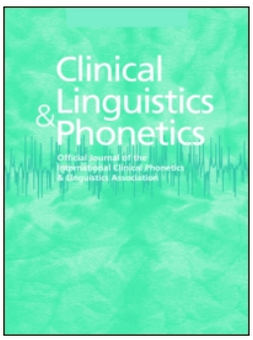

Clinical Linguistics & Phonetics

ISSN: 0269-9206 (Print) 1464-5076 (Online) Journal homepage: http://www.tandfonline.com/loi/iclp20

Robust contour tracking in ultrasound tongue image sequences

Kele Xu, Yin Yang, Maureen Stone, Aurore Jaumard-Hakoun, Clémence Leboullenger, Gérard Dreyfus, Pierre Roussel & Bruce Denby

To cite this article: Kele Xu, Yin Yang, Maureen Stone, Aurore Jaumard-Hakoun, Clémence Leboullenger, Gérard Dreyfus, Pierre Roussel & Bruce Denby (2016): Robust contour tracking in ultrasound tongue image sequences, Clinical Linguistics & Phonetics, DOI: 10.3109/02699206.2015.1110714

To link to this article: http://dx.doi.org/10.3109/02699206.2015.1110714

Published online: 20 Jan 2016.

CLINICAL LINGUISTICS & PHONETICS
http://dx.doi.org/10.3109/02699206.2015.1110714

Taylor & Francis
Taylor & Francis Group

Robust contour tracking in ultrasound tongue image sequences

Kele Xu[a,b,*], Yin Yang[c], Maureen Stone[d], Aurore Jaumard-Hakoun[a,b,*], Clémence Leboullenger[a,b,*], Gérard Dreyfus[b], Pierre Roussel[b,*], and Bruce Denby[e]

[a]Faculty of Engineering, Université Pierre et Marie Curie, Paris, France; [b]Signal Processing and Machine Learning (SIGMA) Lab, ESPCI ParisTech, Paris, France; [c]Electrical and Computer Engineering Department, University of New Mexico, Albuquerque, New Mexico, USA; [d]Vocal Tract Visualization Lab, University of Maryland Dental School, Baltimore, Maryland, USA; [e]Cognitive Computing and Applications Lab, Tianjin University, Tianjin, China

ABSTRACT

A new contour-tracking algorithm is presented for ultrasound tongue image sequences, which can follow the motion of tongue contours over long durations with good robustness. To cope with missing segments caused by noise, or by the tongue midsagittal surface being parallel to the direction of ultrasound wave propagation, active contours with a contour-similarity constraint are introduced, which can be used to provide 'prior' shape information. Also, in order to address accumulation of tracking errors over long sequences, we present an automatic re-initialization technique, based on the complex wavelet image similarity index. Experiments on synthetic data and on real 60 frame per second (fps) data from different subjects demonstrate that the proposed method gives good contour tracking for ultrasound image sequences even over durations of minutes, which can be useful in applications such as speech recognition where very long sequences must be analyzed in their entirety.

ARTICLE HISTORY
Received 15 February 2015
Revised 12 October 2015
Accepted 17 October 2015

KEYWORDS
Active contour model; automatic re-initialization; contour tracking; similarity constraint; tongue

Introduction

Ultrasound imaging is currently an attractive imaging modality for real-time human tongue movement (Stone, 2005), with applications in a variety of fields, such as the study of speech disorders and three-dimensional (3D) tongue modeling. Accurate, robust tongue contour extraction, however, remains a challenging problem for ultrasound sequences of long duration, due to acoustic effects such as shadows, speckle noise and low signal-to-noise-ratio (SNR). Despite significant research efforts, manual re-initialization is still usually needed, which is impractical for large-corpora systems.

A variety of processing techniques can be used to track the contour of the tongue, for example, active contour models (Akgul, Kambhamettu, & Stone, 1999; Li, Kambhamettu, & Stone, 2005); active appearance models (Roussos, Katsamanis, & Maragos, 2009); machine learning-based tracking (Fasel & Berry, 2010; Jaumard-Hakoun et al., 2015) and ultrasound image segmentation-based approaches (Tang, Bressmann, & Hamarneh,

CONTACT Bruce Denby denby@ieee.org School of Computer Science and Technology, Building 55, Tianjin University, 135 Yaguan Road, Haihe Education Park, Jinnan District, Tianjin 300350, China.
*Present address: Langevin Institute, ESPCI ParisTech, Paris, France.

2012). More recently, some researchers have also proposed using physical properties of the tongue, contained in a realistic 3D model of the tongue, to help guide the contour extraction process (Wrench and Balch, 2015; Xu et al., 2014, 2015).

Due to their ability to be guided by constraint forces, active contours may be particularly useful for contour tracking in images of the tongue. Indeed, tongue contour tracking using energy-minimization-based active contours, or 'Snakes', has been used extensively in previous research. In the contour-tracking algorithm proposed by Akgul et al. (1999), the snake model was used on ultrasound tongue images for the first time, introducing gradient information in the definition of an external energy term. By including an intensity-related constraint, Li et al. (2005) proposed a new contour-tracking system, named *EdgeTrak*, which works very well for sequences in which the entire contour always remains visible. If a part of the contour disappears in some images, however, due to poor acoustic coupling or a decrease in reflected energy, the obtained contour can become erroneous and require manual re-initialization to get back on track. This can become problematic for applications where long speech sequences are to be analyzed.

To help cope with low SNR in ultrasound images, some researchers have proposed using other imaging modalities (e.g., X-rays) to obtain prior tongue shape information (Roussos et al., 2009). However, these modalities may use different frame rates, and registration between different modalities can also be difficult, making such an approach impractical (the use of X-ray is also nowadays banned). In this article, we explore the idea of extracting prior shape information from the ultrasound image sequence itself, included as an extra force to guide the evolution of the hypothetical tongue contour to better handle images with missing/vague contours. To alleviate the problem of accumulated tracking error over long sequences, which can necessitate manual re-initialization, we also propose an automatic procedure, based on the complex wavelet structural similarity (CW-SSIM) index (Sampat, Wang, Gupta, Bovik, & Markey, 2009). The resulting algorithm has been tested on both synthetic data and real ultrasound image sequences from multiple subjects. Results obtained demonstrate that the proposed method can improve robustness of active contours against missing segments and has the ability to re-initialize the tracking automatically without manual intervention. Such an automatic tracking approach can act as a complement to hand-scanning and more traditional (but more labor-intensive) contour extraction tools, such as EdgeTrak and AAA (Articulate Instruments Ltd., 2012) and can be a valuable improvement for research in areas where longer sequences must be analyzed, such as speech production, speech recognition in silent speech interfaces (Denby et al., 2010), swallowing disorders (Sonies et al., 2003) and the like.

