# Testing AutoTrace

## A machine-learning approach to automated tongue contour data extraction

Gustave Hahn-Powell[1], Diana Archangeli[1,2]

[1]Department of Linguistics, University of Arizona, USA
[2]Linguistics, University of Hong Kong, HK

## Introduction

- Ultrasound is efficient, noninvasive, portable, and (relatively) inexpensive
- Used in experimental work (8; 11; 1; 7), in field (9; 10; 3; 2), clinical (4) and pedagogical settings (15; 16), and for ASR (12)

### *Problem*

The time it takes to identify and hand label the tongue's surface in each image is prohibitive.

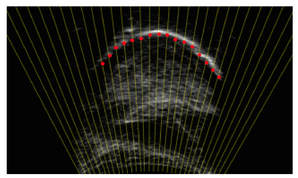

Figure: Ultrasound image of tongue contour

- How can the error from automated systems (AutoTrace) be minimized?
- Retraining networks with the inclusion of hand-corrected images improves performance (20), (6)

### *Questions*

- Can retraining improvements be extended beyond the initial study (6)?
  1. *Is a network trained on multiple speakers effective on data from another set of speakers?*
  2. *Can a network trained on data from one language be adapted to work effectively with data from another language?*

## Background

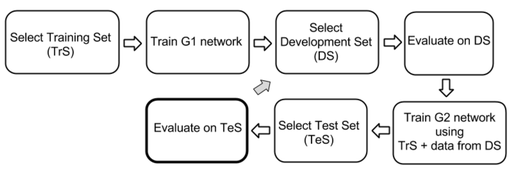


Figure: Network training procedure (see (6) & (18))

- (6) and (18) found:
  - A high-entropy TrS gives better performance than a randomly-selected TrS
  - A second training (G2) (the TrS + DS) resulted in improved performance
  - A third training (TrS + DS + additional DS) failed to improve performance

## Design

- SonoSite TITAN; C-11/7-4 11-mm broadband curved array transducer
- Transducer created a 90° fan-shaped 5 mm-thick mid-sagittal section
- Image at the same depth for all studies
- Modified test design (below)

### *G1 Network*

- TrS uses data from Harvard Sentences (HS) (17)
- Phonetically-balanced; 12 subjects

Table: Data comparison between studies

| | Language | Speakers | Images | FPS | G1 size (best) |
|---|---|---|---|---|---|
| Current | English | 12 | 33,000+ | 30 | 850 |
| Previous | Italian | 1 | 3,200 | 25 | 450 |

- Previous work: G1 size is variable
- Our G1: 850 images from HS
  - 800 with high entropy
  - 50 with lowest entropy

### *Development and Test Sets*

- Four datasets for development & testing
- DSs & TeSs balanced for contour variety (see (14; 6))
- Each DS & TeS consists of 100 unique images, proportionally:
  - Top 80% high-entropy
  - Bottom 20% low-entropy

| Language | Abbrev. | Type | ≈Images | Speakers |
|---|---|---|---|---|
| English (HS) | HS | (see (17)) | 33,000 | 30 |
| English (palatal) | EN | (see (Sung)) | 2,000 | 7 |
| Korean (palatal) | K | (see (Sung)) | 1,200 | 11 |
| Scottish Gaelic | SG | (see (3)) | 10,000 | 4 |

Table: Data pool for DS and TeSs

### *G2 Networks*

- Evaluation metric: Mean Sum of Distances (MSD) (see (13))

Figure: Mean Sum of Distances

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

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

- Two G2 networks for each DS
  - DS errors MSD>10
  - DS errors MSD≥5
- TeS labeled by G1 and G2 for comparison

