EasyAlign: an automatic phonetic alignment tool under Praat

Jean-Philippe Goldman

Department of Linguistics, University of Geneva, Switzerland
jeanphilippegoldman@gmail.com

Abstract

We provide a user-friendly automatic phonetic alignment tool for continuous speech, named EasyAlign. It is developed as a plug-in of Praat, the popular speech analysis software, and it is freely available. Its main advantage is that one can easily align speech from an orthographic transcription. It requires a few minor manual steps and the result is a multi-level annotation within a TextGrid composed of phonetic, syllabic, lexical and utterance tiers. Evaluation showed that the performances of this HTK-based aligner compare to human alignment and to other existing alignment tools. It was originally fully available for French, English. Community's interests for its extension to other languages helped to develop a straight-forward methodology to add languages. While Spanish and Taiwan Min were recently added, other languages are under development.

Index Terms: Praat, HTK, phonetic alignment, phonetic segmentation

1. Introduction

Phonetic alignment (or phonetic segmentation) determines the time position of phone, syllable, and/or word boundaries in a speech corpus of any duration on the basis of the audio recording and its orthographic transcription.

Aligned corpora are widely used in various speech applications including automatic speech recognition, speech synthesis, as well as prosodic and phonetic research. Unlike corpus-based text-to-speech systems which require a high level of alignment precision, studies may require less precision. Because of this, automated transcriptions can greatly enhance preparation of data for research purposes. Though segmentation can be completed manually or automatically, an accurate fully manual approach may require as many as 800 times real-time; 13 hours for a one-minute recording [1]. The processing time is a major drawback for manual labelling, especially when faced with very large spontaneous speech corpora. Thus, an automatic phonetic alignment tool with quick performance is highly desirable. Besides, it is consistent and reproducible. But, although an alignment tool can save time, speech, especially spontaneous speech, has many unpredictable phonetic variations that can decrease the accuracy of the transcription process. Even with precise computational tools and data preparation, automatic systems can make errors that a human would not. Thus, post-processing detection of major segmentation errors is needed to improve accuracy.

In fact, automatic approaches are never fully automatic nor straightforward and instantaneous. It is a matter of compromise among time, aimed precision and computational skills. The question then lies in what is the degree of accuracy needed for (semi-)automatic segmentation.

To build an automatic tool, both computational skills and data preparation are required before the automatic tool can do its job.

Various computational methods have been developed for phonetic alignment. Some have been borrowed from the automatic speech recognition (ASR) domain. However, the alignment process is much easier than speech recognition because the alignment tool need not determine what the segments are but only their locations. For this reason, HMM (Hidden Markov Models)-based ASR systems are widely used in a forced-alignment model for phonetic segmentation purposes.

Another approach combines a text-to-speech system (TTS) and a Dynamic Time-Wrapping (DTW) algorithm. In this case, synthetic speech is generated from the orthographic or phonetic transcription and compared to the corpus as in [2]. The DTW will find the best temporal mapping between the two utterances using acoustic feature representation.

In [3], the two techniques are compared and it turns out that the second system is often more accurate than HMM but may encounter some errors that account for its lower overall evaluation. A hybrid system based on these two techniques in cascade (first HMM then TTS+DTW) is presented in [4], where results improved. These results were compared to two additional techniques, i.e., artificial neural networks and Classification and Regression Trees. The hybrid HMM-based aligner had the best results by far. In [5], some contour detection techniques borrowed from image processing also give interesting results. All of these existing systems require preliminary training and a command line interface is usually required.

The presented system, named EasyAlign, relies on namely HTK [7], a well-known HMM toolkit. It should be seen as a friendly layer under Praat [6] which facilitates the whole alignment process. This Praat plug-in consists of a group of tools to successively perform utterance segmentation, grapheme-to-phoneme conversion and phonetic segmentation. The whole process starts from a sound file and its orthographic (or phonetic) transcription within a text file or already in Praat's TextGrid format.

EasyAlign has initially been developed for French and English. Then some interests of users helped to develop a full methodology to easily add new languages. Spanish and Taiwan Min could be added with few efforts, while Portuguese and Slovak are under development.