Active contour model with similarity constraint

An active contour model is a spline function obtained by minimizing an energy function that is intended to fit the spline to edges present in the image while retaining a reasonably regular shape (Kass, Witkin, & Terzopoulos, 1988). Suppose we have an active model of the tongue contour that can be represented by n discrete points $\mathbf{V} = [\mathrm{v}_1, \mathrm{v}_2, \cdots, \mathrm{v}_n]$, with the total energy for snakes defined as:

$$E_{\text{total}} = E_{\text{int}} + E_{\text{ext}} \tag{1}$$

where E_{int} is the total internal energy (sum over the n contour points of the local internal energies defined in relation (2)) and E_{ext} is the total external energy (sum over the n contour points of the local external energies defined in relation (2)). In this article, we follow the definition of local internal energy and local external energy used by Akgul et al. (1999):

$$\begin{cases} E_{\text{int}}(\mathrm{v}_j) = \alpha\left(1 - \overrightarrow{\mathrm{v}_{j-1}\mathrm{v}_j} \cdot \overrightarrow{\mathrm{v}_j\mathrm{v}_{j+1}} \big/ \left|\overrightarrow{\mathrm{v}_{j-1}\mathrm{v}_j}\right|\left|\overrightarrow{\mathrm{v}_j\mathrm{v}_{j+1}}\right|\right) + \beta\left|\left|\mathrm{v}_j - \mathrm{v}_{j-1}\right| - d\right| \\ E_{\text{ext}}(\mathrm{v}_j) = 1 - \left|\nabla \mathbf{I}(\mathrm{v}_j)\right| / \mathrm{K} \end{cases} \tag{2}$$

where $j = 2, \cdots, n-1$, and α, β are the weighting parameters for the internal energy; d is the average distance between two consecutive points on the contour; v_j is the jth point of the active contour; $\nabla \mathbf{I}$ is the gradient of the image intensity; and K is a normalization constant.

Although *EdgeTrak* used dynamic programming in the optimization process, in practical applications, the performance of the active contour model is prone to corruption by missing boundaries due to movement of the tongue or to speckle noise. To help cope with this problem, a contour sequence similarity constraint is proposed in this work. In an ultrasound image sequence with a high frame rate (in our experiment, 60 fps), a contour extracted in a previous frame should normally be very similar to that obtained in the current frame, i.e., the deviation of contours extracted from adjacent frames should not exceed a certain threshold. Even when the local deformation of the tongue is large, a previous contour can act as a predictor to help regularize the movement of the active contour, since the true motion of the tongue must be physically reasonable.

Before the contour-similarity constraint can be added to the active contour model, an appropriate similarity measure must be defined. A classical method of measuring the similarity between two contours is to calculate distances between corresponding points. However, this is not suitable in our case as there is no straightforward way to identify actual corresponding physical tissue points in contour tracking in ultrasound tongue image sequences. Also, existing techniques to compare contours in the absence of strict point correspondence, such as mean sum of distances (MSD) (see 'Materials and methods' section), would be difficult to integrate into an energy-based active contour approach. Here, we explore the use of the rank of a matrix formed from a set of contours to measure the similarity of the contours of a contour sequence, as in (Zhou, Huang, Duncan, & Yu, 2013). Let a set of m consecutive contours be represented by vectors $\mathbf{c}_j$, for $j = 1, 2, \cdots, m$ (the size of the vector is $2n \times 1$). Each $\mathbf{c}_j = \left[\mathbf{c}_j^x, \mathbf{c}_j^y\right]^T$ is obtained by concatenating vectors $\mathbf{c}_j^x$ and $\mathbf{c}_j^y$, where $\mathbf{c}_j^x$ is the orthogonal projection of vector $\mathbf{c}_j$ on the x axis, and $\mathbf{c}_j^y$ is the orthogonal projection on the y axis. For an image sequence of a sufficiently high frame rate, it can be assumed that $\mathbf{c}_j$ is generated from $\mathbf{c}_{j-1}$ via an affine transformation. The vectors form a matrix $\mathbf{C} = \left[\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_j, \cdots, \mathbf{c}_m\right]$ (of size $2n \times m$, where n is the number of discrete points and m is the number of consecutive contours), which is a low-rank matrix for any m. It was proved in Zhou et al. (2013) that $rank(C) \leq 6$.

To enforce similarity between the contours extracted from adjacent frames, a constraint term is added to the energy defined in (1), resulting in a new energy E_{sim} for each contour:

$$E_{\text{sim}} = E_{\text{int}} + E_{\text{ext}} + \gamma\ rank(\mathbf{C}) \tag{3}$$

where γ is a weighting parameter. Now, the rank of a matrix is a discrete quantity, which makes the optimization of E_{sim} difficult. One way to solve this problem is to use the nuclear norm regularized linear least-squares technique to rewrite the constraint term as (Zhou et al., 2013):

$$E_{\text{sim}} = E_{\text{int}} + E_{\text{ext}} + \gamma\|\mathbf{C}\|_* \tag{4}$$

where $\|\mathbf{C}\|_*$ is the nuclear norm of $\mathbf{C}$ (the sum of the singular values of the matrix $\mathbf{C}$). As the nuclear norm of a matrix is a good approximation to the rank of the matrix, it is common to substitute the rank minimization problem by the minimization of the nuclear norm, as has been widely done in low-rank modeling such as in Candès, Li, Ma, and Wright (2011). The accelerated proximal gradient algorithm of Zhou et al. (2013) is used for the optimization in this article, as follows: the active contours are evolved at each iteration using image-based forces, after which the contour group similarity regularization is imposed by means of singular value thresholding. More detail can be found in Nesterov (2007) and Zhou et al. (2013). Note that for the $m - 1$ frames at the beginning of the sequence, no contour-similarity constraint is added.

Compared to extracting the contour from a single frame, the new force included with the active contour model acts as prior information that influences the movement of the contour. In the proposed algorithm, the internal energy is used to keep the continuity of the contour; the external energy is used to attract the active contour to the real contour in the ultrasound tongue image, while the similarity constraint acts as an additional force to limit the degrees of freedom of the movements of the active contour. When several adjacent contours are clear in the ultrasound tongue image sequences, the external energy will be the dominant term; whereas if dramatic deformations of the tongue occur in some frames, the similarity-constraint force dominates. The value of γ was chosen by hold-out validation: (i) using different values of γ, contours were extracted by the above algorithm from frames belonging to a subset of the hand-labeled data, (ii) the MSD defined in relation (6) was computed for each frame of the subset, and (iii) the value of γ that resulted in the minimum value of the mean MSD was selected for subsequent contour extraction from the whole data set. As can be seen in Figure 1 (a) for synthetic contour data (with random speckle noise added) and Figure 1(b) on some real ultrasound test data for frame when the tongue goes parallel to the ultrasound beam, the similarity constraint makes the active contour more robust and physical.

