



Can an initial phoneme-based algorithm improve automatic naming latency detection during picture naming tasks? A feasibility study

Eli Rickert^{a,1}, Sven Altermatt^{a,1}, Sunghea Park^b, Marc Stawiski^a, Katrin Petra Kuntner^b, Sandra Widmer Beierlein^b, Anja Blechschmidt^b, Claire Reymond^c, Markus Degen^a, Simone Hemm^{a,*}

^a Institute for Medical Engineering and Medical Informatics, School of Life Sciences, University of Applied Sciences and Arts Northwestern Switzerland FHNW, Muttens, Switzerland

^b Institute for Special Education and Psychology, School of Education, University of Applied Sciences and Arts Northwestern Switzerland FHNW, Muttens, Switzerland

^c Institute Digital Communication Environments, Academy of Art and Design, University of Applied Sciences and Arts Northwestern Switzerland FHNW, Basel, Switzerland

ARTICLE INFO

Keywords:

Automatic naming latency detection
Aphasia
Picture naming task

ABSTRACT

Picture naming tasks are widely used to evaluate language processing in aphasia. Automatically detected naming latencies (aNL) often differ from the gold standard, manually detected latencies (mNL). Initial phonemes (InitP) influence aNL detections. The study aimed to develop and explore an InitP-based optimization approach to enhance automatic naming latency detection. An algorithm was developed to optimize several parameters of the word speech envelope for 16 InitPs using German training data from 134 healthy subjects and 31 persons with aphasia. These parameters were applied to test data to assess InitP influence. The resulting naming latencies (FHNW_NL), detected by the newly developed algorithm, were compared to manually detected latencies (mNL) and to speech onsets identified by Kaldi and Chronset.

The mean signed difference between FHNW_NL and mNL falls within the ± 15 ms inter-rater variability interval. Differences from Kaldi and Chronset to mNL were over twice and four times higher, respectively, than FHNW_NL. Persons with aphasia showed higher variances than healthy subjects. This study shows that InitP-based optimization of naming latency detection parameters improves aNL detection accuracy. This method could also benefit other research and clinical fields.

1. Introduction

Picture naming exercises represent a common approach in evaluating language processing and particularly word finding ability by matching a picture to its corresponding name [1,2]. These exercises also assess phonological encoding with speech production as the final outcome of the process. Aphasia is a language processing disorder caused by damage to the brain regions involved in language comprehension and production, due to events such as stroke or traumatic injury [3]. The ability to match a picture with its name and to convert speech sounds can be disrupted by aphasia [4]. Therefore, picture naming tasks are a specialized tool for evaluating language processing abilities in persons with aphasia. Such naming tasks are applied during aphasia

diagnosis and therapy [5,6] in validated language tests (e.g., Aachen Aphasia Test [7]).

Naming latency (NL) refers to the time interval between the presentation of a picture and the onset of the correct spoken response. NL has been used as an indirect predictor of the neurological mechanisms underlying picture naming, encompassing visual recognition, conceptual-lexical activation for word retrieval, and phonological encoding and articulation for speech production [8,9]. Slow NL in picture naming can indicate higher competition for a word selection in the stage of visual object recognition and multiple elaborations of names in the stage of word retrieval [10]. In particular, individuals with aphasia commonly experience difficulties in picture naming and exhibit slower NLs compared to those without aphasia [11–14]. This delay is primarily

* Corresponding author at: Institute for Medical Engineering and Medical Informatics, School of Life Sciences, University of Applied Sciences and Arts Northwestern Switzerland, Muttens, Switzerland, Hofackerstr. 30, 4132 Muttens, Switzerland

E-mail address: simone.hemm@fhnw.ch (S. Hemm).

¹ Shared first authorship.

<https://doi.org/10.1016/j.bspc.2025.108851>

Received 7 May 2025; Received in revised form 25 August 2025; Accepted 5 October 2025

Available online 3 November 2025

1746-8094/© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

attributed to the core symptom of word finding difficulties, which significantly influence NLs across different types of aphasia.

Manual NL (mNL) detection remains the gold standard due to its accuracy, but the process is time-consuming, involving detailed analysis of the speech onset using visualized acoustic signals in waveforms and spectrograms. This challenge led to the development of the voice key, an electronic tool to detect speech onset [15], which records signals when sound intensity exceeds a set threshold. Although voice keys are intended to capture low-amplitude signals, they remain unreliable, particularly in case of the low sound intensities of unvoiced initial phonemes (InitP). Another approach is automated naming latency (aNL) detection algorithms, such as the open-source tools “Chronset” [16] and the “SayWhen” algorithm [17], which aim to streamline and expedite this process. In addition to diagnostic applications, therapy could also benefit from such algorithms through mobile applications with automated feedback concerning correctness and naming latency. These would increase the accessibility to high-intensity home training and enhance training effectiveness [18,19]. However, current algorithms are primarily designed for research analysis rather than direct integration into mobile applications for patient feedback. They have been reported to be resource-intensive in computation time due to frequency analysis [16] or prone to generating large data sets requiring extensive manual review due to potential mismeasurements [17].

To address this gap, a patient-friendly prototype application for tablets has been developed [20] to give objective feedback on the correctness of responses. This application utilizes speech recognition feedback based on the Kaldi automatic speech recognition toolkit² (see Rickert et al., 2021: conference presentation³; [21]) but does not yet integrate NL detection. Automated speech recognition tools like Kaldi can provide the detected word and speech onset [22], but they are not specifically designed for NL measurement and thus are not expected to be precise.

