EVALUATING FULL AUTOMATION OF FORMANT EXTRACTION IN THE GERMAN PLAPPER CORPUS

Robert Fromont[1], *Jennifer Hay*[1,2], *Daniel Duran*[3], *Allie Osborne*[1,2], *Melanie Weirich*[4], *Miriam Oschkinat*[5], *Stefanie Jannedy*[5]

[1]*New Zealand Institute of Language, Brain and Behaviour, University of Canterbury,*
[2]*Department of Linguistics, University of Canterbury,* [3]*Bielefeld University, Germany,*
[4]*Friedrich-Schiller-University Jena, Germany,*[5]*Leibniz-Centre General Linguistics (ZAS Berlin)*
robert.fromont@canterbury.ac.nz

Abstract: We have collected over thirteen thousand recordings of German sentence readings elicited via the *Plapper* app as part of a citizen science project. However, manual correction of transcripts and phone alignments is a bottleneck for sociolinguistic study of these recordings, with less than 10% of recordings processed so far, as data collection continues. We evaluate the possibility of fully automating the transcription and alignment of the *Plapper* corpus. Two automated transcription processes were tried and their target-word error rates compared; BAS ASR Web Service, and using the text prompt read by the participants as the transcript. Furthermore two forced alignment systems were evaluated: *WebMAUS* and the *Montreal Forced Aligner*. Their outputs were compared with manually corrected alignments, using both overlap rate to compare them temporally, and correlation of resulting formant measurements. We found that using prompt text as a transcript is more accurate than ASR, and that both forced aligners produced sufficiently trustworthy alignments, with resulting formant measurements that correlated highly with those generated from manually corrected alignments.

1 Introduction

The increasing ubiquity of high quality mobile devices presents an opportunity for sociophonetics research to collect large volumes of speech, via apps designed to facilitate contributions to citizen science projects. The *Plapper* project [1, 2] is one such project. The *Plapper* project collects, among other things, readings of German sentences.

To prepare for analysis, Automatic Speech Recognition (ASR) tools and forced alignment systems can facilitate a first pass over the data to obtain approximate transcripts and phone alignments. Although current practice generally prescribes a second pass to manually correct transcripts and alignments, this manual intervention creates a bottleneck when the volume of data is large, as data collection quickly outpaces manual correction.

Coto-Solano [3] has suggested that recent improvements in these technologies may render manual correction unnecessary. Indeed, software to fully automate this process for US English has been in development for over a decade [4, 5].

We seek to determine whether manual correction is necessary for the *Plapper* project, specifically:

1. Do automated transcripts have low enough error rates and
2. Do uncorrected vowel forced alignments provide sufficiently accurate formant measurements to obviate manual correction?

2 Data

The *Plapper* corpus consists of an ever-growing collection of short recordings elicited from German speakers using the *Plapper* app installed on their own devices. Participants read sentences designed to cover Standard German vowel and consonant contrasts. All target vowels are contained in nouns, carry stress and occur in accented positions in the sentence.

To date, the corpus contains 13,509 recordings of 2,207 speakers, with a total duration of 31 hours and 39 minutes. The mean sentence recording length is 8.4 seconds.

The resulting recordings are uploaded anonymously to cloud servers in Germany. A first-pass transcript has been obtained using BAS Web Services' ASR [6, 7]. For a subset of the recordings, transcripts have been manually corrected, with particular attention to nouns containing the target vowels. Then force-alignment was carried out using the BAS Web Services' *WebMAUS*, and the resulting alignments were laboriously hand corrected, with particular attention to target vowels.

We have used 1,173 manually corrected recordings of 348 speakers with a total duration of 2 hours and 37 minutes, to evaluate the impact of fully automating the transcription and alignment process.

3 Method

Our methods and metrics for evaluating automatic transcription and alignment are explained in the following two sections. We evaluated two options for each automated process.

3.1 Transcription

The BAS ASR Web Service was already being used for first-pass transcription to bootstrap the manual transcription process, so evaluation of the accuracy of the ASR transcripts was simply a matter of comparing the original ASR texts with the corrected versions of the transcripts.

