PEASYV: A PROCEDURE TO OBTAIN PHONETIC DATA FROM SUBTITLED VIDEOS

Adrien Méli[1], Nicolas Ballier[2]

[1]Université Paris Cité, CLILLAC-ARP, [2] Université Paris Cité, CLILLAC-ARP & LLF
adrienmeli@gmail.com, nicolas.ballier@u-paris.fr

ABSTRACT

This paper presents a procedure to easily obtain phonetic data from subtitled YouTube videos in English. Called PEASYV (Phonetic Extraction and Alignment of Subtitled YouTube Videos) and based on a UNIX shell script, this procedure makes it possible to obtain fine-grained phonetic and phonological data from any subtitled videos posted on YouTube. It uses two aligners, P2FA [1] and SPPAS [2], to generate Praat [3] TextGrids for each video. These TextGrids contain tiers specific to each aligner, with segmental, syllabic and lexical alignments. SPPAS tiers also contain prosodic information. After describing the procedure and the generated files, a case study of a YouTube channel on English pronunciation is presented, in order to assess the quality of the alignment: the formants of the monophthongs from the channel's 452 videos, amounting to more than 179 hours of spontaneous speech, are analysed. The potential alignment errors, along with their prevention, are then discussed.

Index Terms: English vowels, corpus phonology, formant extraction, automatic alignment

1. INTRODUCTION

Several forced aligners have been developed, among others MAUS (Munich AUtomatic Segmentation) [4] and WebMAUS [5], Prosodylab-aligner [6], LaBB-CAT [7] and the Montreal Forced Aligner (MFA) [8] (see [9] for a comparison of MAUS, FAVE, LaBB-CAT and MFA and [10] for a comparison of FAVE, MAUS and MFA and for a discussion of the acoustic models used in the forced aligners). Using manually aligned data from Trinidadian English [10] has compared Forced Alignment and Vowel Extraction (FAVE), Munich Automatic Segmentation (MAUS), and the Montreal Forced Aligner (MFA) and their performances in automatically segmenting Trinidadian speech. Forced aligners presuppose a text when deep neural network acoustic modeling is available for YouTube video transcription [11]. Several corpora have been made available as text transcriptions of YouTube videos [12]. We wish to explore the possibility of exploiting this automatic generation of captions for videos in English on Youtube to capture phonetic datasets online.

The emergence of tools such as YouGlish[1], that acts as a concordancer on a selection of the subtitled videos posted on YouTube, may be used for the observation of spontaneous spoken data in an ecological environment or at least as an alternative to the observer's paradox [13]. As time-stamped transcriptions of speech, subtitles have the potential to serve as data for phonetic and phonological purposes. The issue arises of the reliability of such a source. This study examines whether subtitle time stamps can be used for finer-grained segmental alignment. It focuses on English monophthongs because they can be easily represented in the F_1/F_2 space: the accuracy of the alignment can therefore be visually checked. The rest of the paper is organised as follows. Section 2 presents an outline of the procedure. Section 3 offers a case study investigating the reliability of PEASYV by visually inspecting the vocalic productions of a female speaker on a YouTube channel teaching British pronunciation. Section 4 discusses the results and explores future venues of research.

2. DESCRIPTION

This section outlines the workflow (Section 2.1) to align the subtitled video automatically, and then describes the resulting files (Section 2.2).

2.1. Workflow

PEASYV can be installed on any computer running a major Linux distribution. Execution on Windows or MacOS has not been tested. The following programmes are required: `yt-dlp` [14], `ffmpeg` [15], `R` [16], `Praat` [3], `SPPAS` and `P2FA`. The workflow is controlled by a UNIX-compatible shell script that takes a text file as argument. Each line of the text file contains a link to a subtitled YouTube video. Optional fields on the same line are the speaker's gender and a description of the variety of English they speak. The script then loops over each line of the text file and executes a series of subscripts to perform the actions in the list below.

