, $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_T]'$ (e.g. an articulator trajectory). That is, $y_i = f(x_i)$. For the moment, consider the simplified case where $x_i \in \mathfrak{R}$ and $y_i \in \mathfrak{R}$. The set of \mathbf{x} trajectories meeting the constraint are all trajectories that can be written as $\mathbf{x} = \mathbf{C}\mathbf{v}$, $\mathbf{v} = [v_1 \ v_2 \ \dots \ v_n]'$, $v_i \in \mathfrak{R}$, and \mathbf{C} is a constant matrix with T rows and n orthogonal columns, $T > n$. Call this set ξ . Note that bandlimited signals can be represented in this fashion – simply let the columns of \mathbf{C} be sines and cosines. We define a third signal, $\hat{\mathbf{x}} = [\hat{x}_1 \ \hat{x}_2 \ \dots \ \hat{x}_T]'$, where $\hat{x}_i = g(y_i)$. Our goal is to find a $g(\cdot)$ such that for every $\mathbf{x} \in \xi$ we get $\hat{\mathbf{x}} \in \xi$. The proof shows that, under very weak constraints on \mathbf{C} , if we can find such a $g(\cdot)$, then $\mathbf{x} = a\hat{\mathbf{x}} + b$. Which is inversion to within an affine transform. The proof can be extended to allow x_i and y_i to be vectors.

2.1. Proof

By construction, \mathbf{C} has rank n but more than n rows. This implies that some rows of \mathbf{C} are linear combinations of other rows. Without loss of generality, suppose we order the columns of \mathbf{C} and the rows of \mathbf{v} so that rows 1 to n are the independent rows of \mathbf{C} . Then, for $i > n$:

$$c_{ij} = \sum_{k=1}^n \alpha_{ik} c_{kj} \quad (1)$$

Algebraic manipulation shows that, for $i > n$:

$$x_i = \sum_{k=1}^n \alpha_{ik} x_k \quad (2)$$

For all $\mathbf{x} \in \xi$. Similarly, for $i > n$:

$$\hat{x}_i = \sum_{k=1}^n \alpha_{ik} \hat{x}_k \quad (3)$$

for all $\hat{\mathbf{x}} \in \xi$. Let $h(x_i) \equiv g(f(x_i))$, so

$$\hat{x}_i = h(x_i). \quad (4)$$

Substituting 2 into 4 gives, for $i > n$:

$$\hat{x}_i = h\left(\sum_{k=1}^n \alpha_{ik} x_k\right) \quad (5)$$

Substituting 4 into 3 gives, for $i > n$:

$$\hat{x}_i = \sum_{k=1}^n \alpha_{ik} h(x_k) \quad (6)$$

Equating 5 and 6 gives:

$$h\left(\sum_{k=1}^n \alpha_{ik} x_k\right) = \sum_{k=1}^n \alpha_{ik} h(x_k) \quad (7)$$

Note that x_k values are independent for $1 \leq x_k \leq n$, and so take on all values if \mathbf{x} takes on all values in the set ξ .

Eq. 7 is actually many equations, one for every $i > n$. If any one of those equations meets the requirement that at least 2 α_{ik} values are not 0 and $\sum_{k=1}^n \alpha_{ik} \neq 0$, then Eq. 7 constrains $h(\cdot)$ to be affine as long as it is at least piecewise continuous. This is proven by defining a new function:

$$\tilde{h}(x) \equiv h(x) - h(0) \quad (8)$$

and showing that $\tilde{h}(\cdot)$ can be transformed into Cauchy's functional equation [11]:

$$\tilde{h}(x+y) = \tilde{h}(x) + \tilde{h}(y) \quad (9)$$

Cauchy proved that if $\tilde{h}(\cdot)$ is at least piecewise continuous, then $\tilde{h}(\cdot)$ is linear, implying that $h(\cdot)$ is affine, or linear in cases where we can prove that $h(0) = 0$.

2.2. Importance of the proof

The extension of the proof to the case where x_i and y_i are vectors and the dimensionality of y_i is greater than the dimensionality of x_i is directly relevant to inverting the mapping from articulator positions to acoustics. This extension is also relevant for dealing with mappings that

are not memoryless, although a discussion of this point is beyond the scope of this paper.

For many purposes, including speech processing, inverting to within an affine transformation is a significant step. For example, if the unobservable input signal is one-dimensional, then an affine transformation is simply a change in amplitude and the addition of D.C. offset, either of which can be easily removed.