Furthermore, InitPs can affect automated detections of NL. Each InitP possesses distinct phonetic and acoustic properties, leading to variability in detected speech onset times influenced by factors such as the manner and position of articulation and vocal fold vibration [23,24]. For instance, the automatic detection of InitPs, such as voiceless fricative /f/, is often impeded by their weak energy profiles. Voiceless fricatives are characterized by high-frequency energy and lack the low-frequency voicing found in voiced sounds, which can result in delayed or missed automatic speech onset detection [23,25]. In our previous study [24], the impact of InitP on an extended threshold-based aNL detection algorithm for picture naming tasks was investigated. The algorithm evaluated the speech envelope, taking into account 1) a threshold above which the signal was considered for analysis, 2) the time the signal had to stay above this threshold related to the minimal word length and duration and excluding short noises and 3) the time the signal was allowed to stay below the threshold before the end of the word (accounting for longer words where the signal may drop below this threshold for a certain period of time). The parameters were optimized and the algorithm evaluated with audios from German speaking healthy subjects. Results showed a strong positive correlation between mNL and aNLs (ibid); however, their differences underscored the impact of InitPs, in terms of their articulation manners and positions, on the speech envelope detection (ibid). This highlights the need to adapt the aNL detection algorithm to account for InitPs.

The aim of the present study was to develop an algorithm, for automatic naming latency, FHNW_NL⁴, detection, that considers initial phonemes. Moreover, it requires low computational resources by avoiding frequency analysis, with the longer-term objective of integration into an app. The performance was evaluated on test data and compared to the gold-standard mNL detection and existing speech onset detection tools. Precisely, an InitP-based optimization approach across different InitPs (e.g., /t/, /p/, /j/ and /h/) and InitP groups (e.g., Affricative, Plosive and Vowel) was added to the previously published algorithm from Park et al. [24]. The signal envelope was refined, introducing additional time-domain based optimization parameters, and optimizing the detection parameters for each available InitP separately. Additionally, the algorithm was programmed to begin searching for speech onset at a specific reference time, which can be provided by an automatic speech recognition tool. Training was conducted using available data from German-speaking healthy subjects as well as from some persons with aphasia. The study also compared the naming latencies from the new algorithm (FHNW_NL), Kaldi (Kaldi time), and Chronset (Chronset_NL) to the manual latency, investigating which of these approaches most closely aligns with mNL detection and whether FHNW_NL outperforms Kaldi and Chronset in this regard.

2. Material

Audio data from healthy subjects and persons with aphasia during naming tasks were available from two different studies: study 1 evaluating the name agreement for graphic representations in Swiss and Standard German with healthy subjects [20] and study 2 investigating differences in image naming performance between persons with and without aphasia ([14], [ClinicalTrials.gov](https://clinicaltrials.gov/ct2/show/study?term=NCT05164380) Identifier: NCT05164380). None of the datasets were specifically collected for the purposes of the current study. Speech material including in total 134 healthy subjects and 31 persons with aphasia (Table 1) was selected for the present study.

The used dataset comprised 9029 audio recordings in Swiss and Standard German from correctly named images for 86 words. Only InitPs with at least 100 available audio recordings were considered. The available recordings were divided into two data sets (Table 2). The first data set consisted of the audio data used for the InitP-specific optimization of the algorithm’s parameter set i.e., they were used for training (Training data). The second data set comprised new audio recordings of known and new words not included in the optimization procedure but matching the optimized InitPs (test data). Within this test data set, a subset of audio files for which both, Kaldi time and Chronset_NL, were available, was selected (test subset), allowing for a precise evaluation and comparison of the different algorithms’ performances.

The training data used for the optimization of the parameter set for the FHNW_NL (for algorithm details, see method section 3) were

Table 1

Description of populations of healthy subjects and persons with aphasia (mean age $\pm$ SD; subject number (female/male)). Data from the two studies [14,20] was used for setting up the optimization and testing.

	Study 1	Study 2
HS N(f/m)	102 (62/40)	32 (16/16)
mean age $\pm$ SD	41.8 $\pm$ 17.6	59.3 $\pm$ 12.6
PwA N(f/m)	N/A	31 (14/17)
mean age $\pm$ SD		58.6 $\pm$ 13.8

N/A: not available; HS: Healthy Subjects; PwA: Persons with Aphasia; SD: Standard Deviation; N: Subject number; f: Female; m: Male.

² <https://kaldi-asr.org/>.

³ Rickert, E., Wyss, S., Altermatt, S., Degen, M., Widmer Beierlein, S., Raymond, C. M., & Hemm, S. (2021, November). App for naming exercises with automatic feedback for aphasia patients. 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society October 31 – November 4, 2021.

⁴ FHNW is the acronym for University of Applied Sciences and Arts Northwestern Switzerland (Fachhochschule Nordwestschweiz).

Table 2
Training and test audio data.

	Number of Audios (Number of different words)
	Overall
Training data	1603 (39)
Test data	7426 (77)
└ Test subset	1329 (34)
Total	9029 (86)

composed of the correctly recognized audios of the 39 available words in our Kaldi models [21]. Details about the number of words and audios for healthy subjects and persons with aphasia for the different InitPs and InitP groups can be found in Table 3. Words used for the analysis are listed in the appendix (Tables A.1, B.1 and C.1).

3. Methods

In this section, we introduce the different naming latency detection tools used in the present paper, i.e., in more detail the newly developed, InitP-based detection algorithm and then the existing detection methods Kaldi and Chronset, and their application and performance evaluation.

3.1. Initial phoneme-based naming latency detection (FHNW_NL)

3.1.1. Algorithm description

The algorithm is designed to detect the onset of spoken responses—referred to as naming latencies—within short audio recordings. It is an enhancement of the previously presented extended threshold algorithm [24] and was implemented in Matlab⁵. As illustrated in Fig. 1, it builds on a threshold-based method and operates within a fixed detection window that spans from 100 ms before to 1000 ms after a given reference time. This reference time (Fig. 1B) corresponds to the speech onset as detected by an automatic speech recognition system.

