Methods for Phonetic Scraping of Youtube Videos

Adrien Méli [1], Steven Coats [2], Nicolas Ballier [1 3]
[1] CLILLAC-ARP / [3] LLF / Université Paris Cité, F-75013 Paris, France
[2] Faculty of Humanities, University of Oulu, FI-90014, Finland
adrienmeli@gmail.com, nicolas.ballier@u-paris.fr, steven.coats@oulu.fi

Abstract

This paper discusses two pipelines for the automatic collection of automatic speech recognition (ASR) transcripts and audio content from YouTube videos and subsequent phonetic analysis: PEASYV (Phonetic Extraction and Alignment of Subtitled YouTube Videos) and YTPP (YouTube Phonetics Pipeline). The pipelines differ somewhat in terms of processing steps as well as the tools used for forced alignment, but produce comparable results. The two pipelines may be useful for large-scale collection of acoustic data for phonetic analysis.

1 Introduction

Widespread availability of high-quality audio and rapid advances in the quality of ASR transcripts have opened new doors for data collection in phonetics. This paper presents two systems designed to collect transcript and audio data from YouTube for the purposes of phonetic forced alignment and analysis: PEASYV (Phonetic Extraction and Alignment of Subtitled YouTube Videos) and YTPP (YouTube Phonetics Pipeline). The pipelines make use of open-source libraries collect data from YouTube, align the transcripts with the audio tracks, and analyze the acoustic data therein. While both pipelines make use of yt-dlp for data collection, PEASYV aligns audio with text by means of the Penn Forced Aligner (p2f) and SPPAS (SPeech Phonetization Alignment and Syllabification, Bigi 2012), and YTPP uses the Montreal Forced Aligner (McAuliffe et al., 2017a). For Acoustic analysis, for example of F1 and F2 formant values, both pipelines ultimately use Praat (Boersma and Weenink, 2023). YTPP is Python-based and its code is available (see Section 4 below).

The rest of the paper is structured as follows. Section 2 discusses a few papers in which forced aligners are compared. Section 3 provides details on PEASYV, and Section 4 introduces YTPP. In Sections 3 and 4, as proof of concept, we demonstrate analyses of an example YouTube video using the two pipelines. Section 5 provides a brief summary and future outlook, including caveats that may be relevant for the automatic harvesting of phonetic data from YouTube and other platforms.

2 Forced aligner comparisons

Forced alignment of speech, or the exact matching of an audio transcript with an audio file, is a necessary prerequisite for the phonetic analysis of acoustic segments such as phrases, words, or phones. A number of software tools have been developed for forced alignment, for example the Munich Automatic Segmentation System (MAUS), which has a web-based implementation (Kisler et al., 2017). Many are based on HTK, the Hidden Markov Model Toolkit (Young et al., 2006), or Kaldi (Povey et al., 2011). The Penn Forced Aligner is based on HTK, while the Montreal Forced Aligner builds on Kaldi. The SPPAS aligner is derived from Julius (Lee and Kawahara, 2019).

MacKenzie and Turton (2020) compared alignments for British English speech produced by composite tools that build upon HTK and Kaldi: They found that while both underlying algorithms produce acceptable alignments, the Montreal Forced Aligner (built upon Kaldi) performed somewhat better than the Penn Forced Aligner (built upon HTK). Similarly, Gonzalez et al. (2020) compared several aligners for Australian speech, finding them to be suitable even when using default models trained on American English. They found a Kaldi-based aligner to be slightly better than HTK-based aligners.