2. EasyAlign

EasyAlign is freely available system, made of Praat scripts but also relies on 2 external components: 1. a grapheme-to-

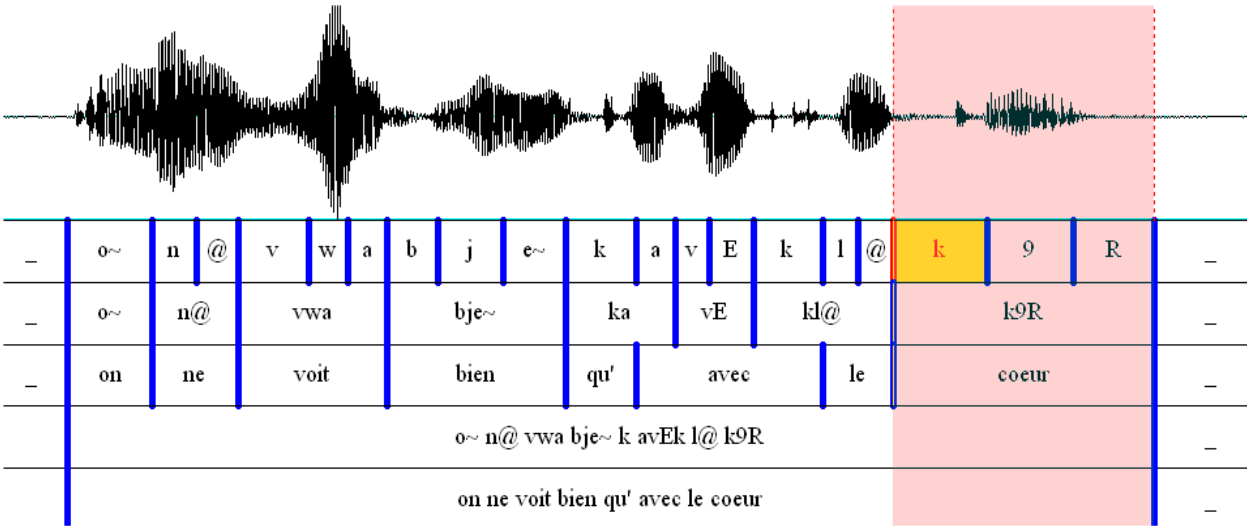

Figure 1*: the full resulting TextGrid with 5 tiers from bottom to top: ortho, phono, words, syllables, phones* for the sentence "On ne voit bien qu'avec le coeur" (one sees well only with the heart)

phoneme conversion system and 2. an acoustic tool for the alignment at the segment level. It is distributed as a self-installable plug-in, with additional tools and the already trained acoustic models of phones.

The whole process to segment a speech file is as follows: from a speech audio file and its corresponding orthographic transcription in a text file, the user has to go through 3 automatic steps; manual verifications and adjustments can be done in-between to ensure even better quality. The result is a multi-tier TextGrid with phones, syllables, words and utterance segmentation as in Figure 1.

More precisely, these three steps are:

1. macro-segmentation at utterance level
2. grapheme-to-phoneme conversion
3. phone segmentation.

Providing a TextGrid already segmented into utterances with an orthographic and/or a phonetic transcription speeds up the process as the first macro-segmentation step is possible and can be skipped. Each step is explained in details below and Figure 3 summarizes the whole process.

2.1. Utterance segmentation

As the data to align can be a long sequence of continuous speech, the automatic phonetic alignment process requires a major preliminary step, i.e. macro-segmentation into utterances or any kind of major speech units. The two main reasons are: 1. recognition tools are not designed to process unlimited-length recordings and 2. it is easier to scroll and make use of a large corpus if such major units (i.e. about utterance-sized) exist. Existing transcription may be various formats:

- as a unique paragraph or as "one sentence per line"
- with or without punctuation

The newline character and/or the punctuation is used to guess utterances in the transcription. The only particular case is if the transcription is in paragraphs and without punctuation. Then the user has to preformat the text file containing the orthographic transcription into a *one-utterance-per-line* format, i.e. by simply adding a newline character between utterances (which may preferably be separated by an empty pause but can also be connected i.e. without pauses).