To begin, the audio signal (Fig. 1A) is converted to a mono channel and its amplitude envelope is extracted (Fig. 1D1). This envelope is then denoised using wavelet-based filtering to suppress high-frequency fluctuations and non-speech artifacts. From the resulting signal, several features are computed: a moving maximum (capturing short-term peaks in energy) and derived from it a moving slope (measuring the local rate of amplitude change), and a smoothed moving average (providing an estimate of the overall signal trend). These features are normalized and scaled appropriately, see (Fig. 1D1).

The detection algorithm is parameterized with phoneme-specific values (Fig. 1C). These parameters include thresholds applied to the amplitude envelope (scaled by an overall Median Factor) and to the normalized slope of the envelope (Threshold Slope). The overall median factor scales the median of a smoothed envelope, defining the amplitude threshold for speech onset. The threshold slope determines the minimum slope required to indicate rising signal energy. Temporal parameters control how gaps in the envelope or slope are handled: Time Full Gap defines the maximum duration of low-amplitude frames permitted before a segment is considered finished, while Time Slope Gap After Beginning restricts the allowable duration of low-slope frames after an initial grace period. This grace period is set by Time After Beginning, allowing transient signal fluctuations immediately after onset. A final constraint, Min Word Length, enforces a minimum segment duration. These parameters collectively enable robust detection by accommodating brief low-energy fluctuations and natural intra-word pauses common in speech. With preprocessing complete, the algorithm (Fig. 1D3) scans the detection window (Fig. 1D2) in search of a valid onset. It identifies the first point where both the amplitude and slope signals

exceed their respective thresholds (Median Factor, Threshold Slope). Upon detection, it verifies whether this threshold-crossing is sustained for a short minimum duration (Time After Beginning) and ensures that any subsequent dips below the thresholds are brief and remain within predefined limits (Time Full Gap, Time Slope Gap After Beginning). If these conditions are met and the total detected segment duration exceeds the minimum word length (Min Word Length), the algorithm confirms the detection and returns the onset time in milliseconds (Fig. 1E).

If, however, these conditions are not satisfied—for instance, if the signal falls below threshold too quickly or the segment is too short—the algorithm continues its search further along the timeline until either a valid onset is found or the end of the detection window is reached.

In summary, the algorithm combines amplitude and slope-based criteria within a constrained time window to robustly identify the onset of speech. Its design incorporates tolerance for intra-word variability while minimizing false positives due to transient noise or irrelevant fluctuations.

3.1.2. Parameter set training for InitP

To ensure the most accurate aNL detection, the above-mentioned features of the parameter set were optimized for each InitP. The mNL was used as gold standard and reference for the optimization. The Nelder Mead algorithm (fminsearch Matlab function⁶) was applied to minimize the median difference between the mNL and the FHNW_NL detected by the algorithm. This optimization was performed using the training data, which included data from 82 healthy subjects and 28 persons with aphasia (Table 3). The optimization algorithm ran 16 times, once for each available InitP. Ultimately, the phoneme-specific thresholds were consolidated into a 16×6 parameter matrix — representing 16 initial phonemes and 6 parameters each — ready to be used by the FHNW_NL detection algorithm. Results can be found in the appendix E.

3.2. Kaldi time detection

For the detection of the Kaldi time, corresponding to the identified start of the detected word, the Standard German Kaldi model built with the Kaldi ASR toolkit version 5.5 and described in Wyss et al. [21] was used. This model was extended with recordings from an additional audio data collection for the speech model training and a specifically set-up Swiss German model.

3.3. Chronset naming latency detection

To perform naming latency detection with Chronset, the Chronset tool (version: Chronset_July14_2017⁷) was used in conjunction with MATLAB R2014a. The analysis required the installation of essential MATLAB toolboxes, including the Parallel Computing Toolbox (v6.4), the Signal Processing Toolbox (v6.21), and the Statistics Toolbox (v9.0). The audio files were batch-processed in folders containing 20 files each to streamline workflow and ensure consistent handling of data sets. The Chronset_NL was determined for audios of the test subset.

3.4. Manual latency detection

The mNLs of all the audios ($n = 9029$) used for aNL detection (study 1 and 2) were manually determined by trained observers using speech analysis software called PRAAT⁸.

⁵ <https://mathworks.com>.

⁶ <https://www.mathworks.com/help/matlab/ref/fminsearch.html>.

⁷ https://www.bcbi.eu/databases/chronset/chronset_July14_2017.zip.

⁸ <https://www.praat.org>.

Table 3

Number of audios per initial phoneme group and initial phoneme from healthy subjects and persons with aphasia for the training of the algorithm (train), testing (test) and the test subset (subset) for evaluation and comparison of the algorithms. Data from healthy subjects were extracted from both study 1 and study 2, data from persons with aphasia from study 2 only.

InitP Group	InitP	Healthy Subjects						Persons with Aphasia					
		InitP Group Audios			InitP Audios			InitP Group Audios			InitP Audios		
		Train	Test	subset	Train	Test	subset	Train	Test	subset	Train	Test	subset
Affricate	/ts/	35	129	0	35	129	0	0	19	0	0	19	0
Approximant Liquid	/l/	37	435	107	37	435	107	8	74	9	8	74	9
Approximant Vibrant	/r/	116	658	156	116	658	156	21	84	11	21	84	11
Fricatives	/f/				38	182	12				7	17	0
	/h/				35	177	15				10	21	0
	/z/	531	2632	374	139	766	64	82	332	32	26	113	7
	/ʃ/				245	1201	142				30	136	6
	/v/				74	306	141				9	45	19
Nasals	/m/	144	440	109	67	285	27	13	66	9	8	38	0
	/n/				77	155	82				5	28	9
Plosives	/b/				111	714	36				20	104	2
	/g/				79	560	133				19	74	8
	/k/	485	2173	424	115	470	138	85	289	25	16	59	9
	/p/				37	87	66				8	11	5
	/t/				143	342	51				22	41	1
Vowel	/e/	38	81	69	38	81	69	8	14	4	8	14	4
Total		1386	6548	1239	1386	6548	1239	217	878	90	217	878	90

3.5. Data analysis

First, differences between the FHNW_NL, Kaldi time and Chronset_NL to the mNL (FHNW_NL-mNL, KT-mNL, and Chronset_NL-mNL respectively) were calculated for all available data.

