Ultrax: An Animated Midsagittal Vocal Tract Display for Speech Therapy

Korin Richmond and Steve Renals

The Centre for Speech Technology Research,
School of Informatics, University of Edinburgh, United Kingdom
`korin@cstr.ed.ac.uk, s.renals@ed.ac.uk`

Abstract

Speech sound disorders (SSD) are the most common communication impairment in childhood, and can hamper social development and learning. Current speech therapy interventions rely predominantly on the auditory skills of the child, as little technology is available to assist in diagnosis and therapy of SSDs. Realtime visualisation of tongue movements has the potential to bring enormous benefit to speech therapy. Ultrasound scanning offers this possibility, although its display may be hard to interpret. Our ultimate goal is to exploit ultrasound to track tongue movement, while displaying a simplified, diagrammatic vocal tract that is easier for the user to interpret. In this paper, we outline a general approach to this problem, combining a latent space model with a dimensionality reducing model of vocal tract shapes. We assess the feasibility of this approach using magnetic resonance imaging (MRI) scans to train a model of vocal tract shapes, which is animated using electromagnetic articulography (EMA) data from the same speaker.

Index Terms: Ultrasound, speech therapy, vocal tract visualisation

1. Introduction

Speech sound disorders (SSD), whereby a speaker has difficulty producing a given speech sound of their native language clearly and distinctly, are the most common communication impairment in childhood. Approximately 6.5% of all UK children are affected. Having an SSD can affect a child's confidence, and may introduce a barrier to communicating with peers and teachers. This in turn can bring a detrimental effect on learning, as well as social interaction and development. Though speech therapy interventions for SSDs are available, they typically rely heavily on auditory skills; clients must listen to the sounds they produce and try to modify them. This can be problematic where a client does not have strong auditory skills, as is often the case in children with an SSD. Technology to assist in treating SSDs is currently limited. For external, labial articulation a simple mirror is useful, while electropalatography (EPG) can give visual feedback of tongue contact with the roof of the mouth. But crucially, visual feedback of other articulation within the oral cavity is not currently an option during speech therapy. The *Ultrax* project (`http://www.ultrax-speech.org`) is a three year project that aims to address this lack and provide a means to give visual feedback of articulation within the mouth in real time, which will provide a valuable aid in the assessment, diagnosis and treatment of SSDs.

For live feedback, it will be necessary to obtain measurements of articulation within the mouth in realtime and then display this to the user in a simple and intuitive way. Of the various techniques for capturing intraoral articulation, ultrasound is arguably the most promising. By placing a standard medical probe under the chin it is possible to capture tongue movements in a relatively cheap, convenient, and minimally invasive way. However, though ultrasound has for several years provided tongue movement data for research purposes, several drawbacks mean a raw ultrasound display may not be well suited as a speech therapy tool for children. Fig. 1 gives an example snapshot of the tongue. In addition to the potential image quality problems indicated, a critical drawback is the lack of landmarks to provide context for the tongue contour. It is not possible to see passive articulators, such as the hard and soft palate or the teeth. Nor is there even any indication as to which is the front or back of the tongue. Such limitations mean a raw ultrasound display can be difficult to interpret. The goal of *Ultrax* is to use ultrasound not only to capture the movements of the tongue, but to provide an enhanced, diagrammatic display to the user featuring detail not visible in ultrasound data itself.

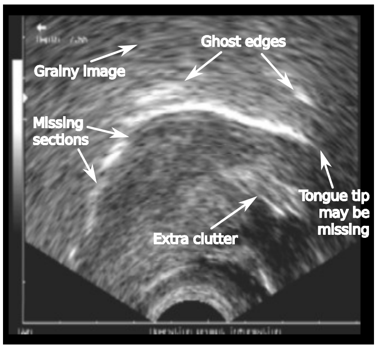

Figure 1: *Typical midsagittal ultrasound image of the tongue.*

Previous work on extracting tongue contours from ultrasound data would seem most relevant to this topic. The most well known example is probably *EdgeTrak* [1], which uses an active contour algorithm that minimises “internal” and “external” error functions that measure contour smoothness and alignment to contrast edges in the image respectively. Another recent and interesting example uses deep belief networks (DBNs) [2]. However, neither of these approaches are in fact suitable for *Ultrax*. Although *EdgeTrak* is useful for semi-automating the process of tongue contour labelling by human experts, it requires too much human intervention and supervision to serve as a fully automatic realtime tongue tracker. Though the DBN approach reportedly runs in realtime with human-like performance, it requires significant amounts of labelled training data beforehand. It is also unclear how it scales to speaker-independent labelling. Most crucially, however, both approaches were developed to label tongue contours for research purposes, whereas *Ultrax* requires realtime visualisation of more of the vocal tract: more of the tongue than may be visible in ultrasound image data, as well as other landmarks such as the passive articulators.

