The GP-LVM for Vocal Joystick Control

Jonathan Malkin†, Neil Lawrence, Jeff Bilmes†*

†Department of Electrical Engineering
University of Washington, Seattle, WA, USA

**Department of Computer Science*
University of Sheffield, Sheffield, UK

UWEE Technical Report
Number UWEETR-2006-0016
October 2006

Department of Electrical Engineering
University of Washington
Box 352500
Seattle, Washington 98195-2500
PHN: (206) 543-2150
FAX: (206) 543-3842
URL: http://www.ee.washington.edu

The GP-LVM for Vocal Joystick Control

Jonathan Malkin†, Neil Lawrence*, Jeff Bilmes†

†Department of Electrical Engineering
University of Washington, Seattle, WA, USA

*Department of Computer Science
University of Sheffield, Sheffield, UK

University of Washington, Dept. of EE, UWEETR-2006-0016

October 2006

Abstract

The Vocal Joystick (VJ) is an assistive device that uses the rich complexity of the human voice to drive a human-computer interface. The system has previously been shown to work well for control of a computer mouse, yet it does not currently make full use of the continuous nature of the vowel space. This work examines the potential use of the Gaussian process latent variable model (GP-LVM), which provides a non-linear, smooth probabilistic mapping from latent to data space. The results show promise for some speakers but do not currently generalize well across speakers. The GP-LVM provides a well-motivated approach, but additional work is still needed to translate the existing potential into an effective new control scheme for VJ.

1 Introduction

The Vocal Joystick (VJ) project [1] at the University of Washington has created a novel human-computer interface device which allows users to control a computer using various vocal parameters. In contrast to traditional automatic speech recognition (ASR), VJ uses both discrete commands as well as continuous aspects of the human vocal system. In this way it is able to provide continuous control suitable for almost any mouse-based task.

Although usable by anyone, VJ is being designed as an assistive device for users with motor impairments. It overcomes some of the limitations of traditional ASR, performing under a range of noise levels, speaking styles, and accents. Even were ASR a solved problem, there would still be use for VJ – natural spoken language is excellent for communicating a concept or idea, but much worse when used for efficient continuous control. For instance, trying to move a mouse cursor from the lower-left to the upper-right corner of the screen, the words "up" and "right" carry little notion of the speed, duration, or distance the user would like the cursor to move. Specifying this additional information separately reduces the efficiency of communication.

Additionally, compared other assistive devices, it is quite inexpensive – a basic microphone is sufficient, compared to the several hundred US dollars of an eye gaze tracking system. For more information on the design goals of the system, see [2].

The machine learning community has recently seen rapid progress in the area of Gaussian processes [3]. The Gaussian process latent variable model (GP-LVM) harnesses the power of Gaussian

processes for dimensionality reduction [4]. The GP-LVM model can find nonlinear low dimensional manifolds embedded in high dimensional space.

There has been a considerable amount of work on the Vocal Joystick system to date. Details of the adaptation system has been reported in [5], and the mapping between vocalizations and actions was described in [6]. We have had public demonstrations in which approximately 100 primary and secondary school students used VJ for the first time to play a video game. It was very well received, and most performed quite respectably.

Having proven the viability of the concept, we are now looking to new applications and to expand the degree of control the system provides. The ultimate goal is to have a wheelchair with a robotic arm, both of which can be controlled by a user via the Vocal Joystick. This work is a first step in that direction, and examines the possibility of a novel control paradigm.

Specifically, we explore the hypothesis that a speaker's vowel space, the space of all vowels the human vocal tract can produce, can be modeled as a manifold embedded in feature space. If we can find such a model which is consistent across speakers, we could then examine various options for control via location on the manifold.

2 Vocal Joystick Control Model

This section describes the control scheme used to determine direction and speed in the current Vocal Joystick, along with some of the advantages and disadvantages of that model.

Our current system derives direction control from linguistic theory. The range of vowel sounds humans can make is traditionally characterized by the first two resonant frequencies of the vocal tract, also known as formants. The frequency of these is primarily determined by the location (or lack of) constrictions in the vocal tract, usually caused by the lips, tongue tip or tongue body. The realizable combinations most people can make is roughly a quadrilateral. A warped version showing a more perceptual display of relative locations can be seen in Figure 1(a).