Automatic re-initialization based on complex wavelet structural similarity (CW-SSIM) measurement

As with other techniques to extract tongue contours in ultrasound image sequences (Li et al., 2005; Tang et al., 2012), the input of our algorithm consists of several points (in our work, at least 12) placed on the surface of the tongue in the first frame of the sequence (NB: As mentioned earlier, there is no fixed relation between the points chosen and physical tissue points). Using this input along with the technique of image similarity measurement, a novel automatic tracking re-initialization is presented in this section. In this way, we may hope to dispense with the need to manually re-initialize the contour finding process due to accumulated tracking error over long sequences.

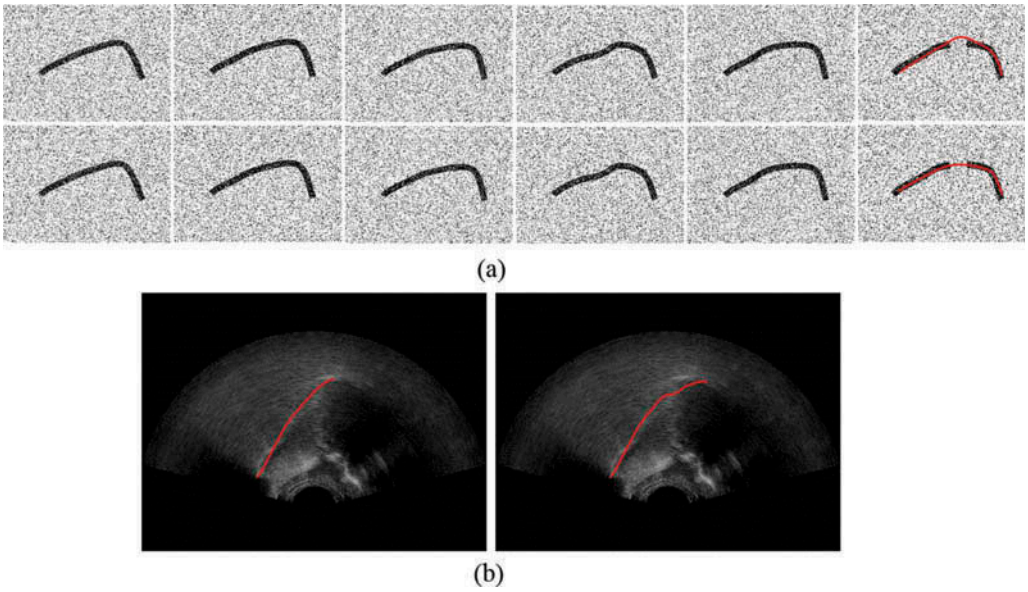

Figure 1. (a) Validation on the synthetic data, each row represents the image sequences and the gray (red) line represents the contour extracted from the image. The gray (red) line in the top row shows the contour extracted without similarity constraint while the one in the bottom row is the one with similarity constraint. In our experiment, m = 6 and n = 24 (see discussion in text). (b) Validation on the ultrasound tongue data. The gray (red) line in the left column shows the contour extracted without similarity constraint while the one in the right column is the one with similarity constraint.

The proposed method works as follows. Before contour tracking is carried out, the image-similarity coefficient (defined below) between the current frame and first frame based on the entire image is calculated. (We note that in our experiment, the first frame always corresponds to the tongue in rest position.) If this coefficient exceeds a set threshold, that is, if the current frame is sufficiently similar to the first frame, the positions of the points on the active contour are re-initialized to those which were input on the first frame of the sequence. This provides a method to prevent accumulation of errors over long sequences, which can lead to erroneous contours, and amounts to a sort of 'automatic re-initialization' of the contour based on prior information.

The simplest and most widely used similarity measure is the mean squared error (MSE), which calculates the mean of squared differences between the intensities of the pixels of the first frame (with manually chosen contour points) and those of the current frame. Due to the multiplicative character of speckle noise, however, the technique does not perform well on ultrasound images. The SSIM-based method (Wang, Bovik, Sheikh, & Simoncelli, 2004) is another kind of widely used image similarity measurement method, which, however, is rather sensitive to local image variations, such as small translations and rotations, which are greatly reduced by the use of a probe-stabilizing helmet, but may still be present to a reduced extent in ultrasound data taken with such a system. In our approach, an image similarity measure based on the CW-SSIM technique (Sampat et al., 2009) is used, which is known to be invariant under small translations, rotations and distortions, and which, as we shall see, provides very satisfactory results on our data.

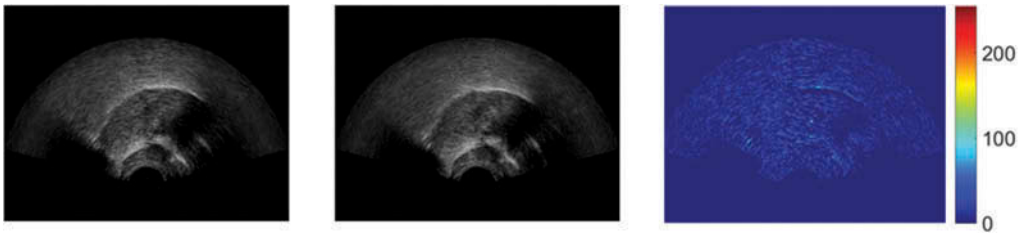

Figure 2. Two frames used to calculate the similarity index, the size of the frame is 320x240. (c) The difference between Frame 1 and Frame 46. (a) Frame number: 1; (b) Frame number: 46; (c) Frame 46 - Frame 1 (colormap).

To implement the CW-SSIM index, the images are decomposed using a complex version of a multi-scale, multi-orientation 'steerable pyramid' decomposition (Simoncelli and Freeman, 1995). We denote by $\mathbf{W}_{\text{img1}} = \{w_{\text{img1},l} | l = 1, \ldots, M\}$ and $\mathbf{W}_{\text{img2}} = \{w_{\text{img2},l} | l = 1, \ldots, M\}$ the complex wavelet coefficients of the two decomposed images at level M of the pyramid decomposition. The CW-SSIM similarity index of the two images is defined as:

$$s(\mathbf{W}_{\text{img1}}, \mathbf{W}_{\text{img2}}) = \frac{2\left|\sum_{l=1}^{M} w_{\text{img1},l} w_{\text{img2},l}^{*}\right| + K}{\sum_{l=1}^{M} |w_{\text{img1},l}|^2 + \sum_{l=1}^{M} |w_{\text{img2},l}|^2 + K} \tag{5}$$

where $w_{\text{img2},l}^{*}$ is the complex conjugate of $w_{\text{img2},l}$, and K is a small positive stabilizing constant. The similarity index is positive, and its maximum value is 1, corresponding to two identical images. As an example, consider the two frames given in Figure 2a and b (from an unpublished dataset from 'Male 1', as defined in the 'Materials and methods' section), in

