Autophon user guide
Swedish
Model: SweFA version 2.0

1 What is forced alignment?

Forced alignment (FA) refers to the automatic process by which speech recordings are phonetically time-stamped with the help of Hidden Markov models or Deep Neural Networks. Autophon uses the latter by means of the *Montreal Forced Aligner* (McAuliffe, Socolof, Mihuc, Wagner, and Sonderegger 2017). The software outputs a time-stamped phonetic annotation, readable in Praat (Boersma and Weenink 2017), that is based on an optimization of two user inputs: (1) the speech recording and (2) a corresponding orthographic transcription. For an FA tool to work for a particular language, an acoustic model must be trained and an accompanying pronunciation lexicon must be built that covers every word in the language. FA is important because it automates something that is resource-intensive when done manually. A typical phonetic annotation can take between 250 and 400 minutes per recorded minute. In a place like Scandinavia – where labor costs are high – this cost has presented a barrier for linguists.

2 How to cite

Any study that makes use of Autophon Swedish should cite the following sources:

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer [Computer software], Version 6.0.36. http://www.praat.org/

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of Interspeech*, 498–502.

Young, N. J. (2019b). SweFA 2.0 – Forced Alignment of Swedish, version 2.0. www.autophon.se

Young, N. J. (2023). Autophon – Automatic phonetic annotation of Nordic languages (web application). www.autophon.se

3 Instructions

Logging in Creating an account and logging in are relatively intuitive. You must verify your email address before you can use Autophon. Wait at least 1 hour for your verification email.

Cost Autophon is free of charge to the public.

Uploading files Files can be uploaded individually, as a group, or in a zip file. Every transcription file must have a corresponding audio file by the same name, e.g., Micke_spontan.TextGrid and Micke_spontan.wav. You can upload/align a maximum of five files, totaling 350 MB, at a time. The following formats are supported:

> **Transcription formats:** Transcription files can be made in Praat (**.TextGrid**) (soon also ELAN **.eaf**) and must only have *one tier each*. Multiple-tier transcriptions will be rejected. If you have a dialogue, extract each tier as a separate file, duplicate the sound file, and give it corresponding names to each extracted tier. TextGrid encoding needs to be either **UTF-8/Unix** or **UTF-16/Windows CRLF**. If you have older Praat files from the mid to early 2000s, they may produce a yellow error box, in which case you should open them on your desktop in a current version of Praat, resave them, and then try uploading them to Autophon again.
>
> **Audio formats:** Audio files can be **AAC(M4A)**, **AC-3**, **AIFF**, **AIFF/24bit**, **AIFF/32bit**, **ALAC**, **FLAC**, **M4R**, **MP3**, **OGG**, **OPUS**, **WAV/8bit**, **WAV/24bit**, **WAV/32bit**, **WAV/A-law**, **WAV/mu-law**, **WMA**. Autophon will automatically consolidate stereo files to mono, which can compromise quality. Therefore, you may wish to convert your audio to mono yourself.

Transcription preparation Transcriptions should be broken down into chunks that are between one and 20 words, and the boundary demarcations should have at least 0.01 seconds of buffer before and after the speech stream. This is illustrated in Figure 1. In natural-speech data, sometimes a researcher will encounter utterances of rapid speech with more than 20 words that have no pauses. In these cases, one must violate one rule in order to appease the other, and it is unclear to us at Autophon what the best approach is. Either force in a break to stick to the 20-word recommendation – violating the 0.01-second buffer – or permit the utterance to go beyond 20 words.

Some users will have transcriptions of single words or phrases that are segmented in Praat down to the exact start and finish time of the speech stream. Autophon will unfortunately perform poorly and not maintain the manual start and finish times. This, however, is something we would be interesting in remedying if a user expresses a need for his/her project.

