📖 **NoFA-en.md**

NoFA -- Norwegian model for forced alignment

Introduction

Forced alignment (FA) refers to algorithms that take an audio file with speech and an orthographic transcription of the speech as input and produce a phonetic transcript where each segment (each phone) is time-aligned with the audio file.

FA is useful in many contexts, including phonetics, phonology, computational linguistics and NLP development. For an FA tool to work for a particular language, an acoustic model and a pronunciation lexicon are needed. Språkbanken at the National Library of Norway distributes an open access pronunciation lexicon for Norwegian Bokmål, the NST lexicon. However, Norwegian has until now lacked a publically available acoustic model for FA.

Nate Young has created a FA model for Norwegian (Bokmål) on behalf of Språkbanken, *NoFA*. This model is made for the Montreal Forced Aligner (MFA). NoFA is trained on Språkbanken's speech database *NB Tale* and the phonetic part of the RUNDKAST database developed at the Norwegian University of Science and Technology (Amdal et al. 2008). We refer to the attached document *Construction of Forced aligner for Norwegian Bokmål. February Progress Report* for details on the development of this model.

The NST lexicon is not ideal for FA, firstly because it contains only a limited degree of pronunciation variation and mostly represents a standard pronunciation of south-eastern Norwegian, secondly because it was developed at the turn of the millennium and thus lacks updated vocabulary. In this release, we provide a version of the NST lexicon that is readable for MFA, but also a slightly expanded lexicon that has been used in the development of the FA model. We would like to stress, however, that the latter lexicon is somewhat experimental and only to a limited extent reflects the pronunciation variation of Norwegian. Both lexica are in Norwegian Bokmål only. Språkbanken is going to develop pronunciation lexica for Norwegian Bokmål and Norwegian Nynorsk with updated vocabulary and dialectal variation. These will be well suited for FA. In addition, users are encouraged to make their own updates and adaptations of the lexica for their own use. If you wish to share updated lexica or update rules with Språkbanken, please contact us at sprakbanken@nb.no. It is also possible to automatically generate pronunciation lexica for words that do not exist in existing lexica using grapheme-to-phoneme conversion. We will return to this below.

Content

- **documentation/** contains relevant information about NoFA, including this document
- **model/** contains the Norwegian acoustic model for FA: *NoFA_1-0.zip*
- **lexicon/** contains lexica formatted for MFA and converted to the NoFAbet transcription standard (see below). *NST_nob.dict* has all the orthography-pronunciation pairs of the NST lexicon. *NoFA-lex.dict* is a slightly modified lexicon made for the NoFA project.

NoFAbet

A dedicated set of ASCII-based transcription conventions, *NoFAbet*, has been developed for this project. NoFAbet is partly based on ARPABET. Each phone is represented by a string of one or more uppercase letters, numbers, and other characters. There is a whitespace between each phone.

Lastebilen ('the lorry') is transcribed as follows in NoFAbet: *L AH2 S T AX0 B II3 L NX0*. All vowels and consonants serving as syllable nuclei have a number attached to them. *1* and *2* signify toneme 1 and 2 respectively. *0* marks an unstressed syllable. *3* is used for secondary stress. Long vowels are transcribed using two letters, e.g. *II3*, while *H* is used to mark short vowels, as in *AH2*. In the case of vowels such as æ, which do not exist in the ASCII set, short and long vowels are only distinguished by the presence or absence of *H*: *AE2* vs. *AEH2*. The letter *X* is used in some syllable nuclei, e.g. syllable nucleus consonants, *NX0*, and schwas, *AX0*. The table below lists X-SAMPA phones, as used in the original NST lexica, and their NoFAbet equivalents:

X-SAMPA	NoFAbet	Example
A:	AA0	b**a**d
{:	AE0	v**æ**r
{	AEH0	v**æ**rt

X-SAMPA	NoFAbet	Example
{*I	AEJ0	s**ei**
E*u0	AEW0	s**au**
A	AH0	h**a**tt
A*I	AJ0	k**ai**
@	AX0	b**e**hage
b	B	**b**il
d	D	**d**ag
e:	EE0	l**e**k
E	EH0	p**e**nn
f	F	**f**in
g	G	**g**ul
h	H	**h**es
I	IH0	s**i**tt
i:	II0	v**i**n
j	J	**j**a
k	K	**k**ost
C	KJ	**k**ino
l	L	**l**and
l=	LX0	
m	M	**m**an
m=	MX0	
n	N	**n**ord
N	NG	e**ng**
n=	NX0	
o:	OA0	r**å**
O	OAH0	g**å**tt
2:	OE0	l**ø**k
9	OEH0	h**ø**st
9*Y	OEJ0	k**øy**e
U	OH0	f***o**rt
O*Y	OJ0	konv**oy**
u:	OO0	b**o**d
@U	OU0	sh**ow**
p	P	**p**il
r	R	**r**ose
d`	RD	reko**rd**
l`	RL	pe**rl**e

X-SAMPA	NoFAbet	Example
l`=	RLX0	
n`	RN	ba**rn**
n`=	RNX0	
s`	SJ	pe**rs**
t`	RT	sto**rt**
r=	RX0	
s	S	**s**il
S	SJ	**sj**u
s=	SX0	
t	T	**t**id
u0	UH0	r**u**ss
u0 j	UH0_J	Anh**ui**
}:	UU0	h**u**s
v	V	**v**ase
w	W	**W**ashington
Y	YH0	n**y**tt
y:	YY0	n**y**

The reason why NoFAbet is chosen instead of X-SAMPA is that the phoneme transcriptions are closer to Norwegian orthography, and it is, therefore, easier to read and manually produce transcriptions using this standard.

