Testing AutoTrace: A machine-learning approach to automated tongue contour data extraction

Jae-Hyun Sung [1]
Jeff Berry [2]
Marissa Cooper [1]
Gustave Hahn-Powell [1]
Diana Archangeli [1,3]

[1] *Department of Linguistics, University of Arizona, USA*
[2] *InsideSales.com*
[3] *Linguistics, University of Hong Kong, HK*

November 2013

So much data, so little time

Manual tracing of tongue contours is impractical

- ...ultrasound images are captured at 30-100+ fps[1]
- ...but an expert takes two seconds or more to trace one frame![2]
- for five minutes of speech recorded at 30fps, it would take an expert **50+ hours** to trace all 9,000 frames!

Can an automated system match the performance of a human expert?

[1] 30 fps for Sonosite Titan
[2] (Berry, 2012:28)

Ongoing Problem

Use of tongue contours

- Stone (2005), Iskarous (2005), Gick et al. (2006), Davidson (2006), Archangeli et al. (2010), Berry et al. (2011)

Automatic methods of contour extraction

- Hueber et al. (2007):
  - used entire image (c.f. RoI approaches)
- Li et al. (2005):
  - EdgeTrak and semi-automatic methods

Overview

AutoTrace

- use a modified Deep Belief Network...
- train network on diverse images...

Deep Belief Networks

Deep Belief Networks (DBN)[3]:

- generative probabilistic graphical models
- State-of-the-art ANN
    - robust, cross-domain success
- similarities to visual cortex[4]

Translational Deep Belief Networks (tDBN)[5]:

- make use of labels during both parts of training

[3] Hinton and Salakhutdinov (2006)
[4] Riesenhuber and Poggio (1999): Serre et al. (2005)
[5] Berry et al. (2012)

Deep Belief Networks (cont'd)

Advantages:

- multilayered design useful for "compositional" vision problems
- considers multiple hypotheses
- Larger set of features akin to relations over relations

Challenges:

- long training time
- weak to underrepresented data
  - solution: use high entropy cases for greater representivity

Method

- specify region of interest (RoI)
- In training, network is fed ultrasound image and corresponding trace

Figure: Input to Output (Berry, 2012:45)

Jump to Errors

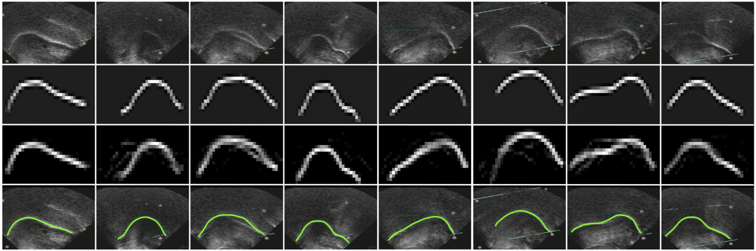


Figure: *

Top: Ultrasound frame (input)

Second: Expert trace (input)

Third: *t*DBN trace (output)

Bottom: *t*DBN trace transformed into *n* coordinates (output)

Then and Now

Training Data:[6]

- Italian
- 74 distinct words
- 3,209 images
- recorded for EMA study
- 25 fps
- one speaker
- not representative

Training Data:

- English
- 1906 distinct words
- 33,000+ images
- recorded for ultrasound
- 30 fps
- eleven speakers
- representative[7]

[6] Tavella (2007)
[7] used Harvard Sentences

Quantifying performance

Evaluation

- use Mean Sum of Distances (MSD) metric to measure performance[8]
  - the *lower* the MSD, the better performance

Figure: Mean Sum of Distances

$$\mathrm{MSD}(U,V) = \frac{1}{2n}\left(\sum_{i=1}^{n} \min_j |v_i - u_j| + \sum_{i=1}^{n} \min_j |u_i - v_j|\right)$$

Figure: *

*Where U and V are vectors representing pixel values at points along the contour of the tongue

[8] Li et al. (2005)

Selecting images for training

"**Informed Undersampling**" [9]

- **Problem**: there are too many image...
- **Solution**: sample from the "**most diverse**" images

Berry (2012)

- training set of most diverse images outperforms a random sampling of the same size
  - MSD from high entropy $<$ MSD from random sampling
    - * ($p < 0.0001$)