The first script generates a TextGrid with a single tier called *ortho*. Each interval of this tier contains one utterance transcription and its boundaries are estimated as follows: each utterance-ending boundary position is calculated on the basis of the next punctuation mark or newline character position within the transcription depending on the transcription length and the duration of the audio file. More precisely, a pause detection tool is used to refine the calculation of the speech duration by omitting the silent parts. Then, if a pause lies "near" the first estimation, the boundary is adjusted to the middle of that pause. By *near*, we mean within an adjustable duration set to one second by default.

To evaluate this task, 10 files with various speaking styles (from slow political discourse to animated dialogue) and with a duration from 1 to 6 minutes, with a total of 27 minutes and 567 utterances, were taken. Depending on the corpus style, its recording quality, the existence of pauses between utterances and finally the length and number of utterances, 63% to 96% of the estimated boundaries were correctly positioned. At this step, the user is required to adjust the few misplaced utterance boundaries within the TextGrid. This manual task takes between 1 to 3 times real-time.

2.2. Grapheme-to-phoneme conversion

The purpose of this step is to create the *phono* tier, which is a duplicate of the *ortho* tier (i.e with the same boundaries) but with a phonetic transcription. It is rather unusual for an HMM-based aligner to require the phoneme sequence as an input, since they usually rely a pronunciation dictionary (including variations of pronunciation per word). Thus it should be designed to automatically detect which variant is pronounced. As mentioned before, spontaneous speech shows more variants than basic phonological rules can predict. Many phonemes can be assimilated or elided. So, it is very difficult to add all predictable phonological variations to a pronunciation dictionary for a word, and it is almost endless to add all the possible phonetic pronunciations that can be

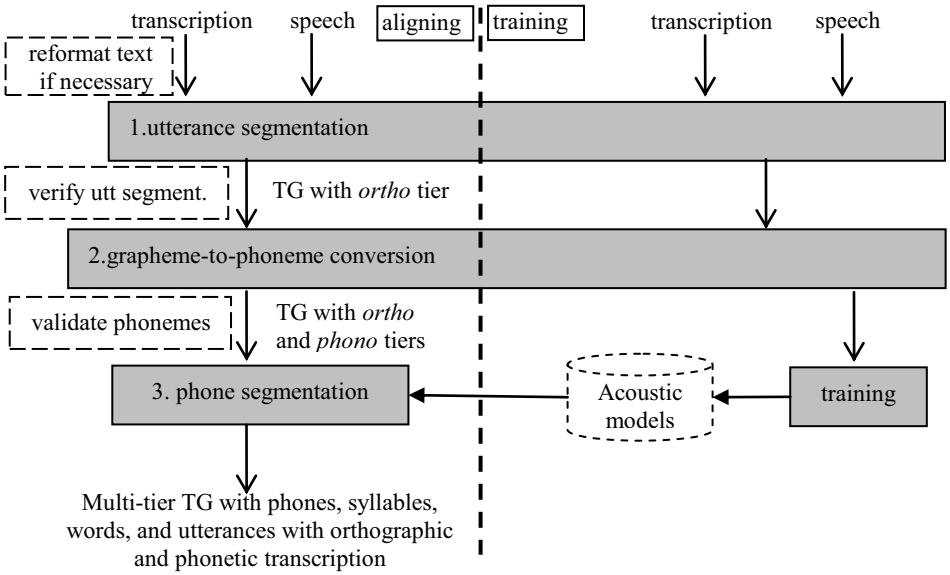

Figure 2: On the left side, the whole process yields a multi-level annotation TextGrid, after 3 automatic steps and manual steps in dashed lines. During the training step, on the right, the same process is followed, excepted that the TextGrid with *ortho* and *phono* tiers are used to train the acoustic models.

found in real corpora. Furthermore, the more pronunciations are added, the more confusion may occur.

