Parameterisation of Speech Lip Movements

James D. Edge, Adrian Hilton and Philip Jackson

1Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK.

j.edge@surrey.ac.uk, a.hilton@surrey.ac.uk, p.jackson@surrey.ac.uk

Index Terms: speech synthesis

1. Introduction

Synthetic talking heads are becoming increasingly used across a wide range of applications: from entertainment (e.g. Computer Games/TV/Films) through to natural user interfaces and speech therapy. Increasingly the techniques used in synthesis are data-driven, that is as it becomes easier to capture articulatory movements (e.g. motion-capture/stereophotogrammetry etc.) In this paper we present techniques for parameterising lip movements captured using a dynamic stereo-capture system, and discuss the properties of the underlying manifold and how these can aid synthesis.

When trying to understand the process of speech production it is useful to project captured data, whichever form that may take (e.g. 3D markers, 2D contours, lip aperture/protrusion etc.), into a lower dimensional space which provides greater fidelity in the underlying structure of this task. This projection fulfils two roles: firstly, to identify several latent parameters which better describe the underlying processes; and secondly to remove the components of the original signal which can be regarded as insignificant or as noise.

The most common method for parameterising speech data (and motion data in general), has been Principal Component Analysis (PCA). This method forms an orthogonal basis by calculating the eigenvectors (principal components) of the covariance matrix. The eigenvalues corresponding to these basis eigenvectors hold the variance that each vector accounts for, and thus a reduced set of vectors can be retained holding the majority of the variance in the original data. PCA has been used as an underlying parameterisation for many applications, principally tracking/image synthesis [Blanz and Vetter 1999] [Cootes et al. 1998] and data-driven animation [Theobold and Wilkinson 2007]. Similar techniques applied to the symmetric distance matrix are termed Multi-Dimensional Scaling (MDS), or Principal Coordinate Analysis (PCO, [Gower 1966]).

Another linear method, particularly popular in source separation, is Independent Component Analysis (ICA). This technique applies the concept of statistical independence to recovering an underlying parameterisation. This is a stronger assumption than that of PCA, but does not produce an orthogonal basis. ICA has been applied to recovering a basis for speech movements, and for identifying and separating speech-task and emotion-task parameters. In [Cao et al. 2003] ICA is employed to separate emotional parameters from speech control parameters, which are later recombined in a synthesis framework. The success of such techniques is arguable. Problematically, ICA decompositions tend to produce non-smooth trajectories which are generally inappropriate for tasks such as synthesis.

More recently a host of non-linear methods have come to the fore, e.g. [Roweis and Saul 2000] [Tenenbaum et al. 2000]. These methods mostly unwrap the data according to an approximation of geodesic distances across an underlying manifold structure. The main problem with PCA is that where non-linear relationships are evident in the underlying data, e.g. when projecting onto a pair of basis vectors a curve is plotted this curve may be better described by a single non-linear vector recovered using a method such as Isomap. Unfortunately, many non-linear methods are not generative (i.e. you can project into the latent space, but not back to the data domain), and those which are often cannot deal with very large datasets. Furthermore, it is only worthwhile performing a non-linear parameterisation if the recovered space is more descriptive of the underlying data than simpler linear methods.

In [Grochow et al. 2004] Gaussian-Process Latent Variable Models (GPLVM), a non-linear and generative dimensionality reduction technique is used to form a low dimensional manifold for motion data. Trajectories in the manifold can then be fitted to real data, such as video. This technique demonstrates one of the main problems with non-linear methods, as in some cases the optimisation of the projective space leads to discontinuities in motion trajectories. Gaussian Process Dynamic Models (GPDM) [Wang et al. 2005] have been designed to remove discontinuities in the generated projection from GPLVM. However, even in cases where dynamics or back-constraints have been used to maintain the contiguity of the projected trajectory, we may have cases where identical frames are projected to different locations in the nonlinear space.

In this paper we address the problem of finding a representation of speech lip movements captured using 3D stereo capture technology. It is often considered that the only role of this stage is to take high-dimensional data and map it into a low-dimensional space, i.e. the simplification of the task space is

1 Recovered nonlinearities in PCA are often referred to as the "horseshoe problem".
the end in itself. However, depending upon the end application there may be other qualities which define how good an embedded space is. For the purposes of our work (i.e. synthesis) we can define several factors which are important:

- **minimal** - as with all dimensionality reduction techniques we aim for a greatly reduced set of variables.
- **interpretable** - the variables should have a logical interpretation when relating back to the physical process of speech production.
- **smooth** - projected speech trajectories should be smooth, i.e. there is some geometric continuity to movement through the task space. Smooth trajectories are more easily synthesised than those with random fluctuations.
- **generative** - at least for the task of synthesis, it is important that trajectories within our simpler task space can be projected back to the original data domain.