[9] Liu et al. (2009)

Test 1: Most-Diverse Training

Training Design

- Selected from a pool of 33,000+ images
- 50 images for testing
- Ranked images
  1. determine RoI
     - calculate an average image (in terms of pixel values in the RoI)
     - calculate average pixel intensity for each pixel in RoI
  2. compare each image to average image and sort results

Distribution of Image Diversity

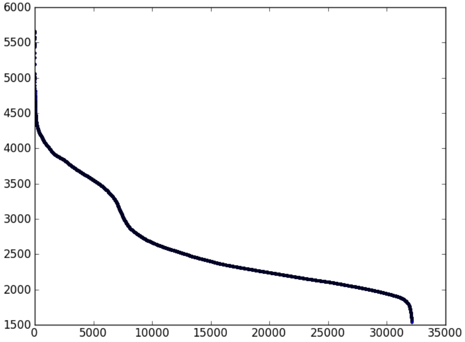


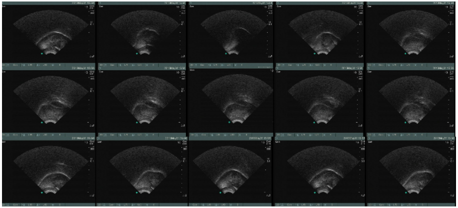
Figure: Plot of distance scores ordered by rank[10]

Jump to Test 4

[10] score = sum of the absolute value of the differences in pixel intensity for each from the average image

Calculating an "average" image...

Figure: Examples of ultrasound frames



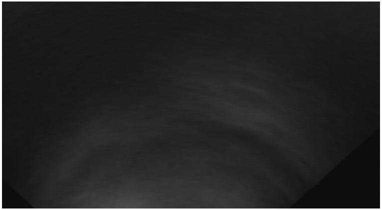


Figure: RoI-constrained averaging of pixel values

Test 1: Most-Diverse Training (cont'd)

Table: Most-Diverse Training Results

Most Diverse	Least Diverse	AutoTrace MSD
200	50	7.52
300	50	7.316
350	0	7.345
400	50	5.940
450	0	6.438
500	50	5.217
550	0	6.300
600	50	5.477
700	50	5.308
800	50	4.778

Table: *

Results: *sampling from both most and least is important; performance positively correlated with size of most diverse set*

Table: *

NB: *Average pixel difference between expert tracers*: 4.077

Test 2: Filtered Most-Diverse Training

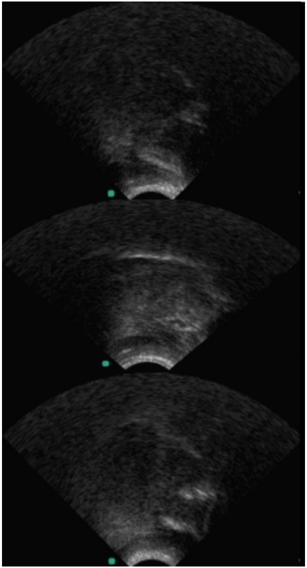
Training Design

- filtering step to remove "bad tongues" from Test 1 model
- 32,182 images remaining after filtering (c.f. 40,600 originally)
- Tested on 100 images

"Bad Tongues"

Figure: *

Removed from training in Test 2



Possible explanations for poor imaging:

- subject was not properly hydrated
- incomplete contact with probe

Test 2: Filtered Most-Diverse Training (cont'd

Table: Filtered Most-Diverse Training Results

Most Diverse	Least Diverse	AutoTrace MSD
200	50	14.962
300	50	14.2965
350	0	13.395
400	50	13.121
450	0	14.023
500	50	13.322
550	0	14.306
600	50	12.961
700	50	12.738
800	50	11.6345
850	50	12.5775
931	50	11.8035

Table: *

Results: *removal of "bad tongues" in training has a negative impact on performance*

Test 3: Retraced Training

Training Design

- errors from the second test were retraced and added to the last training set
- Test set: 100 images

Table: Retraced Training Results

Most Diverse	Least Diverse	AutoTrace MSD
931+75 corrected	50	5.655

Table: *

Results: *retraining gives us a clear performance boost; similar experiment needs to be run on results of Test 1*

Types of Errors I

Overextending traces[11]

