A semi-automatic workflow for orthographic transcription and syllabic segmentation

Riservato Riservato Riservato

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 combines the speed of automatic processing with the quality of manual transcription and segmentation. While recent evaluations suggest that read speech might be efficiently segmented without requiring further manual verification (Vagnini-Holbl & Draxler 2018), in this submission we use highly spontaneous speech as test material. We show virtually no loss of quality (compared with manual output), while reducing manual processing time by an average of 75%.

Two non-native speakers of English were recorded while playing a videogame, using two head-mounted microphones (AKG C544L) connected through an audio-interface (Focusrite Scarlett 6i6) to a computer running an audio-processing software (*REAPER*, Cockos 2018), with a sample rate of 44100 Hz and a bit depth of 24 bit. From the two separated mono channels, we extracted 8 audio files of 1 minute in duration. 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. In the semi-automatic workflow, the second author transformed the audio files into video files, by pairing them with an image. The video files were uploaded to a private channel on *YouTube*. The *YouTube* automatic transcription function was used to obtain a first pass of the orthographic transcription. The automatic transcription contains timestamps for suggested beginning and end of each interpausal unit. The transcription was exported as SBV file and transformed into a TextGrid file via a custom *Praat* script. Errors in the automatic transcription (including imprecise timestamps) were manually corrected in *Praat*. Wave and TextGrid files were further processed using *WebMAUS* (Kisler et al. 2017), using the G2P-MAUS-PHO2SYL pipeline. This pipeline 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 all unused tiers, and leaving only the orthographic transcription and the syllable segmentation. Syllabic boundaries were then manually corrected (separately) both by the second author and by the third author (who performed the manual segmentation). This strategy allowed to counter annotator bias, both when comparing the time necessary to the two workflows (by not presenting the same audio files twice to the same annotator, and thus introducing order effects) and when comparing the precision of the two segmentations (by not evaluating syllable segmentations provided by two different annotators).

Fig. 1 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 (*Manual*) and semi-automatic (*Auto*) workflows. A qualitative inspection suggests that the two workflows yield virtually undistinguishable quality. In order to quantify this assessment, we extracted the timestamps of 1094 syllables for each of the two segmentations (*Manual* and *Auto*, and evaluated distances between matching boundaries across the two segmentations. Fig. 2 shows the distribution of the interval durations between workflows, and indicates that 68% of the semi-automatic boundaries fall within +/-20ms from the correspondent manual boundary. More importantly, Fig. 3 shows processing times for the two workflows, as performed (separately) by the two annotators. Depending on the files, compared to the manual workflow, the automated workflow requires 64%-85% shorter processing times.

Given the small amount of available data, we refrain from providing an evaluation based on inferential statistics. The results are nonetheless encouraging, since they suggest that, compared to the manual workflow, the semi-automatic workflow provides virtually undistinguishable precision, with a substantial time processing reduction. Note that these results are underestimating the efficacy of the semi-automatic workflow, since automatic orthographic transcriptions often needed correction either because of the non-native English of the speakers or because of the use of game-specific words, which are not featured in the *WebMAUS* dictionaries (e.g. ‘necroguards’). Automatic syllable segmentation was perceived to be relatively easy to perform; it often involved corrections relative to sounds with unusual durations – at least when compared to the reference phone models. This is for example the case of sounds which appear to be particularly long due to interactional reasons, such as the frication noise in a turn-opening ‘so’. Applying the semi-automatic procedure on native read speech with lemmatised words would surely yield even better results.

With an inflated ratio of approximately 9 minutes of processing to 1 minute of highly spontaneous non-native speech, this free and semi-automatic workflow for orthographic transcription and syllabic segmentation could be particularly useful for the valorisation of existing speech corpora – provided it is legally possible to upload and process them on the Internet.

