EasyAlign: a friendly automatic phonetic alignment tool under Praat

Jean-Philippe Goldman

Department of Linguistics, University of Geneva, Switzerland
TCTS Lab, Université de Mons - UMONS, Belgium
jeanphilippegoldman@gmail.com

Abstract

We propose a user-friendly automatic phonetic alignment tool for continuous speech: EasyAlign. It is developed and freely distributed as a plug-in of Praat, the popular speech analysis software. 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. The performances of this HTK-based aligner compare to human alignment and to other existing alignment tools. It is currently available for French and English.

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

1. Introduction

The purpose of phonetic alignment (or phonetic segmentation) is to determine the time position of phone 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 like automatic speech recognition, speech synthesis, as well as prosodic and phonetic research. Precision requirement depends on how this alignment will be used: a corpus-based text-to-speech system requires a high level of precision as various segments of speech will be extracted and concatenated to create the speech output, whereas a lower precision is needed if the purpose of segmentation is to find the syllable nuclei - for prominence detection, for example.

The segmentation task can be done manually or automatically. In the fully manual approach, the labeller will use an annotation tool like in Praat [1] or ESPS [2], and look at the signal and spectrogram representations while listening to small parts of speech in order to decide where exactly to place the boundary of each phone. This is still the most accurate method if the annotator spends enough time and concentration on the corpus to segment, and it is also supposed to be perceptively validated. But processing time is a major drawback as the need for large aligned speech corpora keeps growing. Schiel reports in [3] that 800 times real-time is needed all in all for a fully manual phonetic segmentation, i.e. more than 13 hours for a one-minute recording.

Thus, various automatic methods are now used as they are much quicker and their performances are comparable to the manual approach. Besides, their results are reproducible, unlike with manual labelling, and will be consistent throughout a large corpus that would be annotated differently if shared among several human labellers, i.e. using different criteria for segmentation. Unfortunately, whatever the automatic tool chosen, computational skills are often needed and many steps are required to prepare the data before the automatic tool can do its job. Moreover, in spontaneous speech, many phonetic variations occur. Some of these phonologically known variants are predictable and are included in the pronunciation dictionary (although these variations are not always properly detected automatically) but many others are still unpredictable. After all, experience shows that automatic systems sometimes make obvious errors that a human would not.

Some improvement may facilitate the process, like post-processing tools that detect major errors in segmentation by looking at low confidence scores in recognition and point them out to the user. These enhancements lengthen the process and still require manual checking. All in all, the automatic approaches are never fully automatic, nor straightforward, nor instantaneous. It is a matter of compromise between time, precision and computational skills. The real question is how many times real-time it actually takes to segment (semi-)automatically, i.e. formatting the data, executing the tools and correcting the results.

In the first section, we compare existing techniques. The second part describes EasyAlign, its internal details, innovations and performances in terms of agreement with human and automatic alignment and time needed (compared to the corpus duration).

This ergonomic tool should be seen as a friendly layer based on an existing alignment tool, namely HTK [2], and which facilitates the alignment process for a computer science non-specialist. This Praat plug-in consists of a group of tools to successively perform utterance segmentation, grapheme-to-phoneme conversion and phonetic segmentation, respectively. Between the scripts execution, some minor manual verifications and adjustments may be required to ensure better quality. The whole process starts from a sound file and its orthographic (or phonetic) transcription within a text file or in a convenient TextGrid format.

2. Automatic alignment

2.1. Existing techniques

Various techniques have been developed for phonetic alignment. Some of them have been borrowed from the automatic speech recognition (ASR) domain. But the alignment process is much easier than speech recognition as the task is not to guess which words and phonemes are pronounced, but when. For that reason, the famous (Hidden Markov Models) HMM-based ASR systems are widely used in a forced-alignment mode for phonetic segmentation purposes.

Another approach combines a text-to-speech system (TTS) and a Dynamic Time-Wrapping algorithm (DTW). In this case, synthetic speech is generated from the orthographic or phonetic transcription and is compared to the corpus as in [4]. The DTW will find the best temporal mapping between the two utterances using an acoustic feature representation.