In some other systems, human transcribers are allowed to use notation tags in the orthographic transcription to help the following grapheme-phoneme conversion module. But experience has shown that it is rather difficult, even for an expert, to stay focused on detecting audible phonetic variations in an utterance and to transcribe them on a visual orthographic transcription, mainly because the orthographical representation may influence the phonetic perception. Besides, the human transcriber must keep in mind the abilities of the grapheme-to-phoneme conversion engine in order to filter out the predictable variations and annotate only the unpredictable ones.

In our view, speech alignment systems are far from perfect in choosing this correct pronunciation from the available ones in the pronunciation dictionary. Thus EasyAlign proceeds in two steps. A grapheme-to-phoneme conversion provides a phonetic transcription with some major phonological variations. The optional phonemes are marked with a star. Then, the expert annotator can compare the sequence of phonetic symbols with the audible speech of each utterance. The grapheme conversion tool is provided by eLite TTS system [8] and suggests some pronunciation variants.

2.3. Phonetic segmentation

In this final automatic step, the Viterbi-based HVite tool (within HTK) is called to align each utterance to its verified phonetic sequence. For both the French and English languages, this tool was trained on the basis of about 30 minutes of unaligned multi-speaker speech for which a verified phonetic transcription was provided. The acoustic models are monophones with tied states for silence phonemes.

During the alignment, two tiers (*phones* and *words*) are computed. Then two additional calculations are processed: 1) within the *phones* tier, a "PTK-filter" merges a short pause with a following unvoiced plosive (the pause has to be shorter than a settable threshold, 90ms by default), and 2) a syllable tier is generated on the basis of sonority-based rules for syllable segmentation.

The final result is a multi-level annotation TextGrid containing *phones*, *syll*, *words*, *phono*, and *ortho* tiers as shown in Figure 1. The following figure summarizes the whole procedure.

[Preliminary manual step (if the transcription is in a paragraph format and/or without punctuation): the user reformats the transcription file with one utterance per line]

1. Utterance segmentation script: creates a TextGrid with an interval tier *ortho* containing transcription

[Manual step: user verifies the utterance boundaries]

2. Grapheme-to-phoneme conversion: duplicates the *ortho* tier to *phono* tier, generates a phonetic transcription with major variations

[Manual step: the user validates the phonetic transcription]

3. Phoneme segmentation: generates the *phones* and *words* tiers, then the *syllables* tier

Figure 3: manual and automatic steps

2.4. Evaluation

The evaluation of such a semi-automatic system can be seen in two ways: i) its automatic performance, i.e. how robust and accurate the automatic tool is, and ii) its ergonomics, i.e. how the whole process is made easier and how many times real-time it takes.

For both French and English languages, a 15-minute test corpus of spontaneous speech was fully manually annotated by two experts, independently. This represents respectively 9651 and 9357 phonetic segments (including silences). Table 1 shows the agreement between the 3 annotators (2 humans

and EasyAlign). As some segments might be very short, especially in spontaneous speech, the evaluation was done with two thresholds: the 20ms (as mentioned above) and a narrower one set at 10ms.

	French		English	
	20ms	10ms	20ms	10ms
H1 vs. H2	81%	57%	79%	62%
H1 vs. M	79%	49%	77%	50%
H2 vs. M	82%	52%	75%	51%

Table 1 Percentage of boundary time differences below 20 ms and 10ms for human/human and human/machine comparison for French and English

The table shows that the human vs. human 20ms-agreement is surprisingly low despite the expertise of the annotators. The proposed automatic approach gives nice results as, for both thresholds, the performances of EasyAlign are fairly comparable to human/human ones. The system performs slightly better in French.

As for the 10ms threshold, the segmentations by human annotators are closer to each other than compared to alignment produced by EasyAlign. This is probably due to a default configuration setting in the automatic recognition process that rounds boundary positions to the nearest 10ms. This suggests further investigation is needed for a narrower precision.

On one hand, each annotator needed about 2 hours to manually segment the 15-minute test corpus. It must be noted that the task was facilitated as the utterance segmentation and the phonetic transcription were provided. On the other hand, users usually need approximately 5 times real-time to go through the whole process with EasyAlign. Two people replicated the alignment process for the same 15-minute test corpus within about 1 hour.