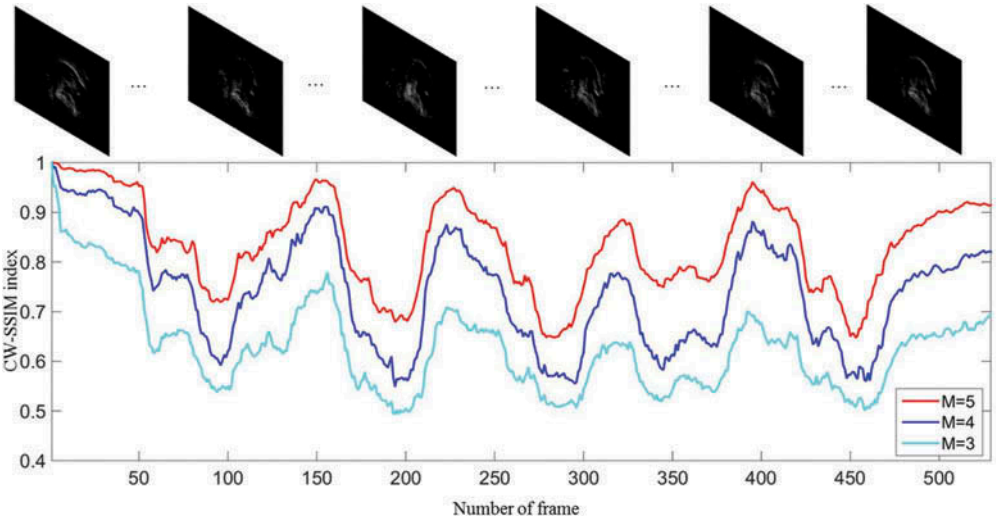

Figure 3. CW-SSIM index of the entire image in an ultrasound image sequence of five utterances of phoneme /k/. Three different levels of decomposition M are shown.

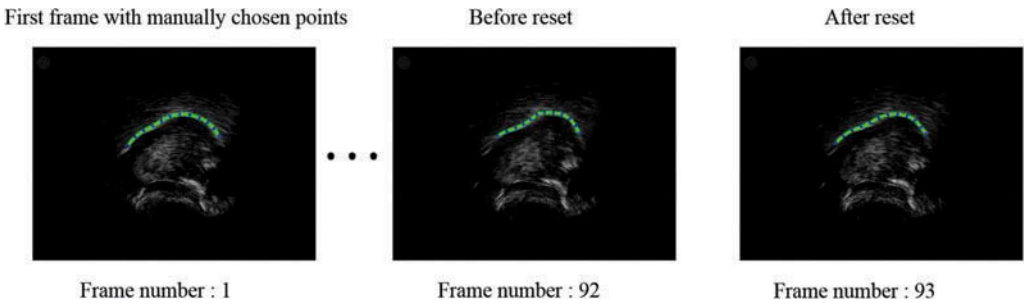

Figure 4. Example of automatic re-initialization in the contour tracking. As the CW-SSIM index between the first frame and the Frame 93 exceeds the threshold, the contour is re-initialized to the original position in the first frame. (The CW-SSIM index between Frame 1 and Frame 92 is 0.79).

which the MSE is 167.32 and the SSIM index 0.69; the CW-SSIM index is 0.86 for four levels of decomposition ($M = 4$) and 0.98 for five levels of decomposition ($M = 5$). More detail can be found in Sampat et al. (2009). Figure 3 (unpublished test data from 'Female 1' as defined in the 'Materials and methods' section) shows the variation of the similarity index during a sequence of five isolated utterances of phoneme /k/. The minima of the similarity indices are seen to occur at the instants at which these utterances are pronounced.

We implemented the re-initialization method based on the code provided by the web site (http://www.mathworks.com/matlabcentral/fileexchange/43017-complex-wavelet-structural-similarity-index-cw-ssim/content/cwssim_index.m), where a MATLAB tool is proposed; more detail can be found in the site (Simoncelli, 2008, MatlabPyrTools).

In this article, to avoid erroneous re-initialization and make a compromise between accuracy and complexity, the CW-SSIM index threshold is set to 0.8, and $M = 4$. In the ultrasound image sequences, if the index exceeds this threshold, the locations of the discrete points will be re-initialized to the positions that were set manually in the first frame, thus improving the contour tracking automatically, without manual intervention. The example given in Figure 4 (from 'Female 2', as defined in the 'Materials and methods' section) shows the automated re-initialization process. Before applying the active contour model on Frame 93, the CW-SSIM index is calculated between 93 and the first frame (with manually chosen contour points). As the index (0.81) exceeds the threshold we set (0.80), the contour is reinitialized as the first frame (in our tests, always the rest position). Without the contour tracking, the execution time of CW-SSIM is about 0.19 seconds for each pair of frames in our tests.

Materials and methods

Four datasets involving five speakers were used in our tests:

(1) An unpublished test dataset from speaker 'Male 1';
(2) An unpublished test dataset of isolated utterances from subject 'Female 1';
(3) Portions of a POLYVAR corpus recorded at our laboratory in 2011 on three speakers – 'Female 2', 'Male 2' and 'Male 3' (Cai et al., 2013);
(4) Portions of a TIMIT corpus recorded in our laboratory in 2010, on speaker 'Male 1' (Cai et al., 2011).

All datasets used a helmet equipped with a 128-element, 1-inch diameter microconvex 4–8 MHz ultrasound probe, running at 60 fps. Details of the helmet used for datasets 1, 2 and 4 can be found in (Cai et al., 2011), and in Al Kork et al. (2014); while the helmets used for dataset 3 are described in (Cai et al., 2013). Our algorithm is implemented using MATLAB 2014b on a Windows 8 desktop with Intel 4-Core 3.7 GHz CPU, 16 GB RAM and ATI Radeon HD 7800 with 6 GB DDR3 VRAM, Dual AMD Filepro 512 GB PCIe-based flash storage. Hand labelling for comparison to automatic tracking results was performed by a single labeller and took about 2 months.

Experimental results

First, we examine the execution time of the proposed algorithm. The size of $\mathbf{C}$ is $2n \times m$, where n is the number of discrete points ($n = 12$) and m is the number of adjacent contours, here set to 10. The singular value decomposition calculation is fast, and the energy minimization using the decomposition is in fact faster than the minimization processing without the contour-similarity constraint; the computationally heaviest part of our work is the calculation of the similarity between the current frame and the first frame (the tongue in rest position) using CW-SSIM. The time performances for the different subjects are very similar, ranging from 220 milliseconds to 239 milliseconds per frame; thus, on average, a sustained rate of roughly 5 Hz can be maintained with the algorithm in its present form. It is interesting to note that offline contour extraction frame rates for EdgeTrak and AAA are quite similar to that obtained here, as long as the sequence under analysis is short enough that long-term drift does not necessitate re-seeding (Alan Wrench, private communication, 2015). On longer sequences, of course – and even more so with hand-scanning – hand re-seeding of contours will slow overall processing time down very dramatically, as compared to the proposed method. In this sense, the approach presented can be viewed as a complementary, more automatic approach, which can be of significant value in applications where long sequences must be analyzed in their entirety, as mentioned earlier. Furthermore, it is conceivable that improvements in the algorithms employed in the proposed method could, in the future, render it real time, even for 60 fps, or higher.

