DOI: 10.17469/O2106AISV000024

FRANCESCO CANGEMI, JESSICA FRÜNDT, HARRIET HANEKAMP, MARTINE GRICE

A semi-automatic workflow for orthographic transcription and syllabic segmentation

Manual orthographic transcription of spontaneous speech is notoriously time consuming, and segmentation at the level of the syllable requires further large amounts of processing time. Automatic orthographic transcription and automatic syllable segmentation, on the other hand, usually yield unsatisfactory precision, especially when applied to spontaneous speech. In this contribution, we report on a semi-automatic workflow that uses freely available software and combines the speed of automatic processing with the quality of manual transcription and segmentation. We apply the procedure to highly spontaneous speech and show (i) no loss of quality compared with manual output and (ii) an average 75% reduction of manual processing time.

Key words: orthographic transcription, forced alignment, syllabic segmentation, spontaneous speech, automation.

1. *Introduction*

The first step in the phonetic analysis of spontaneous speech is often represented by an orthographic transcription and a syllabic (or phone-level) segmentation of the recordings. Both operations can be performed either manually or automatically. When performed by a trained phonetician, fully manual transcription and segmentation can be highly reliable, but they have a high cost in terms of time and effort. For certain speech types (such as an excited conversation between friends), accurate manual transcription and segmentation can require up to one hour for each minute of speech. On the other hand, fully automatic transcription and segmentation can immensely reduce the workload placed on the analyst's shoulders, but they are also comparatively less reliable in terms of output quality. Whereas read speech can be automatically transcribed and segmented with remarkable results (Vagnini-Holbl, Draxler, 2018), the performance of automatic procedures remains unsatisfying for spontaneous speech, where words can bear little resemblance to the canonical forms used in building speech recognition systems. Similarly, performance of automatic procedures can be satisfying for languages with extensively trained models (such as English), but is comparatively poorer for understudied languages, for regional varieties and for foreign accented speech. As a result, researchers often combine automatic and manual workflows when analysing spontaneous speech, for example by providing a manual transcription first, and then using the transcription as input for the automatic segmentation. In this paper, we detail a semi-automatic workflow in

which an orthographic transcription is generated automatically, then manually corrected. The corrected transcription is then used as input for an automatic segmentation tool, whose output is then once more manually verified. This procedure yields an output as precise as a reference manual segmentation, but requires significantly shorter manual processing time (in average, down to 25% of the time required for a fully manual annotation). Crucially, this semi-automatic procedure only uses software which is compatible with all operating systems, freely available, and easy to use.

2. *Method*

2.1 Material

In order to test the semi-automatic transcription and segmentation workflow, we recorded a spontaneous conversation between two speakers. We used two head-mounted microphones (AKG C544L) connected through an audio-interface (Focusrite Scarlett 6i6) to a laptop running a digital audio workstation (*Reaper*, Cockos, 2018), with a sample rate of 44100 Hz and a bit depth of 24 bit. This setup is highly portable and relatively inexpensive, but it allows for professional recording quality and matches the standard practices in phonetics research.

The setting and content of the recording, on the other hand, are comparably less usual. Recordings took place at the first author's home, taking only basic precautions for maximising sound quality: windows were closed and drinking glasses were provided with a straw, but no curtains or rugs were used to reduce reverberation, and there was no physical barrier between the two speakers' microphones. This caused voices bleeding into the other speaker's channel, introduced a certain amount of background noise, and included reverberation into the recordings. Nonetheless, the achieved audio quality was deemed acceptable, especially if considering that the unusual recording setup (at home, having drinks, with no barrier between interlocutors) had a positive effect on the naturalness of the interaction.

To further increase naturalness, the recordings capture two good friends (the first author and another phonetician) as they play a videogame. The game offers an enjoyable task, which is also often performed for mere recreational purposes, and is thus substantially different from most other tasks used in phonetic data collection. Compared to recordings where participants are asked to perform tasks they would normally not perform in their daily lives (such as reading aloud a list of words, complete dialogues based on pictorial context, et cetera), speaker involvement is notably higher in this interaction. As a result, most of the recordings sound spontaneous and relaxed, as attested by the relatively high frequency of colloquialisms and profanities captured by the microphones.