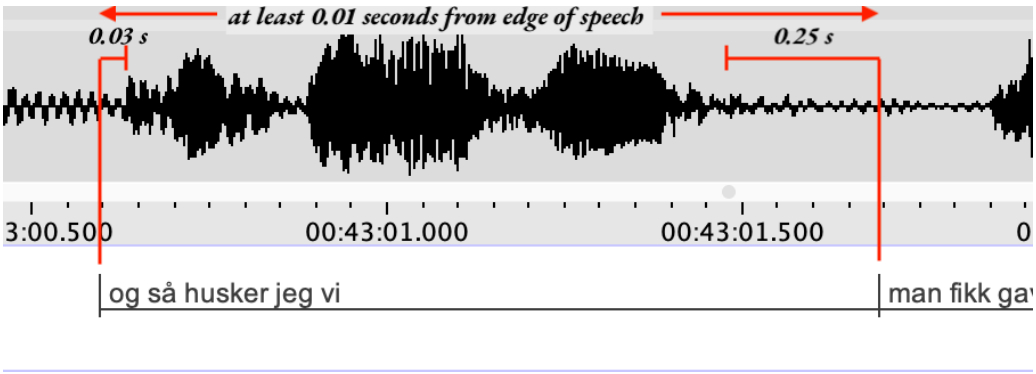


Figure 1: A sample transcription that follows our recommendations: a five-word chunk whereby the start boundary is approximately 0.03 seconds from the speech stream and the end boundary is approximately 0.25 seconds from the speech stream.

Select a language Once the files are uploaded, your files will appear in the topmost upload list where you can proof them before alignment. Autophon will automatically suggest a language for you. You can, of course, change this selection.

Proofing files In the upload list, metrics are provided for your files, including word count, file size, language, and missing words. It is here you can either delete the upload and start over, change the language, add words, or proceed with alignment by clicking "*Align!*".

Missing words Alignment is executed with a language-specific model and a language-specific pronunciation dictionary. These pronunciation dictionaries are finite and do not contain every word in the language. Therefore, the "missing words" column outputs any word in your transcription that is not found in Autophon's dictionary. Specifically, the following process happens: (1) Your transcription is reconciled against Autophon's dictionary. (2) Missing words are extracted. (3) A grapheme-to-phoneme script that was trained on the original dictionary is run on each missing word to offer suggested pronunciations. (4) These are reported in the "missing words" column, available for download. You then have the option to either accept these suggestions or change them and upload your own changes into the user dictionary (see next section).

Your custom pronunciations You may decide that you either (a) do not agree with Autophon's suggested pronunciations in the "missing words" column or (b) do not agree with the pronunciations in the back-end dictionary. In this box you can enter your own pronunciations that will override *both*. Entries must be done in the ASCII-bet specific to the language at hand (DanFAbet – Danish, ARPAbet – English, NoFAbet – Norwegian, etc.). The phoneme key can be accessed in Table 1.

Entries can either be made directly into the dictionary box or they can be uploaded from a txt file. You are limited to 1 million characters. Format entries as LOOKUP–SPACE/TAB–PHONEME–SPACE–PHONEME. Note that each phoneme must be separated by a space and that *every vowel must have a numerical stress attached to it (unstressed vowels take zero)*. Note also that the lookup cannot be two or more words because that will confuse Autophon and make it treat the second word as a phoneme.

If you would like Autophon to optimize and select from more than one pronunciation option, simply enter in multiple pronunciations for the same lookup.

Correct:

jabberwocky D J AEEH1 B EH0 W AO2 K IH0

jabberwocky J AEEH1 B EH0 W AO2 K IH0

Incorrect:

jabberwocky D J AEEH1 B EH W AO2 K IH

jabberwocky D J AEEH BEH WAO2 K IH

jabber wocky J AEEH1 B EH0 W AO2 K IH0

Aligning files Click "*Align!*" to the far right of the upload list to initiate alignment. This can take between a few minutes and an hour, depending on the amount of users accessing the aligner at that moment. You may have to log out and then log back in to see the outputted files.

Downloading the annotations Once alignment is finished, your annotations will be available in the bottom-side download list. You may have to log out and then log back in to see the outputted files. Click on the item to download the folder. The annotations will be in **TextGrid** format and are only readable in Praat (Boersma and Weenink 2017).