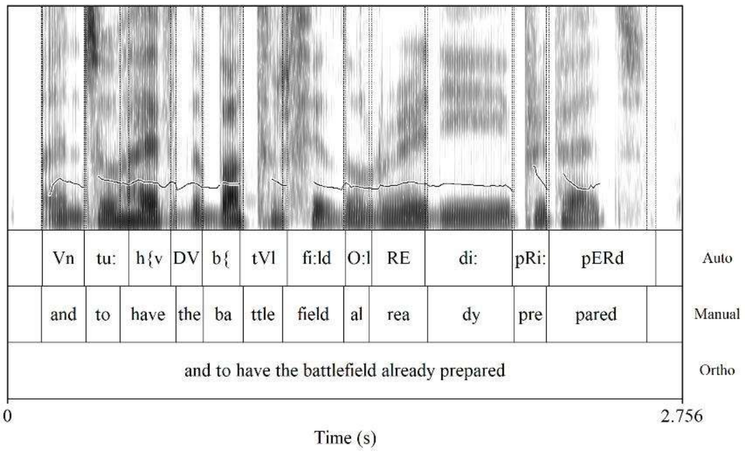
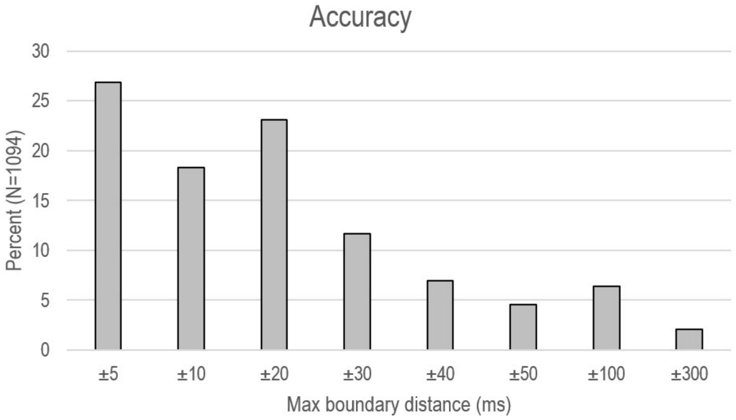


Fig. 1. Spectrogram and pitch track for a sound file excerpt, with orthographic transcription (*Ortho*).

Syllabic annotation is provided separately for the automated workflow (*Auto*, in SAMPA) and for manual segmentation (*Manual*).

Fig. 2. Maximal temporal distance (in milliseconds) between boundaries in the automated and manual workflows. Negative numbers for late automatic boundaries.

45% of automated-workflow boundaries within ±10ms of boundary in the manual workflow; 68% within ±20ms; 91% within ±50ms.



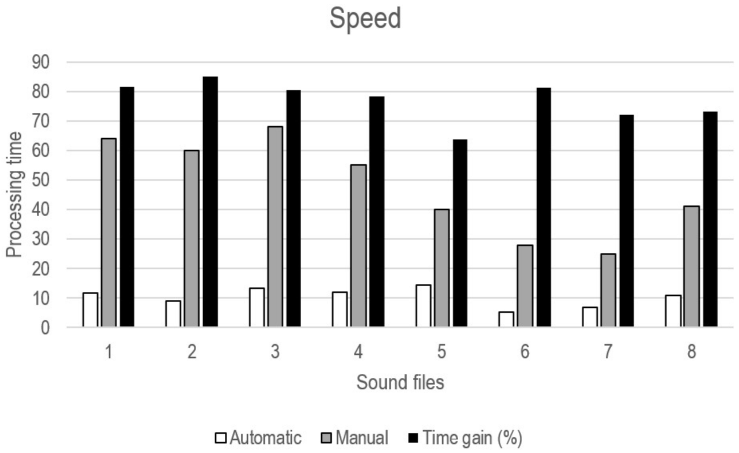


Fig. 3. Processing times (minutes) for each sound file in the Automatic workflow (white bars, ranging from 5 to 14 minutes) and in the Manual workflow (grey bars, ranging from 25 to 68 minutes).

Black bars show the percent time gain when using Automatic workflow (ranging from 64 to 85%).

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

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

Kisler, Reichel, Schiel (2017). Multilingual processing of speech via web services. *Computer Speech & Language* 45:326–347.

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