In the following sections we introduce our technique for constructing a low-dimensional task space which satisfies the properties described above. In Section 3 we describe the construction of the task space and the associated geometric properties of the embedded speech manifold are described. In Section 4 we describe techniques for synthesising trajectories and recovering the mapping back to the original data domain. In Section 5 we briefly discuss clustering of phonetic classes on the generated speech manifold. Finally, in Section 6 we describe directions of future work and the application of our techniques to the problem of speech synthesis.

## 2. Data Capture

The data used in this paper consists of 8 minutes of 3D geometry captured using a commercial stereo face capture system (see [3dMD](https://www.3dmd.com)). Facial geometry is recovered using 2 camera stereo pairs (left/right); an infra-red speckle pattern is projected onto the face and stereophotogrammetry is used to recover the 3D surface. A further two colour cameras also capture the appearance of the subject’s face. The cameras we use operate at 60Hz, and audio is also captured synchronously. The raw data from this system is initially unregistered, i.e. given a point on the surface of the face in the initial frame we do not know the corresponding point in any of the following frames. To solve this correspondence issue, blue markers (both point and contour markers are used) are painted on the subjects face and tracked frame-to-frame in the colour images, these markers are used as the basis of a dense registration of all points in the geometry across all frames in the sequence (details are described in Figure 1: Constructing a model of speech lip movement; the basis \( \tilde{X} \) is an initial PCA projection of the original data, \( Y^* \) is the reduced speech task space.

![Figure 1](image1.png)

Because the colour appearance frames contain the blue markers, the processing described in the rest of this paper is performed on the recovered geometry alone.

The 8 minutes of data consists of a single native British English subject speaking sentences selected from the TIMIT corpus. Sentences were selected to get a good sampling of all phonemes, see table 1. Unfortunately, due to data storage and processing issues it was not possible to get a good sampling of phonemes across all possible contexts. However, it is considered that the data does provide a good sampling of lip dynamics, the description of which is the purpose of this paper.

## 3. Parameterising Speech Lip Movements

### 3.1. Constructing a Task Space Embedding for Speech

In constructing a task space for the data described in the previous section, the registered geometry alone is used. The data consists of a sequence of frames, \( F \), where the \( i^{th} \) frame \( F_i = \{ x_0, y_0, z_0, \ldots, x_i, y_i, z_i, \ldots, x_n, y_n, z_n \} \). A noise reduction step employing standard PCA directly on \( F \) is used to filter out low variance modes. By applying PCA we get a set of basis vectors, \( \tilde{X} \). The EM method for computing principal components [Roweis](https://www.sciencedirect.com/science/article/pii/S009589560310006X) is used here due to the size of the

<table>
<thead>
<tr>
<th>Consonants</th>
<th>p</th>
<th>jh</th>
<th>zh</th>
<th>dh</th>
<th>d</th>
<th>ng</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>72</td>
<td>34</td>
<td>20</td>
<td>81</td>
<td>187</td>
<td>28</td>
</tr>
<tr>
<td>Vowels</td>
<td>a</td>
<td>ae</td>
<td>ay</td>
<td>eh</td>
<td>iy</td>
<td>ua</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>313</td>
<td>57</td>
<td>73</td>
<td>126</td>
<td>72</td>
</tr>
<tr>
<td>Consonants</td>
<td>b</td>
<td>s</td>
<td>f</td>
<td>k</td>
<td>r</td>
<td>h</td>
</tr>
<tr>
<td></td>
<td>79</td>
<td>313</td>
<td>69</td>
<td>133</td>
<td>136</td>
<td>29</td>
</tr>
<tr>
<td>Vowels</td>
<td>ah</td>
<td>ab</td>
<td>ax</td>
<td>ia</td>
<td>ow</td>
<td>uh</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>48</td>
<td>299</td>
<td>65</td>
<td>47</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 1: Frequency of English phonemes in the captured data.
data matrix, $F$, which holds $28,833$ frames $\times 12,784$ xyz coordinates. The first 100 basis vectors are computed, with the first 30 holding over 99% of the recovered variance. The percentage of the total variance accounted for will be lower, but the scree-graph shows that the important features of $F$ are compressed in only a few dominant components (i.e. 95% in the first 10 components shown in fig. 1(b) and 99% in the first 30 components indicating a flattening of the scree-graph.)