Figure 5 shows an example result, analogous to the test of Figure 1 on synthetic data, on 20 contiguous frames from Female 1, in a segment of the data where the contour is rather faint, due to the orientation of the tongue. Gray (red) lines show the results obtained with the contour-similarity constraint; white (yellow) lines show the results obtained without the contour similarity.

The tracking of another example sequence (Female 2) is shown in Figure 6, which lasts more than 3 minutes (over 17,000 frames). The figure shows the contours from a selection of frames in the sequence. No manual re-initialization was made during the tracking, aside from the initial seeding done on Frame 1. On visual inspection of the tracked contours, the algorithm works quite well.

Occasionally, tracking errors do occur, as shown in Figure 7, due to the presence of a high level of noise or other anomaly in a particular region of the image over an extended

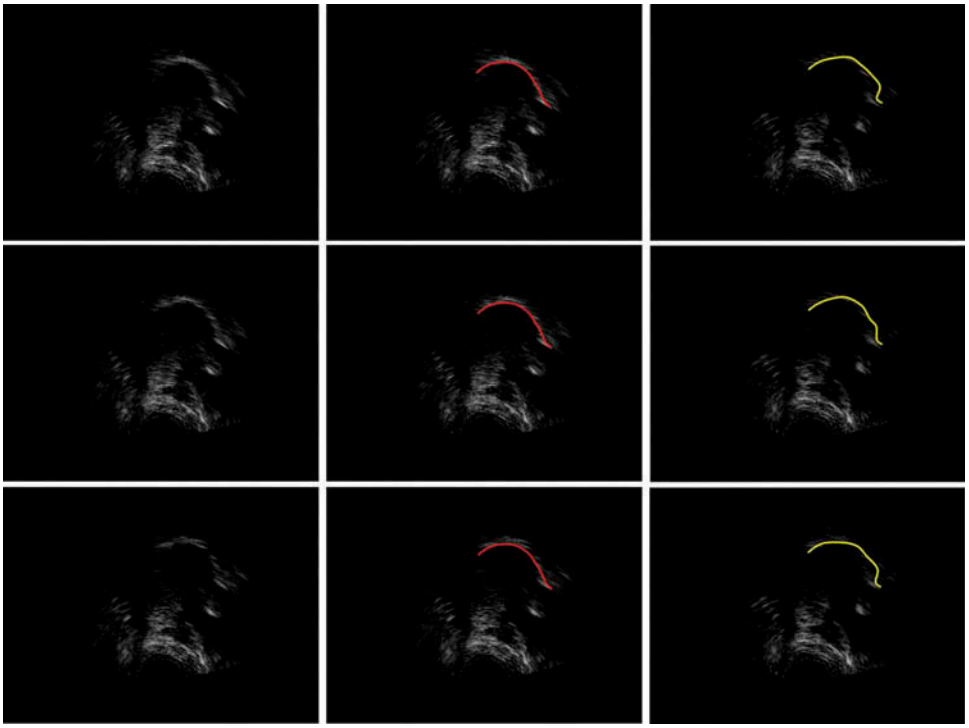

Figure 5. The comparison between the contour-extracted with contour similarity constraint (gray [red] line in second column) and without contour similarity constraint (white [yellow] line in third column). As the number of frames is small, image-similarity-based re-initialization was not necessary here.

period, during which the tongue does not return to the rest position, and the contour is therefore not automatically reset.

To further demonstrate the performance of the algorithm, an additional experiment has been carried out to compare the extracted contours to hand-labeled curves. For validation, MSD is used, defined as:

$$\mathrm{MSD}(\mathbf{U},\mathbf{V}) = \frac{1}{2n}\left(\sum_{i=1}^{n} \min \mathrm{j}\left|\mathrm{v}_i - \mathrm{u}_j\right| + \sum_{i=1}^{n} \min \mathrm{j}\left|\mathrm{u}_i - \mathrm{v}_j\right|\right) \tag{6}$$

where $\mathbf{V}$ is the contour extracted automatically and $\mathbf{U}$ is the result of hand-labelling. Note once again that the points on the compared contours are not physical tissue points but simply representative positions on the contour; MSD simply compares the similarity of two contours U and V by finding, for each point on contour U, the closest point to it on contour V, which, in general, will *not* correspond to a reference point on contour V. As our sequences consist of large numbers of image frames due to the high capture rate (60 fps), it is very laborious to extract contours manually for all frames. Therefore, 4000 frames were chosen randomly for manual contour extraction from the data recording of Female 2, 1000 frames for Male 2 and 2000 frames for Male 3. Compared to the manually extracted contours, the MSD errors