In [5], the two techniques are compared and it turns out that this second system is more accurate than HMM most of the time but that its lower overall evaluation is due to some rough errors. [6] presents a hybrid system based on these two techniques in cascade (first HMM then TTS+DTW), where some results are better, and a comparative test that included another two techniques (artificial neural networks and Classification and Regression Trees). The HMM-based aligner had the best results by far. In [7], 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, i.e. a shell window, is usually required.

2.2. Evaluation

A way of measuring the agreement between two (automatic or human) alignments of the same corpus is to compute the time differences of every phone boundary (with the hypothesis that the phonetic transcriptions are the same) and then, to determine the percentage of differences that are below 20 ms, a commonly agreed threshold.

According to various sources (like [3] and [8]), agreement between human labellers is expected to be between 85% and 95%, depending on the type of corpus (spontaneous or read speech, studio or street recordings) and on the transcribers' training and experience for this task.

The automatic techniques have comparable results to those of human labelling: [9] mentions between 70% and 85% of agreement. Other studies like [10]and more recently [11] based on explicit acoustic features showed above 90% agreement. See also [6] for further investigations on speech-rate-independent measures. In section 3.4, we will base our evaluation on such measurements.

3. EasyAlign

EasyAlign, our freely available system, is also HTK-based but its topmost layer, i.e. the user interface, lies within Praat software. Thus, it hides the in-line commands that require non-trivial computational skills and furthermore provides already trained models of phones. Thus, it is far more ergonomic for a non-specialist. Moreover, it is distributed as a self-installable plug-in, and its tools are directly available from the Praat menus.

Within Praat, 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.

More precisely, these three steps are:

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

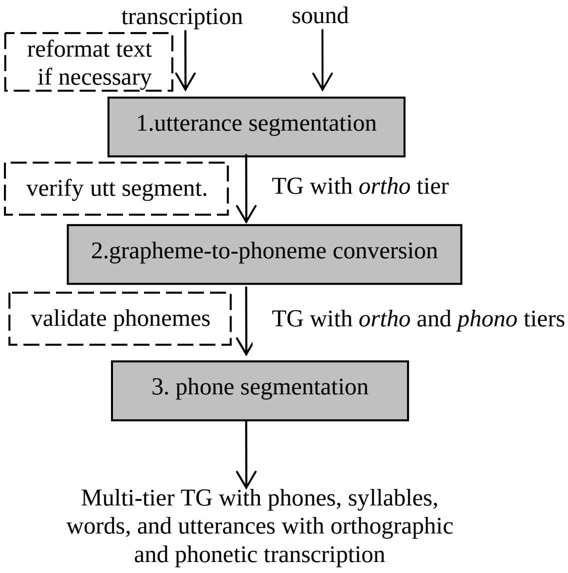
transcription
sound
reformat text if necessary
1.utterance segmentation
verify utt segment.
TG with *ortho* tier
2.grapheme-to-phoneme conversion
validate phonemes
TG with *ortho* and *phono* tiers
3. phone segmentation
Multi-tier TG with phones, syllables, words, and utterances with orthographic and phonetic transcription

Figure 1: the whole process in 3 automatic steps with the manual steps in dashed lines. The TextGrid (TG) Praat format is used as a multi-level annotation

Of course, providing a TextGrid already segmented into utterances together with an orthographic and/or a phonetic transcription speeds up the process. Below, each step is described in detail and a simple user guide is provided in a frame.

3.1. 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. 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 utterances and 2. it is easier to scroll and make use of a large corpus if major units (i.e. about utterance-sized) exist.

We suggest a method that facilitates this step for various formats of transcription:

- as paragraphs or as “one sentence per line”
- with or without punctuation

The newline character and/or the punctuation is/are 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