$F$ can be projected onto the basis $\chi$ to produce the parameterisation $F^e$, i.e. $F_i \times \chi \rightarrow F^e_i$. Broadly, the 1st component of $\chi$ can be described as jaw opening, the 2nd is lip rounding/protrusion, and the 3rd raises/lowers the upper lip. Lower variance components are not as easily contextualised in terms of observed lip-shape qualities. Figure 1(a) shows the first two principal components of $\chi$.

A further projection of the data is performed to maintain the relative dynamic properties of lip movements during speech, i.e. in contrast to the formation of $\chi$ which only takes into account static lip shapes. The first derivative for each frame is approximated as $F_i^{ev} = F_i^e - F_i^{e-1}$ (the parametric displacement of the lips in $1/60^{th}$ of a second.) Each pair $\{F_i^e, F_i^{ev}\}$ describes a distinct state in the principal space of lip movement. As the first derivative is at a different scale the parameters need to be normalised such that $F_i^e$ does not dominate over $F_i^{ev}$. Thus, a matrix $G = \{\alpha(F_i^e - \mu), \beta(F_i^{ev} - \mu')\}$, where $\mu$ and $\mu'$ are the respective means of $F^e$ and $F^{ev}$, is constructed where all parameters are scaled to the range $F_i^e, F_i^{ev} \in [-1, 1]$. This matrix is now processed in a manner similar to MDS/PCG.

A symmetric distance matrix $D$ is formed where each element $D_{ij}$ is the euclidean distance between $G_i$ and $G_j$, i.e. $D_{ij} = \sqrt{(G_i - G_j)^2}$. The matrix $D$ is then decomposed using another iteration of PCA forming a basis $Y$, so for each of the initial frames $F_i$ we have a corresponding projection into the task space $F_i^y$. The first three dimensions of $Y$ account for 93% of the recovered variance in $D$. Figure 1(b) shows that the eigenvalues corresponding to $Y$ are more tightly clustered in the first several components than those of $\chi$. Note that the $F^y$ describe the relative position of the $G_i$, but there is no direct transformation from a point in $Y$ back into $\chi$ (i.e. because $Y$ is a relative embedding.)

It can be seen in fig. 2(a) that the first 3 dimensions of the projected $F^y$ form a manifold $\mathcal{M}$ embedded within the task space $Y$ descriptive of lip movements during speech production. Some useful properties of $\mathcal{M}$ and speech trajectories in this space are described in the following sections.

### 3.2. The Structure of the Speech Manifold

As can be seen from fig. 2(a) the manifold, $\mathcal{M}$, constructed as described in the previous section, forms a paraboloid-type structure in 3D (the first 3 dimensions of $\rightarrow Y$.) $\mathcal{M}$ is symmetric about a plane which coincides with zero-velocity, and either side of this plane we have the opening and closing halves of a speech cycle. In the opening half of the cycle lips move from closed states towards more open states, and in the closing half the lips move from more open states towards more closed states.

The vector orthogonal to the zero-velocity plane corresponds to maximum change in velocity (i.e. maximum acceleration) and parallel to the zero-velocity plane we find a vector which corresponds to maximum change in shape (i.e. lip shape transitions from closed $\rightarrow$ open, see fig. 2(c).) We can define non-linear vectors which globally describe maximum change in lip shape, $S$, and maximum change in lip velocity, $V$. Of course the evident non-linear structure of the speech manifold means that it is more appropriate to define a local tangent space for each $F^y_i$, with $s_i$ defining a local shape-like vector, and $v_i$ defining a local velocity-like vector, with $s_i \times v_i$ defining the local orientation of the manifold (i.e. direction of thickness.)

Traces of speech movements will produce elliptical paths on $\mathcal{M}$. This is natural as $S$ is effectively an ordering of lip shapes according to lip opening (the first component of $\chi$, which will dominate over other components in $F^y$ when computing the distance matrix $D$.) The same is true of $V$, i.e. the first component of $F^{ev}$ will dominate over components with smaller variance, leading to an ordering of $V$ according to that first component.