The script:

1. Downloads the video and its subtitles using `yt-dlp`;
2. Converts the video to a main 11kHz mono `.wav` file;
3. Creates a main TextGrid and a main PitchTier for that sound file;
4. Uses the time stamps of the downloaded subtitles to create matching intervals on a tier in the TextGrid;
5. Prints the subtitles in the corresponding intervals on the tier;
6. Splits the main files (*i.e.* the main sound file, the TextGrid and the PitchTier) into subfiles of the same types;
7. Inputs those subfiles into the two aligners, SPPAS and P2FA, successively;
8. Reintegrates the automatically aligned subfiles into the original main TextGrid

The main sound file is converted to a single-channel sound file sampled at 11,025Hz to comply with the requirements of the aligners. The PitchTier is generated following the procedure described in Hirst [17], which uses the interquartiles values obtained from Praat's default values. These interquartile values are obtained using a separate *R* script. Regardless of whether this extra step is carried out, the obtained prosodic tiers (*c.f.* next section) will only provide accurate information if the video features the same speaker. The reasoning behind step 6 in particular will be discussed in Section 4.

2.2. Resulting files

For each video, the procedure generates a TextGrid file which contains eleven tiers. The first tier is the transcription tier, with intervals corresponding to the time stamps in the original subtitle file. The next two tiers are specific to SPPAS, and deal with prosodic analysis: *(i)* the first prosodic tier is a tier with the target points calculated by the MOMEL algorithm (Modeling Melody, *c.f.* Hirst [18]); *(ii)* the second one is a tier with codes from the International Transcription System for Intonation (INTSINT). The remaining tiers include the phonemic and lexical alignments performed by each aligner, and four syllabic tiers (two for each aligner). The syllabic tiers were generated from the Longman Pronunciation Dictionary (*LPD*, [19]). The reason why there are four syllabic tiers is that although both aligners use the Carnegie Mellon University dictionary [20], they transcribe phonemes differently: SPPAS uses SAMPA, whereas P2FA uses ARPABet. Because the *LPD* offers transcriptions in the International Phonetic Alphabet (IPA), the final TextGrid contains one syllabic tier using the *LPD*'s IPA transcription for each aligner, and one syllabic tier for each aligner's prefered transcription system.

Another set of optional files is generated by PEASYV:

- a textual transcript of the video;
- two spreadsheets, one for each aligner.

The first file is a simple word list created to reference lexical usage and frequency. The two spreadsheets are structured in the same way and provide data for phonetic analysis. Each line corresponds to a vocalic nucleus. The following information is collected (the list is non-exhaustive): the video's metadata, the word and syllable in which the vowel appears, the preceding and succeeding phonemes, whether or not they belong to the same word as the vowel, the vowel's and syllable's durations, and formant readings for the first four formants at each centile of the vowel's duration.

This article contends that the relevance and accuracy of PEASYV can be verified by conducting phonetic analyses of vowels: the assumption is that if the vowels are correctly located on the F_1 /F_2 vocalic trapezoid, then the automatic alignment is also correct. Whether that is the case is explored in the next section.

3. A CASE STUDY

The purpose of this section is to assess the quality of the alignment performed by PEASYV. A cursory phonetic study of vowels is provided using raw, unnormalized F_1/F_2 values in Hertz: if these values are reasonable, then the data may be considered sound for future, more refined, research. The recordings come from a YouTube channel, EnglishLikeANative, which purports to offer lessons on British pronunciation: the channel was created by a female teacher offering advice on how to improve pronunciation in order to sound more British. Apart from a few exceptions, she is the sole speaker in her videos. The vowels she produces are therefore expected to be close to standard Southern British English (SBE) values, such as those presented in Deterding [21].

	SPPAS	P2FA
Vowels	973,780	821,125
Monopththongs	775,042	664,264
Diphthongs	198,738	156,861

Table 1: Per-aligner number of vowels extracted from YouTube channel EnglishLikeANative.