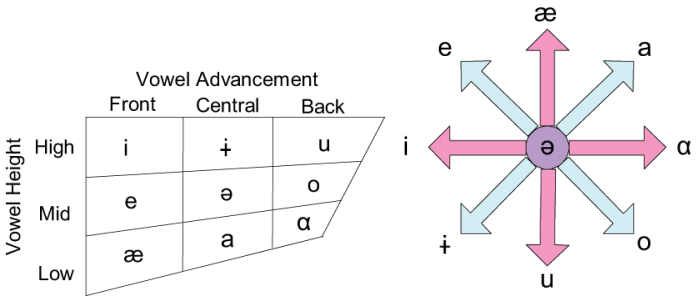
We currently have a VJ version that moves primarily along the four cardinal directions and one that adds the four main diagonals as well. In each case, the cardinal directions are determined by the corners of the vowel space, with this scheme selected for maximum discriminability and separation in production. A graphical depiction of this mapping appears in Figure 1(b). Note that this represents a rotation and reflection compared to Figure 1(a).

Formants have proven quite difficult to track reliably, especially in real-time. As a result, VJ currently uses a multi-layer perceptron (MLP) [7] to train the vowel classes. A speaker-dependent model is arrived at via adaptation [5]. A more detailed look at the various direction control models appears in [8].

A comprehensive treatment of speed control appears in [6]. The basic idea is that the rate of cursor movement is related to the loudness of a vowel. This can be complex due to differences in the energy level at the microphone for different vowels.

It is quite possible for humans to make sounds which smoothly interpolate between the different vowel categories shown in Figure 1(a). Having tried a Gaussian mixture classifier along with a multi-layer perceptron in early versions, we found that performance was superior with the MLP. The drawback is that using such a discriminatively trained classifier, one output tends to dominate the others, even when using a softmax over the classifier outputs to produce posterior probabilities for each class [7].

Additionally, we are using vowel quality for a 2-D control, yet in classifying only the outermost points in the vowel space we are not truly taking advantage of the 2-D nature of the space. As a result, we need to use another control signal, currently loudness, to specify velocity.



(a) (b)

Figure 1: (a) Relative vowel locations in the human vowel space. All labels are in IPA format. (b) Vowel-to-direction mapping used in the current version of Vocal Joystick.

3 GP-LVM Overview

The Gaussian process latent variable model [4] is a relatively new probabilistic method for obtaining a reduced dimension representation of a data set. In its linear form, it can be seen as solving the dual problem to that of probabilistic principal components analysis (pPCA) [9], but it further generalizes to non-linear models as well.

Many dimensionality reduction schemes assume that observations $\mathbf{Y}$ are the result of some function $f(\cdot)$ on an underlying set of points $\mathbf{X}$ plus a noise term ϵ. For instance, in the case of pPCA, the assumption is that there is a linear weight matrix $\mathbf{W}$ so the resulting problem to solve is $\mathbf{Y} = \mathbf{X^T W} + \epsilon$. The weights are found by marginalizing over the latent space variable $\mathbf{X}$ and taking the maximum likelihood solution.

By contrast, the linear form of the GP-LVM instead marginalizes over the weights to find the maximum likelihood solution for the latent variables. When generalizing to non-linear models, this can be seen as finding the set of hidden functions that map from the latent space to the observed space. In this case, both the latent positions and function parameters are optimized: $P(\mathbf{Y}|\mathbf{X}) = \prod_{\mathbf{i=1}}^{\mathbf{d}} \prod_{\mathbf{n=1}}^{\mathbf{N}} \mathbf{p}(\mathbf{y_{in}}|\mathbf{f_{in}})\mathbf{p}(\mathbf{f}|\mathbf{X})$, where y_{in} is the i^{th} dimension (of d) of the n^{th} sample, and likewise for function f_{in} mapping from latent to data space.

By using a Gaussian Process (GP) prior over the functions, this model yields an elegant and tractable procedure that has been shown to work well on a variety of data sets [4]. Additionally, since one of the key drawbacks to such a model is its computation complexity, Lawrence has also published several approximation algorithms to decrease training time [10].

The GP-LVM, unlike most approaches to non-linear dimensionality reduction, finds a smooth mapping that keeps far apart points in latent space if they were far apart in data space. Since local distance preservation is also desirable at times, it is possible to extend the model to include back constraints, a smooth mapping from data space to latent space. Further details appear in [11].