For FHNW_NLs these differences were analyzed on the InitP group and InitP level for training and test data. Within the test data, the performance of the FHNW_NL algorithm was further investigated for the subgroups of healthy subjects and persons with aphasia. Median, mean and data distribution for the different InitP groups were compared.

For the test subset, i.e., all test audios for which Kaldi time and Chronset_NL were available, the obtained differences of FHNW_NL, Kaldi time and Chronset_NL respectively to the mNL were compared. All results were analyzed with respect to the inter-rater variability of the mNL detection (± 15 ms, unpublished data see [appendix D](#)). The Wilcoxon test was used to investigate statistically significant differences because the data were not normally distributed. $P < 0.05$ was considered significant.

Signed means and standard deviations of the differences to mNL were calculated and compared overall and for the test and training data ([Table 2](#)). In addition, absolute differences were determined to quantify the precision of the aNLs [16].

Details regarding the specific words used for training and testing, as well as those included in the test subset, are provided in [Tables A.1, B.1 and C.1 in the appendix](#).

4. Results

4.1. Comparison of FHNW naming latencies in training and test data

[Fig. 2](#) illustrates the differences between FHNW_NL and mNL for both training and test data. The boxes, representing the interquartile range (i.e., 50 % of the values), as well as the medians and means for both training and test data, remain largely within the inter-rater variability range (yellow-shaded area). Significant differences ($p < 0.05$) on the InitP group level exist for the overall comparison between training and test data and for fricatives and plosives ([Fig. 2A](#)). Within the fricative InitP group ([Fig. 2B](#)), /z/ and /ʃ/ show the highest number of outliers in the test data, which significantly differ from the training data (p

< 0.05). Similarly, within the plosive group ([Fig. 2C](#)), /k/, /g/, and /b/ exhibit numerous outliers in the test data, though a significant difference between the test and training data is observed only for /k/.

4.2. Comparison of FHNW naming latencies for healthy subjects and persons with aphasia

When splitting the test data from [Fig. 2](#) in persons with aphasia ([Fig. 3](#), pink boxes) and healthy subjects ([Fig. 3](#), green boxes) differences can be observed. Most boxes, means, and medians of healthy subjects remain within the inter-rater variability interval while boxes of persons with aphasia generally show a tendency towards positive values, i.e., a late detection with the aNL compared to the mNL. Data from persons with aphasia show a slightly larger distribution than those from healthy subjects. All groups except affricates show significant differences ($p < 0.05$) between healthy subjects and persons with aphasia. The highest difference can be observed in the voiceless fricative /ʃ/ and the voiceless plosive /k/ InitPs ([Fig. 3B](#) and [Fig. 3C](#)) for which differences of test data from training data have already been reported ([Fig. 2](#)). Due to the limited data from persons with aphasia, observations must be taken with care.

4.3. Comparison between FHNW_NL, Kaldi time and Chronset_NL

The comparison of FHNW_NL, Kaldi time, and Chronset_NL to mNL for the subset of test data shows greater differences for Chronset_NL and Kaldi time than for FHNW_NL ([Fig. 4](#)). In contrast to NLs identified by Chronset and Kaldi, more than 50 % of the FHNW_NLs (boxes) remain within the inter-rater variability range in all InitP groups. Data were significantly different in all InitP groups except fricatives and nasals between Kaldi and FHNW_NL, and in fricatives between Chronset and FHNW_NL ($p < 0.05$). Results from Kaldi and Chronset were always significantly different. These results are in accordance with the results from [Figs. 2 and 3](#), where FHNW_NLs show a higher variability for fricatives.