In total, 452 videos were aligned using both SPPAS and P2F, amounting to 172 hours and 39 minutes of recording. The combined storage space of all TextGrids and sound files is 3.5G and 15G respectively. 452 vidÃ©os Table 1 summarizes the number of vowels automatically obtained with the two aligners after running PEASYV on all the videos of the channel. Section 3.1 details the results obtained for SBE monophthongs, and Section 3.2, those for SBE diphthongs.

3.1. Monophthongs

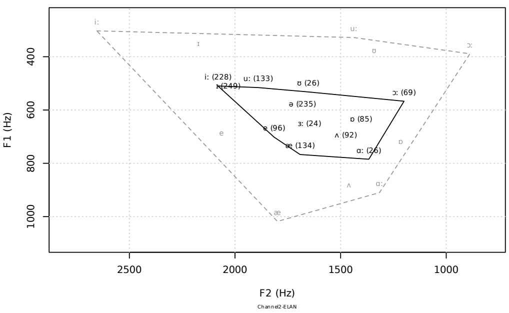

Figure 1: Convex hulls of F_1/F_2 values (in Hz) of all SBE monophthongs. Grey dotted line: Deterding; continuous black line: SPPAS. In brackets: numbers of occurrences (in thousands).

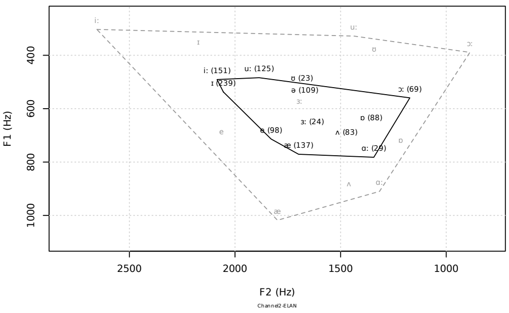

Figure 2: Convex hulls of F_1/F_2 values (in Hz) of all SBE monophthongs. Grey dotted line: Deterding; continuous black line: P2FA. In brackets: numbers of occurrences (in thousands).

One way to obtain an assessment of the accuracy of the alignment carried out by PEASYV is to plot the vowels in the F_1/F_2 space for each aligner, and to compare the resulting values with those from controlled studies. Figure 1 and figure 2 show the mean mid-temporal values in Hertz of the first two formants of each monophthong in colours, along with the values reported in Deterding [21] in grey. The numbers between brackets indicate the number of occurrences of each monophthong in thousands. To facilitate visualization, the convex hulls of the plotted monophthongs have been drawn, in black for SPPAS- or P2FA-aligned values, and in a grey dotted line for Deterding's values. The x-axis (F_2) and the y-axis (F_1) have been inverted to emulate the phonological vocalic trapezoids. As can be seen from these two figures, for both aligners the obtained values plot a consistent vocalic space, albeit more reduced than Deterding's (but *c.f.* section 4 for discussion).