3 PEASYV: Phonetic Extraction and Alignment of Subtitled YouTube Videos

PEASYV is a modular tool for phonetic analysis of YouTube content. The workflow of the tool is automatically managed by shell scripts providing the

sequence of commands described in Figure 1. Subtitled videos are scraped by `yt-dlp`. The downloaded video is then converted to a `wav` file using `ffmpeg`, and the subtitles file is converted to a preliminary TextGrid using `praat` (Boersma and Weenink, 2023). The time stamps from the subtitles serve as boundaries for the TextGrid, and the created intervals are labeled with the subtitles themselves. The sound file and the TextGrid are then split into short files extracted from the intervals. These short sound files, usually lasting under three seconds, are then fed into two forced alignment tools, SPPAS (Bigi, 2012) and the Penn Phonetics Lab Forced Aligner (P2FA, p2f). Both aligners use the Carnegie Mellon University dictionary (CMU, Weide 1994) for grapheme to phoneme correspondences[1]. This procedure contains potential cascading alignment errors and increases accuracy. The resulting short TextGrids are then concatenated back into the main TextGrid, and syllabic tiers, one for each aligner, are created following the syllabification of the *Longman Pronunciation Dictionary* (LPD, Wells 2008). Extra steps are taken regarding prosodic annotation but their description falls beyond the scope of this article (*cf.* Méli and Ballier 2023 for further details). The resulting main TextGrid features segmental, syllabic, and lexical tiers for both aligners, and a Momel (Hirst and Espesser, 1993; Hirst, 2007) and INTSINT tier for SPPAS;[2] two "matching" tiers have also been added (see below). Finally, vocalic data is collected in separate `csv` spreadsheets, one for each aligner.

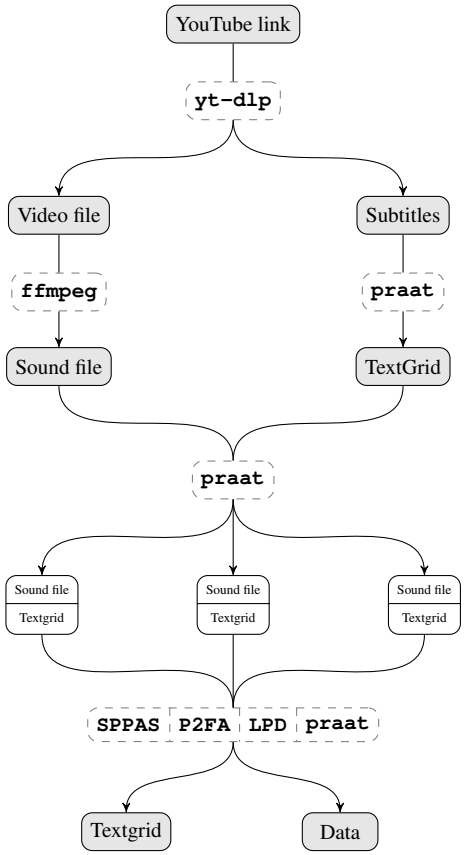


Figure 1: The PEASYV workflow.

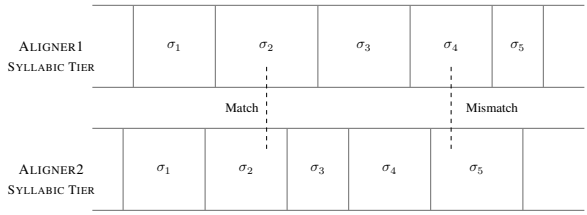


Figure 2: Schematic representations of a "MATCH" (left) and "MISMATCH" (right) case on a PRAAT TextGrid.

Because PEASYV uses two aligners based on two different speech recognition engines (Julius and HTK), assessing the degree of agreement of the generated alignments may arguably provide some insight into their reliability, if not their accuracy. This can be done by comparing local discrepancies and measuring the frequencies of these discrepancies. One way to do this is by flagging vocalic datapoints on a given aligner according to whether the other aligner matches these datapoints. PEASYV implements one such system, and its characteristics are represented in Figure 2. PEASYV uses the LPD-based syllabic tiers as references. The midpoint of σ_1's duration on Aligner1's tier, marked by a vertical dashed line, falls within the boundaries of σ_1's duration on Aligner2's tier. When collecting the phonetic data (*e.g.* formants) corresponding to σ_1 as aligned by Aligner1, the vowel will be marked as "matching". Conversely, σ_4's midpoint on Aligner1's tier, marked by the dashed line on the right, falls outside σ_4's interval on Aligner2's tier: it will therefore be marked as "mismatching".