We thus require additional data to build a vocal tract (VT)

This work is supported by EPSRC Healthcare Partnerships Grant number EP/I027696/1 (“Ultrax”).

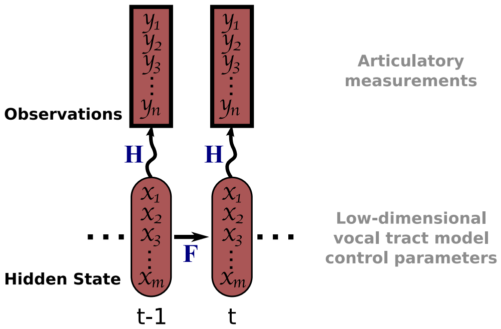
y_1 y_2 y_3 … y_n
Observations
H
H
F
x_1 x_2 x_3 … x_m
Hidden State
t-1
t
Articulatory measurements
Low-dimensional vocal tract model control parameters

Figure 2: *At the heart of our approach is a latent space model*

model that may be animated in realtime. A prime candidate source for this extra anatomical data is magnetic resonance imaging (MRI). A good example of work that has used MRI data to build an animatable VT model is that of Badin et al. [3]. They used MRI and computerised tomography (CT) data together with "guided" principal components analysis (PCA) to build a model of VT shapes that may be animated, or driven, by movements of a handful of fleshpoints on a subject's articulators recorded using electromagnetic articulography (EMA) data. This work is similar to the goals of *Ultrax*—although our end goal is to animate a VT display using ultrasound data—and we have also chosen to use EMA data initially. There are several reasons for this. Practically, we can proceed with data that is already available [8, 9], and thus avoid waiting to acquire MRI and ultrasound data matched for subject. More importantly, EMA arguably offers a simpler articulatory representation, in contrast to high dimensional ultrasound image data. It has already been shown that whole tongue and VT contours can be predicted with considerable accuracy from a small number of fleshpoints[3, 4, 5]. Therefore, EMA data offers a simplification that allows us to build a proof-of-concept system. The separate problem of dealing with high dimensional ultrasound data is thus put aside for now.

In the remainder of this paper, Section 2 gives a general overview of the approach we propose. Section 3 describes the prototype we have built using this approach. In Section 4 we present an experimental evaluation of this prototype.

2. Overview of proposed approach

There are two main components at the heart of our approach to a realtime animated midsagittal VT display. The first is a dimensionality-reducing model of midsagittal VT shapes. The role of this model is to map from a small number of control parameters to a representation of a snapshot VT configuration that may in turn be displayed to a user. This could be a linear model, similar to the PCA described by [3] for example, or nonlinear, as in [5] where radial basis function (RBF) networks predicted whole tongue contours from a small number of fleshpoints.

The second component is a latent space model. For a full understanding of this family of models, the reader may refer to the large body of tutorial material available (e.g. [6] provides excellent, easily-accessible coverage). Here, we aim just to convey the basic principles, as depicted in Fig. 2. Such models maintain a probability distribution over hidden variables at each time t. These are termed hidden variables because they cannot be directly observed or measured, but can only be inferred. The evolution of the hidden state from time $t-1$ to t is governed by an arbitrary function F, designed by the modeller to impart any knowledge about the behaviour of the system of hidden variables. In lieu of any more sophisticated knowledge, the identity transform may be used, which implies the hidden state is expected to stay the same through one time step (though added noise means the state can evolve smoothly over time).