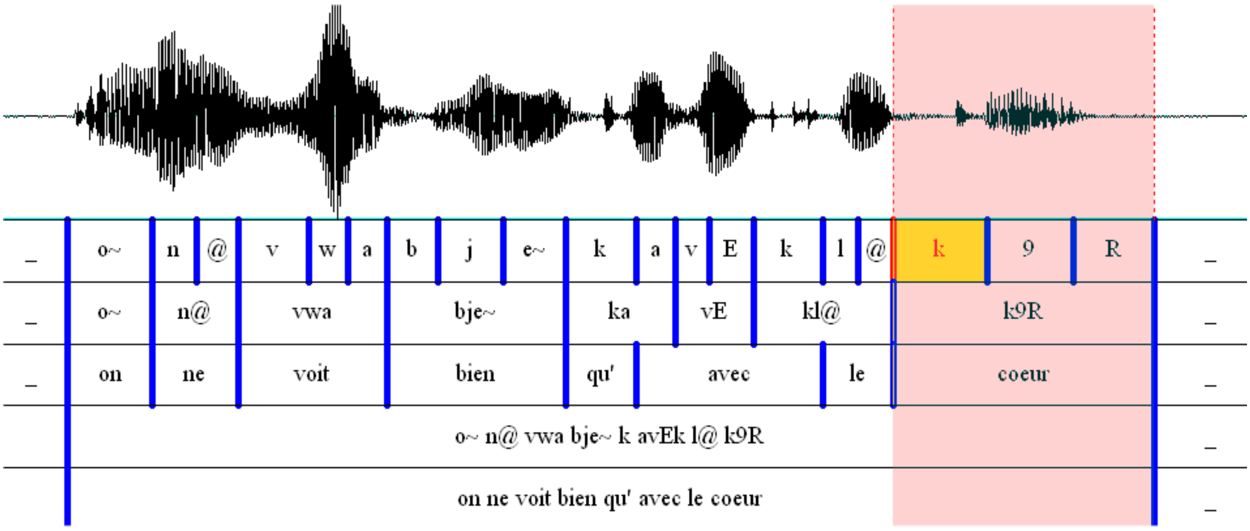
_ o~ n @ v w a b j e~ k a v E k l @ k 9 R _
_ o~ n@ vwa bje~ ka vE kl@ k9R _
_ on ne voit bien qu' avec le coeur _
o~ n@ vwa bje~ k avEk l@ k9R _
on ne voit bien qu' avec le coeur _

Figure 2*: the full resulting TextGrid with 5 tiers from bottom to top: ortho, phono, words, syllables, phones*

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 lively 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 presence or absence of pauses between utterances and 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.

3.2. Grapheme-to-phoneme conversion

The purpose of this step is to create the *phono* tier, which is similar to the *ortho* tier (i.e with the same boundaries) but with a phonetic transcription. This will be used by the speech recognition engine.

It is rather unusual for an HMM-based aligner to require the phoneme sequence as an input, since their implementation is usually based on a sequence of orthographic words, a pronunciation dictionary (including variations of pronunciation per word) and stochastic models of phonemes. 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 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 showed 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 [12] and suggests some pronunciation variants.

3.3. 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 2. The following frame sums up the whole procedure:

→ **Preliminary manual step** (only if transcription is in a paragraph format and without punctuation): the user reformats the transcription text 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

3.4. Evaluation

The evaluation of such a semi-automatic system consists of two measures: i) its automatic performances, 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 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 a 5-minute subset within respectively 21 and 32 minutes.

To complete this evaluation, comparing EasyAlign to other automatic approaches would be a plus.

4. Discussion

These results show that our system has relatively good performances. Moreover, the overall good feedback from many EasyAlign users (researchers as well as students) is promising. This semi-automatic, speaker-independent, corpus-independent phonetic alignment tool working under Praat can be easily extended and trained for any language 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 for French and English (i.e. phonetic conversion and HMM-models). Scripts for training new acoustic models are available upon request.

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

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 the German, Dutch and Hebrew languages.

5. Acknowledgements

The author would like to thank the Programme Wist2, Convention n°616422, financed by the Walloon Region (Belgium) Project EXPRESSIVE: *Système automatique de diffusion vocale d'informations dédicacées: synthèse de la parole expressive à partir de textes balisés*

6. References

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

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

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

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

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

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

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

[8] Bürki, A et al. "Alignement automatique et analyse phonétique: comparaison de différents systèmes pour l'analyse du schwa", Traitement Automatique des Langues, vol.49 n°3

[9] Sjölander, K., "Automatic alignment of phonetic segments, Working Papers", 49:140-143, Lund University, 2001.

[10] Hosom, J.-P., "Automatic Phoneme Alignment Based On Acoustic-Phonetic Modeling", PhD Thesis, OGI of Science and Technology, 2000

[11] Hosom J.-P. "Speaker-independent phoneme alignment using transition dependent stares", Speech Communication (51) 2009

[12] 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