This experimental feature makes it possible to filter out potential alignments errors and obtain more reliable measurements, especially for sizeable datasets. In contrast with other forced aligners, PEASYV also enables direct comparisons, on the same TextGrid, of two aligners, and provides syl-

[1] The transcriptions of the CMU are however different: SPPAS uses a version of SAMPA, P2FA ARPAbet.

[2] "Momel" stands for "Modelling melody", "INTSINT" for "INternational Transcription System for INTonation".

labic tiers for future analyses.

3.1 Results

Table 1 presents the total number of vowels aligned by SPPAS and P2FA respectively. 27.4% of all 2661 SPPAS-aligned vowels appear in syllables whose mid-temporal values are not included within the corresponding P2FA-generated intervals (*i.e.* they are flagged as "mismatching"). 30.8% of the 2743 P2FA-aligned vowels are "mismatching".

	SPPAS	P2FA
Vowels:	2661	2743
– in matching syllables:	1933	1899
– in mismatching syllables:	728	844

Table 1: Per-aligner counts of vowels.

The PEASYV-generated vocalic spaces for monophthongs in a video chosen for test purposes with the identifier _P7_69FeqnU are represented in Figure 3. The formant values of each monophtong are plotted in the F1/F2 space. The ellipses encompass the values within one standard deviation of all the measurements for each monophthong. The label of each monophthong is located at the center of each value (*i.e.* the mean F1/F2 values of the vowel's measurements), and the number next to it gives the number of tokens detected for that vowel. The top row (*i.e.* Figures 3a and 3b) features all monophthongs, while the bottom row (*i.e.* Figures 3c and 3d) only features matching monophthongs (*cf.* previous section and Figure 2).

3.2 Discussion and Prospects

Cursory visual inspection of Figure 3 shows that restricting the data to matching cases yields ellipses which are more clearly defined and less overlapping than using all vowels, regardless of whether their alignment on a given aligner matches that of the other aligner. This is particularly clear with SPPAS-aligned mid front vowels and back vowels. One striking characteristic is the great variation that formant measurements for vowel /uː/ undergo compared to its token count. We contend that the matching procedure may be a simple way to filter out outliers and improve the quality of the extracted data, although no ground truth alignment has been prepared. Of course, the quality of PEASYV-generated data is highly dependent on the original quality of the subtitles. Future research will have to establish whether transcriptions based on automatic speech recognition systems such as Whisper yield more reliable data.

PEASYV is meant to be deployed on a website[3] where links to subtitled videos can be uploaded and generated TextGrids can be downloaded. The source code may also be made partially available for deployment on Linux servers. PEASYV will hopefully be useful to study less common varieties of English. Corpora of Nigerian and Ugandan English are under way.

4 YTPP: YouTube Phonetics Pipeline

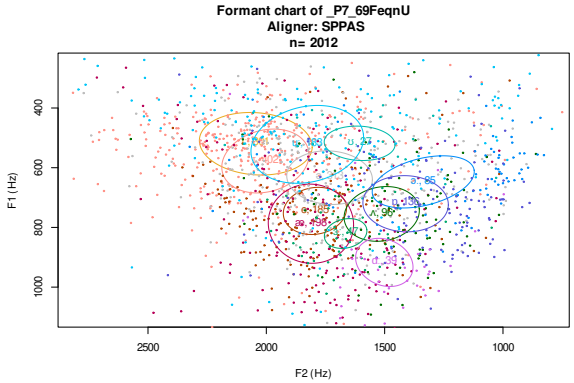
The YouTube Phonetics Pipeline is a Python-based series of scripts for the automatic extraction of audio (or video) content from YouTube and other streaming services. Its main characteristics are described in Coats (2023c). Like PEASYV, YTPP makes use of the open-source yt-dlp library for harvesting YouTube's automatic speech recognition transcripts and audiovisual content; transcripts are then aligned using the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017a). The output from the aligner, in the TextGrid format, is then sent to Parselmouth-Praat (Jadoul et al., 2018; Boersma and Weenink, 2023), a Python port of functions from Praat. This approach allows for the automated analysis of vowel formants, pitch, prosody, or other acoustic parameters within the functionality of a Jupyter notebook. The basic methods of YTPP are available in a Colab environment.[4] Because YTPP is developed in a Jupyter environment, it is fully modifiable, and data can be analyzed statistically or visualized for exploratory analysis with widely used libraries, according to user needs. Transcript data for several publicly available corpora has been collected using the basic approach employed by YTPP (Coats, 2023a).