4 GP-LVM for VJ Control

As mentioned in Section 2, the current version of VJ does not make full use of the 2-D nature of the human vowel space suggested by linguistic theory. One of our goals is to address that shortcoming and to free up the use of loudness for another aspect of control. Since there is reason

to believe that the underlying vowels are indeed generated from an underlying manifold, we believe a dimensionality reduction method such as the GP-LVM could allow smooth, continuous control throughout the entire vowel space.

To be clear, we have not yet built a working system using the GP-LVM. We are instead exploring the potential of such a model. A prerequisite for the use of any dimensionality reduction method in VJ is that it must cluster samples from unique sounds from a single speaker well and also provide a smooth path between those clusters. For this work, we focus on the first portion of that requirement.

The features we use for vowel quality estimation or Mel Frequency Cepstral Coefficients (MFCCs). Information on the theory behind these features can be found in [12]. Essentially, they attempt to capture information on the shape of the human vocal tract. They are calculated by taking the magnitude of the frequency spectrum of a window of speech, typically about 25ms. The inverse Fourier transform of the log of a warped version of the spectrum provides the features.

Most ASR applications shift the window over which MFCCs are calculated by 10ms each time, resulting in a series of partially overlapping frames. Additionally, derivatives of these features, termed deltas, can be used to incorporate context information into each frame.

A large data collection effort has accompanied the Vocal Joystick project [13]. The data for this work came from five speakers consisting of one author of this paper, an experienced VJ user, and four from the database, two male and two female. In each case, sounds at each speaker's normal volume with a level pitch contour were used. These yielded samples with approximately 200 frames per vowel. Mean subtraction and variance normalization were performed independently for each speaker. Because the variances of the energy and delta energy were quite small, they were removed from the set of features.

The GP-LVM was run using the fully independent training conditional [14, 15] and 200 active points [10]. For each speaker, we generated models both including and not including deltas under three conditions, giving a total of six models per speaker. The first two models were a GP-LVM with no back constraints and with MLP back constraints [11], both of which used PCA to initialize the model. The third model also used MLP back constraints but did not use PCA for that initialization.

For comparison, we included results of running simple linear projection methods, PCA (not pPCA) and linear discriminant analysis (LDA). Results of PCA appear in Figure 2 and those of LDA are in Figure 3. Results of the different GP-LVM methods, no back constraints (baseline), back constraints with PCA initialization (GP-LVM$_{PCA}$) and back constraints with random (non-PCA) initialization (GP-LVM$_{rand}$), are in Figures 4, 5 and 6, respectively.

All images are plotted with the same legend: green circle for æ, red cross for a, blue plus for ɑ, red down triangle for o, blue left triangle for u, yellow diamond for ɨ, cyan asterix for e, and green up triangle for ə. For figures from the GP-LVM, the greyscales in the background indicate the precision with which the underlying manifold was expressed in data-space for that latent point; there is no analogous value for the non-probabilistic linear maps of PCA and LDA.

5 Discussion and Future Work

The results for any single user often look interesting; the vowels are well separated in most cases. Although many plots have numerous scattered points, based on additional analysis with PCA not presented here, these appear to be a result of the onset of speech when the sound is known to be less stable.

Based on the results, no particular method appears obviously well-suited to VJ control across users. This holds true even after allowing for possible rotations and reflections of the images. One

possible reason for this is that the latent mapping for each user was trained independently. The introduction of constraints across users may help provide a more universal model, but this would come at the cost of a very large increase in computation time.

For the experienced VJ user, and this paper's author, S1, the results with back constraints and PCA initialization (Figure 5(a)) show excellent separation and tight clustering, arranged in a bull's eye pattern. This suggests a control strategy of a central sound with other sounds evenly spaced around it, which could potentially allow for arbitrary directional control. That this result was not seen in results from other speakers leaves open the question of whether this is truly a "natural" control strategy, or whether it is a result of much experience with the current VJ control model. By contrast, Figure 5(b) shows a nearly linear model; it is not clear how to develop a control model from such a plot.

The general trends that emerge, however, are that the back constraints have an obvious effect to reduce scattering within a vowel cluster, especially when deltas are included. Overall, since these vowels are primarily in steady-state, deltas should be of little use, and this seems supported – the precision is more evenly distributed in the GP-LVM models without deltas and with PCA the results are more likely to be separated. LDA shows little effect from the deltas, which means the procedure has effectively decided that they provide little use in discrimination. Examining the coefficients used in the linear models confirms this intuition.