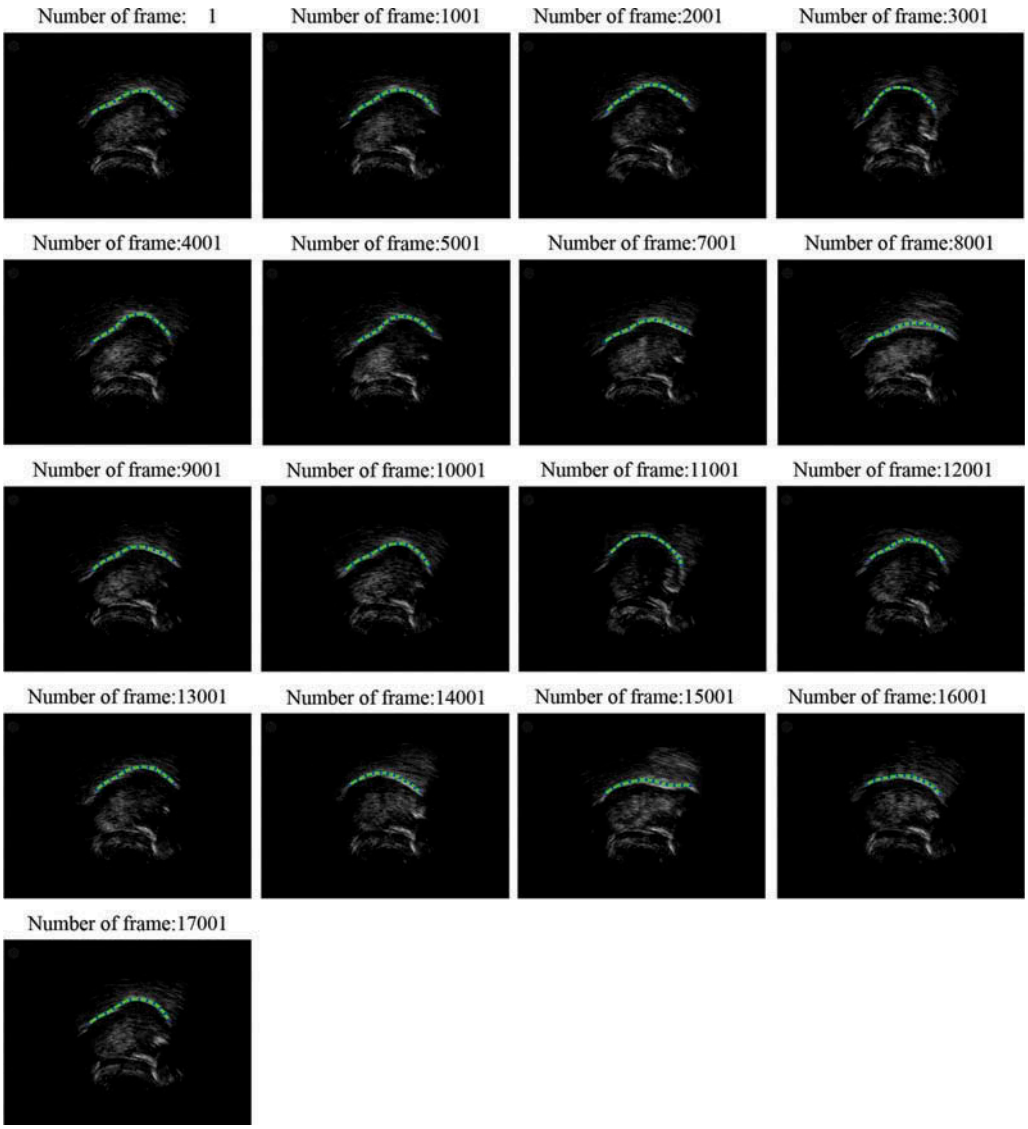

Figure 6. The results (Female 2) of contour tracking in the sequences of long duration (lines are the contour tracked in each frame, while the dots represent the points to represent the curve). To keep the original result, no contour extrapolation is made.

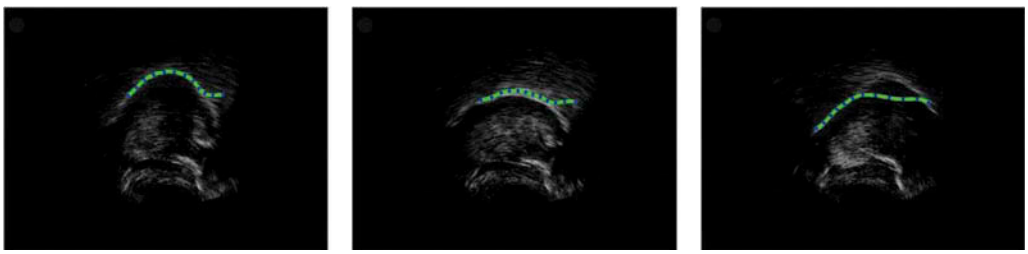

Figure 7. Examples for poor tracking (Female 2).

Table 1. Errors by using different methods for different subjects (for Female 2, 400 contours were extracted manually, for Male 2, 1000 contours were extracted manually, while 2000 contours were extracted manually for Male 3).

	MSD mean errors and standard deviation (pixels, 1 pixel = 0.295 mm)		
Methods	Female 2	Male 2	Male 3
Similarity constraint + CW-SSIM	**3.36 ± 0.86**	**3.65 ± 1.02**	**2.96 ± 0.95**
Similarity constraint + SSIM	4.09 ± 2.01	4.29 ± 2.93	5.84 ± 3.40
No similarity constraint + CW-SSIM	18.96 ± 1.08	16.46 ± 1.29	18.64 ± 1.07
No similarity constraint + SSIM	22.43 ± 2.68	19.43 ± 3.47	21.27 ± 4.55
Similarity constraint	4.52 ± 2.53	5.52 ± 3.41	6.45 ± 2.87

The standard deviation is also given in this table. The values in bold are the best results we obtained by comparing different methods.

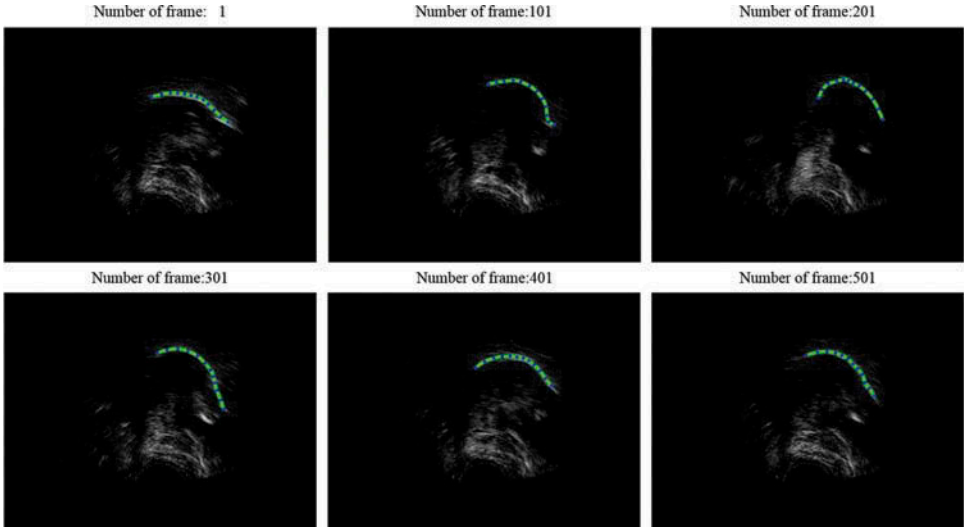

Figure 8. Some examples of results for Female 1.

between the contours extracted automatically by different methods are given in Table 1.

It can be observed that the proposed method, with contour-similarity constraint and automatic re-initialization, has the best performance (smallest mean and smallest standard deviation of MSD error). Some other examples of tracking results for Female 1 are given in Figure 8, showing a variety of tracking qualities. Some results for Male 1, Male 2 and Male 3 are given in Figures 9–11. No manual re-initialization was performed on any dataset during the entire tracking procedure. The performance demonstrates the algorithm versatility on the different subjects. We note that since Male 3 had undergone a laryngectomy, no hyoid bone or shadow of the hyoid bone can be observed in the image sequences of this speaker.