In order to test our semi-automatic workflow on maximally challenging material, we introduced a further obstacle beyond the sub-standard recording conditions and the colloquial speech style: none of the two speakers is a native speaker of the language used in the interaction, which is thus a mixture of Italian-accented and Hebrew-accented English. Moreover, by referencing to in-game characters and

actions which bear custom-made names (e.g. *Taurukh* for a centaur-like creature), speakers often used lexical items that are absent from automatic speech recognition dictionaries, and are thus treated as non-words.

In summary, despite the use of relatively standard recording equipment, we voluntarily hindered the performance of the semi-automatic workflow through poor recording conditions, extreme spontaneity of the interactional situation, use of non-native English, and abundance of non-words. On the other hand, this resulted in speech samples that feel remarkably genuine, with both speakers (including the experimenter) reporting having lost track of the scientific purpose of the recordings early on during the gaming session. Nonetheless, in order to avoid confounds, for the following analyses we will not employ speech uttered by the first author, and will only use 8 sound files (of 1 minute each) extracted from the 29 minutes long recording of the Hebrew-accented English speaker.

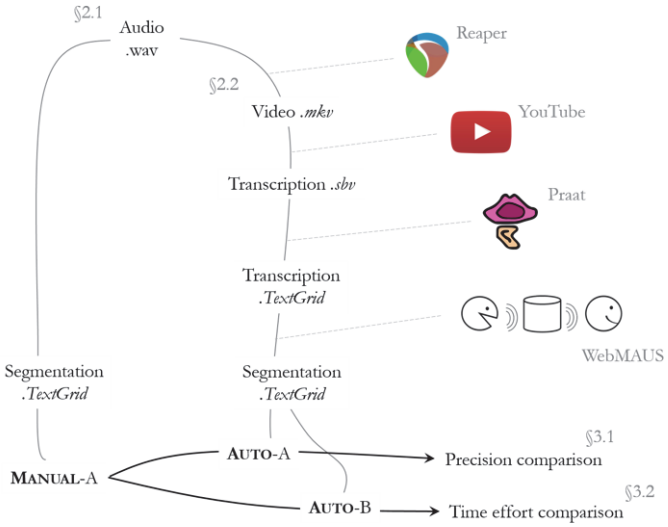
2.2 Method

Each of the 8 audio files was submitted to both a manual and a semi-automatic workflow. In the manual workflow, the third author used *Praat* (Boersma, Weenink, 2018) to create orthographic transcription and manual syllable segmentation following standard best practices (Machač, Skarnitzl, 2013). We will refer to these files as MANUAL-A (manual annotation, annotator A, i.e. the third author).

In the semi-automatic workflow, the second author prepared audio files for upload onto *YouTube*. Since *YouTube* only accepts video files, audio files were combined with an image in *Reaper* (Cockos, 2018). The file was rendered with uncompressed (maximal) audio quality. With the consent of the speaker, the video files were uploaded to a private channel on *YouTube*, and a first pass of the orthographic transcription was obtained using the *YouTube* automatic transcription function. The automatic transcription contains timestamps for suggested beginning and end of each interpausal unit. The transcription was exported in one of the available subtitles formats (sbv), and transformed via script into a *Praat* annotation file (TextGrid). Errors in the automatic transcription (including imprecise timestamps) were manually corrected in *Praat*. Audio and annotation files (wav and TextGrid) were then further processed using *WebMAUS* (Kisler, Reichel & Schiel, 2017), one of the tools available on the webpage of the Bavarian Archive for Speech Signals. *WebMAUS* is a web-based forced alignment service, which takes audio and orthographic transcription files as input, and returns phone-level segmentation as output. We used the so-called *G2P-MAUS-PHO2SYL pipeline*, which provides a phonologisation of the orthographic transcription (*G2P*), a phonetic segmentation (*MAUS*), and the reconstruction of syllables based on the phonetic segmentation (*PHO2SYL*); see Kisler et al. (2017) for details on these three modules. The output TextGrids were further processed by eliminating the tiers generated by *WebMAUS* as intermediate steps, and by only saving the orthographic transcription and the syllable segmentation. Using these simplified TextGrids as starting point, the second author manually corrected the syllabic boundaries automatically placed by

WebMAUS, thus completing the semi-automatic workflow. We will refer to these files as AUTO-B (semi-automatic annotation, annotator B, i.e. the second author). In addition, the simplified TextGrids were also submitted to the third author for manual correction, but only after she had already provided the final TextGrids for the manual workflow. We will refer to these files as AUTO-A (semi-automatic annotation, annotator A, i.e. the third author).