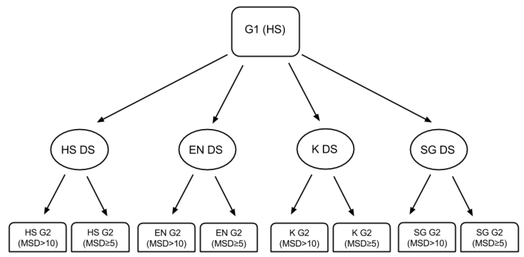


Figure: G2 ancestry

### *Retraining criteria*

- Too much retraining increases error (6)
- Cautious inclusion of retraining data
- Retraining with different MSD thresholds

## Experiments

### *Experiment 1: Additional subjects*

- G1 network retrained with HS & EN DSs
- Two G2 networks for each DS above
- Each G2 network tested on respective TeS (MSD metric)

Table: Experiment 1 Results (in MSD)

| | HS | HS *size* | EN | EN *size* |
|---|---|---|---|---|
| Original | 43.09 | 850 | 14.38 | 850 |
| G2 (≥10) | 13.08 | 896 | 5.03 | 925 |
| G2 (≥5) | 13.57 | 917 | 4.30 | 946 |

### *Experiment 2: Additional languages*

For this study, we again used the original G1 network. Retraining was based on K and SG DSs. In addition to the HS G2 networks, this produced two K G2 networks and two SG G2 networks:

Table: Experiment 2 Results (in MSD)

| | HS | HS *size* | K | K *size* | SG | SG *size* |
|---|---|---|---|---|---|---|
| Original | 43.09 | 850 | 31.30 | 850 | 13.63 | 850 |
| G2 (≥10) | 13.08 | 896 | 5.72 | 948 | 5.40 | 928 |
| G2 (≥5) | 13.57 | 917 | 5.47 | 950 | 4.81 | 943 |

## Discussion

- All G2 networks show improvement over G1
- Results consistent with those reported in (6; 5)
- On average a lower MSD with ≥5 networks than with ≥10 networks
  - except for the HS G2 performance

### *Correction requirements*

- End goal is human-level performance
  - (13) reported an inter-annotator discrepancy of 2.9 MSD
- TeS traces with MSD>2.9 are Candidates for Correction (CC)

Table: TeS CC

| | CC | % <2.9 MSD |
|---|---|---|
| English (HS) | | |
| G1 | 94 | 6 |
| HS G2 (≥10) | 85 | 15 |
| HS G2 (≥5) | 83 | 17 |
| English (palatals) | | |
| G1 | 100 | 0 |
| EN G2 (≥10) | 74 | 26 |
| EN G2 (≥5) | 60 | 40 |

| | CC | % <2.9 MSD |
|---|---|---|
| Korean (palatals) | | |
| G1 | 100 | 0 |
| K G2 (≥10) | 81 | 19 |
| K G2 (≥5) | 75 | 25 |
| Scottish Gaelic | | |
| G1 | 99 | 1 |
| SG G2 (≥10) | 82 | 18 |
| SG G2 (≥5) | 72 | 28 |

- As average MSD decreases, CCs decrease
  - Not the case for the HS G2 ≥5 network
    - slightly *higher* average MSD (13.57>13.08)
    - slightly *fewer* CC (83<85)
- (6; 5) reported a 41% reduction of error between a between a G1 and a G2 network (TrS+100 images)
  - Our greatest reduction of error is similar
    - 40% for EN
    - others are in the range of 27-25%

### *Cross-linguistic performance*

- Greatest performance gain with retraining with EN data (≥5 EN G2 network)
  - language bias from TrS (English and EN (English)?
  - retraining focused network on palatals (focus of study by (Sung))
- Improvements on SG and K data similar
  - little difference between ≥10 MSD and the ≥5 MSD conditions
  - Gains not as great as EN

## Conclusions

Our findings confirm that the model developed in (6) is effective when used with multi-speaker and multi-language data.

- With retraining, a tDBN trained on one language can be adapted to data from other speakers of the same language as well as data from other languages
  - Our best G2 network reduced errors by 40% over the G1 network

**168th Meeting of the Acoustical Society of America, 27 October 2014, Indianapolis, USA**