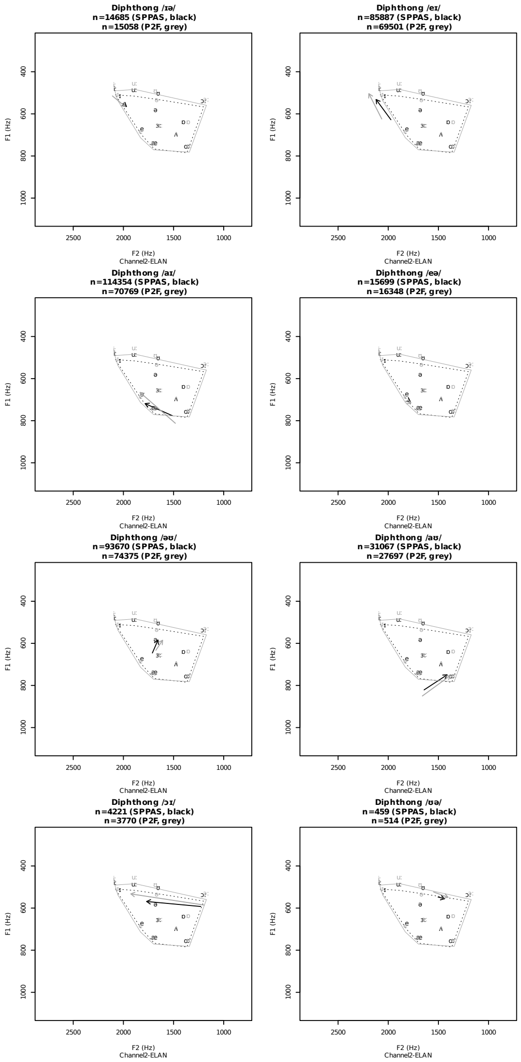

Figure 3: Diphthongs in the F_1/F_2 space compared to the trapezoids of monophthongs (SPPAS: black dotted line; P2FA: continuous grey line) as aligned by SPPAS (in black) and P2FA (in grey).

3.2. Diphthongs

Figure 3 plots the eight diphthongs of SBE in the F_1/F_2 space. SPPAS-aligned formant values are represented in black, and P2FA-aligned values are shown in grey. The numbers of occurrences of each diphthong for each aligner is indicated on top of each diphthong's pane. The diphthongs are visualized with arrows: the starting point of the arrow is the mean F_1/F_2 values in Hertz at 20% of the diphthong's duration, while the pointed end of the arrow corresponds to the mean formant values at 80% of the diphthong's duration. The convex hulls obtained for monophthongs and presented in Section 3.1 were also plotted in order to assess the consistency of the results. The visual inspection of the glides confirms that the formant values extracted by Praat from the automatically generated TextGrids are in keeping with the female speaker's monophthongs, which themselves are located in a realistic vocalic space.

4. DISCUSSION

This section discusses the specifics and motivations of the procedure (Section 4.1) and then explores the validity of the obtained phonetic data (Section 4.2).

4.1. Procedure

One major obstacle to overcome when automatically aligning recordings of variable duration is containing the scope of misalignments and preventing a potential domino effect, where a misaligned segment might cause all succeeding segments to be misaligned too. The design of PEASYV makes it impossible for such cascading misalignments to occur: step 6 of the workflow presented in Section 2.1 entails that the two aligners are only fed extracts that are a few seconds long, possibly even shorter: in this study, the 404 videos were split into 301,511 short recordings. On average, these subfiles lasted 0.8 second. It is worth noting that such sectioning of the main file into smaller files has no consequences on intonational computations, *i.e.* MOMEL and INTSINT: the main PitchTier is created before, being generated following Hirst [18]. This also means that intonational data is only accurate for videos featuring one speaker, as PEASYV in its current state has no system of speaker detection: the interquartile values of the time-stamped pitch contours are therefore calculated across the entire recording.

4.2. Results

The results provided here are based on raw Hertz values: adding intermediary calculations such as vocalic normalization was deemed potentially counterproductive to assessing the reliability and accuracy of PEASYV. For monophthongs, the vocalic trapezoids feature accurately located vowels. The reduced size of the trapezoids compared to Deterding's most likely reveals hypoarticulation: this is to be expected not only because Deterding's reference values were elicited from potentially hyperarticulated /hVd/ templates, but also because the study's values are extracted from connected speech. However, the mean values of /uː/ admittedly prompt further research: the high F_2 may be indicative of an absence of lip-rounding, which might be the result of an abundance of hypoarticulated high-frequency words (such as "too"). For diphthongs, not only are the trajectories across the vocalic space consistent with phonological representations, the findings are also in keeping with more recent studies describing the progressive disappearance of centralizing diphthongs /eə/, /ɪə/ and /ʊə/ (*c.f.* Hannisdal [22] or Lindsey [23]), or advocating changes in transcriptions, for instance from /eə/ to /ɛː/ (*cf.* Collins [24]).

4.3. Availability and future research

PEASYV for the time being is only accessible on request to adrienmeli@gmail.com, along with the files and scripts used for this study. Deployment on a webpage, and access to the scripts on a github page, are planned for the near future.