This proof explains why MIMICRI can invert a large class of functions. To use MIMICRI, the transformed trajectories are first vector quantized, which divides the space of y_i values into a set of Voronoi regions. MIMICRI then maps each Voronoi region to a Gaussian PDF over a *continuity map*. The mapping is done to minimize the difference between the sequence of PDF means and the most probable smooth path calculated from the PDFs. The result is that the mapping from Voronoi regions to PDF means is a non-parametric, discrete approximation to $g(\cdot)$. Using more Voronoi regions allows MIMICRI to better approximate any function.

3. Simulations

Some of the assumptions necessary for the proof will not be met in practice. For example, we will not be able to observe the acoustics for all possible articulator trajectories. This simulation shows that MIMICRI inverts multidimensional functions using a small set of data.

3.1. Mapping Inversion

3.1.1. Data

We studied 5 data sets, each comprising 100, 2-second duration \mathbf{x} trajectories through 2-dimensional space. Trajectories were sampled 50 times per second. \mathbf{C} was chosen to contain frequencies from 0 to 12Hz, with an additional column allowing a trend to be added to the signal. The trajectories for each data set were created using \mathbf{v} selected from a uniform distribution over a hypercube with corners at ± 0.5 . The sampled positions were transformed using a Cartesian-to-polar coordinate transformation. Note that this is a discontinuous function. The roughly Gaussian distribution of the sampled points and the distribution of the transformed points for one data set are shown in Figure 2.

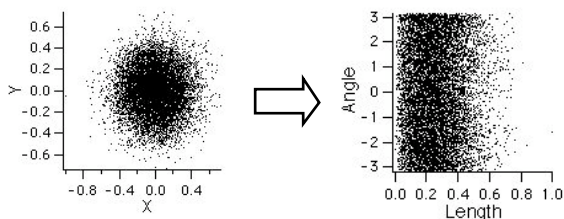


Figure 2: Path points before and after transformation

3.1.2. Procedure

The nonlinearly transformed points in each of the five data sets were vector quantized using between 2 and 256 Voronoi regions. MIMICRI was run on sequences of Voronoi regions. The covariances of all PDFs were constrained to be equal to the identity matrix. Estimated

trajectories were the most probable paths through the continuity map given a sequence of Voronoi regions.

To the extent that MIMICRI inverts $f(\cdot)$ (the Cartesian-to-polar coordinate transformation) we expect our estimated trajectories to be similar to the randomly generated \mathbf{x} trajectories. However, the estimates will be off by some arbitrary affine transformation. To remove the affine transformation, we apply Procrustes analysis [12] to rotate, scale, translate, and reflect the estimated trajectories to best fit the original trajectories, and then calculate correlations between the original trajectories and the estimated trajectories. We would not be able to use Procrustes analysis in a realistic scenario because we would not be able to observe \mathbf{x} . Nonetheless, this method for finding correlations does allow us to determine the extent to which MIMICRI inverted $f(\cdot)$. Correlation, measured by the Pearson r score, is a commonly used measure of similarity, but may not be as familiar as the signal-to-noise ratio. When \mathbf{x} is the signal, $\hat{\mathbf{x}}$ is the signal plus noise, and the noise is zero mean and uncorrelated with the signal, then the signal to noise ratio is:

$$\frac{S}{N} = \frac{r^2}{1-r^2} \quad (10)$$

3.1.3. Results

Correlations between the original trajectories and the transformed trajectories ranged from 0.26% to 2%, showing that the transformation adversely affected the signal. Figure 3 shows that the correlations between the estimated and actual \mathbf{x} trajectories increase rapidly with the number of Voronoi regions, as we would expect.

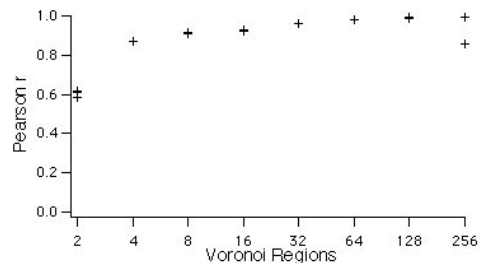


Figure 3: Correlations between estimated and actual paths

Although there are 5 points for each number of Voronoi regions, the results are very consistent and the points overlap each other. For 256 Voronoi regions, the 4 highest correlations ranged from 99.4% to 99.5%, but one correlation of 86% was observed. Aside from the single poor solution, the 256-region MIMICRI improved the signal-to-noise ratio by more than 50 db over doing nothing.

3.2. Manifold inference

3.2.1. Data and procedure

We also simulated a problem where the dimensionality of the data needed to be reduced. In this problem, the unobservable signals were distributed as before, but the transformation was the Cartesian-to-polar transformation

followed by a “Swiss-roll” transformation, which embeds 2-D data in 3-D space. The distribution of transformed data is illustrated in Figure 4.

