# Where do the improvements come from in sequence-to-sequence neural TTS?

*Oliver Watts*
*Gustav Eje Henter*
*Jason Fong*
*Cassia Valentini-Botinhao*

## Motivation

- Attention-based sequence-to-sequence (seq2seq) systems lead to improved quality over statistical parametric speech synthesis (SPSS)
- Which elements of the new paradigm contribute most to these gains?
- Propose a functional mapping between seq2seq *subnets* and SPSS *modules* to step gradually from SPSS → S2S
- In addition to many subsidiary questions:

What is the impact of learning the front end (**T**ext encoder)?
What is the impact of conditioning on acoustic **H**istory?
What is the impact of jointly-learned alignments (**A**ttention)?

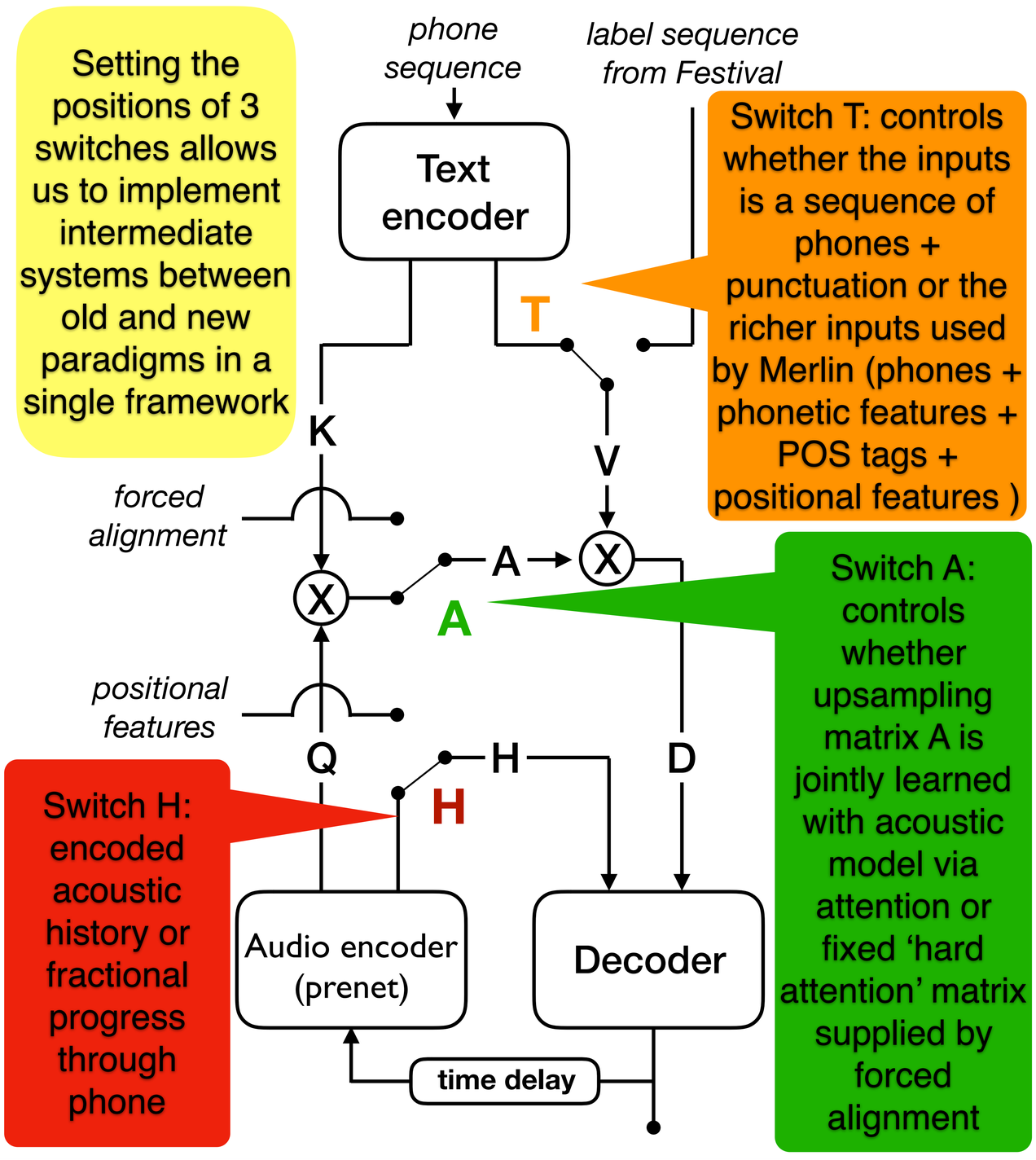
## Systems built

| System | Codebase | Frame hop | Dynamic feats. | Sig. gen. | Acoust. loss | Front-end | Feedback | Alignment | SSRN |
|---|---|---|---|---|---|---|---|---|---|
| M | Merlin | 5 ms | $\Delta + \Delta^2$ | WORLD | L2 | Fixed | As in [5] | Fixed | N/A |
| MM | " | 12.5 ms | " | " | " | " | " | " | " |
| W2 | DCTTS | 50 ms | None | " | " | " | Rel. pos. in phone | " | "W2" |
| W2T | " | " | " | " | " | Learned | " | " | " |
| W2H | " | " | " | " | " | Fixed | Acoustic | " | " |
| G2 | " | " | " | G-L | " | " | Rel. pos. in phone | " | "G2" |
| G1 | " | " | " | " | L1 + BCE | " | " | " | "G1" |
| G1H | " | " | " | " | " | " | Acoustic | " | " |
| G1TH | " | " | " | " | " | Learned | " | " | " |
| G1HA | " | " | " | " | " | Fixed | " | Learned | " |
| G1THA | " | " | " | " | " | Learned | " | " | " |

| | |
|---|---|
| M | Merlin codebase |
| W | WORLD |
| G | Griffin-Lim |
| 2 | L2 loss |
| 1 | L1 + binary cross |
| **A** | **A**ttention |
| **T** | Learned **T**ext analysis |
| **H** | Acoustic **H**istory feedback |

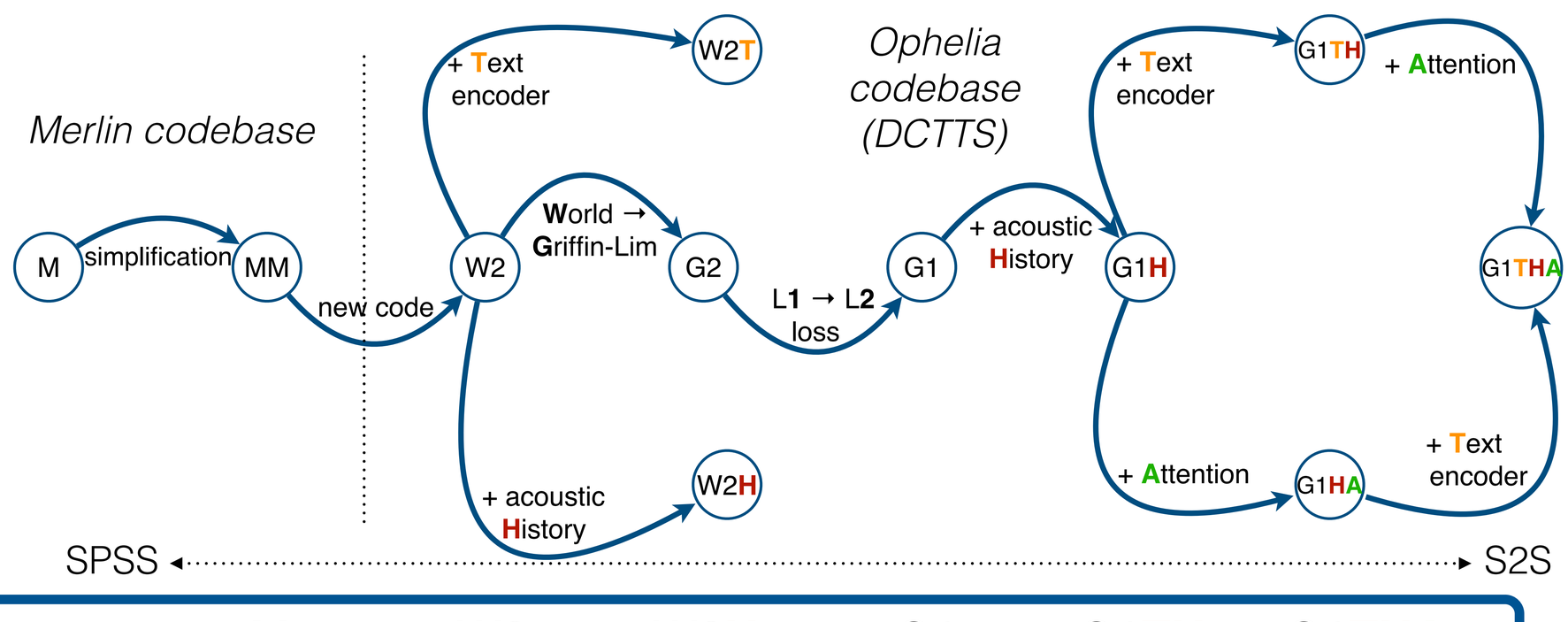
The systems built represent one way of flipping the switches **T**, **A** & **H** to step from Merlin to DCTTS. Several partial paths are added to increase the number of useful comparisons we can make.

## SPSS → S2S-TTS




*The systems built allow us to pose these questions:*

| Type | ID | Question: ("What is the impact of…") | | Relevant system contrasts |
|---|---|---|---|---|
| Primary | Q1 | …learning the front end? | **T** | W2 vs. W2T<br>G1H vs. G1TH<br>G1HA vs. G1THA |
| | Q2 | …acoustic feedback replacing positional feedback? | **H** | W2 vs. W2H<br>G1 vs. G1H |
| | Q3 | …jointly-learned alignments replacing fixed alignments? | **A** | G1H vs. G1HA<br>G1TH vs. G1THA |
| Secondary | Q4 | …Merlin simplifications to ease stepping toward DCTTS? | | M vs. MM |
| | Q5 | …using the DCTTS architecture and codebase? | | MM vs. W2 |
| | Q6 | …DCTTS waveform generation replacing World? | | W2 vs. G2 |
| | Q7 | …the DCTTS loss function replacing L2 loss? | | G2 vs. G1 |
| Interaction | Q8 | Does acoustic feedback interact with the acoustic feature type? | | W2→W2H vs. G1→G1H |
| | Q9 | Does front-end learning interact with learning to align? | | G1H→G1HA vs. G1TH→G1THA |




## Evaluation

- Training data: LJSpeech - 24 hours of read speech
- MUSHRA-like test; G1THA as reference, no anchor
- 24 paid native English listeners
- Each listener rated two sets of 10 synthesised Harvard sentences, every set phonetically balanced

## Main findings

**T** Learning the front-end always improves quality (with or without attention, but more so with attention)

**H** Acoustic feedback has a strong beneficial impact

**A** Attention (compared to fixed alignment)

- 'breaks' without a learned front-end
- helps a bit (but not significantly) with a learned fror