YTTP was used to extract F1 and F2 formant values for monophthongs from the YouTube test video noted above in Section 3. Figure 4 depicts the vowel space for the video _P7_69FeqnU, entitled "Sentence Stress and Intonation in English" from the *Pronunciation with Emma* channel, using an acoustic models trained on UK English, a pronunciation dictionary for UK English, and a phoneset meant to represent UK English.[5] As in Figure 3,

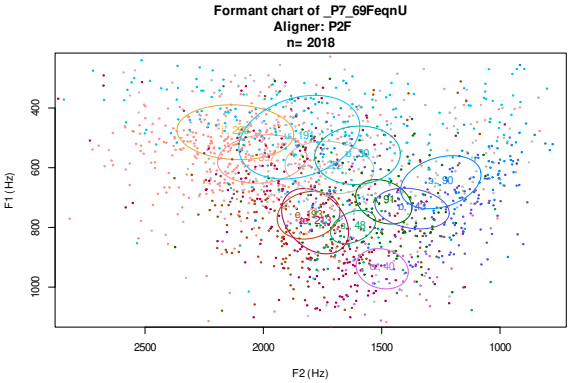
[3] Current information is at https://www.adrienmeli.xyz/peasyv.html

[4] https://github.com/stcoats/phonetics_pipeline

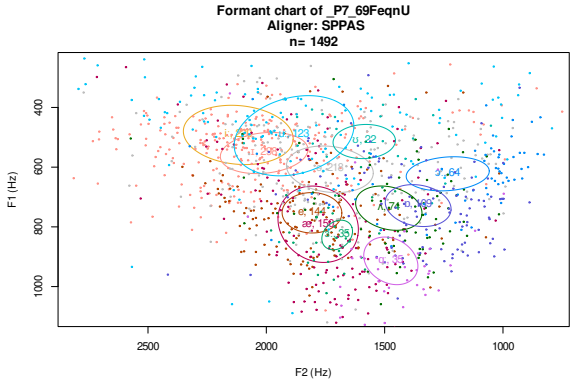
[5] English (UK) MFA dictionary v2.2.1; English MFA acoustic model v2.2.1, https://mfa-models.



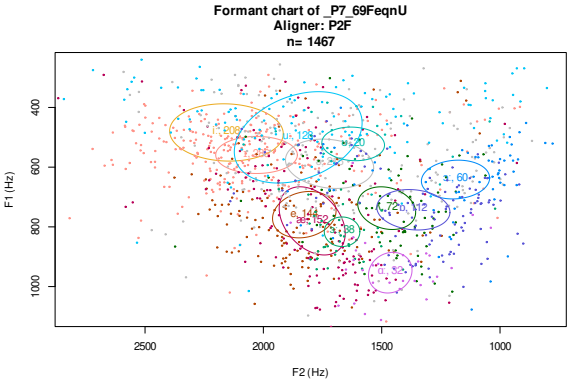
(a) SPPAS-aligned tokens.



(b) P2F-aligned tokens.



(c) Matching SPPAS-aligned tokens.



(d) Matching P2FA-aligned tokens.

Figure 3: PEASYV flowcharts of video _P7_69FeqnU.

the centers of the circles represent the mean measurement values for the monophthong vowels in F1/F2 formant space and the ellipses values within one standard deviation of the mean values; the IPA symbol for each vowel is followed by the number of vowel tokens detected by the aligner in the video.[6]

Figure 4 differs somewhat from Figure 3, not only due to different plotting software being employed, but also due to differences between the acoustic models and phonemic representations in the three systems under consideration. Nevertheless, the figures suggest that the speaker in the video, as the name of her channel suggests, has vowels that correspond to standard English pronunciation norms. Future work may undertake more careful comparison of these (and other aligners) by controlling for the acoustic models employed by the different algorithms and the underlying graphemic representations.