However, these recordings are short and have very predictable content; the participant is shown a written sentence by the app, and they have to read it aloud while being recorded. There is one sentence per recording, and we know which sentence it was. Thus a second possible mechanism for transcribing the recordings exists: just assume the participant faithfully read the text prompt they were shown, and use that text as the transcript.

ASR transcripts inevitably contain errors in relation to what was spoken, although recent technological advances have diminished error rates significantly. The text prompt is also likely to contain errors, as participants can misread the text, insert or repeat words, etc. The questions for us are:

i are there few enough errors for us to simply trust the transcripts without manually checking them, and

ii which automated transcription method (ASR vs. use of transcripts (both without any manual correction)) produces lower error rates?

3.1.1 Evaluation Metric

One of the standard measures for transcript accuracy is Word Error Rate (WER), computed with the following formula:

$$WER = 100 \times \frac{S+D+I}{N}$$

Where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of words in the reference (human) transcript.

We used a modified version of the Wagner & Fischer algorithm [8] to compute an edit path between manual and automatic transcript words, in order to count substitutions, deletions, and insertions. However, for our purposes here, there are two nuances.

Firstly, in order to ensure keywords are reliably matched, the only 'substitutions' our implementation allowed were words that were spelled the same but the case of some letters were different (e.g. "Bären" (*bears*) vs. "**b**ären"). Such cases are not errors for our purposes, as the keyword is still identifiable, so our error rates exclude S.

Secondly, the manual reference transcripts were edited versions of the BAS ASR Web Service's transcripts, and correction was focussed on the keywords containing target vowels. However correction of *non-keywords* was not prioritised, and so many of these words in the reference transcripts are unchanged even though they don't correspond to the words the participant uttered. In order to avoid a resulting bias toward ASR transcripts when comparing error rates with prompt-text transcripts, we exclude non-keywords, and compute error rates using keyword deletions/insertions only.

3.2 Alignment

Various forced alignment systems have been evaluated in the literature (e.g [9, 10]). In addition to the BAS *WebMAUS* service, the *Montreal Forced Aligner* (MFA) [11] is consistently reported to have high accuracy ratings. Both systems are easy to deploy using our corpus management software [12], so we evaluate both *WebMAUS* and MFA to determine whether either or both systems may produce phone alignments that are accurate enough to not require manual inspection and adjustment.

The alignments are compared temporally with our manually corrected segments, but we are ultimately interested in the impact inaccuracies might have on the acoustic measurements we will use for our sociophonetic research; i.e. F1 and F2 measurements of the target vowels.

As with transcription, manual phone alignment correction was focussed on target vowels, so our comparisons are restricted to target vowels only.

So we are evaluating four possible fully-automated workflows:

1. using BAS ASR Web Service for transcription and:
   (a) *WebMAUS* for forced alignment or
   (b) MFA for forced alignment, or
2. using the prompt text as a transcript and:
   (c) *WebMAUS* for forced alignment or
   (d) MFA for forced alignment.

3.2.1 Evaluation Metrics

To evaluate automatic alignments temporally, we use overlap rate (OvR) [13], which is a measure of how much two intervals overlap, independent of their absolute durations. OvR is a value between 0 (no overlap at all) and 1, (perfect overlap – the same start and end times). OvR is calculated as follows:

$$OvR = \frac{CommonDur}{DurMax} = \frac{CommonDur}{DurRef + DurAuto - CommonDur}$$

Where *CommonDur* is the duration in common between the automatically aligned and manually aligned segments, $DurRef$ is the duration of the manually aligned segment, and $DurAuto$ is the duration of the automatically aligned segment. $DurMax$ is the maximum duration of the sound file covered by the pair of segments.

OvR is a number that relates to a single pair of intervals, so each phone from the manual alignment must be paired up with the corresponding phone in each automatic alignment to determine the OvR of that phone. Again we use a modified version of the Wagner & Fischer algorithm to pair manually aligned phones with their automatically aligned counterparts. As with words, there may be insertions and deletions, as the dictionaries used by forced aligners can include multiple possible pronunciations per word, and the choice it makes may differ from that made by a human annotator. For example, "Beeren" (*berries*) might be transcribed /be:rn/ with one vowel in the manual alignments, but with two vowels by MFA: /beːʁən/. Furthermore, as can be seen from these examples, *WebMAUS* output and manual corrections use SAMPA phoneme labels but MFA uses IPA labels; this difference is taken into account by our implementation of the edit path algorithm for phones.