Going forward, one logical first step is to try an LDA-based initialization to see how that alters the final result. Another is to try the side constraints alluded to earlier. This would, however, alter the GP-LVM model such that it would no longer be fully unsupervised. In additional to pure vowels, VJ project has collected a large number of vowel combinations where speakers slowly transitioned between two different vowels. While the use of such combinations would complicate visualization of the results, such a procedure could help provide structure to the space between the existing vowel clusters.

Intractability is a significant issue with many of the proposed options for continued work. Gaussian processes are quite powerful, but their computational cost has been the biggest barrier to more widespread success on large data sets. This certainly applies to training, but even for use in VJ, we need to look at options for implementing this as a VJ control method, probably via a particle filter to allow for real-time performance. Despite these challenges, we remain optimistic that advancements both in computational power and, more crucially, our understanding of Gaussian processes, will soon make the GP-LVM a viable control option for the Vocal Joystick.

References

[1] J.Bilmes et al., "The Vocal Joystick: A voice-based human-computer interface for individuals with motor impairments," in *Human Lang. Tech. Conf. and Conf. on Empirical Methods in Nat'l Lang. Proc.*, 2005.

[2] J.Bilmes et al., "The Vocal Joystick," in *IEEE ICASSP*, Toulouse, France, 2006.

[3] C.E.Rasmussen and C.K.I.Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2006.

[4] N.Lawrence, "The Gaussian process latent variable model," Tech. Rep. CS-06-03, Sheffield University, Dept. of Computer Science, 2006.

[5] X.Li and J.Bilmes, "Regularized adaptation of discriminative classifiers," in *IEEE ICASSP*, Toulouse, France, 2006.

[6] J.Malkin, X.Li, and J.Bilmes, "Energy and loudness for speed control in the Vocal Joystick," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico, 2005.

[7] C.Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[8] X.Li, J.Malkin, S.Harada, J.Bilmes, R.Wright, and J.Landay, "An online adaptive filtering algorithm for the Vocal Joystick," in *Proc. Int. Conf. on Spoken Language Processing*, Pittsburgh, PA, 2006.

[9] M.Tipping and C.Bishop, "Probabilisitic principal components analysis," *Journal of the Royal Statistical Society*, vol. 6, no. 3, pp. 611–622, 1999.

[10] N.Lawrence, "Large scale learning with the Gaussian process latent variable model," Tech. Rep. CS-06-05, Sheffield University, Dept. of Computer Science, 2006.

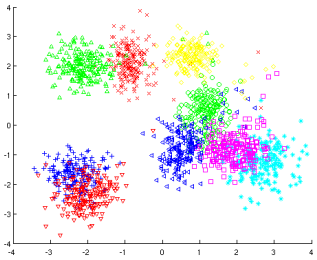
[11] N.Lawrence and J.Quiñonero-Candela, "Local distance preservation in the GP-LVM through back constraints," in *Int'l Conference on Machine Learning*, Pittsburgh, PA, 2006.

[12] X.Huang, A.Acero, and H.-W.Hon, *Spoken Language Processing*, Prentice Hall PTR, 2001.

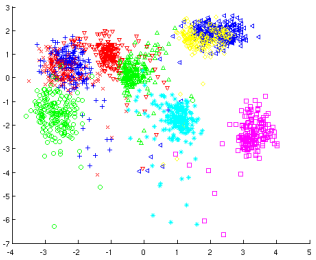
[13] K.Kilanski, J.Malkin, X.Li, R.Wright, and J.Bilmes, "The Vocal Joystick data collection effort and vowel corpus," in *Interspeech*, Pittsburgh, PA, Sept. 2006.

[14] E.Snelson and Z.Ghahramani, "Sparse Gaussian processes using pseudo-inputs," in *Advances in Neural Information Processing Systems*, Cambridge, Massachusetts, 2006, The MIT Press.

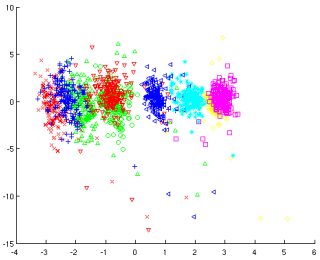
[15] J.Quiñonero-Candela and C.E.Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *Journal of Machine Learning Research*, pp. 1939–1959, 2005.