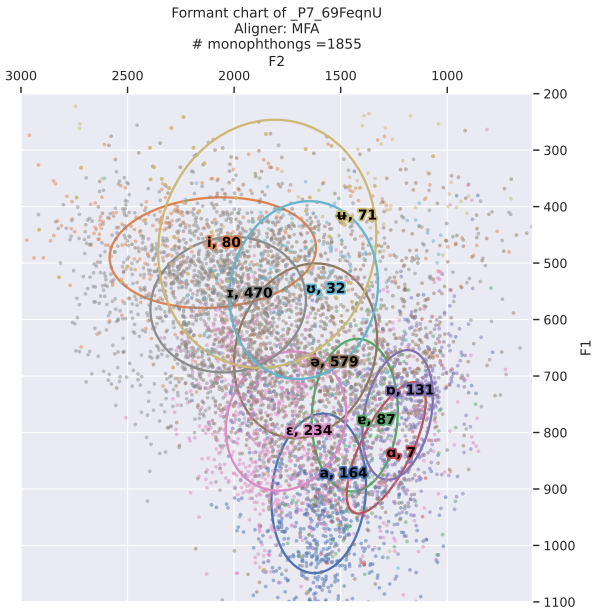


Figure 4: YTPP formant chart for the video _P7_69FeqnU

readthedocs.io/en/latest/acoustic/index.html. MFA's functionality includes a variety of acoustic models, dictionaries, phonesets, and other options.

[6] In this example, the script has set the number of measurements per phone at 9, at equally spaced intervals within the total duration of the phone, but formant intensity could not be registered at all measurement intervals due to acoustic quality. The number of measurements per phone can be changed in the script.

Python plotting functionality can also be used to generate Praat-style charts of sound intensity and frequency, as in Figure 5.

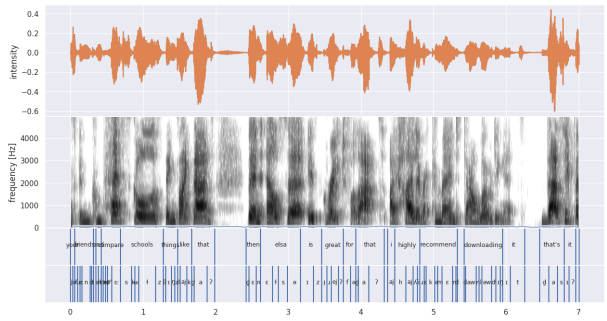


Figure 5: Sound intensity and frequency for an excerpt of _P7_69FeqnU

5 Discussion, caveats, and outlook

No longer must the phonetician travel to distant locales with a tape recorder and painstakingly interview informants: both PEASYV and YTTP offer researchers in phonetics and acoustic analysis the means for the automatic and extraction and analysis of hundreds or thousands of hours of speech.

PEASYV output grids include the results of two aligners: the overlap method described above may help to identify and extract segments more accurately, especially for audio files with acoustic background noise. PEASYV also includes syllabification information, making it potentially useful for automated studies of lexical stress patterns or other prosodic features.

YTPP utilizes the MFA aligner, which is more recent and possibly more accurate than HTK- or Julius-based aligners (see the citations above). In addition, YTPP is available and can already be used "out-of-the-box" for data collection and analysis tasks. Its code is fully available and customizable.

The pipelines both offer the means to collect and analyze online speech recordings, but two considerations should be noted pertaining to the accuracy of ASR transcripts and the legal contexts in which online data collection can be undertaken.