4.4. Overall comparison of signed means

[Table 4](#) summarizes the signed means and standard deviations of the

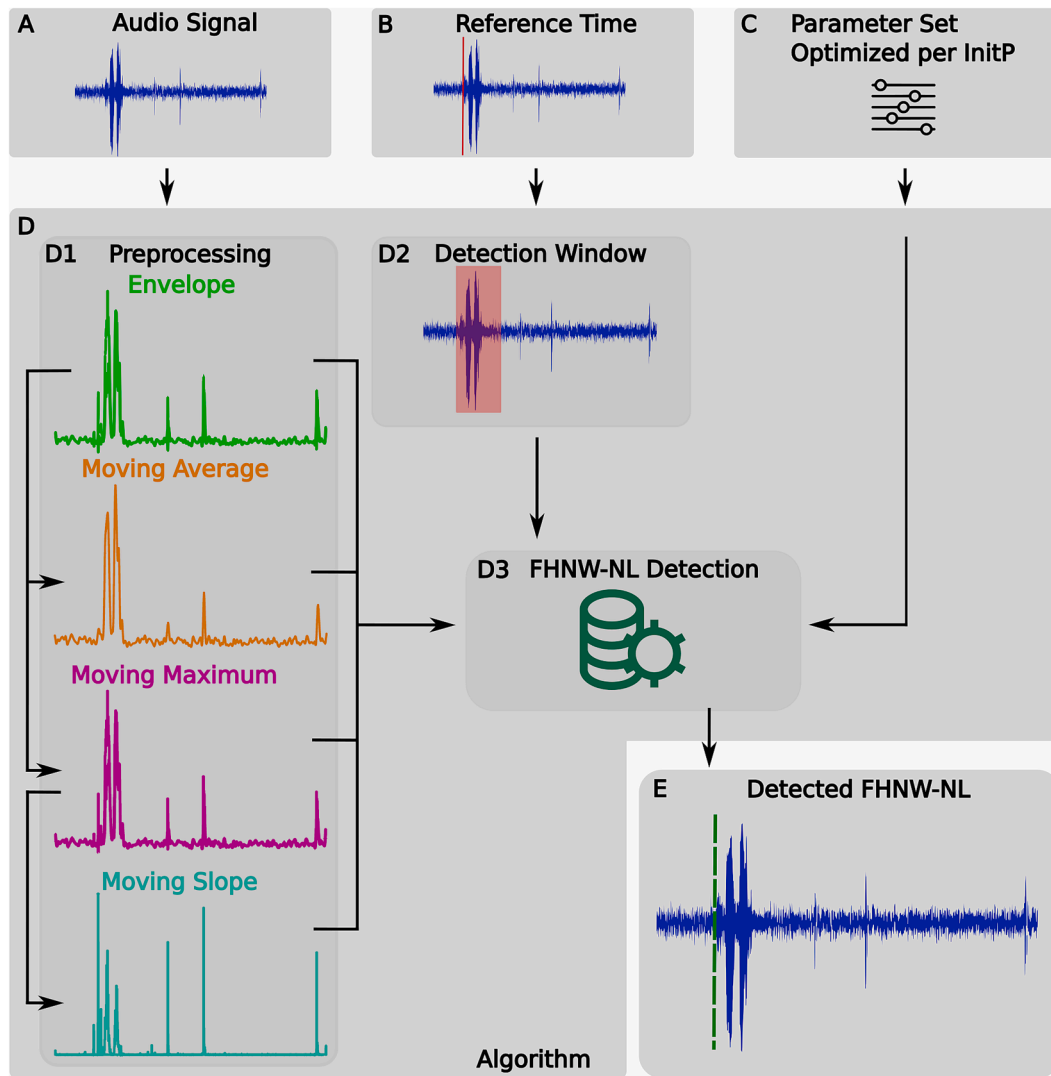


Fig. 1. Presented automatic naming latency detection process. A. Original audio signal as input for the algorithm. B. Reference time (manual naming latency or from automatic speech recognition tool) as the second input. C. Parameter set optimized for the corresponding initial phoneme as the third input for the algorithm. D. Different algorithm steps: D1. Processed signal results – The envelope from the audio signal from A is determined. Moving average and moving max are then derived from the envelope, and the moving slope is derived from the moving max. D2. Definition of the detection window based on the reference time from B (100 ms before to 1000 ms after the reference time) in which the algorithm searches for the beginning of the word. D3. Execution of automatic naming latency algorithm. E. Detected automatic naming latency from the algorithm, FHNW_NL.

differences between FHNW_NL, Kaldi time and Chronset_NL respectively and mNL for the training and test data sets from Figs. 2 and 4. The signed means for FHNW_NLs are in a similar range, regardless of whether the whole test data or the test subset are considered. However, the means of the test data are higher compared to the training data. The signed difference between Kaldi to mNL for the test subset is approximately 16 times higher than that of FHNW_NL, while for Chronset it remains in a similar range. In terms of absolute differences, means were more than twice as high for Kaldi and approximately four times as high for Chronset than for FHNW_NL. Standard deviations of Kaldi were about twice as high, and Chronset's nearly six times as high as that of FHNW_NL.

5. Discussion

The main objective of the presented study was to investigate the potential of an initial phoneme-based approach for automatic naming latency in naming exercises. Results for test data were compared on an

InitP level with those for training data and subdivided into healthy subjects and persons with aphasia. FHNW_NLs were compared to manually detected naming latencies, Kaldi times from the automatic speech recognition tool, and Chronset_NLs from the NL detection tool Chronset. The present feasibility study shows encouraging results: we demonstrate that using an InitP-based training of several NL detection parameters and starting from a reference such as Kaldi time can provide naming latencies close to mNLs, which are commonly referred to as the gold standard. In the test subset, FHNW_NL produced the smallest variance relative to the manual latency reference, markedly outperforming both Kaldi time and Chronset_NL.

In the following, the inter-rater variability, the automatic naming latency detection, and possible limitations are discussed.

5.1. Different speech onset detection approaches

Different approaches have been proposed for aNL detection over the last decades with the aim of reducing manual measurements of NL.