Once phones are paired, OvR is calculated and then the mean OvR across target vowels provides a single measure of temporal accuracy that can be used to evaluate alignments.

To compare acoustic measurements, Praat [14] was used to measure F1 and F2 at the midpoint of each vowel, using a 5,000Hz formant ceiling for male speakers, and 5,500Hz for all others, for each manual and automatic alignment. As it is common practice to remove outliers after such acoustic measurement, we also perform outlier removal, discarding tokens with a measurement more than 2.5 SD from the mean.

We compute Pearson's product-moment correlation between the manual measures and the automatic ones, and also plot vowel space ellipses for visual comparison.

4 Results

4.1 Transcription

Error rates for each transcription method are shown in Table 2, which shows that using the prompt text as a transcript has a lower error rate (1.1%) than using ASR (3.6%), but both error rates are very low.

4.2 Alignment

4.2.1 Overlap Rate

Overlap rates (for target vowels only) are shown in Figure 1, which shows that the inter-quartile ranges are narrower when using MFA for alignment. The mean overlap rates are listed in the first two columns of Table 3, which shows that both aligners achieve a high rate of overlap with manual alignments; *WebMAUS* has an overlap rate of 0.803 when transcribing via ASR and 0.798 when using the prompt text as the transcript. MFA's overlap rates are slightly higher, at 0.829 with ASR transcripts, and 0.828 with prompt transcripts.

4.2.2 Acoustic Measurement

F1 and F2 were measured at the midpoint each alignment interval for 7,758 target vowel tokens that were common across all four conditions. Outliers more than 2.5 standard deviations from the mean F1/F2 values were then discarded, which resulted in 4% of tokens being excluded.