Figure 1 - *Diagram of workflows, including annotators and comparisons*



2.3 Performance comparisons

The use of three segmentations (MANUAL-A, AUTO-A and AUTO-B) aimed at countering annotator bias in the comparison of the two workflows (cf. Figure 1). When focussing on the *time effort* required by the two workflows, we compare MANUAL-A with AUTO-B, since a comparison between MANUAL-A and AUTO-A would introduce order effects (i.e. presenting the same audio files twice to the same annotator would lead to an advantage for any workflow performed last). Similarly, when comparing the *precision* of the two workflows, we compare MANUAL-A with AUTO-A, in order to avoid the unescapable differences between the segmentation habits of the two annotators. This approach has the additional advantage of avoiding the pitfalls of promoting a manual annotation to the rank of reference segmentation (cf. Cosi, Falavigna & Omologo, 1991).

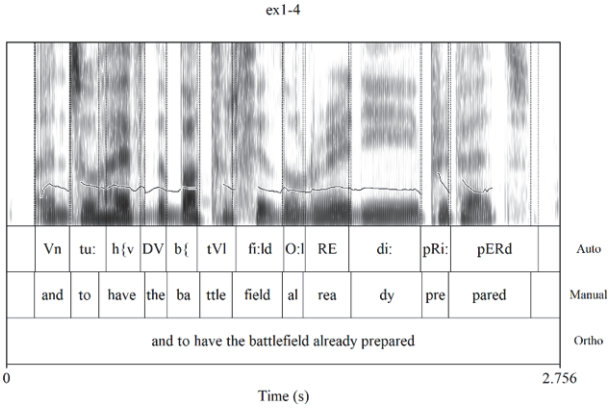
Since the focus of the present paper lies with a first description of the semi-automatic workflow as tested on a limited speech database, we will refrain from using inferential statistics in the comparisons reported below.

3. *Results*

3.1 Precision comparison

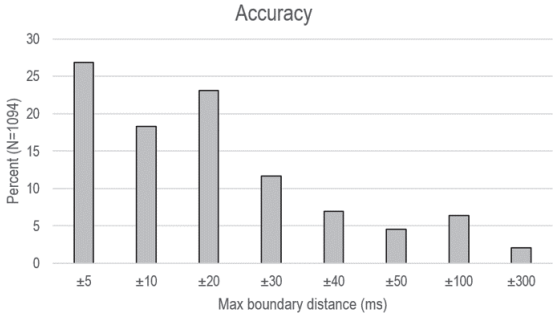
In order to assess the precision of the two workflows, we compared the position of syllabic boundaries in the two segmentations provided by the third author (Manual-a vs. Auto-a). Figure 2 shows spectrogram and pitch track for a small portion of one of the test sound files. In the annotation panel are visible the orthographic tier (Ortho) and syllabic tiers for both manual workflow (Manual) and semi-automatic workflow (Auto, in SAMPA). A visual inspection suggests that the two workflows yield virtually undistinguishable quality.

Figure 2 - *Example of manual and automatic segmentations*



In order to quantify this assessment, we extracted the timestamps of 1094 syllables for each of the two segmentations (Manual and Auto), and calculated distances between matching boundaries across the two segmentations. Figure 3 shows the distribution of maximal temporal distance (in milliseconds) between boundaries in the automated and manual workflows, using negative numbers for late automatic boundaries. The figure shows that 45% of automated-workflow boundaries fall within ±10ms of boundary in the manual workflow; 68% of boundaries fall within ±20ms of boundary in the manual workflow, indicating highly comparable performances. Approximately half of automated-workflow boundaries fall at the maximum level of precision available for automatic (and human) segmentation (±10ms), while less than 9% of boundaries fall in a ±50ms window.