SweFA	*IPA*	*ex.*
Vowels		
AA	ɑː	*lat*
AH	a	*lass*
AE	ɛː	*nät*
AEH	ɛ̝	*lätt*
EE	eː	*leta*
EH	ɛ	*lett*
II	iː	*dis*
IH	ɪ	*disk*
OA	oː	*lås*
OAH	ɔ	*lott*
OE	øː	*söt*
OEH	œ	*dörr*
OO	uː	*sot*
OH	ʊ	*rott*
UU	ʉː	*lus*
UH	ɵ	*ludd*
YY	yː	*typ*
YH	ʏ	*flytta*
Diphthongs		
AJ	a͡j	*tajming*
AU	a͡u	*Gaude*
EJ	ɛ͡j	*Facebook*
EU	ɛ͡u	*Europa*
OJ	o͡j	*bojkottera*
Consonants		
B	b	*bil*
CH	tʃ	*cok*
D	d	*dal*
DH	ð	*that's it!*
F	f	*fil*
G	g	*gas*
H	h	*hal*
J	j	*jag*
JH	dʒ	*Jaffar*
K	k	*kål*
L	l	*lös*
M	m	*mil*
N	n	*nål*
NG	ŋ	*ring*
P	p	*pil*
R	r	*ris*
RD	ɖ	*bord*
RL	ɭ	*Karl*
RN	ɳ	*barn*
RS	ʂ	*fors*
RT	ʈ	*fart*
S	s	*sil*
SJ	ɧ	*sjuk*
T	t	*tal*
TH	θ	*thriller*
TJ	ɕ	*tjock*
V	v	*vår*
W	w	*wolla*
Z	z	*guzz*
Lexical stress and pitch accents		
AA0	ɑː	*cirkus*
AA1	¹ɑː	*cirkus*
AA2	ˌɑː	*cirkusdirektör*
AA3	²ɑː	*flytta*
AA4	ɑːꜜ	*flytta*

Table 1: Phoneme key: SweFAbet, IPA, and lexical examples

4 Phoneme key

Autophon will output two versions of the same TextGrid for every file you align: (1) in SweFAbet and (2) in IPA. SweFAbet is the ASCII-based phoneme coding that is specific to the Swedish language and resembles CMU's ARPAbet[1]. The key is located in Table 1.

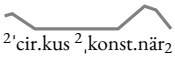
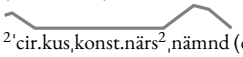
	SMALL accent (*usually ω stress*)	BIG accent (*usually ϕ head*)	Example
Accent 1	H L*	L* H	 ¹'cir.kus (circus)
Accent 2	H* L	H*L H	 ²'konst.när$_2$ (artist)
Accent 2 *compounds*	H* L	H*L *H	 ²'cir.kus ²ˌkonst.när$_2$ (circus artist)
			 ²'cir.kusˌkonst.närs^{2}ˌnämnd (circus artist committee)

Table 2: Description of pitch accents for Stockholm Swedish as described by (Myrberg and Riad 2015, p. 116): SMALL and BIG accents 1, 2, and compound accent 2. Examples are shown with orthodox contours. ω stands for word accent, and ϕ stands for phrase accent (Table from (Young 2019a, p. 39)).

Every SweFAbet vowel is followed by a numerical code that denotes suprasegmental information. XX0 refers to lexically unstressed vowels; XX1 – lexically stressed accent 1 vowels; XX2 – secondary stress in compound words, which is always accent 2; XX3 – lexically stressed accent 2 vowels; XX4 – the posttonic vowel immediately after lexically stressed accent 2 vowels (because accent 2 has a delayed peak). The superscript and subscript denotations are adopted from Myrberg and Riad (Myrberg and Riad 2015, p. 116) and are explained more transparently by Table 2, borrowed from Young (Young 2019a, p. 39).

5 Performance and metrics

SweFA version 2.0 was trained on Stockholm Swedish and standard Central Swedish – both spontaneous and read-aloud speech – and meets most of the benchmarks established in the forced-alignment literature. Its accuracy is measured here by comparing alignments of approximately 1 000 phonemes in spontaneous speech, each from nine adult male speakers from the Stockholm region. The alignments for the older SweFA version 1.0 (Young and McGarrah 2023) and the current SweFA version 2.0 (Young 2019b) are compared in Table 3 against manual segmentation.

