Autophon user guide
Norwegian Bokmål
Model: NoFA version 1.0

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 Norwegian 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. (2020). NoFA 1.0 – norsk modell for forced alignment, version 1.0. https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-59/

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.

Select a language Once the files are uploaded, you will be asked to pick a language. After you click "OK", your files will appear in the topmost upload list where you can proof them before alignment.

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).

User dictionary Here, you can upload your own pronunciations that will override Autophon's suggested pronunciations in the "missing words" column. Entries must be done in the ASCII-bet specific to the language at hand (DanFAbet for Danish, ARPAbet for English, NoFAbet for Norwegian, SweFAbet for Swedish). The phoneme key can be accessed in Table 1. Entries can either be made directly into the dictionary box or they can be uploaded directly from a txt file. Entries should always be in the following format: WORD–SPACE–PHONEME–SPACE–PHONEME.

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).

4 Phoneme key

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

NoFA	*IPA*	*ex.*
Vowels		
AA	ɑː	*bad*
AH	ɑ	*hatt*
AE	æː	*vær*
AEH	æ	*vært*
EE	eː	*lek*
EH	ɛ	*penn*
II	iː	*vin*
IH	ɪ	*sitt*
OA	oː	*rå*
OAH	ɔ	*gått*
OE	øː	*løk*
OEH	œ	*høst*
OO	uː	*bod*
OH	u	*fort*
UU	ʉː	*hus*
UH	ʉ	*russ*
YY	yː	*ny*
YH	y	*nytt*
UX	ɤ	*girl (English)*
AX	ə	*behage*
Diphthongs		
AEJ	æj	*sei*
AEW	æw	*sau*
AJ	ɑj	*kai*
OEJ	œj	*køye*
OJ	ɔj	*konvoy*
OU	ou	*show*
Consonants		
B	b	*bil*
D	d	*dag*
DH	ð	*this (English)*
DJ	dʒ	*George (English)*
F	f	*fin*
G	g	*gul*
H	h	*hes*
J	j	*ja*
K	k	*kost*
KJ	ç	*kino*
L	l	*land*
M	m	*man*
N	n	*nord*
NG	ŋ	*eng*
P	p	*pil*
R	r	*rose*
RD	ɖ	*rekord*
RL	ɭ	*perle*
RN	ɳ	*barn*
RT	ʈ	*stort*
S	s	*sil*
SJ	ʂ	*sju*
T	t	*tid*
TH	θ	*thin (English)*
TSJ	tʃ	*church (English)*
V	v	*vase*
W	w	*Washington*
X	x	*ach (German)*
Z	z	zigzag (English)
Syllabic consonants		
LX	l̩	*djeve**l**sk*
MX	m̩	*lands**om**fattande*
NX	n̩	*avtale**n***
RLX	ɭ̩	*vars**el***
RNX	ɳ̩	*ture**n***
RX	r̩	*søke**r***
Paralinguistic features		
EXH		<*exhale*>
INH		<*inhale*>
NHES		<*nasal*>
VHES		<*vowel*>
LG		< *laughter* >
Lexical stress		
AA0	ɑː	*adv**a**rer*
AA1	ˈˈ1ɑː	***a**dvarer*
AA2	ˈˈ2ɑː	***a**dvarsel*
AA3	ˌɑː	*adv**a**ring*

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

Every NoFAbet vowel is followed by a numerical code that denotes suprasegmental information. XX0 refers to lexically unstressed vowels; XX1 – toneme 1 primary lexical stress; XX2 – toneme 2 primary lexical stress; XX3 – secondary lexical stress.

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

5 Performance and metrics

NoFA version 1.0 was trained on Norwegian Bokmål from the RUNDKAST (Amdal, Strand, Almberg, and Svendsen 2008) and NB Tale corpora – both spontaneous and read-aloud speech – and is quite accurate. Its accuracy is measured here by comparing alignments of 1000 phonemes, each from one speaker from three different dialect regions, respectively: Southeastern (Oslo), Western (Bergen), and Central (Trøndelag). The alignments are compared in Table 2 with manual segmentation.

Dialect	Speaker	*n* boundaries	median onset difference (ms)	pct 20 ms or less	pct 10 ms or less
Southeastern (Oslo)	Anniken Hauglie	1000	10	79%	49%
Western (Bergen)	Audun Lysbakken	1000	10	77%	48%
Central (Trøndelag)	Trine Skei Grande	1000	12	72%	40%

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 20 ms or less	percentage of aligner boundaries that fall within 20 milliseconds of manual g.s. boundaries
pct 10 ms or less	percentage of aligner boundaries that fall within 10 milliseconds of manual g.s. boundaries

Table 2: Accuracy metrics for NoFA 1.0

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.

References

Amdal, I., Strand, O. M., Almberg, J., & Svendsen, T. (2008). RUNDKAST: An Annotated Norwegian Broadcast News Speech Corpus. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. http://www.lrec-conf.org/proceedings/lrec2008/pdf/486_paper.pdf

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. (2020). NoFA 1.0 – norsk modell for forced alignment, version 1.0. https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-59/

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