This local and global structure of $\mathcal{M}$ leads to several results. Firstly it is impossible to travel on the zero-velocity plane. This is obvious seeing as this would imply changing lip shape with zero velocity! Thus when moving between the opening and closing halves of a speech cycle the trajectory will cross the plane of symmetry at right angles. Furthermore, at all other positions on the manifold, except on the plane of symmetry, the trajectory cannot travel parallel to $v_i$. This follows as to travel in the local tangential direction $v_i$ implies maximum acceleration and minimum change in shape.

These properties show that the speech manifold is highly structured. In fact the trajectory can be described using the local tangent space described above, $\{s_i, v_i\}$. If you construct an offset between two consecutive frames, $z_i = (F^y_i - F^y_{i+1})$, then $z_i, s_i$ and $z_i, v_i$ will categorise each frame into four different classes: opening/acceleration, opening/deceleration, closing acceleration, and closing/deceleration.

### 3.3. Distances on the Speech Manifold

Euclidean distances between pairs of points in the first several dimensions of $Y$ approximate euclidean distances between pairs of points in $G$. However, the recovered structure of the $\mathcal{M}$ indicates that a direct line between two non-consecutive states $F^y_i$ and $F^y_{i+n}$ does not give a good measure as to the time it would take to travel between them. In fact even a standard geodesic computation does not perform well because the manifold itself is directed, i.e. there is a temporal ordering of states, and thus distances are asymmetric. The natural way of understanding this is that given the lips are in a state $F^y_i$ with velocity $F^{ev}_i$, the number of possible following states will be constrained by the physical properties of the lips (e.g. muscle properties, inertia etc.)

This evident asymmetry in distances can be accounted for using a modified geodesic computation. The $F^y$ are initially grouped into states using K-means. A transition matrix, $A$, is then constructed with $A_{jk} = 1$ iff there is a transition from a member of state $j$ to a member of state $k$, otherwise $A_{jk} = 0$. From this adjacency structure Dijkstra’s algorithm can be used to calculate the minimum distance (i.e. number of transitions $\approx$ time in frames) between two states. Also, minimum distance paths between states can be calculated by propagating across the graph structure. There is no upper bound on the distance...
The structure of $M$ embedded in the task space $Y$. Colour indicates density in each projection, from most dense (red) to least dense (blue). The global shape, $S$, global velocity, $V$, local shape, $s_i$, and local velocity, $v_i$, are shown. The zero velocity plane is shown as a dotted line. Black arrows show and example trajectory on $M$.

Directed geodesic distances from points on $M$: colour indicates geodesic similarity, from most similar (red) to least similar (blue); arrows indicate the direction of possible future events; dashed line delineates past and future events; dotted line shows the zero velocity plane. Images show the three possible states of the system: opening, zero velocity, and closing. Note that when the current state has zero velocity the future states are ambiguous, i.e. future states could be further opening or closing of the lips.

Lip shapes indexed by increasing $S$, the lip shape vector.

Figure 2: Properties of the visual speech manifold $M$.

4. Trajectory Synthesis

The speech manifold, $M$, as discussed in the previous section, is useful in visualising what is happening in terms of the dynamics of speech production for the lips. However, interesting applications can be tackled if the inverse problem can be solved, i.e. can trajectories be derived directly from the structure of $M$. The observations in Section 3.3 indicate that there is a temporal ordering of states in $M$ which can potentially aid the generation of speech trajectories. An obvious application for this is speech synthesis, supposing that we define a set of states that the lips must pass through to articulate an utterance - the complete trajectory can be inferred from $M$. This can be seen as a form of data-driven interpolation lying between traditional spline-based techniques (e.g. Cohen and Massaro 1993) and trajectory concatenation (e.g. Kshirsagar and Magnenat-Thalmann 2003).

4.1. Calculating Paths Between Pairs of Points

By examining minimum distance plots on $M$, see fig. 2(b), the causal nature of the visual speech manifold becomes apparent. It can be clearly seen that given a current state $F^y_i$ the possible set of future states is a subset of $F^y$, i.e. those for which $\text{dist}(F^y_i, F^y_p) = 1$, $F^y_p \in F^y$. A hemisphere of possible future events is evident at each point on the manifold, and likewise if the adjacency matrix is constructed using links between current and previous states (i.e. $A_{jk} = 1$ iff there is a transition from a member of state $k$ to a member of state $j$) a hemisphere of possible past events can be seen. This indicates that a temporal ordering exists at each point on $M$ pointing in the general direction of future events in the speech cycle. For states with zero-velocity, there are two directions of possible future states, this is due to the fact that with the lips in a state with zero velocity the lips can proceed into a further opening or closing cycle equally. Also, for states closer to the boundary of $M$ the set of possible next states is reduced (i.e. there is a smaller set of possible future events.)
ability in a manner similar to dynamic programming. This is efficient as, where phone mid-points are used as targets, the distance in time between targets is rarely greater than 10 frames.