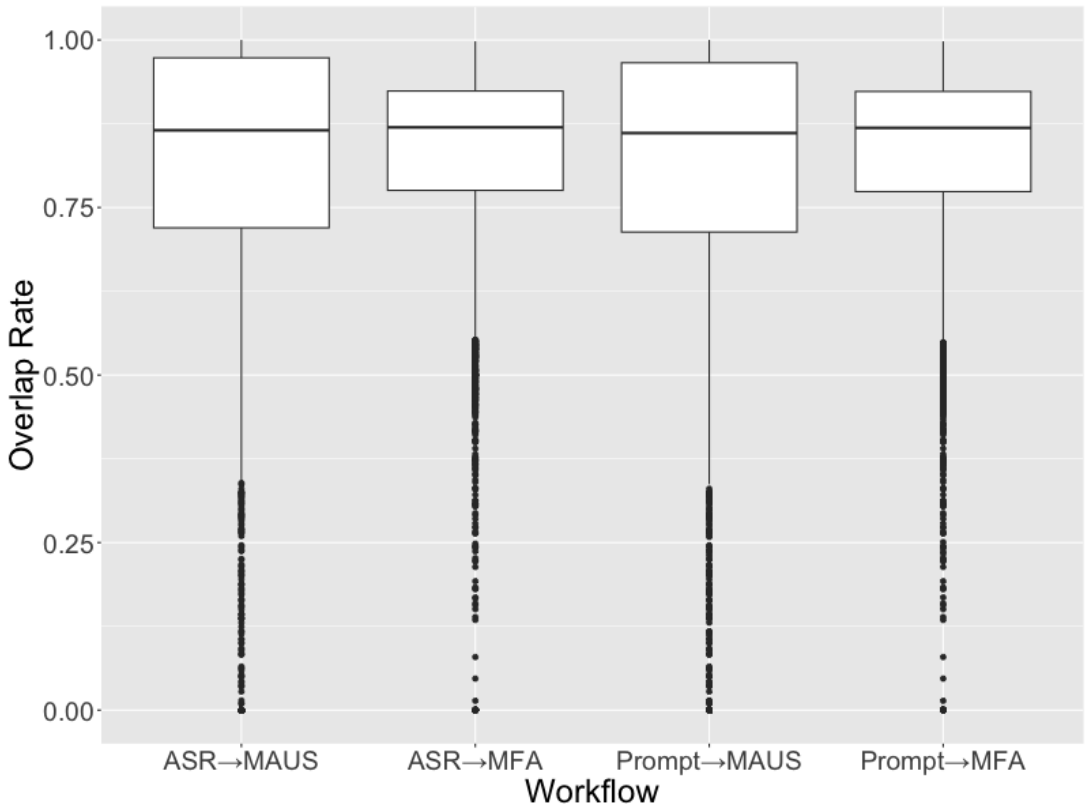


Figure 1 – Target vowel overlap rates for each transcription→alignment workflow

For the remaining formant measurements, Pearson's product-moment correlation was computed for each set of automatic alignments in relation to the manually corrected ones. The results are shown in Table 3. For the F1 measurements, correlations were highest using MFA alignments, with ASR-based transcripts having a 0.961 correlation, and prompt-based transcripts, 0.959. With *WebMAUS* alignments, the correlations were 0.939 and 0.938 respectively. For the F2 measurements, correlations were highest using *WebMAUS* alignments, with prompt-based transcripts having a 0.917 correlation, and ASR-based transcripts, 0.916. With MFA alignments, the correlations were 0.915 and 0.914 respectively.

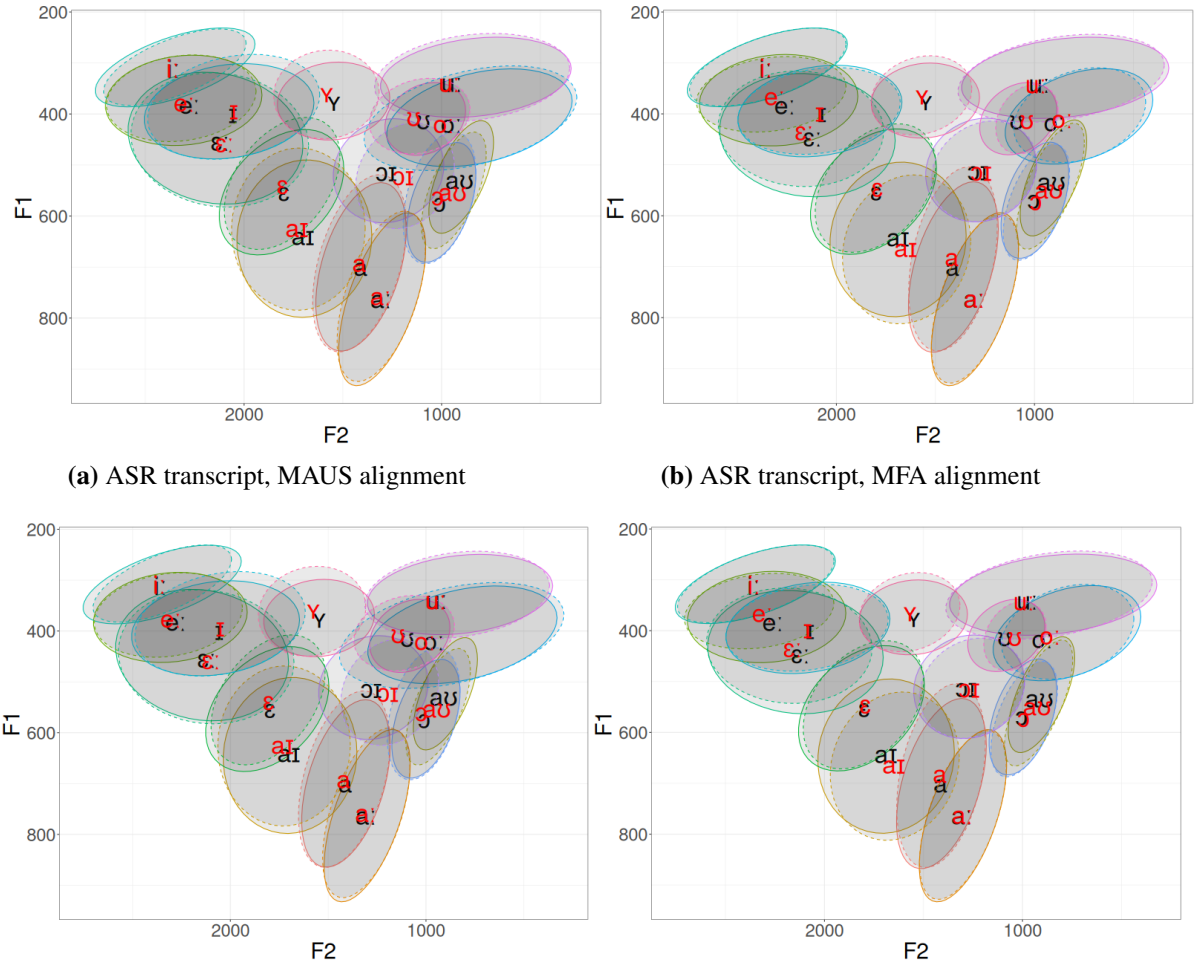
Figure 2 (b) shows vowel-space ellipses for the best correlation for F1; BAS ASR Web Service transcripts with MFA alignments. For F2 the best correlation resulted from prompt text transcripts with *WebMAUS* alignments, the vowel-space ellipses of which are shown in Figure 2 (c). To complete the four workflows tested, Figure 2 (a) shows ellipses for BAS ASR Web Service transcripts with *WebMAUS* alignments, while Figure 2 (d) shows text transcripts with MFA alignments. In all these figures, for each vowel the ellipsis from automatic alignments (marked with dashed lines) is generally very close to that of from the corrected alignments (solid lines). The token counts for each vowel and automated workflow are shown in Table 1