Figure: *

tDBN Overextended Trace

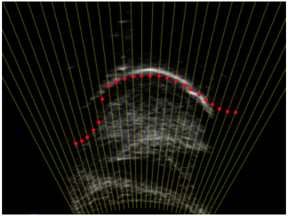
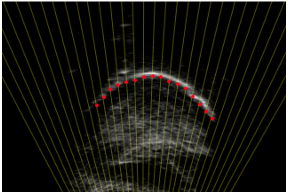


Figure: *

Manually Corrected Trace



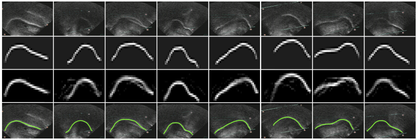
Such corrections can be performed quickly, although a better solution might be to limit the span of a given trace by surrounding pixel intensity or distance from the center.

Types of Errors (cont'd)

Dual traces [12]

Figure: *

tDBN Input and Output



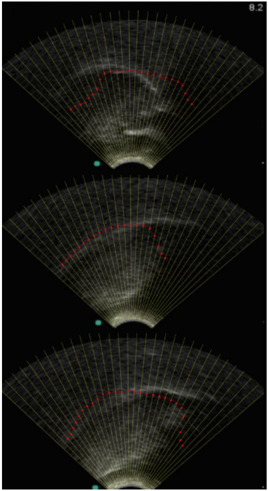
▸ Enlarged image

See image 3,3 for an example of dual tracing caused by an inconsistency in the refresh and frame rates (i.e. a shortcoming of the equipment)

[12](Berry, 2012:45-46)

Types of Errors (cont'd)
Misaligned Traces

Figure: *
*t*DBN Misaligned traces



AutoTrace fails to find a complete contour. Possible explanations:

- a lack of relevant exemplars in training
- tongue outside of RoI

Questions

What we've learned...

- Informed undersampling works
- Previous work scales
- Human-level performance/reliability attainable

What's left...

- Error Analysis
- Preprocessing
  - normalize images for brightness? Filter?
- Measuring diversity
  - alternative entropy metrics?

Many thanks to...

National Science Foundation

- Grants 1059266 and 1244687

James D. McDonnell Foundation

- (grant awarded to Diana Archangeli)

References I

Archangeli, D., Berry, J., Ji, S., Josephs, K., Hunt, N., Fisher, M., and Carnie, A. 2010. Atr in scottish gaelic tense sonorants: A preliminary report.

Berry, J. 2012. *Machine learning methods for articulatory data*. PhD thesis, The University of Arizona.

Berry, J., Fasel, I., Fadiga, L., and Archangeli, D. 2012. Training deep nets with imbalanced and unlabeled data. In *Interspeech*.

Berry, J., Ji, S., Fasel, I., and Archangeli, D. 2011. Articulatory reduction in mandarin chinese words. In *INTERSPEECH*, pages 2809–2812.

Davidson, L. 2006. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *Journal of Acoustical Society of America*, 120:407–415.

Gick, B., Campbell, F., Oh, S., and Tamburri-Watt, L. 2006. Toward universals in the gestural organization of syllables: A cross-linguistic study of liquids. *Journal of Phonetics*, 34(1):49–72.

Hinton, G. E. and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

References II

Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., and Stone, M. 2007. Eigentongue feature extraction for an ultrasound-based silent speech interface. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 1, pages I–1245–I–1248.

Iskarous, K. 2005. Detecting the edge of the tongue: A tutorial. *Clinical linguistics & phonetics*, 19(6-7):555–565.

Li, M., Kambhamettu, C., and Stone, M. 2005. Automatic contours tracking in ultrasound images. *Clinical Linguistics and Phonetics*, 19:545–554.

Liu, X.-Y., Wu, J., and Zhou, Z.-H. 2009. Exploratory undersampling for class-imbalance learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(2):539–550.

Riesenhuber, M. and Poggio, T. 1999. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025.

Serre, T., Wolf, L., and Poggio, T. 2005. Object recognition with features inspired by visual cortex. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 994–1000. IEEE.

References III

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

Tavella, M. 2007. Simultaneous recording of phono-articulatory parameters during speech production. Master's thesis, Universita‘ degli Studi di Genova.