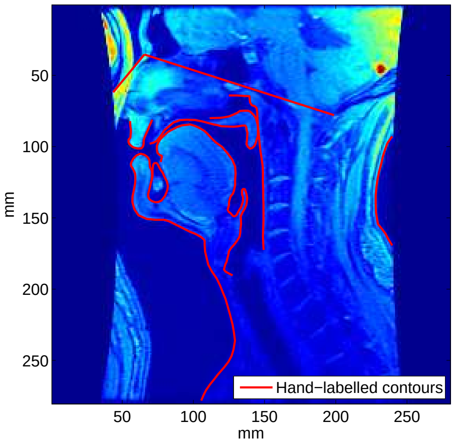

Figure 3: *Hand-traced contours for MRI scan of [ʃ] phone.*

The hidden state is related to the observable variables via a second function H. As indicated in Fig. 2, we propose to use the hidden state to represent the set of control parameters of the dimensionality-reducing VT shape model, hence H takes the form of this mapping model. In addition, though, H must also include the extra step of transforming the mapped VT shape to match the vector form of the observation data for a given time frame. For example, to drive the animation using EMA data, points corresponding to the EMA coil locations used must be selected in the mapped VT shape. To use ultrasound data, the mapped VT shape must likewise be converted to the same representation of features that may be extracted from an ultrasound image. There are many possible ways this might be achieved, but using EMA fleshpoints is conceptually simpler, and offers a way at this initial stage to establish the viability of our proposed approach to animating a VT display with articulatory data. Finally, having implemented suitable F and H mappings, standard recursive Bayesian estimation algorithms can perform state estimation for a given sequence of articulatory observations. In principle, this yields the optimum state sequence to match the articulatory observations, but by using the VT model above we can in effect animate the entire VT display.

There are many variants of latent space model to choose from: for a linear VT shape model, a simple Kalman filter will suffice; if either F or H are nonlinear, we can use the Unscented Kalman Filter [7]; moving beyond Gaussian state or observation distributions would require a particle filter, and so on. Thus, model choice depends largely on how F and H are implemented.

3. Vocal tract model

Here, we describe our first implementation of the general approach outlined above. Since the aim was to pilot the approach in a similar way to [3, 4] (i.e. with speaker-dependent EMA and MRI), we could use the publicly available *mngu0* corpus [8, 9]. This contains 3D MRI scans for 28 phones, and around 1300 utterances recorded using EMA for the same subject, with coils on the upper and lower lips (UL & LL), the lower incisor (LI), and three coils on the tongue (T1, T2 & T3, from tip to dorsum).

3.1. Contour extraction

The first step was to parameterise a range of VT configurations. We selected the midsagittal slices from the *mngu0*

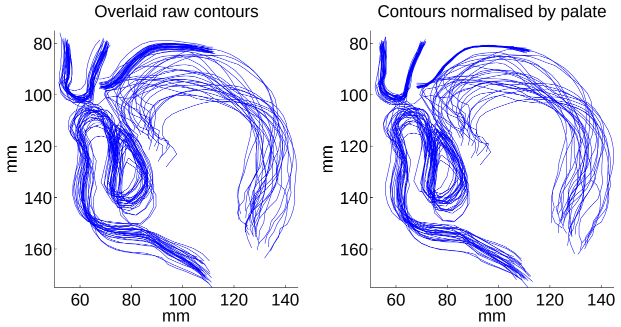

Figure 4: *Contours extracted from* mngu0 *MRI overlaid (left), and those normalised by hard palate position (right).*

3D MRI scans, and traced selected contours using a custom-designed graphical user interface (GUI). Fig. 3 indicates the hand-labelled contours for a scan of example phone [ʃ]. We chose to trace the outline of several structures (though not all are used here): the upper and lower lips; the jaw, chin and neck; the hard palate, soft palate and velum; the tongue and epiglottis; and the rear pharynx wall. The tracing process was semi-automated using an algorithm based on Gradient Vector Flow [10], which allowed to place a spline roughly in the vicinity of the curve to be traced in our custom GUI and then automatically "attract" the spline to the curve. Though this made labelling faster, hand-tweaking of spline points was often needed to achieve a perfect match. No constraint was imposed on the number of spline points that could be used for a given curve, but care was taken to maintain consistent start and end points in different scans. This was often difficult. For example, it proved easy to identify a consistent landmark to anchor one end of the pharynx wall, but for some articulations it was hard to differentiate the tip of the tongue from the mouth floor.

In addition, we labelled three reference points that could be located relative to the skull reasonably reliably, indicated by the two straight lines in Fig. 3. These reference points offer one way to perform head-movement correction to normalise away slight changes in head pose between different scans.

3.2. Normalisation

The raw contour data was normalised in two important ways. First, the splines were "resampled" to find a fixed number of spline points on each contour, in order to make modelling easier. Second, all splines were shifted to correct for changes in head pose. We evaluated two methods for this. For the first, we automatically adjusted the hard palate in each frame to minimise the translation and rotation relative to a selected reference palate (in terms of least squares error), then adjusted all other splines in that frame accordingly. The effect of this can be seen by comparing the plots in Fig. 4. The second method was similar, but used two of the available reference points to translate and rotate each set of contours. Following informal empirical evaluation, we decided to use the palate-based method.