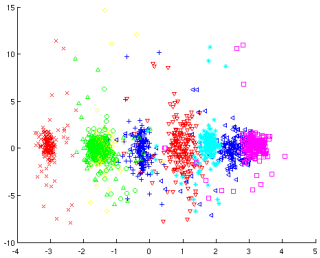
(a) S1 PCA, deltas



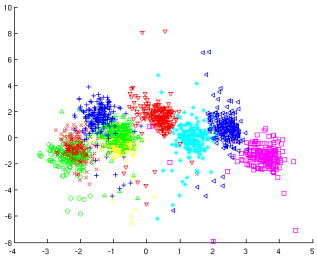
(b) S2 PCA, deltas



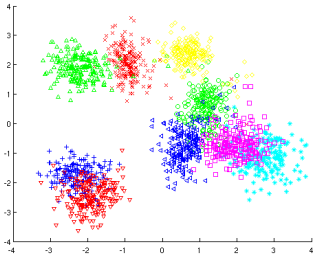
(c) S3 PCA, deltas



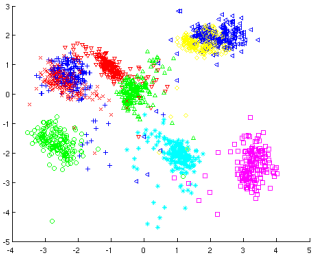
(d) S4 PCA, deltas



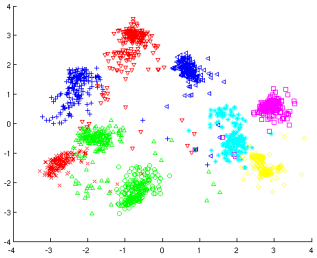
(e) S5 PCA, deltas



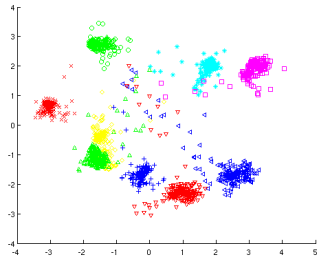
(f) S1 PCA, no deltas



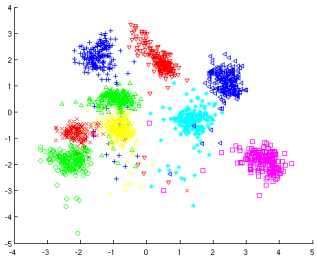
(g) S2 PCA, no deltas



(h) S3 PCA, no deltas

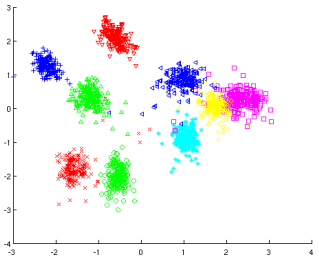


(i) S4 PCA, no deltas

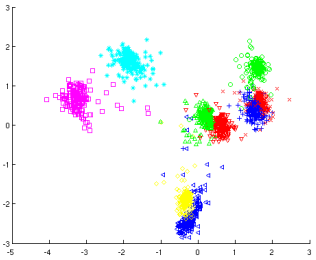


(j) S5 PCA, no deltas

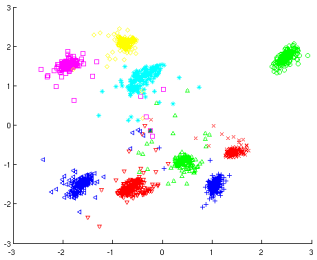
Figure 2: PCA results for speakers 1 through 5.