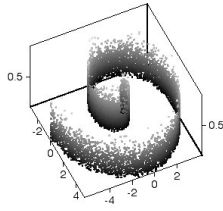


Figure 4: Path points after “Swiss Roll” transformation

3.2.2. Results

As in the previous experiment, MIMICRI accurately recovered the original paths. Using 256 Voronoi regions gave 98% correlations for 4 of the 5 data sets, with one data set giving a lower 91% correlation.

3.3. Contextual recovery of information

3.3.1. Data and procedure

We performed a third simulation to show that MIMICRI can use context to recover information lost in a many-to-one mapping. In this study, the distribution of points in the original trajectories was the same as in the previous two experiments. The many-to-one transformation was simple: points within a radius of 0.6 from the origin were left unchanged and points outside that radius were all transformed to the point [10,10]. Approximately 1.5% of the points were outside the radius. Figure 2 shows that the points outside the radius vary considerably in position.

After running MIMICRI with the covariances set to identity, an extra iteration was run in which the covariance matrices were constrained to be some factor times the identity matrix, where the factor varied between PDFs. We report correlations for only those points outside the radius, but the estimated trajectories were rotated, translated, and scaled to maximize the correlation over all points.

3.3.2. Results

For 256 Voronoi regions, the correlations between the estimated and actual positions of points that were originally outside the radius, i.e. were all mapped to point [10,10], varied between 98.1% and 99.2%. Correlations for points inside the radius were higher. The mapping found by MIMICRI was interesting: Voronoi regions inside the radius were mapped to PDFs with small covariances and the Voronoi region that included point [10,10] was mapped to a PDF with large covariances. Thus, the point [10,10] had little effect on the probability of a path, and context was used to determine the estimated trajectory. Similar results were found when points inside the radius were cubed and points outside were mapped to [10,10].

4. Discussion

This is the first paper to provide an explanation for previously cited empirical observations that MIMICRI recovers the positions of the articulators from acoustics for vowels. It has often been argued that the mapping from

articulator positions to acoustics is many-to-one, and so not invertible. However, we have shown that at least one many-to-one mapping can be inverted using a blind algorithm. Since many-to-one mappings decrease the information in a signal, MIMICRI processing increased the information. It may be that applying MIMICRI to acoustics could increase the information content back to the amount of information in the articulator positions.

This work shows that the smoothness of articulator motions makes them special – we can blindly invert a large class of non-linear functions of smooth trajectories. The implication is that articulator positions recovered from acoustics using MIMICRI should be relatively invariant to nonlinear transformation of the acoustics.

5. Acknowledgements

The authors thank Leonid Gurvits for finding the related proof that applying a nonlinear warping to the set of all low-pass signals will force at least one signal to have energy above the frequency cutoff.

6. References

- [1] J. Frankel and S. King, "ASR -- Articulatory Speech Recognition," *Proc. Eurospeech*, pp. 599-602, 2001.
- [2] C. S. Blackburn and S. Young, "Enhanced speech recognition using an articulatory production model trained on X-ray data," *Computer Speech and Language*, vol. 15, pp. 195-215, 2001.
- [3] S. Roweis, "Data Driven Production Models for Speech Processing," Dissertation, Pasadena, CA: California Institute of Technology, 1999, pp. 238.
- [4] D. Nix, "Machine Learning Methods for Inferring Vocal-Tract Articulation from Speech Acoustics," Dissertation, Boulder, CO: University of Colorado, 1998, pp. 134.
- [5] J. Hogden, "A maximum likelihood approach to estimating speech articulator positions from speech acoustics," *J. Acoust. Soc. America*, vol. 100, pp. 2663, 1996.
- [6] J. Hogden, "Speech processing using maximum likelihood continuity mapping." U.S.A. Patent # 6,052,662: Assigned to the University of California, 2000, pp. 1-22.
- [7] D. Kimber, "Geometric Methods for Nonparametric Modeling of Dynamical Systems," Dissertation, Stanford, CA: Stanford University, 1994, pp. 184.
- [8] S. Das, R. Bakis, A. Nadas, D. Nahamoo, and M. Picheny, "Influence of background noise and microphone on the performance of the IBM TANGORA speech recognition system," *Proc. ICASSP*, vol. 2, pp. 71-74, 1993.
- [9] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323-2326, 2000.
- [10] J. Tenenbaum, V. d. Silva, and J. Langford, "A Global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319-2323, 2000.
- [11] J. Aczel and J. Dhombres, *Functional Equations in Several Variables with Applications to Mathematics, Information Theory, and the Natural and Social Sciences.*, vol. 31. Cambridge University Press, 1989.
- [12] Lederman, "Orthogonal Procrustes Analysis," in *Handbook of Applicable Mathematics*, E. Lloyd, Ed. New York: John Wiley & Sons, 1984, pp. 761-781.