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

		SweFA 1.0				SweFA 2.0 (current)			
		n boundaries	Median onset difference (ms)	Pct 10 ms or less	Pct 20 ms or less	*n* boundaries	Median onset difference (ms)	Pct 10 ms or less	Pct 20 ms or less
Received Stockholmian	August	1 000	13	39	70	1 176	10	48	78
	Jan-Axel	1 000	16	35	60	1 116	12	40	71
	Joseph	1 000	13	40	69	1 239	11	47	77
Working-class Stockholmian (Ekensnack)	Nils	1 000	13	39	66	1 240	10	49	78
	Paul	1 000	14	33	61	1 179	11	44	73
	Per	1 000	15	37	62	1 200	12	42	72
Stockholm Multiethnolect (Rinkeby Swedish)	Antonio	1 000	14	35	63	1 241	12	42	72
	Max	1 000	12	42	71	1 184	11	47	77
	Hayder	1 000	13	36	65	1 255	12	44	72
	all	9 000	13	37	65	10 830	11	45	74

Key

n boundaries	number of boundaries tested against the manual gold standard (g.s.)
median onset difference (ms)	median difference between aligner boundaries and manual g.s. boundaries
pct 10 ms or less	percentage of aligner boundaries that fall within 10 milliseconds of manual g.s. boundaries
pct 20 ms or less	percentage of aligner boundaries that fall within 20 milliseconds of manual g.s. boundaries

Table 3: Accuracy metrics for SweFA version 1.0 (Young and McGarrah 2023) and the current SweFA version 2.0 (Young 2019b)

6 Data security

Everything you upload is encrypted and sent to a server in Frankfurt, Germany, that is run by *Digital Ocean*. Transcriptions and sound files are deleted immediately after alignment, which significantly reduces the chance of a data breach. On the other hand, finished TextGrids are stored in your account for as long as you like. Once, however, you delete them, they will be removed from our server permanently.

7 Features and limitations

What Autophon is

Autophon is a frontend web application for the Nordic languages that uses the Montreal Forced Aligner (MFA) (McAuliffe, Socolof, Mihuc, Wagner, and Sonderegger 2017) as a core component of its backend. The language-specific models and pronunciation dictionaries were constructed by Nate Young. The language-specific models are trained on various corpora, and the pronunciation dictionaries are usually adaptations of existing dictionaries available online.

The main advantages of using Autophon are:

1. Autophon is a web app, which means it is OS-agnostic.
2. As a web app, Autophon also requires less computational knowledge, which expands access to more researchers and students.
3. Autophon takes nearly all types of transcription and sound formats. It runs your transcriptions and sound files through a converter to ensure that they can be aligned with the MFA backend.
4. Autophon offers an easy way to add the pronunciation of missing words by integrating grapheme-to-phoneme algorithms into the user interface.
5. Autophon allows the user to have access to the latest language training models without needing to constantly check for updates.

What Autophon is *not*

Important limitations are:

1. Autophon is no magic bullet. You need a highly accurate orthographic transcription for it to work. And even then, you may not be satisfied with the results.

2. Autophon varies in its accuracy, and this accuracy depends on the language. Accuracy metrics are provided above.
3. Autophon will be slower to implement core MFA updates because it consists of layers and layers of code packed around MFA.

Acknowledgements

Numerous individuals helped make SweFA possible. Michael McGarrah and Joe Fruehwald assisted me with developing version 1.0. Kaosi Anikwe has been an invaluable backend and frontend developer. I also have the following programmers to thank, all of whom worked tirelessly to build a robust backend and UX: Ismail Raji Damilola, Nabil Al Nazi, Zamanat Abbas Naqvi, and Santiago Recoba.

References

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer [Computer software], Version 6.0.36. http://www.praat.org/

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of Interspeech*, 498–502.

Myrberg, S., & Riad, T. (2015). The prosodic hierarchy of Swedish. *Nordic Journal of Linguistics*, *38*(2), 115–147.

Young, N. J. (2019a). *Rhythm in late-modern Stockholm – Social stratification and stylistic variation in the speech of men*. Department of Linguistics, Queen Mary, University of London. ISBN: 978-91-7699-210-4. http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-178897

Young, N. J. (2019b). SweFA 2.0 – Forced Alignment of Swedish, version 2.0. www.autophon.se

Young, N. J. (2023). Autophon – Automatic phonetic annotation of Nordic languages (web application). www.autophon.se

Young, N. J., & McGarrah, M. (2023). Forced alignment for Nordic languages: Rapidly constructing a high-quality prototype. *Nordic Journal of Linguistics*, *46*(1), 105–131. https://doi.org/10.1017/S033258652100024X