3.3. Dimensionality reducing vocal tract shape model

To begin simply, we implemented a linear dimensionality-reducing VT model using PCA, applied to the normalised contours shown in Fig. 4 (x- and y-coordinates of 132 spline points in total). This is similar to [3], though simpler in fact, since they employed a step-wise "guided" PCA to derive linear components corresponding to biomechanically plausible movements (or "degrees of freedom") of individual articulators. For our application though, independent control over individual articulators is less important, so we chose to begin with a linear model where coordinated movements of articulators may be effected by a single principal component (PC) vector[1].

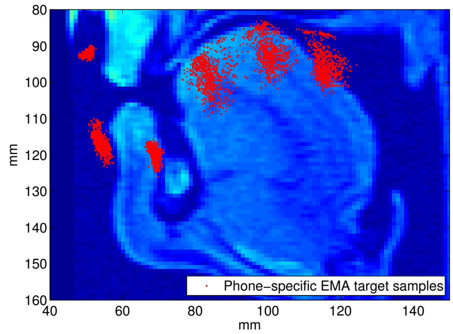

Figure 5: *Scatter plot of sample EMA points shifted and rotated to MRI coordinate space "by eye".*

3.4. Achieving EMA-driven animation

The final step needed to animate the PCA VT shape model was to match the EMA data to the VT shapes. First, the EMA coil coordinates needed to be translated into the MRI coordinate system. A scatter plot of EMA points at multiple instances of the 28 phones was overlaid on a selected MRI image. These were adjusted to find a reasonable match by eye, thus identifying a suitable translation of the EMA data into the MRI coordinate space, as indicated in Fig. 5. Note this could not be performed exactly since EMA does not give the exact coordinates of the fleshpoint to which a coil is attached, due to the construction of the coils. This unfortunately may have introduced error, but was difficult to avoid. Second, to use the mapping within the latent space model, we needed to identify points on the VT contours corresponding to the EMA coil locations on the subject's articulators. For this we overlaid a phone-specific scatter plot of EMA points onto each corresponding set of VT shape contours, and labelled points on those contours to match the EMA positions. This was again done by eye, and is a further potential source of error. These points were appended to the vectors of contour spline points prior to the PCA described in Section 3.3.

With our initial implementation of the H mapping in Fig. 2 complete, we opted to use the simple identity transform for F. Since both these are linear, we could then use a standard Kalman filter to animate the VT display.

4. Evaluation

Evaluating the system described above is unfortunately not straightforward. Whereas [4] or [5] had a large set of tongue contours available for evaluation purposes, the only ground-truth data available here are the coordinates of the 6 EMA coils. Our objective evaluation is thus limited to comparing the EMA coil positions with the corresponding points on the VT contours identified in Section 3.4. We first evaluated the effect of varying hidden state size (i.e. number of PCA components). Fig. 6 gives results in terms of root mean square (RMS) error expressed in millimetres, calculated over 20 test utterances. Generally, increasing the hidden state size reduces the error, reaching approximately zero when the hidden state size is twelve. This

[1] see http://homepages.inf.ed.ac.uk/korin/ultraxis2012 for video visualising selected principal components, as well as natural utterances.

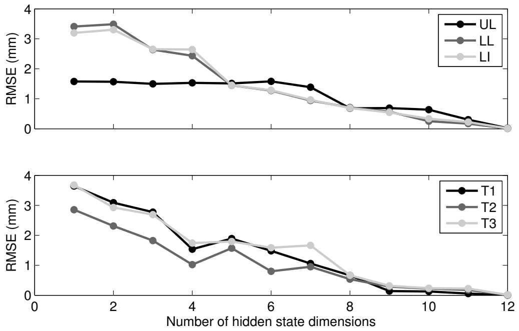

Figure 6: *The effect of varying hidden state dimensionality on the performance of tracking each EMA coil.*

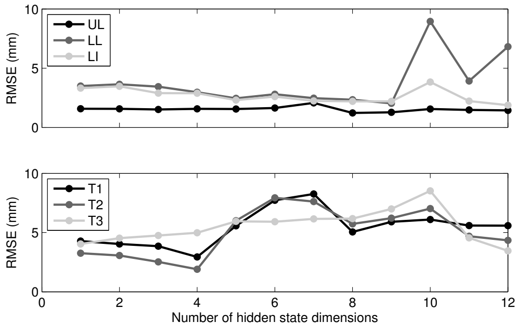

Figure 7: *The effect of varying hidden state dimensionality on the accuracy of "held-out" EMA point prediction.*

indicates the Kalman filter is working successfully to track the target EMA points, but is otherwise unsurprising.