3. Adding a new language

After developing EasyAlign for several languages, a straightforward methodology has been built up to welcome any demand of its extension to a new language. In other words, the needs are simply 1. a grapheme-phoneme conversion system that can be called from Praat and 2. at least 1 hour of multi-speaker speech data with its transcription, for acoustic training. After the integration of the phonetisation system, the training data is process through the first two steps, i.e. 1.utterance segmentation which is language-independent and 2. grapheme-phoneme conversion. Then a training step produces the acoustic models according to the phoneme inventory provided by the phonetic transcription as shown on the right side of Figure 2. Few minutes of manually aligned data are needed to evaluate these acoustic models.

Taiwan Min and Spanish were recently added with minimal effort.

For Taiwan Min [9], the training data consisted of 3 hours of monolingual speech from conversational dialogues, with 3 males and 3 females. The evaluation data consisted of 5 extra minutes of each of these 6 speakers. The 20ms and 10ms thresholds methodology gave only 52,% and 30.9% of accuracy. Several reasons could explain these lower results, but a way to increase the performance would be to take advantage of the 3 hours of training data and train acoustic models of triphones instead of monophones.

Therre hours of Spanish speech recordings were also used to train acoustic models and a grapheme-phoneme conversion system has been integrated[10] [11]. Evaluation is currently undergoing.

4. Discussion

The results showed the good performances of our system. Moreover, the overall good feedback from many EasyAlign users (researchers as well as students) is promising. This automatic, speaker-independent, corpus-independent phonetic alignment tool working under Praat can be easily extended for other languages on the basis of a few-minute-long corpus with its phonetic transcription.

EasyAlign is freely available online and comes with a tutorial and a demo. The whole system exists now for French, English and Spanish (i.e. phonetic conversion and HMM-models), while a grapheme-phoneme conversion system must be added for Taiwan Min.

Some extensions are under development like increasing its usability and its performances (to a narrower precision) as well as grapheme-phoneme conversion and acoustic training for other languages.

EasyAlign can be downloaded from this link:

http://latlcui.unige.ch/phonetique/easyalign

5. References

[1] Schiel, F., Draxler, C. "The Production of Speech Corpora Bavarian Archive for Speech Signals", Munich, 2003

[2] Malfrère, F., Dutoit, T., High-Quality "Speech Synthesis for Phonetic Speech Segmentation", Proceedings of Eurospeech, 1997

[3] Kominek, J. and Black, A., "Evaluating and correcting phoneme segmentation for unit selection synthesis", Proceedings of Eurospeech'03, 2004

[4] Sérgio G. Paulo and Luis C. Oliveira, "Automatic Phonetic Alignment and Its Confidence Measures", 4th EsTAL, 36-44, Springer, 2004

[5] J.P.H. van Santen and R. Sproat, "High accuracy automatic segmentation", Proceedings of EuroSpeech99, Budapest, Hungary, 1999

[6] Boersma, P., Weenink, D., "Praat: doing phonetics by computer", http://www.praat.org, accessed in Mar 2010

[7] Young, S. et al. "The HTK book" Cambridge University Engineering Department, http://htk.eng.cam.ac.uk/, acc. in Mar.2010

[8] Beaufort, R. and Ruelle, A. "eLite: système de synthèse de la parole à orientation linguistique". In proc. XXVIe Journées d'Etude sur la Parole,pp. 509-512, Dinard, France, 2006

[9] Fon, J. "A Preliminary construction of Taiwan Southern Min spontaneous speech corpus (Technical report No. NSC-92-2411-H-003-050). Taipei: National Science Council.

[10] Llisterri, J. & Mariño, J.B. (1993). Spanish adaptation of SAMPA and automatic phonetic transcription, Proyecto Esprit 6819. Informe SAM-A/UPC /001/V1 (February 1993).

[11] Moreno, A. & Mariño, J.B. (1998). Spanish dialects: phonetic transcription, Proc. ICSLP'98, Sydney, Australia (November 1998), pp. 189-192.