Figure 3 - *Precision comparison*

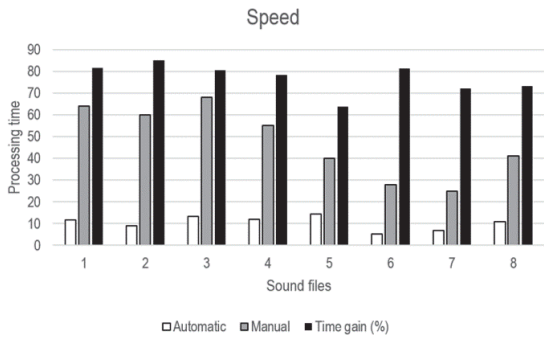


3.2 Time effort comparison

Figure 4 shows the crucial comparison between the processing times required by the two workflows, as performed separately by the two annotators (Manual-a vs. Auto-b). Processing times are expressed in minutes for each of the 8 test files (on the x-axis), separately for the Automatic workflow (white bars) and in the Manual workflow (grey bars). The 8 test files were excerpted from relatively varied moments in the interaction, with some samples featuring 1 minute of excited monologue, and other samples capturing 1 minute of a relaxed dialogue. As such, they can be more or less challenging to the manual and automatic annotator. It is thus unsurprising that the processing times vary greatly from sample to sample. In the Manual workflow, processing times vary from 25 to 68 minutes, in line with the expectations for orthographic syllabic segmentation of highly spontaneous non-native speech. Processing times are equally varied for the Automatic workflow, ranging from 5 to 14 minutes.

Crucially, however, the percent time gain when using the Automatic workflow is less variable, as shown by the black bars. Depending on the files, the advantage can range between 64% and 85%, suggesting that, in average, the automated workflow requires a fourth of the processing time required for manual annotation.

Figure 4 - *Time effort comparison*



4. *Discussion*

The preliminary evidence presented above suggests that, compared to the manual workflow, the semi-automatic workflow provides virtually undistinguishable precision, with a substantial processing time reduction. These results are particularly encouraging, considered that the materials used in this test were maximally challenging due to the spontaneous nature of the interaction, the familiarity between the speakers, the use of non-native English, the sub-standard recording conditions, and the use of a large number of non-words (see §2.1). Applying the semi-automatic procedure on native read speech with lemmatised words would surely require less manual processing time.

Manual correction of the orthographic transcription posed no particular challenge, with the exception of false starts and non-words. Syllable segmentation seemed to be more affected by the spontaneous nature of the interaction. For example, manual correction was required for several cases of long frication noise in a turn-opening 'so', a common phenomenon in conversational speech.

Despite these challenges, the semi-automatic workflow for orthographic transcription and syllabic segmentation presented here reduced processing time by approximately 75%. Importantly, it only required multiplatform and free software, and it was applied to extremely challenging speech, recorded with inexpensive and portable equipment.

Bibliography

Boersma, P., Weenink, D. (2018). *Praat: doing phonetics by computer* [Computer program]. Retrieved from www.praat.org.

Cockos (2018). *Reaper* [Computer program]. Retrieved from www.reaper.fm.

Cosi, P., Falavigna, D. & Omologo, M. (1991). A preliminary statistical evaluation of manual and automatic segmentation discrepancies. In *Proceedings of Eurospeech*, Genova, 1991, 693-696.

Kisler, T., Reichel, U. & Schiel, F. (2017). Multilingual processing of speech via web services. In *Computer Speech & Language*, 45, 326-347.

Vagnini-Holbl, C., Draxler, C. (2018). Comparing acoustic measurements from manual and automatic segmentations. Paper presented at *Phonetik und Phonologie im deutschsprachigen Raum*, Vienna, 6-7 September 2018.

Acknowledgments

This work was partly supported by the German Research Foundation through the Collaborative Research Centre "Prominence in Language" (SFB 1252).