INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# NeMo Forced Aligner and its application to word alignment for subtitle generation

*Elena Rastorgueva, Vitaly Lavrukhin, Boris Ginsburg*

NVIDIA Corporation, United States

{erastorgueva, vlavrukhin, bginsburg}@nvidia.com

## Abstract

We present NeMo Forced Aligner (NFA): an efficient and accurate forced aligner which is part of the NeMo conversational AI open-source toolkit. NFA can produce token, word, and segment-level alignments, and can generate subtitle files for highlighting words or tokens as they are spoken. We present a demo which shows this functionality, and demonstrate that NFA has the best word alignment accuracy and speed of alignment generation compared with other aligners.

**Index Terms**: forced alignment, speech recognition, subtitle generation

## 1. Introduction

In the speech domain, alignment is the mapping of text to when it is spoken in audio. Forced alignment specifically is "a technique to take an orthographic transcription of an audio file and generate a time-aligned version"[1]. Forced alignment can be applied to speech processing tasks such as: dataset segmentation for speech corpus creation, dataset analysis, phoneme duration extraction for Text-To-Speech.

For the task of generating word-by-word subtitles, we typically have 3 requirements: (1) the reference text is some highly accurate ground truth text that we provide, (2) we require every part of the reference text to map to some part of the audio, and (3) alignments are non-overlapping. Typically 'forced alignment' implies all of these conditions are met. NFA and Montreal Forced Aligner (MFA) [1] meet all of the above requirements. There are several other commonly used aligners which may be suitable for this task if we relax our constraints. Gentle[2] violates assumption (2) by removing text which it cannot align with high confidence. CTC segmentation [2] can be made to meet assumption (2) by not removing any low-confidence alignments, but it violates assumption (3). WhisperX[3] violates assumption (3) as it requires the reference text to be transcriptions from the Whisper model[4].

We created a tool called NeMo Forced Aligner (NFA) which applies Viterbi decoding to the log-probabilities outputted by CTC [3] models in NeMo[5]. NFA generates very good alignments, which we will demonstrate quantitatively in this paper by comparing the accuracy and speed of NFA with other forced and non-forced aligners. NFA is available in NeMo [6].

[1] https://montreal-forced-aligner.readthedocs.io/en/latest/user_guide/index.html

[2] https://github.com/lowerquality/gentle

[3] https://arxiv.org/abs/2303.00747

[4] https://cdn.openai.com/papers/whisper.pdf

[5] https://github.com/NVIDIA/NeMo

[6] https://github.com/NVIDIA/NeMo/tree/main/tools/nemo_forced_aligner

## 2. NeMo Forced Aligner

NeMo Forced Aligner contains an efficient PyTorch-based implementation of Viterbi forced alignment.

The reference text by default is the text provided by the user, though NFA has a flag which can be set to instead use predicted text from a NeMo CTC-based ASR model (in this case we use the same model for generating the predicted text and for Viterbi decoding, to save computation time).

As NFA does Viterbi decoding over the input sequence of tokens, the forced alignment produced is at the token level. NFA also produces alignments for words (i.e. space-separated substrings) and user-specified segments: by default a 'segment' is the entire input text except for the first and final 'blank' token (this allows us to trim any initial and final silence), but a user can also introduce separators such as "|" in the reference text, which will be interpreted as segment boundaries. These word and segment boundaries are obtained by grouping together the alignments of their constituent tokens.

NFA outputs the alignments in the format of CTM files and ASS subtitle files. In the ASS subtitle files, words/tokens in the same segment appear at the same time, and word/tokens are highlighted at the times when the alignment dictates that they were spoken.

## 3. Demo description

The demo utilizes Gradio[7] to present an interface where the user can test NFA's alignments for various languages. The user can select the language spoken in the audio, upload or record an audio file, and type the reference text into a text field (or leave it empty, in which case NFA will use the ASR model used for alignment to generate a transcription which will be used as a reference text). The demo passes the inputs to NFA, which saves the results of the alignment into some ASS subtitle files. These files are combined with the input audio to generate a video which highlights the text at the time it is aligned to (Figure 1).

## 4. Experiments

In order to compare the speed and accuracy of the various aligners mentioned, we obtained alignment predictions from each aligner and compared them with the word alignments of the AMI corpus [4] (specifically the test set in single-channel Mixed Headset format). This experiment follows the methodology of the WhisperX paper. Where relevant, we conducted experiments with both the "ground truth" transcript and ASR predicted text as the reference text. The "ground truth" AMI transcript was created by joining together the words in the pro-