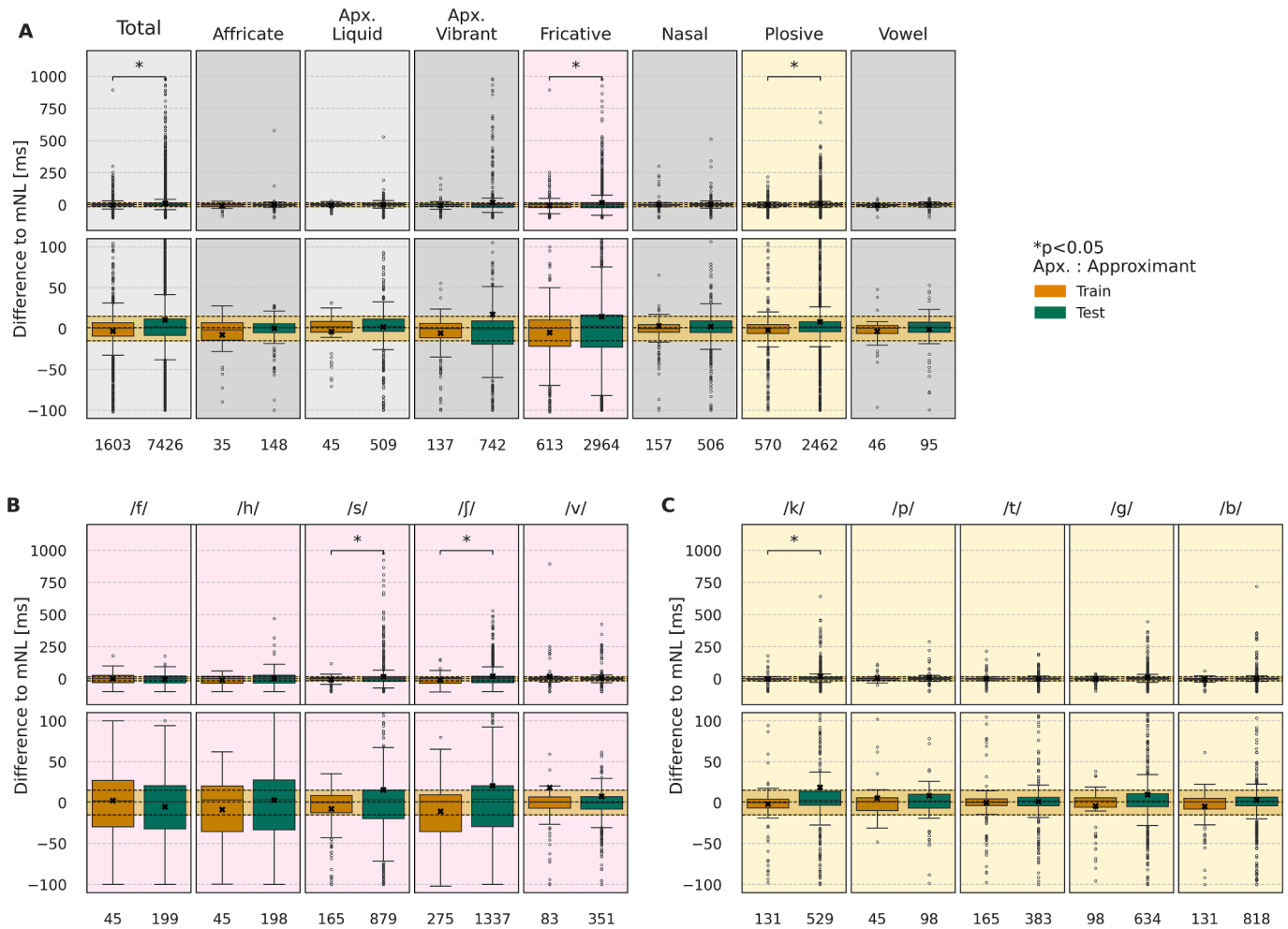


Fig. 2. Comparison of differences between automatic naming latency and manual naming latency for training and test data for the different initial phoneme groups (A) and different initial phonemes within the fricative (B) and plosive group (C). A, B and C comprise an upper part with overall distribution of all data points including outliers and a lower part with a zoom from -100 to $+100$ ms. The yellow area represents the standard deviation of the inter-rater variability interval in manual naming latency detection. The number of audios included in the data analysis of the respective subgroups are indicated at the bottom of each subfigure. * $p < 0.05$.

Boxplot details: Lower and upper edges of the boxes depict the boundaries between the 25th and the 75th percentile of the data, representing the interquartile range. The lines inside the boxes represent the median of the respective data, the crosses indicate the mean values. Whiskers above and below the boxes were set to 1.5 times the interquartile range. All data points outside this range are considered outliers. mNL: manual naming latency.

However, an optimal solution, especially for InitPs with low signal energy like voiceless InitPs, has not yet been identified. For voice key threshold-based aNL detection considering signal amplitude, some phonemes reach the target amplitude faster than others [23]. Even with an additional threshold for the detection of low amplitude audio signals (e.g., voiceless InitPs), acoustic onsets of a syllable cannot reliably be detected by voice key [15]. More complex software tools have been presented in the literature. They use the Fast Fourier transform spectrum with an amplitude threshold integrated in the semi-aNL detection tool “SayWhen” [17] or multiple spectral features from the audio waveform (Amplitude, Wiener entropy, spectral change, amplitude modulation, frequency modulation, harmonic pitch) for speech onset detection in “Chronset” [16]. In the latter case, speech onset is considered when four features are simultaneously above thresholds determined by optimization. To detect unvoiced onsets, they shift speech onset to the moment when the audio signal amplitude is above a predefined threshold within a specific time window before the detected onset. The tool flags data for manual correction in case of unusually short or long onset times. Another approach combines automatic speech recognition and a naming latency detection tool to first segment the audio signal into phonetic parts, identify a window of interest, and then apply aNL detection: [26]

provided this approach with the semi-aNL detection “AlignTool” based on PRAAT and the automatic speech recognition toolkit MAUS⁹ (Munich Automatic Segmentation) [27,28]. These existing aNL algorithms are time- and resource-intensive due to their dependence on spectral information. Therefore, we proposed an algorithm based exclusively on the time-based waveform which is more lightweight.

Accordingly, our previously proposed extended threshold-based aNL detection algorithm considered three parameters of the speech wave envelope [24] without spectral analysis, aiming to provide an algorithm with low computing time and the possibility to integrate it in a patient app. Even though the aNL showed a strong positive correlation with the mNL, aNLs confirmed sensitivity to different InitP groups linked to different positions in articulation. The use of the speech envelope revealed an influence of the signal energy of the InitP. Initial high signal energy was often responsible for early detection by the algorithm. To reduce this sensitivity, further detection parameters have been added to the optimization in the present study, and training was performed individually for each InitP. The slope threshold, which depends heavily

⁹ <https://www.bas.uni-muenchen.de/Bas/BasMAUS.html>.