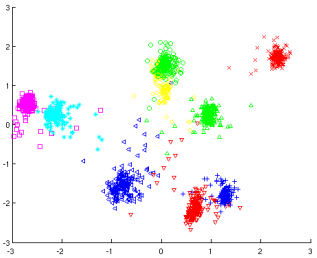
(a) S1 LDA, deltas



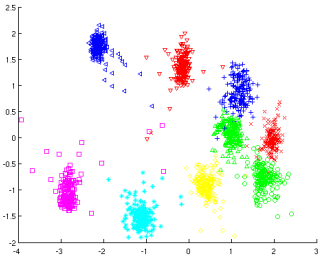
(b) S2 LDA, deltas



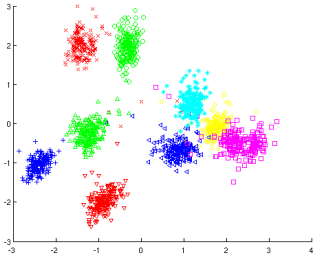
(c) S3 LDA, deltas



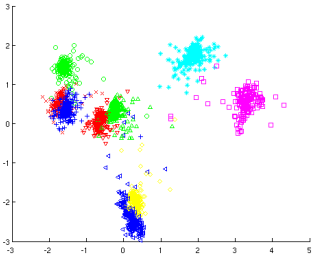
(d) S4 LDA, deltas



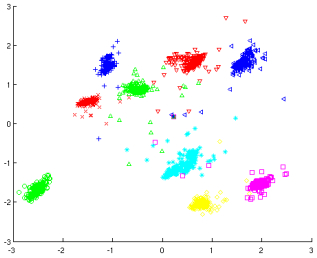
(e) S5 LDA, deltas



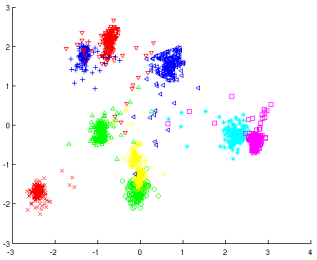
(f) S1 LDA, no deltas



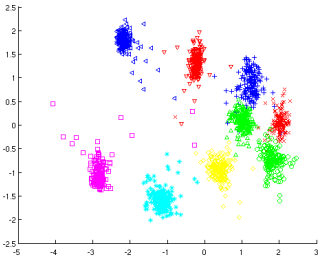
(g) S2 LDA, no deltas



(h) S3 LDA, no deltas

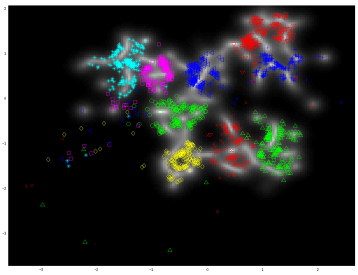


(i) S4 LDA, no deltas

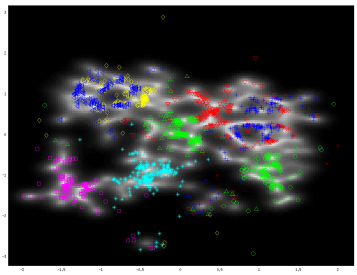


(j) S5 LDA, no deltas

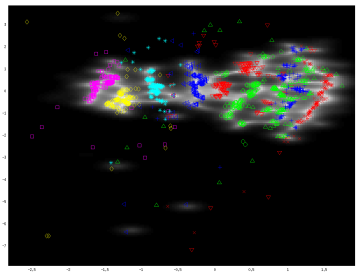
Figure 3: PCA results for speakers 1 through 5.