5 Conclusions

For the analysis of vowel formants in the *Plapper* corpus, we conclude that fully automating transcription and forced alignment with no manual correction introduces a level of noise that is acceptable to us given the amount of manual labour otherwise required.

Our speech recordings are short read sentences, and using the text prompt as a transcript is a very simple and accurate method for transcription, but the BAS ASR Web Service also provides acceptably accurate transcripts for recordings of this nature.

Both *WebMAUS* and MFA provide sufficiently accurate vowel alignments, which result



(a) ASR transcript, MAUS alignment

(b) ASR transcript, MFA alignment

(c) Prompt transcript, MAUS alignment

(d) Prompt transcript, MFA alignment

Figure 2 – Vowel space ellipses, showing manually corrected (solid line, black labels) vs automatic alignments with outliers removed (dashed lines, red labels).

Table 1 – Vowel token counts for each automated workflow.

Vowel	ASR→MAUS	ASR→MFA	Prompt→MAUS	Prompt→MFA
a	314	311	312	311
aɪ	798	790	797	790
aʊ	287	291	287	291
aː	1187	1194	1187	1193
eː	307	309	306	309
iː	63	63	63	63
oː	557	533	560	533
uː	797	805	793	805
ɔ	345	341	345	341
ɔɪ	362	361	362	361
ɛ	598	612	598	612
ɛː	745	730	745	731
ɪ	554	550	554	550
ʊ	376	375	377	375
ʏ	162	159	163	159

Table 2 – Keyword Error Rates for each transcription method - C = Correct, D = Deleted, I = Inserted (Substitutions are treated as Correct)

Transcript	C	D	I	Keyword WER
BAS ASR Web Service	8010	177	114	3.6
Prompt Text	8161	26	68	1.1

Table 3 – Mean Overlap Rates, and Pearson's product-moment correlations to corrected-alignment formants, for each automated workflow. Correlations were computed after outliers more than 2.5SD from the mean were removed (p<0.001 in all cases).

Transcript	Mean OvR		Correlation			
			F1		F2	
	MAUS	MFA	MAUS	MFA	MAUS	MFA
BAS ASR Web Service	0.803	0.829	0.939	0.961	0.916	0.915
Prompt Text	0.798	0.828	0.938	0.959	0.917	0.914

in formant measurements that correlate highly with those that would be produced by manual correction.

We recognize that these methods may not be suitable for data of a different nature, e.g. spontaneous speech, longer recordings, other languages, etc. and we have not evaluated consonant alignments and corresponding acoustic measurements. However this approach may speed up the analysis of other similar corpus projects, with high quality recordings of short read sentences in languages supported by these tools, by eliminating the manual-correction bottleneck.