An example of a trajectory generated by interpolating phone mid-points, for the sentence "Herb's birthday occurs frequently on Thanksgiving", on the speech manifold is shown in fig. 3. As can be seen, where the phone mid-points strongly define the trajectory this method produces a good approximation to the original data. Where the interpolation fails this is generally due to two reasons: firstly, the discretisation of the manifold means that the trajectory may not exactly be matched; secondly, where there are higher frequency components to the original signal (i.e. far higher frequency than the targets being interpolated.) The second issue is the more important and implies that the number of required targets may be greater than the number of phone mid-points in the target utterance. Even so this form of trajectory synthesis performs far better than a direct interpolation of phone targets.

4.2. Calculating the Inverse Mapping

It is important to provide a mapping back from the embedded manifold $F^y \rightarrow F^x$ for the parameterisation to be useful for many applications, particularly for synthesis. There is no direct mapping because $\mathcal{Y}$ is constructed from the distance matrix $D$, and only preserves relative distances between states. In $\mathbb{R}^n$ barycentric coordinates can be defined using a simplex consisting of $n + 1$ points. Thus, in 3D for any given point in the embedded space $\mathcal{Y}$, the projection of a point $p^y$ back into $\mathcal{X}$ can be determined using the bounding tetrahedron $\Delta^y_{ijkl}$ Delaunay tetrahedralisation can be used to find an appropriate structure for the speech manifold to provide this mapping, and similar techniques can be used even when the number of dimensions exceeds 3.

Given $p^y$ is in the tetrahedron $\Delta^y_{ijkl}$ then the projected point $p^x$ is calculated as a barycentric combination of the surrounding vertices.

$$p^x = F^x_i \cdot B_i + F^x_j \cdot B_j + F^x_k \cdot B_k + F^x_l \cdot B_l$$

The barycentric weights, $B_i$, are defined as the ratio of sub-volumes within the tetrahedron,

$$B_i = \frac{V_\Delta(p^y, F^y_i, F^y_j, F^y_k, F^y_l)}{V_\Delta(F^y_i, F^y_j, F^y_k, F^y_l)}, \quad B_j = \frac{V_\Delta(p^y, F^y_i, F^y_k, F^y_l)}{V_\Delta(F^y_i, F^y_j, F^y_k, F^y_l)}$$

$$B_k = \frac{V_\Delta(p^y, F^y_i, F^y_j, F^y_l)}{V_\Delta(F^y_i, F^y_j, F^y_k, F^y_l)}$$

where $V_\Delta$ is the volume of a tetrahedron spanning four vertices.

$$V_\Delta(p_0, p_1, p_2, p_3) = \frac{|(p_0 - p_3) \times (p_1 - p_3) \times (p_2 - p_3)|}{6}$$

Likewise, the mapping from points in $\mathcal{Y}$ back to the space of the original data $F$ can be constructed by replacing the $F^x_i$ in the barycentric combination with the original frames $F_i$.

5. Phonetic Clustering on the Speech Manifold

Of further interest is how lip shapes corresponding to phone mid-points cluster on $\mathcal{M}$. An example of clustering for bil-

---

3The same temporal resolution as the original data.
it is evident from captured trajectories that phonetic centres are
rôle of context in the realisation of target extrema. Furthermore,
trajectory. Naturally this is obfuscated by coarticulation, the
clusters - between which interpolation generates the synthetic
sis becomes a matter of selecting key targets from each of the
stances onto the speech manifold, and that these cluster, synthe-
sented, which demonstrate the separation of phonetic groups
initial results for clustering of phone in-
dynamics are included within the parameterisation. [Wang et al. 2005] Wang, J.M., Fleet, D.J., and Hertzmann, A., Gaus-
model. This is a form
abials can be seen in fig. This group demonstrates an ex-
ample of dynamics separating groups which are similar when
only lip-shape is taken into account. The nasal /m/ cluster is
located mostly in the closing half of the speech cycle whereas
the voiced/voiceless /b/ and /p/ clusters are located mostly on
the opening half of the cycle. Spatially all three classes cluster
about narrow/closed lip shapes, but are widely spread across all
velocities - likely according to the surrounding context. This fits
with the plosive nature of /p,b/, and the nasal stop description
of /m/. This indicates that the nasal /m/ group should be treated
separately from the plosive group /p,b/, which closely overlap,
whilst in many cases where dynamics are not taken into account
these are all treated as one group (e.g. in viseme models.)