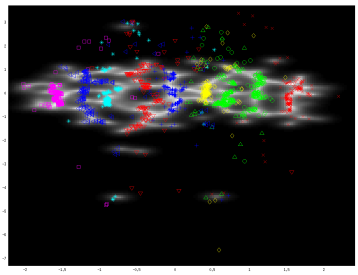
(a) S1 GP-LVM, deltas



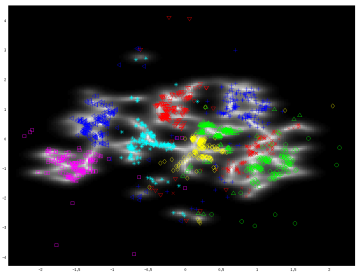
(b) S2 GP-LVM, deltas



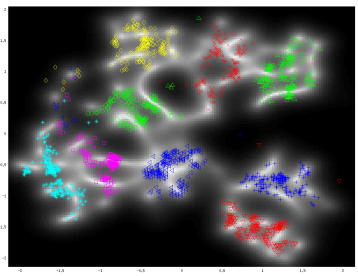
(c) S3 GP-LVM, deltas



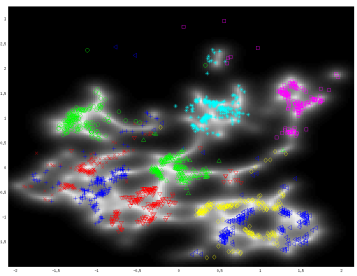
(d) S4 GP-LVM, deltas



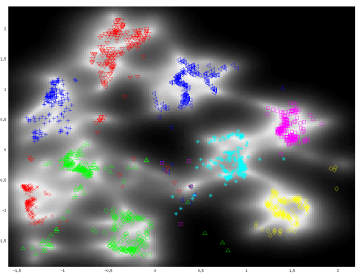
(e) S5 GP-LVM, deltas



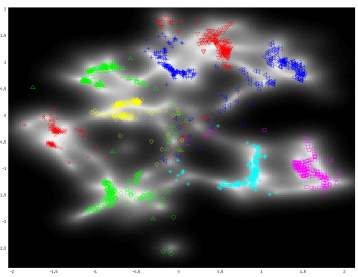
(f) S1 GP-LVM, no deltas



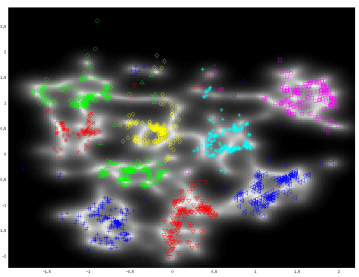
(g) S2 GP-LVM, no deltas



(h) S3 GP-LVM, no deltas

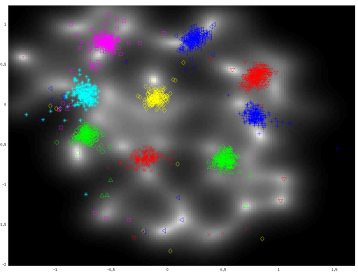


(i) S4 GP-LVM, no deltas

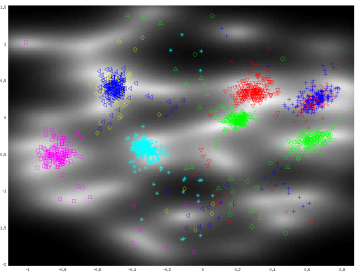


(j) S5 GP-LVM, no deltas

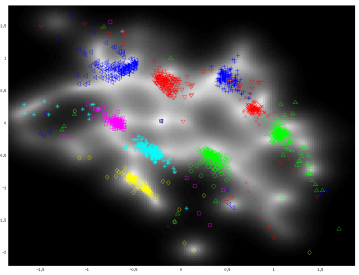
Figure 4: GP-LVM results without back constraints for speakers 1 through 5.