We then performed a more challenging evaluation, whereby each of the EMA coils was left out of the observation vector in turn, and the accuracy of predicting that coil using observations of only the remaining five coils to drive animation was tested. These results are shown in Fig. 7. Generally, these results are slightly worse than those in Fig. 6. This is to be expected since in this task no observations were available to the Kalman filter that corresponded directly to each held-out target point. In addition, less observation data was available. It has been shown that predicting tongue contours from two tongue fleshpoints is less accurate than when three or more are used [4, 5]. Another important result is that using more PCA components in the hidden state does not necessarily increase accuracy. This is especially true for the front two tongue points, two of the most salient articulators, for which the optimum state dimensionality appears to be four. Informally viewing several animated utterances supported this finding. This is an intriguing result, for which one possible explanation is that using only five or six EMA point observations does not bring adequate information to constrain the model when extra degrees of freedom above four are available.

Finally, to gain some impression of the significance of the errors in Fig. 7, we compared the proposed approach (using four hidden state dimensions) with two linear mappings to perform the same leave-one-out prediction task. These results are shown in Fig. 8. "Linear Mapping 1" was trained on the same 28 sets of EMA target points as used for the PCA in Section 3.3. "Linear Mapping 2" was trained using all recorded EMA coil positions for the test sentences, and gives a rough lower bound on the error possible. Predicting coil positions via the (linear) MRI-derived shape model is on average only 0.4mm worse than using a more favourable direct linear mapping. Based on these results, the approach seems promising and worth pursuing further.

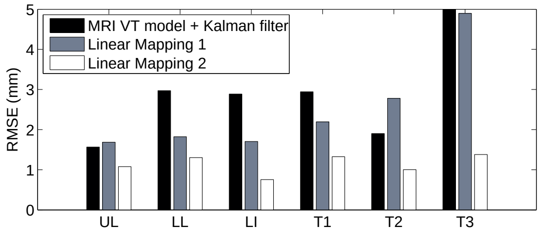

Figure 8: *Error when predicting held-out coil position using the proposed approach (mean 2.9mm) compared with that of Linear Mappings 1&2 (mean 2.5mm and 1.1mm respectively).*

5. Conclusion

In this paper, we have used MRI scans to train a model of vocal tract shapes and then animated it using EMA data separately collected from the same speaker. Our approach to this problem was based on combining a latent space model with a dimensionality reduction model of vocal tract shapes. We carried out experiments on the *mngu0* corpus, which contains both MRI and EMA data from a single speaker, matching the MRI and EMA coordinate systems by hand. The system was evaluated by prediction of the position of an EMA coil, on a leave-one-out basis. We find the results of this pilot are encouraging.

6. References

[1] M. Li, C. Kambhamettu, and M. Stone, "Automatic contour tracking in ultrasound images," *International Journal of Clinical Linguistics and Phonetics*, vol. 19, no. 6/7, pp. 545–554, 2005.

[2] I. Fasel and J. Berry, "Deep belief networks for real-time extraction of tongue contours from ultrasound during speech," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, aug. 2010, pp. 1493 –1496.

[3] P. Badin, F. Elisei, G. Bailly, and Y. Tarabalka, *An Audiovisual Talking Head for Augmented Speech Generation: Models and Animations Based on a Real Speaker's Articulatory Data*, ser. LNCS. Berlin, Heidelberg, Germany: Springer-Verlag, 2008, no. 5098, pp. 132–143.

[4] T. Kaburagi and M. Honda, "Determination of sagittal tongue shape from the positions of points on the tongue surface," *The Journal of the Acoustical Society of America*, vol. 96, no. 3, pp. 1356–1366, 1994.

[5] C. Qin, M. Carreira-Perpiñán, K. Richmond, A. Wrench, and S. Renals, "Predicting tongue shapes from a few landmark locations," in *Proc. Interspeech*, Brisbane, Australia, Sept. 2008, pp. 2306–2309.

[6] D. Simon, *Optimal state estimation: Kalman, H [infinity] and nonlinear approaches*. John Wiley and Sons, 2006.

[7] S. Julier and J. Uhlmann, "A new extension of the Kalman filter to nonlinear systems," in *Int. Symp. Aerospace/Defense Sensing, Simul. and Controls*, vol. 3. Spie Bellingham, WA, 1997, p. 26.

[8] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Proc. Interspeech*, Florence, Italy, August 2011, pp. 1505–1508.

[9] I. Steiner, K. Richmond, I. Marshall, and C. D. Gray, "The magnetic resonance imaging subset of the mngu0 articulatory corpus," *The Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. EL106–EL111, January 2012.

[10] C. Xu and J. Prince, "Snakes, shapes, and gradient vector flow," *Image Processing, IEEE Transactions on*, vol. 7, no. 3, pp. 359 –369, March 1998.