5.1 Authorship and Contributorship Statement

- Robert Fromont: Conceptualization, Methodology, Software, Visualization, Writing, Review & editing
- Jennifer Hay: Conceptualization, Methodology, Software, Visualization, Funding acquisition, Review & editing
- Daniel Duran: Conceptualization, Data Curation, Review & editing
- Allie Osborne: Conceptualization, Methodology, Software, Data Curation, Review & editing
- Melanie Weirich: Conceptualization, Funding acquisition, Review & editing
- Miriam Oschkinat: Conceptualization, Data Curation, Review & editing
- Stefanie Jannedy: Conceptualization, Funding acquisition, Writing, Review & editing

We gratefully acknowledge the funding by the *German Research Foundation* to the Collaborative Research Center CRC1412-*Register*, project C02 to Jannedy & Weirich, grant # 416591334.

References

[1] JANNEDY, S.: *Plappern für die Wissenschaft – Eine App gegen die Datenlücke. Bericht über das Forschungsjahr 2020/2021 : ZAS ; Impressionen*, pp. 115–126, 2022. URL https://www.leibniz-zas.de/fileadmin/media/Dokumente/Jahresberichte/JB2020_21_Jannedy.pdf.

[2] WEIRICH, M., D. DURAN, and S. JANNEDY: *Gender and age based f0-variation in the German Plapper Corpus*. In *Interspeech 2024*, pp. 1565–1569. ISCA, 2024. doi:10.21437/Interspeech.2024-1592.

[3] COTO-SOLANO, R.: *Computational sociophonetics using automatic speech recognition*. *Language and Linguistics Compass*, 16(9), p. e12474, 2022. doi:10.1111/lnc3.12474.

[4] REDDY, S. and J. STANFORD: *Toward completely automated vowel extraction: Introducing darla*. *Linguistics Vanguard*, 1, 2015. doi:10.1515/lingvan-2015-0002.

[5] LIU, H., B. MACWHINNEY, D. FROMM, and A. LANZI: *Automation of language sample analysis*. *Journal of Speech, Language, and Hearing Research*, 66(7), pp. 2421–2433, 2023. doi:10.1044/2023_JSLHR-22-00642.

[6] KISLER, T., U. REICHEL, F. SCHIEL, C. DRAXLER, B. JACKL, and N. PÖRNER: *BAS Speech Science Web Services - an Update of Current Developments*. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, and S. PIPERIDIS (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3880–3885. European Language Resources Association (ELRA), Portorož, Slovenia, 2016. URL `https://aclanthology.org/L16-1614/`.

[7] KISLER, T., U. REICHEL, and F. SCHIEL: *Multilingual processing of speech via web services*. *Computer Speech & Language*, 45, pp. 326–347, 2017. doi:10.1016/j.csl.2017.01.005.

[8] WAGNER, R. A. and M. J. FISCHER: *The string-to-string correction problem*. *J. ACM*, 21(1), p. 168–173, 1974. doi:10.1145/321796.321811.

[9] GONZALEZ, S., J. GRAMA, and C. E. TRAVIS: *Comparing the performance of forced aligners used in sociophonetic research*. *Linguistics Vanguard*, 6(1), p. 20190058, 2020. doi:10.1515/lingvan-2019-0058.

[10] MACKENZIE, L. and D. TURTON: *Assessing the accuracy of existing forced alignment software on varieties of British English*. *Linguistics Vanguard*, 6(s1), p. 20180061, 2020. doi:10.1515/lingvan-2018-0061.

[11] MCAULIFFE, M., M. SOCOLOF, S. MIHUC, M. WAGNER, and M. SONDEREGGER: *Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi*. In *Proc. Interspeech 2017*, pp. 498–502. 2017. doi:10.21437/Interspeech.2017-1386.

[12] FROMONT, R. and J. HAY: *LaBB-CAT: an Annotation Store*. In *Proceedings of Australasian Language Technology Association Workshop*, pp. 113–117. Australasian Language Technology Association, 2012.

[13] PAULO, S. and L. C. OLIVEIRA: *Automatic Phonetic Alignment and Its Confidence Measures*. In J. L. VICEDO, P. MARTÍNEZ-BARCO, R. MUŃOZ, and M. SAIZ NOEDA (eds.), *Advances in Natural Language Processing*, pp. 36–44. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. doi:10.1007/978-3-540-30228-5_4.

[14] BOERSMA, P. and D. WEENINK: *PRAAT, a system for doing phonetics by computer*. *Glot international*, 5, pp. 341–345, 2001.