Note that we have chosen to make a few simplifications when converting the original NST lexicon to the NoFAbet version. Firstly, the original lexicon has markings of syllable boundaries, and in the case of multi-word expressions, word boundaries and phrase accent. We have chosen to omit this, since MFA doesn't use this information. Secondly, the original NST lexicon has its own representation of retroflex *s*, *s`*, which differs from the postalveolar fricative they transcribe as *S*. We have chosen to use *SJ* for both phones.

How to run NoFA

1. Install MFA. The instructions are found here.
2. Prepare your data. Put audio files and corresponding transcriptions in a directory. The transcriptions have to be in Praat *TextGrid* format. The audio files need to be *wav* files with a sampling rate of at least 16 kHz. The audio files and the transcription files must have the same name, e.g. *myrecording.wav* and *myrecording.TextGrid*.
3. Validate the files (can be omitted). While in the root directory of MFA, write the command *bin/mfa_validate_dataset path/to/input path/to/lexicon path/to/acoustic/model*. The script generates a list with the words in the transcriptions which are not in the lexicon. In the FA transcriptions, these words will be marked as unknown unless they are added to the lexicon. The validation script will also report on possible errors in the audio files and text files. See Se here for more details.
4. Run the aligner. *bin/mfa_align path/to/input path/to/lexicon path/to/acoustic/model path/to/output*. FA transcriptions will appear in the output directory in the form of TextGrid files.

For further information, see MFA's homepage.

Evaluation

Nate Young has tested how accurate the time alignments in NoFA are. The test data are transcribed audio files from Norwegian parliamentary proceedings of three speakers of three different dialects: Oslo (Anniken Hauglie), Bergen (Audun Lysbakken), and Trøndelag (Trine Skei Grande). Each data set contained between 1000 and 1500 phones. The test shows how big the divergence is between the time codes NoFA produces and manually adjusted time codes. The test parameters are median deviations in milliseconds (ms) for phone onset (i.e. the start time of a phone), the percentage of predicted phone onsets that are within 20 ms of the actual onset, and the percentage of predicted phone onsets that are within 10 ms of the actual onset. The results are as follows:

Dialect	Median onset displacement	Percentage within 20 ms	Percentage within 10 ms
Oslo	10	79,2%	48,8%
Bergen	10	76,8%	47,8%
Trøndelag	12	72,3%	40,2%

According to Nate Young, the results for Oslo and Bergen are within the gold standard for FA systems, and for Trøndelag, they are close to the gold standard. See the report *Construction of Forced aligner for Norwegian Bokmål, Final Report* for details.

Lexicon limitations

Lexica compatible with MFA are text files where each line contains the orthographic form of a word and a phonetic transcription of the word. The orthographic form and the transcription are separated by a whitespace. If a word has several alternative pronunciations, these will be on separate lines with the same orthographic form, but with different transcriptions. MFA will only produce transcriptions that are present in the lexicon. If a word is not in the lexicon, MFA will insert an *unknown* tag. If the word is in the lexicon but with a different pronunciation than the one found in the audio file, the transcript will not fully reflect what is being said.

The NST lexicon is created around the turn of the millennium. Consequently, newer additions to the vocabulary are not in the lexicon. In addition to this, you can freely create new compounds in Norwegian. Therefore, most transcripts will probably contain words that are not in the lexicon. If you want these to be taken into account by MFA, you have to add them to the lexicon. You can either enter words and transcriptions manually, or you can use *grapheme-to-phoneme conversion* (G2P). Språkbanken has trained a model for the G2P system Phonetisaurus using the NST lexicon. It can be found here.

The NST lexicon is based on the south-east dialect, and MFA will, therefore, produce better transcripts for this than other dialects. The NST lexicon also information about contractions, etc. that you find in fast speech. In *NoFA-lex.dict* we have tried to remedy this to some extent. However, users who need accurate transcriptions are encouraged to update the lexicon themselves. Språkbanken plans to develop pronunciation lexica with dialectal variation for Norwegian Bokmål and Nynorsk. When these are available, they can be used with MFA.

License and contact information

The NoFA model and auxiliary resources are public domain with the license CC0. There are, therefore, no restrictions on their use. See here for MFA's license.

This resource is shared as is, without a warranty of any kind. We would, nevertheless, be interested in hearing from users if they find errors, have suggestions for improvement or questions. You can reach us at sprakbanken@nb.no.

Sources

- I. Amdal, O.M. Strand, J. Almberg, T. Svendsen. 2008. RUNDKAST: An Annotated Norwegian Broadcast News Speech Corpus. Proceedings of LREC 2008