[7] https://arxiv.org/pdf/1906.02569.pdf

Table 1: *Results of alignment on AMI test mixed headset. The results for MFA are marked with a * since only 5 out of 16 audio files were aligned successfully. 'bs=1' and 'bs=4' indicate a batch size of 1 and 4 respectively.*

| Aligner | Source of reference text | Model for alignment | Precision (%) | Recall (%) | RTF |
|---|---|---|---|---|---|
| NFA (bs=1) | ground truth | ConformerCTCMedium | **98.35** | **98.35** | 149 |
| NFA (bs=1) | ground truth | CN1024gamma0.25 | 97.56 | 97.56 | 219 |
| NFA (bs=4) | ground truth | CN1024gamma0.25 | 97.56 | 97.56 | **308** |
| CTC Segmentation | ground truth | ConformerCTCMedium | 98.17 | 98.17 | 152 |
| CTC Segmentation | ground truth | CN1024gamma0.25 | 94.28 | 94.28 | 126 |
| Gentle | ground truth | default | 94.47 | 78.72 | 13 |
| MFA* | ground truth | `english_mfa` | 82.93* | 83.76* | 5* |
| NFA (bs=1) | ConformerCTCMedium | ConformerCTCMedium | 85.65 | **71.09** | 152 |
| NFA (bs=1) | CN1024gamma0.25 | CN1024gamma0.25 | 82.49 | 60.17 | 219 |
| NFA (bs=4) | CN1024gamma0.25 | CN1024gamma0.25 | 82.49 | 60.17 | **308** |
| WhisperX v3 | Whisper `large_v2` | `VOXPOPULI_ASR_BASE_10K_EN` | **90.26** | 70.12 | 36 |

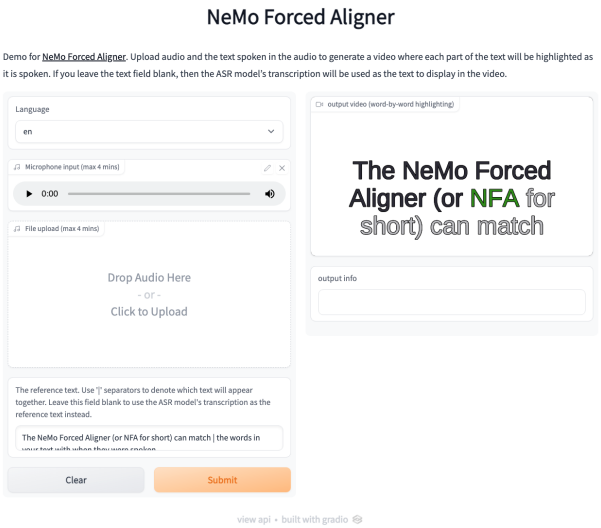


Figure 1: *NFA gradio demo.*

vided XML files and lowercasing the resulting text.

We show precision & recall metrics, where a *true postive* is when a predicted alignment and a true alignment match, a *false positive* is when a predicted alignment does not have a matching true alignment, and a *false negative* is when a true alignment does have a matching predicted alignment. A predicted and true alignment 'match' if they occur within 200ms of each other (this value is in following with the description in the WhisperX paper), and if they have the same text (both texts were preprocessed by lowercasing, removing any digits, removing punctuation except for apostrophes, and removing any remaining spaces).

We also show the Real Time Factor (RTF) for producing the alignments. All aligners were run on a system with a Intel(R) Core(TM) i9-10900X CPU @ 3.70GHz and 125.5 GiB RAM. For GPU-based aligners, a single NVIDIA Quadro RTX 8000 GPU was used. For CTC Segmentation we used the NeMo-integrated version [5]. The NeMo ASR models used are Citrinet[8] and Conformer CTC[9]. For the latter, we restricted attention context size to 64x64. For the NFA runtimes, both CTM and ASS files were generated, but the ASS files used NFA's automatic resegmentation to make sure approximately 2 lines of text would appear on the screen at any time, and not all of the text for the whole meeting.

[8] `https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_citrinet_1024_gamma_0_25`

[9] `https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_ctc_medium`

## 5. Results

Table 1 shows the alignment precision and recall of various recent aligners on the AMI test set Mixed Headset data.

Within the context of using the ground truth as reference text for alignment, NFA can obtain the best precision & recall and the best RTF.

Within the context of using ASR model predictions as the reference text for alignment, NFA is significantly faster than WhisperX but slightly less accurate. This is expected because aside from using a much smaller model for transcription, in its current implementation, NFA generates a transcription for the entire audio file and aligns it all at once, whereas WhisperX transcribes and aligns smaller sections of the audio at a time.

## 6. Conclusion

NFA is the best aligner for the task of word-by-word subtitle generation due to the high accuracy and speed of its alignment generation, as well as its ability to use a provided ground truth text as reference text, and produce non-overlapping word alignments.

## 7. References


[1] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Interspeech*, 2017.

[2] L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, "CTC-segmentation of large corpora for german end-to-end speech recognition," in *Speech and Computer*. Springer International Publishing, 2020.

[3] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[4] I. Mccowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska Masson, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus," in *Int'l. Conf. on Methods and Techniques in Behavioral Research*, 2005.

[5] E. Bakhturina, V. Lavrukhin, and B. Ginsburg, "A toolbox for construction and analysis of speech datasets," in *NeuRIPS, Track on Datasets and Benchmarks*, 2021.