One field of research to fine-tune the alignments is that of subtitle accuracy, as the quality of the alignment is highly dependent on that of the subtitles. Research is being carried out to assess the discrepancy between automatic and human-supervised subtitles. Another field is expansion of PEASYV to languages other than English, such as French, using SPPAS.

5. CONCLUSION

This study contends that PEASYV is a robust, low-resource procedure to align subtitled videos and collect phonetic data. The reliability of PEASYV was demonstrated with a study of all the vowels produced by a female speaker on a YouTube channel teaching British pronunciation. The mean formant values extracted from the files generated by PEASYV are consistent with established results and ongoing changes in SBE.

6. REFERENCES

[1] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus." *Journal of the Acoustical Society of America,*, vol. 123(5), pp. 5687-5690, 2008.

[2] B. Bigi, "Sppas: a tool for the phonetic segmentations of speech," in *The Eighth International Conference on Language Resources and Evaluation*, 2012, pp. 1748–1755.

[3] P. Boersma and D. Weenink. (2022) Praat: doing phonetics by computer [computer program]. version 6.3.03, retrieved 17 december 2022 from http://www.praat.org/.

[4] F. Schiel, "Automatic phonetic transcription of non-prompted speech," in *Proc. of the ICPhS 1999. San Francisco, August 1999*, 1999, pp. 607–610.

[5] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[6] K. Gorman, J. Howell, and M. Wagner, "Prosodylab-aligner: A tool for forced alignment of laboratory speech," *Canadian Acoustics*, vol. 39, no. 3, pp. 192–193, 2011.

[7] R. Fromont and J. Hay, "Labb-cat: An annotation store," in *Proceedings of the Australasian Language Technology Association Workshop 2012*, 2012, pp. 113–117.

[8] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[9] S. Gonzalez, J. Grama, and C. E. Travis, "Comparing the performance of forced aligners used in sociophonetic research," *Linguistics Vanguard*, vol. 1, no. open-issue, 2020.

[10] P. Meer, "Automatic alignment for new englishes: Applying state-of-the-art aligners to trinidadian english," *The Journal of the Acoustical Society of America*, vol. 147, no. 4, pp. 2283–2294, 2020.

[11] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 368–373.

[12] S. Coats, "Dialect corpora from youtube," *Language and Linguistics in a Complex World*, vol. 32, pp. 79–102, 2023.

[13] W. Labov, "Some principles of linguistic methodology," *Language in Society*, pp. 97–120, 1972.

[14] D. yt dlp, "yt-dlp," https://github.com/yt-dlp/yt-dlp, 2022.

[15] F. Developers, "ffmpeg tool (version be1d324)[software]," *URL: http://ffmpeg. org*, 2021.

[16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: http://www.R-project.org/

[17] D. Hirst, "The analysis by synthesis of speech melody: from data to models," *Journal of Speech Sciences*, vol. 1, pp. 55–83, 01 2011.

[18] ——, "A praat plugin for momel and intsint with improved algorithms for modelling and coding intonation," in *Proceedings of the XVIth International Conference of Phonetic Sciences*, vol. 12331236, 2007, pp. 1223–1236.

[19] J. Wells, *Longman pronunciation dictionary*. London: Pearson Longman, 2008.

[20] R. Weide, "CMU Pronouncing Dictionary," 1994. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[21] D. Deterding, "The formants of monophthong vowels in Standard Southern British English pronunciation," *Journal of the International Phonetic Association*, pp. 47–55, 1997.

[22] B. R. Hannisdal, *Variability and change in Received Pronunciation: A study of six phonological variables in the speech of television newsreaders*. [Doctoral dissertation, The University of Bergen], 2006.

[23] G. Lindsey, *English after RP: Standard British pronunciation today*. Palgrave, 2019.

[24] B. Collins and I. Mees, *Practical phonetics and phonology: A resource book for students*, ser. Routledge English language introductions. Routledge, 2013. [Online]. Available: https://books.google.de/books?id=faVJTQIw9eQC

[1] https://youglish.com/