5.1 ASR Accuracy

While ASR has made great advances in recent years, many ASR transcripts of videos on YouTube (and other platforms) contain errors due to issues such as poor audio quality, out-of-vocabulary lexical items, or strongly accented speech not accounted for in the training data. Despite this, given sufficient quantities of data, transcript errors in phonetic analysis pipelines such as PEASYV and YTPP may tend to cancel each other out: Coto-Solano (2022), for example, found that even pipelines that utilize error-ridden transcripts are generally able to accurately capture the formant values of a given speaker.

5.2 Legal context

While content from YouTube and other streaming platforms is generally owned by the content creator and/or the platform, use of copyrighted content for non-profit purposes such as academic research is generally permitted in most jurisdictions. In the US, for example, the "Fair Use" provisions of copyright law (U.S.C. Title 17, § 107) permit re-use of copyrighted material for research purposes; other Anglophone jurisdictions have similar laws.

In the EU, Directive 2019/790 of the European Parliament and of the Council instructed member states to pass legislation allowing the re-use of copyrighted content for purposes of scientific research or teaching; the directive has since been implemented by most member state legislatures (see also the discussion in Coats).

We expect that legislation will continue to permit fair and reasonable use of copyrighted materials for non-profit research purposes and that researchers who follow the appropriate ethical guidelines will be able to make use of PEASYV and YTPP for data collection.

5.3 Outlook

A paradigm shift in data collection and analysis practices in the language sciences is underway, and PEASYV and YTPP represent potentially valuable tools for researchers in a wide variety of linguistic subfields. Future work with the pipelines may include, as noted above, more detailed comparison of aligners and of outputs; the development of interoperability with other data formats (for example, PolyglotDB McAuliffe et al. 2017b, an SQL-based system with a Python API for the organization of speech data and alignments); and the creation of searchable online databases that include aligned audio content. In a broader perspective, it is hoped that the tools will help researchers to collect and study the rich acoustic variation of the speech signal.

References

Brigitte Bigi. 2012. SPPAS: a tool for the phonetic segmentations of speech. In *The Eighth International Conference on Language Resources and Evaluation*, pages 1748–1755.

Paul Boersma and David Weenink. 2023. Praat: doing phonetics by computer [Computer program]. Version 6.3.10, retrieved 1 June 2023 http://www.praat.org/.

Steven Coats. 2023a. Dialect corpora from Youtube. In *Language and linguistics in a complex world*, pages 79–102. De Gruyter.

Steven Coats. 2023b. A new corpus of geolocated ASR transcripts from Germany. *Language Resources and Evaluation*.

Steven Coats. 2023c. A pipeline for the large-scale acoustic analysis of streamed content. In *Proceedings of the 10th International Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2023)*, page 51–54. Mannheim: Leibniz-Institut für Deutsche Sprache.

Rolando Coto-Solano. 2022. Evaluating word embeddings in extremely under-resourced languages: A case study in Bribri. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4455–4467, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Simon Gonzalez, James Grama, and Catherine E Travis. 2020. Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard*, 6(1):20190058.

Daniel Hirst and Robert Espesser. 1993. Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix*, pages 75–85.

Daniel J Hirst. 2007. 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*, pages 1233–1236.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347.

Akinobu Lee and Tatsuya Kawahara. 2019. julius-speech/julius: Release 4.5.

Laurel MacKenzie and Danielle Turton. 2020. Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard*, 6(s1):20180061.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017a. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Michael McAuliffe, Elias Stengel-Eskin, Michaela Socolof, and Morgan Sonderegger. 2017b. Polyglot and speech corpus tools: A system for representing, integrating, and querying speech corpora. In *INTERSPEECH*, pages 3887–3891.

Adrien Méli and Nicolas Ballier. 2023. PEASYV: A procedure to obtain phonetic data from subtitled videos. *Proceedings of the International Congress of Phonetic Sciences*, pages 3211 – 3215.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Robert L. Weide. 1994. The CMU Pronouncing Dictionary. *http://www.speech.cs.cmu.edu/cgi-bin/cmudict*.

John C. Wells. 2008. *Longman Pronunciation Dictionary*. Pearson Longman, London.

Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. 2006. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK.