HTK: Train-and-Align

You can use 'HTK' to train new speaker-specific acoustic models on your speech data, and then to force align the data on those models. You may decide to do this if:

- You can't share your data with third parties and so can't use WebMAUS.
- Your data isn't US English (or similar) and so you can't use HTK with the P2FA pre-trained models.
- You have at least 5 minutes' speech for each speaker.

The general process is illustrated below:

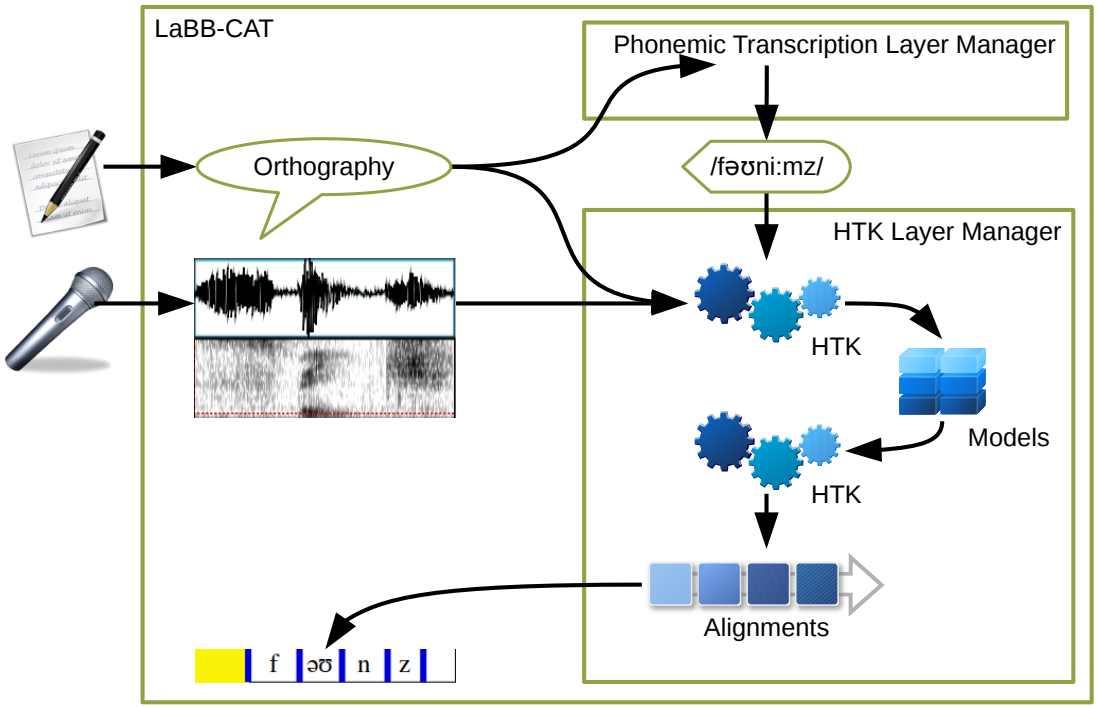

Figure 1: Pronunciations are generated from transcripts, and then combined with the recordings to train acoustic models, which are then used to compute phone-level alignements, which are saved to LaBB-CAT's database

Prerequisites

In order to be able to force-align transcripts to the word and/or segment level, you first need the following:

1. Transcripts that are aligned at the utterance level (i.e. there's a known time-point every 20 or so words), whch have been uploaded into LaBB-CAT
2. A WAV file for each transcript, on the LaBB-CAT server
3. A phonemic transcription word layer, that has at least one pronunciation for every word. If there are some lines/utterances that contain words with missing pronunciations, those lines will be ignored by the HTK Layer Manager.

Depending on your speech data, there are several ways to obtain phonemic transcriptions for words:

- Lexical tagging
  - CELEX - for British English, German, Dutch, using one of the CELEX layer managers.
  - CMU Pronouncing Dictionary - for US English, using th CMU Pronouncing Dictionary layer manager.
  - Unisyn - for various English varieties, using the Unisyn layer manager.
  - Define your own lexicon, and use the Flat File Dictionary layer manager to integrate it into LaBB-CAT.
- Inferring pronunciation from orthography
  - Spanish, using the Spanish Phonological Transcriber layer manager
  - Bas Web Service: G2P - for various languages.
  - Define your own simple mapping rules from orthography to phonology, using the Character Mapper layer manager.

If the speech corpus includes data in more than one language, it is possible to ensure that the utterances are phonemically tagged in a way that's sensitive to the language of the specific utterance, using the *language* layers and attributes, and auxiliary layer managers.

Whichever method you choose, you need a phonemes 'word layer' on which each word token is tagged with its pronunciation, before you can proceed with the forced-alignment steps below.

Procedure for HTK Forced Alignment

The broad steps for getting forced-alignments from HTK are:

1. Install HTK on the same computer that LaBB-CAT is installed on
2. Install the HTK Layer Manager, which integrates LaBB-CAT with HTK
3. Create and configure a new HTK layer in LaBB-CAT

4. Pick a speaker/participant in your database
5. Fill in the missing pronunciations for that participant
6. Run forced alignment
7. Repeat steps 4-6 for all the participants in your database

Install HTK

As mentioned above, HTK is a 3rd-party tool that you must download and install from the Cambridge University website.

1. Register at http://htk.eng.cam.ac.uk/register.shtml
2. Download the version of HTK that is appropriate for the computer that LaBB-CAT is install on:
   For Windows systems, there are pre-compiled .exe files that you can download. For Unix-like systems, you need to download the source code, which you will then install following the provided instructions (you may also need to install the xorg-dev package before it will successfully compile).
3. Unzip (for Windows) or compile and install (for Unix-like systems) the downloaded files on the computer that LaBB-CAT is installed on.

Install the HTK Layer Manager

The HTK Layer Manager is a LaBB-CAT module that integrates LaBB-CAT with HTK.

1. In LaBB-CAT, select the *layer managers* option on the menu, which gives you a list of the layer managers already installed.
2. At the bottom of the page, follow the *List of layer managers that are not yet installed* link.
3. Look for *HTK* in the list, and press it's *Install* button.
4. You will see a form with boxes for filling in information.
   - *HTK Path* must be set to the location where the HTK files are installed on your system. If this is already filled in, it's probably correct. If it's blank, you have to enter the full path for the HTK programs:
     - On Windows systems, this is where you unzipped the HTK .exe files - e.g. something like `C:\Downloads\HTK`
     - On Unix-like systems, this is probably `/usr/local/bin`, but you can verify this by entering which HCopy at a command shell prompt.
   - *HTK Working Folder* will already have a default value, which is probably best left as-is
5. Press *Install Layer Manager*

Create the HTK layer

Once you've installed HTK and the HTK Layer Manager, you need to create a new layer for triggering and controlling forced alignment. This layer will itself contain a timestamp for each line/utterance it has force-aligned (and so it's a 'phrase' layer), but during that process, the word and phone alignments will also be set on other layers.

1. In LaBB-CAT, select the *phrase layers* option
2. At the top of the page, there's a blank form for creating a new layer; fill in the following details:
    - *Layer ID*: HTK
    - *Type*: *Text*
    - *Manager*: *HTK Manager*
    - *Alignment*: *Time Intervals*
    - *Generate*: *Always*
3. Press *New*.
   You will see the layer configuration page.
4. Check the online help if you want information about all the options, however, most likely the default options are approriate, except:
    - *Pronunciaton Layer*: this is the layer that provides the phonemic transcriptions for all the words; ensure you select the phonemes layer you created above.
      NB If you have created this layer but it doesn't appear here as an option, it's probably because the 'layer type' of your pronunciation layer is not set to 'Phonological', which will need to be changed in order for it to appear as an option here.
      NB In the list of options there's also a layer called "pronounce"; this is a system layer for manually-added pronunciations, and you would only select that layer here if you have manually annotated pronunciations against *every single word* in your transcripts. You probably haven't done that, so you don't want to select the "pronounce" layer here.
    - *Use P2FA models*: ensure this option is **un-ticked**.
5. Press *Save*.
6. If you are confident all your transcripts include all pronunciations for all words, you can press *Regenerate* to force-align your whole corpus now. However, most likely you'll need to proceed per-speaker, described below, in order to fill in missing pronunciations.

Per-speaker Alignment

To start a forced-alignment process per-speaker, you need to first select a speaker who will be aligned. Then you will fill in any missing pronunciations. After that, HTK will automatically force-align their utterances.

(1) In LaBB-CAT, select the *participants* optoin on the menu

(2) Tick a speaker, and press the *All Utterances* button
(3) Click *List*
(4) Once the paginated list of utterances appears, press the *HTK* button below.
Basically you need to fill in the boxes with the pronunciations and click *Save Pronunciations*.

> **Note**
>
> - You don't have to fill them all in at once, you can do a few, and click *Save*, which will save your work and list what's left.
> - You don't have to fill them all in, you can leave some empty and continue with the HTK forced-alignment by clicking *Start* (HTK will ignore any lines where the remaining unknown words appear, but the ones you filled in will be included).
> - Some of the boxes will be initially filled in with a suggestion from the lexicon layer manager - these may or may not be correct, and aren't saved until you save them.
> - The pronunciations have to be in the 'DISC' format - i.e. one character per phoneme, with no spaces. There's a 'helper' link on the right of each pronunciation box - if you click it, it expands into a list of clickable phonemes - just the ones that aren't ordinary letters, and diphthongs etc.
> - The *search* button lets you look up the lexicon for similar words - this probably won't help for place names, but for words like "tarseal", you can click the *lookup* button, enter "tar seal" in the box as two separate words, and you'll get back the DISC pronunciation of each word, with clickable buttons to copy the given pronunciation into the box. This is useful for digits and numbers too, which may not be in the lexicon - so for "1", search for "one" and copy the pronunciation.
> - If you click on the word itself, the transcript for the first instance of that word is opened, in case you want to listen to it, or in case it's actually just a typo and you want to correct the transcript.
> - If you're using CELEX, when you specify the pronunciations, it's recommended to put syllable separators (-) and primary stress markers (') too - e.g. for "tarseal" you can put *t#sil* but it would actually be better to put *t#-'sil*. These markers are entered into the dictionary even though they're stripped out for HTK, and they may come in handy later (e.g. the syllable separators are used by the CELEX layer manager to count syllables).

When you add pronunciations this way, they're added to the dictionary and all the instances of those words in LaBB-CAT are updated with the pronunciations - not just the participant you're looking at, but all participants in the database. So you only have to come up with a pronunciation for each word once.

(5) Once you've filled in all the missing pronunciations, forced alignment will start automatically. If you want to start forced alignment before you've entered all pronunciations, click the *Start* button at the bottom of the page.

You should see a progress bar while the forced alignment is running. It will take a few minutes to complete.

Once HTK has produced the word and segment alignments, it:

- sets the start/end times of the words on the transcript layer accordingly,
- adds new phone annotations to the *segments* layer with the alignments of the phones, and
- saves a timestamp in the *htk* layer.

When the layer manager has finished, you'll see a message saying "Complete - words and phones from selected utterances are now aligned."