Conclusion and perspectives

In this article, a new automatic contour-tracking algorithm is proposed for ultrasound tongue image sequences. Based on the active contour model, a novel method is presented

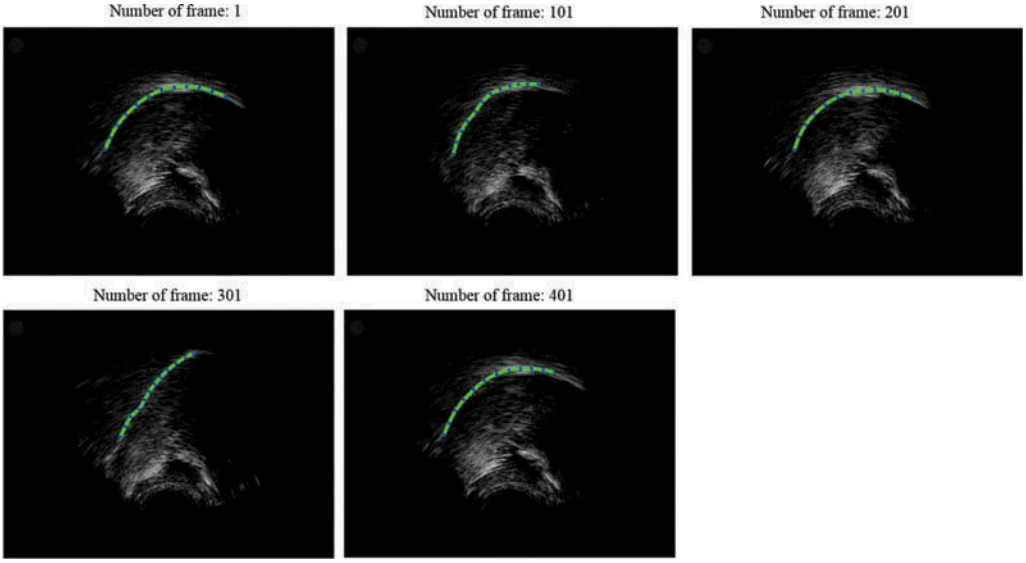

Figure 9. Some examples of results for Male 1.

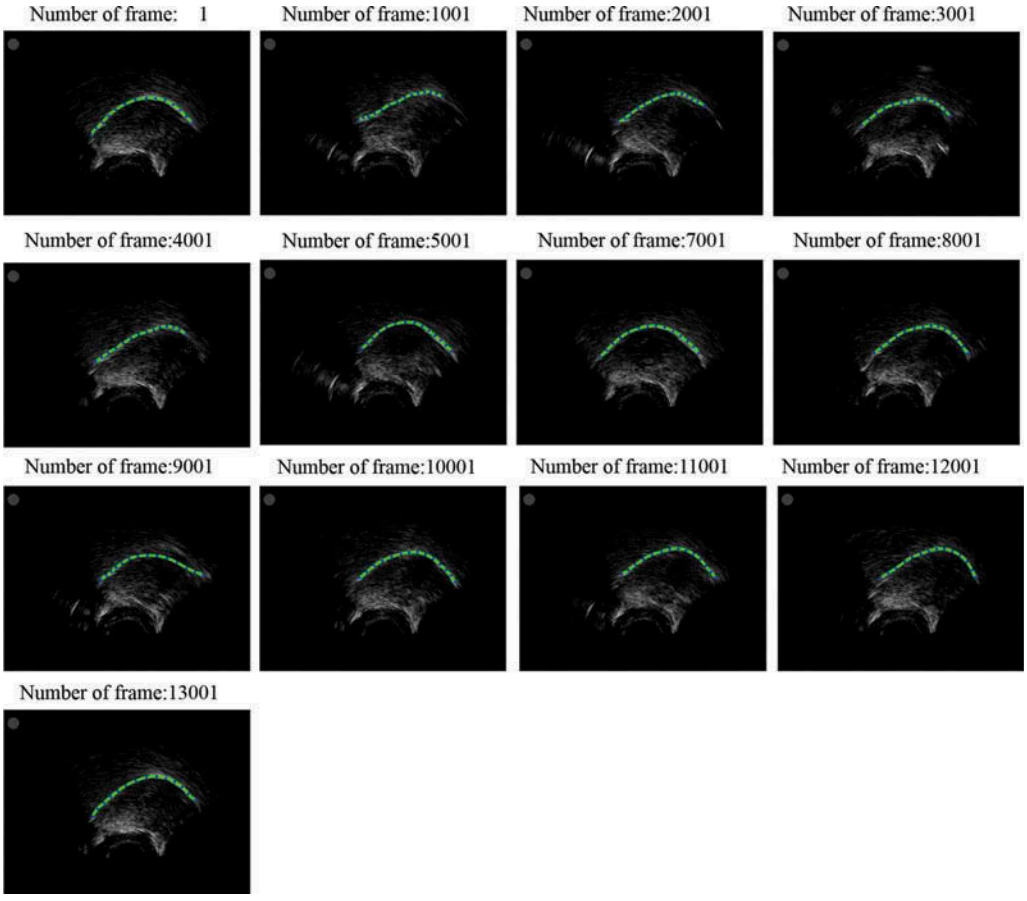

Figure 10. Some examples of results for Male 2.

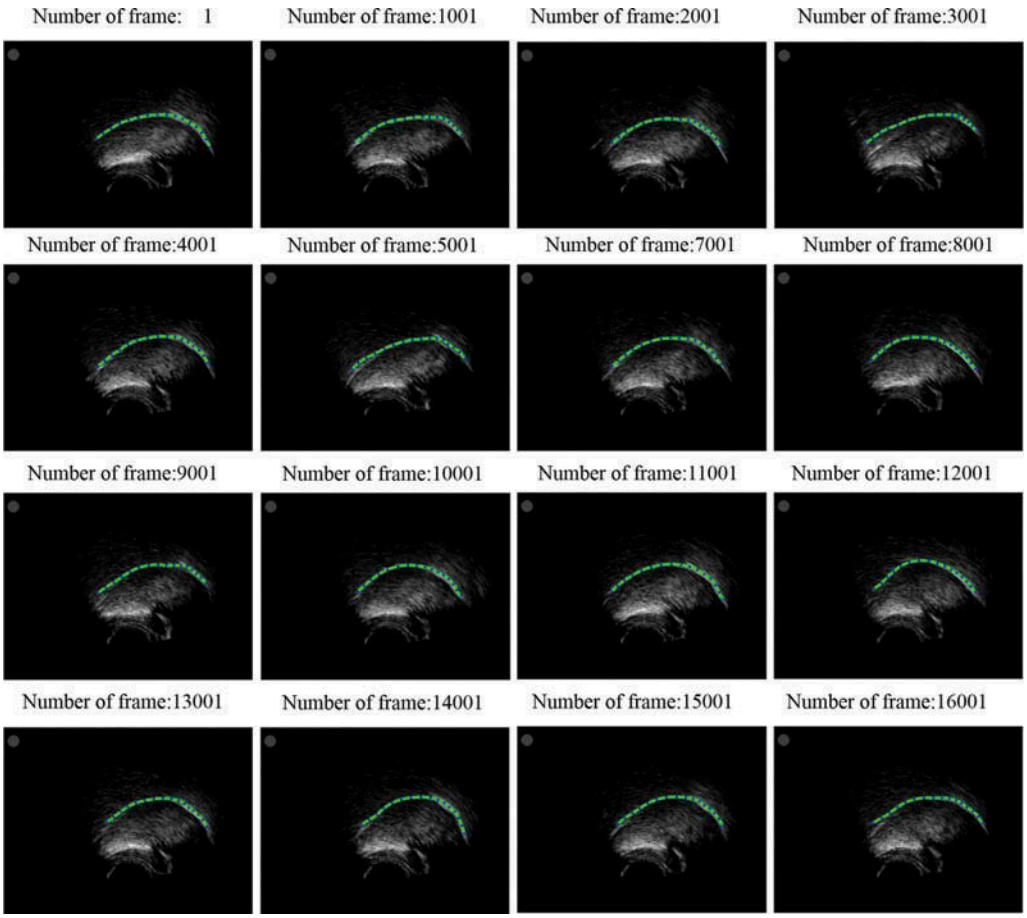

Figure 11. Some examples of results for Male 3.