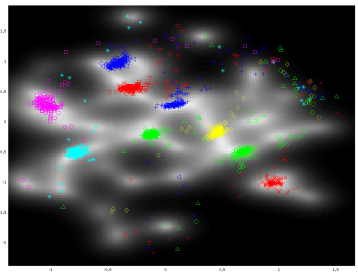
(a) S1 GP-LVM$_{PCA}$, deltas



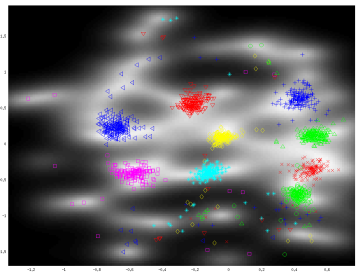
(b) S2 GP-LVM$_{PCA}$, deltas



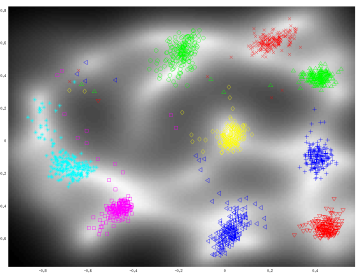
(c) S3 GP-LVM$_{PCA}$, deltas



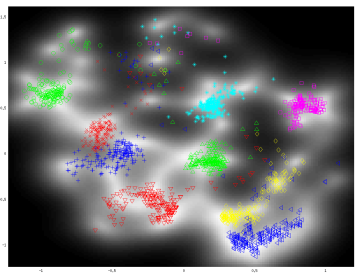
(d) S4 GP-LVM$_{PCA}$, deltas



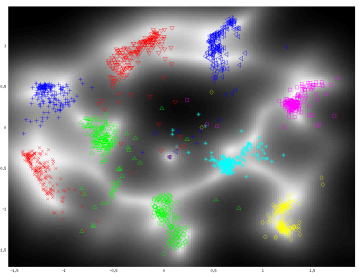
(e) S5 GP-LVM$_{PCA}$, deltas



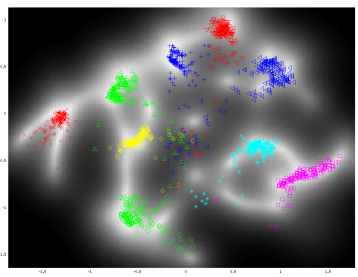
(f) S1 GP-LVM$_{PCA}$, no deltas



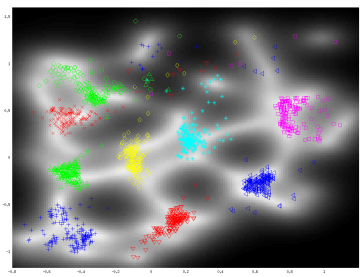
(g) S2 GP-LVM$_{PCA}$, no deltas



(h) S3 GP-LVM$_{PCA}$, no deltas

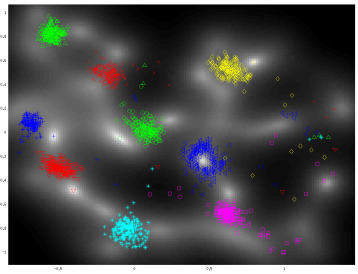


(i) S4 GP-LVM$_{PCA}$, no deltas

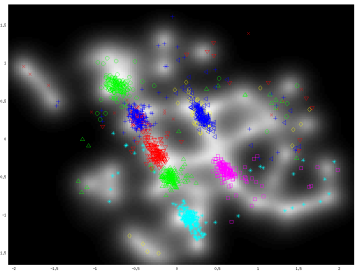


(j) S5 GP-LVM$_{PCA}$, no deltas

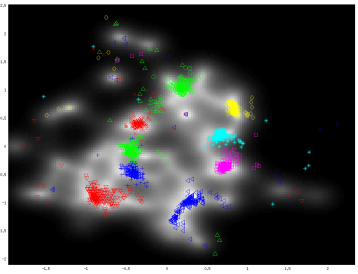
Figure 5: GP-LVM results with back constraints and PCA initialization for speakers 1 through 5.



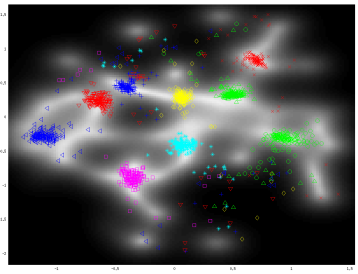
(a) S1 GP-LVM$_{rand}$, deltas



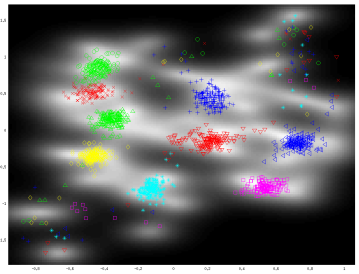
(b) S2 GP-LVM$_{rand}$, deltas



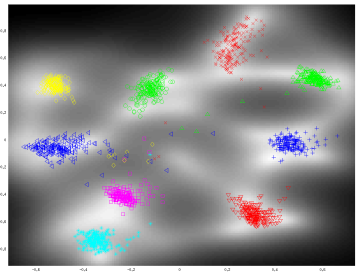
(c) S3 GP-LVM$_{rand}$, deltas



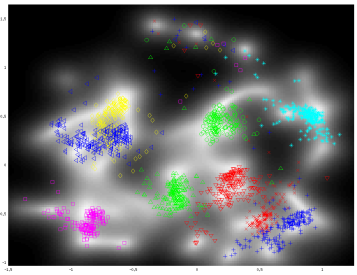
(d) S4 GP-LVM$_{rand}$, deltas



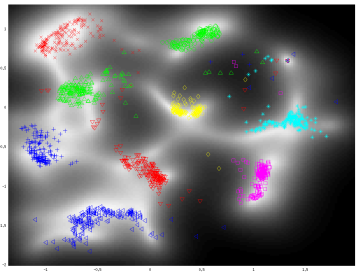
(e) S5 GP-LVM$_{rand}$, deltas



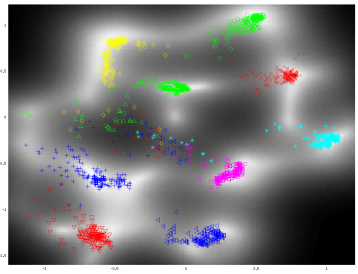
(f) S1 GP-LVM$_{rand}$, no deltas



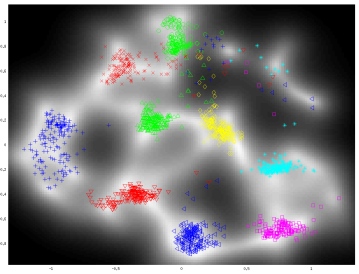
(g) S2 GP-LVM$_{rand}$, no deltas



(h) S3 GP-LVM$_{rand}$, no deltas



(i) S4 GP-LVM$_{rand}$, no deltas



(j) S5 GP-LVM$_{rand}$, no deltas

Figure 6: GP-LVM results with back constraints and random initialization for speakers 1 through 5.