6. Conclusions

In this paper we have presented a novel parameterisation of 3D
lip speech movements. This parameterisation is a constructed
space in which lip movements can be visualised as elliptical
paths on a non-linear manifold. Geodesic distances on the mani-
fold, calculated according to the temporal ordering of the orig-
inal data, allow us to extract an underlying causal structure
to the task space. By exploiting this ordering of data, trajec-
tories between discrete states on the speech manifold can be
generated using a finite-state Markov model. This is a form of
data-driven target interpolation which is not constrained by
a pre-defined mathematical model of speech trajectories (such
as spline-based models), or by the discrete set of pre-captured
fundamental units (such as with concatenative models.) Some
initial results for clustering of phone mid-points are been pre-
sented, which demonstrate the separation of phonetic groups
when dynamics are included within the parameterisation.

The presented techniques are intended to aid the task of
speech synthesis. Given that we can project sets of phone in-
stances onto the speech manifold, and that these cluster, synthe-
sis becomes a matter of selecting key targets from each of the
clusters - between which interpolation generates the synthetic
trajectory. Naturally this is obfuscated by coarticulation, the
rôle of context in the realisation of target extrema. Furthermore,
it is evident from captured trajectories that phonetic centres are
not the only key features within speech utterances. Future work
will concentrate upon the identification and selection of key tar-
gets from the speech manifold.

7. References


for the synthesis of 3D faces, Proceedings of ACM Siggraph,

[Cao et al. 2003] Cao, Y., Faloutsos, P., and Pighin F., Unsupervised
learning for speech motion editing, Proceedings of Symposium

[Cohen and Massaro 1993] Cohen, M.M., and Massaro, D.W., Mod-
eling coarticulation in synthetic visual speech, In: Magenat-
Thalmann, N., Thalmann, D. (Eds.), Models and Techniques in

[Cootes et al. 1998] Cootes, T.F., Edwards, G.J., and Taylor, C.J., Ac-
tive appearance models, Proceedings of Symposium on

[Edge et al. 2008] Edge, J.D., Hilton A., Nadotka, N., 3D video face
capture and registration, BMVA Symposium on 3D Video

and vector methods used in multivariate analysis, Biometrika 53,
1966, pp.325-338.

[Grochow et al. 2004] Grochow, K., Martin, S.L., Hertzmann, A.,
Popovic, Z. Style-based inverse kinematics, Proceedings of ACM

[Kshirsagar and Magnenat-Thalmann 2003] Kshirsagar, S., and
Magnenat-Thalmann, N., Visyllable based speech animation,

[Roweis 1997] Roweis, S., EM algorithms for PCA and SPCA, Pro-

[Roweis and Saul 2000] Roweis, S., and Saul, L., Nonlinear dimen-
sionality reduction by locally linear embedding, Science 290
(5500), 2000, pp.2323–2326.

[Saltzman et al. 2006] Saltzman, E., Kubo, M., and Tsao, C., Con-
trolled variables, the uncontrolled manifold method, and the task-
dynamic model of speech production, Dynamics of Speech Pro-

[Tenenbaum et al. 2000] Tenenbaum, J.B., de Silva, V., and Langford,
J.C., A global geometric framework for nonlinear dimensionality

A real-time speech-driven talking head using active appearance

ian process dynamical models, Proceedings of NIPS’05, 2005,
pp. 1441-1448.

Figure 4: Clustering of bilabial consonants centres on the
speech manifold $M$: /b/ centres are shown in red, /p/ centres are
shown in green, /m/ centres are shown in yellow. Ellipses show
2 s.d. distribution of states, squares show mean state. The /p,b/
centers are mainly located on the opening half of the speech
cycle, and the nasal /m/ clusters are mainly within the closing
half of the cycle.