in this article to address the problem of missing or faint contours with a contour-similarity constraint. Moreover, an image similarity-based automatic re-initialization technique is also proposed in this article, which can increase the robustness of the tracking and can be useful for other tongue contour-tracking systems. The algorithm requires no manual re-initialization, and the performance is encouraging. It serves as a complementary approach to hand-scanning and existing semi-automatic scanners (e.g., EdgeTrak, AAA) and can be an important tool for applications where analysis of long sequences is important, such as speech production, speech recognition, swallowing research (Sonies et al., 2003) and the like.

Funding

This work was partially funded by the European FP7 i-Treasures project (FP7-ICT-2011-9-600676-i-Treasures). The authors also thank the China Scholarship Council (CSC).

Declaration of interest

The authors report no conflicts of interest.

References

Akgul, Y. S., Kambhamettu, C., & Stone, M. (1999). Automatic extraction and tracking of the tongue contours. *IEEE Transactions on Medical Imaging*, *18*(10), 1035–1045.

Akgul, Y. S., & Kambhamettu, C. (1999). A new multi-level framework for deformable contour optimization. IEEE international conference on computer vision and pattern recognition, June 1999, Collins, Colorado, USA.

Al Kork, S.K., Jaumard-Hakoun, A., Adda-Decker, M., Amelot, A., CrevierBuchman, L., Chawah, P., Dreyfus, G., Fux, T., Pillot, C., Roussel, P., Stone, M., Xu, K., & Denby, B. (2014). A multi-sensor helmet to capture rare singing, an intangible cultural heritage study. International seminar on speech production. May 2014, Cologne, Germany.

Articulate Instruments Ltd. (2012). *Articulate assistant advanced user guide: Version 2.14*. Edinburgh, UK: Articulate Instruments Ltd.

Cai, J., Hueber, T., Manitsaris, S., Roussel, P., Crevier-Buchman, L., Stone, M., Pillot-Loiseau, C., Chollet, G., Dreyfus, G., & Denby, B. (2013). Vocal tract imaging system for post-laryngectomy voice replacement. IEEE international conference on international instrumentation and measurement technology conference (I2MTC) (pp. 676–680), May 2013, Minneapolis, Minnesota, USA

Cai, J., Denby, B., Roussel, P., Dreyfus, G., & Crevier-Buchman, L. (2011). Recognition and real time performance of a lightweight ultrasound based silent speech interface employing a language model. Annual conference of the international speech communication association (InterSpeech), August 2011, Florence, Italy

Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? *Journal of the ACM*, *58*(3), 1–37.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., & Brumberg, J. S. (2010). Silent speech interfaces. *Speech Communication*, *52*(4), 270–287.

Mehul. (2013). Complex-wavelet structural similarity index. Retrieved 15 October, 2014, from http://www.mathworks.com/matlabcentral/fileexchange/43017-complex-wavelet-structural-similarity-index-cw-ssim/content/cwssim_index.m

Fasel, I., & Berry, J. (2010). Deep belief networks for real-time extraction of tongue contours from ultrasound during speech. IEEE international conference on pattern recognition (ICPR) (pp. 1493–1496), August 2010, Istanbul, Turkey.

Jaumard-Hakoun, A., Xu, K., Dreyfus, G., Roussel, P., Stone, M., & Denby, B. (2015). Tongue contour extraction from ultrasound images based on deep neural network. Proceedings of the 18th international congress of phonetic sciences (ICPhS), August 2015, Glasgow, Scotland.

Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: active contour models. *International Journal of Computer Vision*, *1*(4), 321–331.

Li, M., Kambhamettu, C., & Stone, M. (2005). Automatic contour tracking in ultrasound images. *Clinical Linguistics & Phonetics*, *19*(6–7), 545–554.

Nesterov, Y. (2007). Gradient methods for minimizing composite objective function. Technical Report-CORE. Universite Catholique de Louvain, Louvain, Belgium.

Roussos, A., Katsamanis, A., & Maragos, P. (2009). Tongue tracking in ultrasound images with active appearance models. IEEE international conference on image processing (pp. 1733–1736), November 2009, Cairo, Egypt.

Sampat, M. P., Wang, Z., Gupta, S., Bovik, A. C., & Markey, M. K. (2009). Complex wavelet structural similarity: a new image similarity index. *IEEE Transactions on Image Processing*, *18*(11), 2385–2401.

Simoncelli, E. P. (2008). MatlabPyrTools. Retrieved 15 October, 2014, from http://www.cns.nyu.edu/~lcv/software.php

Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: a flexible architecture for multi-scale derivative computation. IEEE international conference on image processing (ICIP), October 1995, Washington, DC, USA.

Sonies, B. C., Chi-Fishman, G., & Miller, J. L. (2003). Ultrasound imaging and swallowing. In *Normal and Abnormal Swallowing* (pp. 119–138). New York: Springer.

Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics & Phonetics*, *19*(6–7), 455–501.

Tang, L., Bressmann, T., & Hamarneh, G. (2012). Tongue contour tracking in dynamic ultrasound via higher-order MRFs and efficient fusion moves. *Medical Image Analysis*, *16*(8), 1503–1520.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wrench, A. A. (2015). Private communication.

Wrench, A. A., & Balch, P. (2015). Towards a 3D tongue model for parameterising ultrasound data. Proceedings of the 18th international congress of phonetic sciences (ICPhS), August 2015, Glasgow, Scotland.

Xu, K., Yang, Y., Jaumard-Hakoun, A., Adda-Decker, M., Amelot, A., Crevier-Buchman, L., Chawah, P., Dreyfus, G, Fux, T., Pillot-Loiseau, C., Al Kork, S., K., Stone, M., & Denby, B. (2014). 3D tongue motion visualization based on ultrasound image sequences. Annual conference of the international speech communication association (InterSpeech) September 2014, Singapore (pp. 1482–1483).

Xu, K., Yang, Y., Jaumard-Hakoun, A., Leboullenger, C., Dreyfus, G., Roussel, P., Stone, M., & Denby, B. (2015) Development of a 3D tongue motion visualization platform based on ultrasound image sequences. Proceedings of the 18th international congress of phonetic sciences (ICPhS), August 2015, Glasgow, Scotland.

Zhou, X., Huang, X., Duncan, J. S., & Yu, W. (2013). Active contours with group similarity. IEEE conference on computer vision and pattern recognition (CVPR), June 2013, Portland, Oregon, USA (pp. 2969–2976).