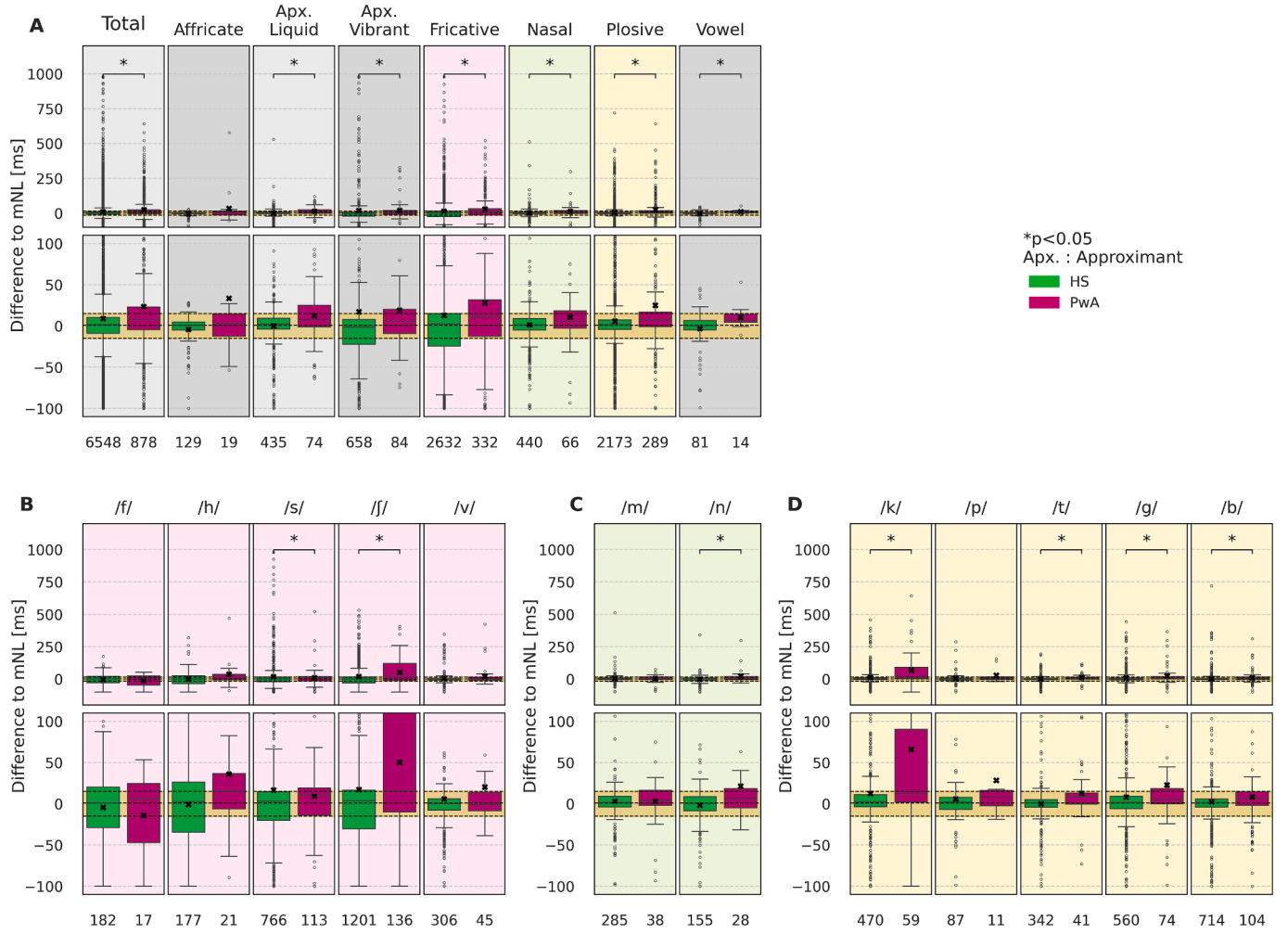


Fig. 3. Comparison of differences between FHNW_NL and manual naming latency between healthy subjects and persons with aphasia in the test data set. Negative values indicate an earlier detection, positive a later detection with respect to the corresponding manual naming latency. Detailed numbers of subjects, words and initial phoneme group InitPs can be found in Table 3. A. Overall and InitP group-level distribution of all determined differences, including all outliers. B. Boxplots for InitP Group Fricatives, C. Nasals and D. Plosives. Significant differences between healthy subjects and persons with aphasia are indicated. The yellow area represents the standard deviation of the inter-rater test in manual naming latency detection. Numbers at the top present the number of audios available for the respective subgroup. For a detailed explanation about the boxplots see Fig. 2. * $p < 0.05$.

on the signal envelope and in consequence on the InitP, is likely responsible for the improved onset detection. Implementing two thresholds allows for a lower amplitude threshold than in a simple amplitude-based approach. Additional parameters for word end and silence help avoid transient noise detection. Integrating more words into the training should further enhance onset detection and reduce outliers.

Furthermore, we decided to use a detection window around a reference time, similar to [26] in which the aNL is searched for. Depending on availability, Kaldi time or mNL were used. The overall improvement from Kaldi time to FHNW_NL (Fig. 4) confirms the potential of integrating onset times from an automatic speech recognition tool as the reference time. An additional advantage of using such a tool is that the targeted word can be identified even if the patient says more than the expected word during picture naming. If the word cannot be found within this detection window based on the InitP specific thresholds, it could be noted as aNL not available.

In this paper we included speech data from both healthy participants and participants with aphasia; but the presented approach may be highly relevant for other applications, such as psycholinguistic language production experiments with healthy speakers, especially when collecting data in online settings [29,30], which should not be resource-intensive.

5.2. Comparison of output of speech onset detection approaches

A comparison of the detected speech onsets by *AlignTool*, *SayWhen*, and *Chronset* [26] revealed differences depending on the complexity of the words. The aNL was determined by each tool for data from [15] including audio from 24 subjects for two groups of 20 words, simple words with the InitP /s/ and complex words with the InitP /s/ followed by a /t/ or /p/ phoneme. *Chronset* performed best for the simple words (signed mean difference ($mean\ aNL - mean\ mNL$): 8 ms) but was as inaccurate (signed mean difference: 144 ms) as *SayWhen* (signed mean difference: 86 ms) for complex words. Our study which is composed of simple words is with -3 ms close to this reported result for *Chronset*. Results with *AlignTool* were nearly similar for both word groups (signed mean difference: -36 ms (simple words) and -30 ms (complex words) respectively). The results demonstrate an improvement compared to voice keys, which has been reported to trigger over 100 ms after the mNL for the InitP /s/ [23]. The authors proposed the introduction of multiple features and the use of a reference time to improve aNL detection.

The potential of such an approach is confirmed by our results: the mean of the signed differences of 11 ms (overall test data) and 2 ms (test subset) between FHNW_NL and mNL is lower than or close to those

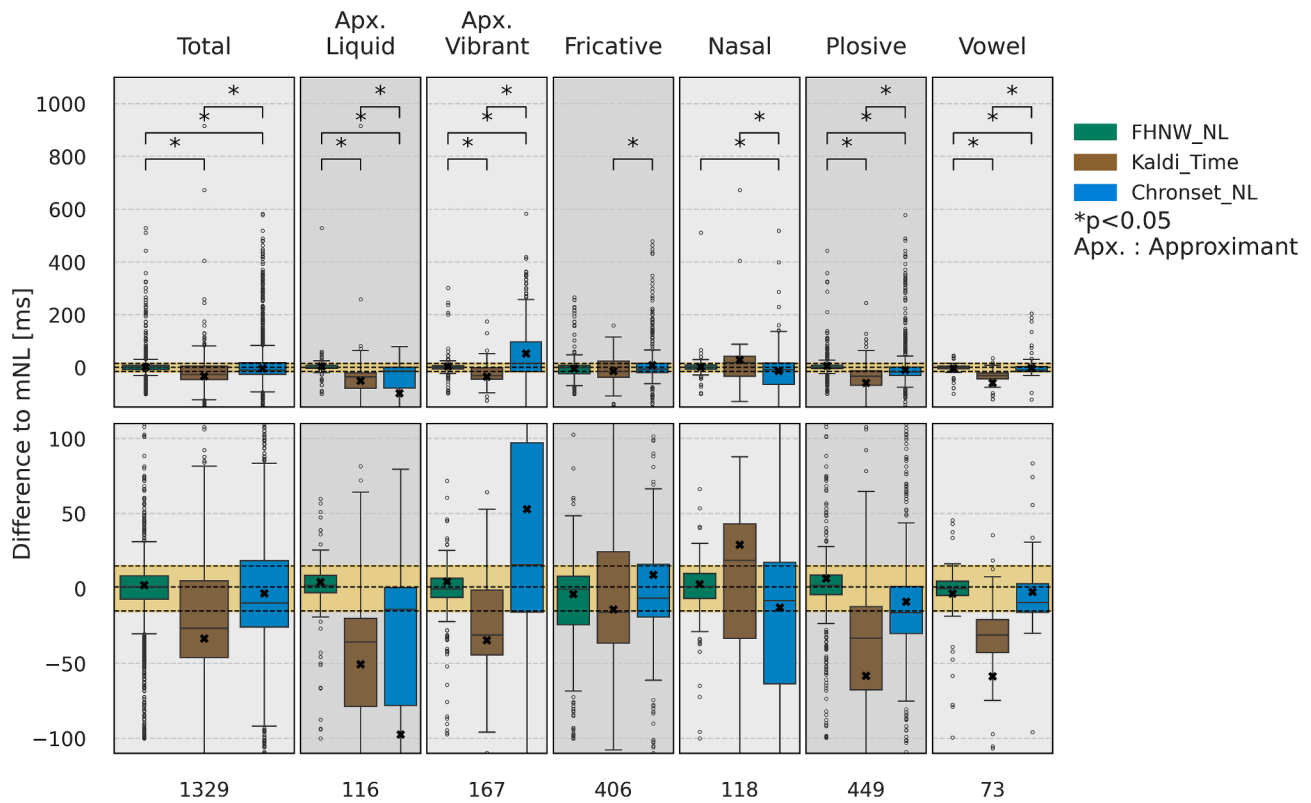


Fig. 4. Overall comparison of the differences in manual naming latency in milliseconds from FHNW_NL (green), Kaldi time (brown) and Chronset_NL (blue) for the test subset (see Table 3). A. Distribution of all determined differences on the overall and for the different InitP groups including all outliers. B. Zoom into A from -150 to 150 ms. Negative values indicate earlier detection, while positive values correspond to later detection compared to the corresponding manual naming latency. The yellow area represents the standard deviation of the inter-rater variability test in manual naming latency detection. Detailed explanations of the boxplots can be found in Fig. 2. $*p < 0.05$.

Table 4

Mean and standard deviations of signed and absolute differences between FHNW_NL, Kaldi time and Chronset_NL to the manual naming latency.

Training and test data			Test subset (n = 1329)	
Data	FHNW_NL – mNL mean $\pm$ SD [ms]	FHNW_NL – mNL mean $\pm$ SD [ms]	Model	mean $\pm$ SD [ms]
Training data(n = 1603)	-3.1 ± 43.1	20.6 ± 38.0	FHNW_NL – mNL	2.1 ± 51.0
Test data(n = 7426)	10.5 ± 76.8	33.4 ± 70.0	FHNW_NL – mNL	23.3 ± 45.4
Overall(n = 9029)	8.1 ± 72.2	31.1 ± 65.6	Kaldi time – mNL	-33.4 ± 109.9
			Kaldi time – mNL	57.2 ± 99.6
			Chronset_NL – mNL	-3.4 ± 296.1
			Chronset_NL – mNL	99.9 ± 278.8

FHNW_NL: automatic Naming Latency obtained with new algorithm; mNL: manual Naming Latency; KT: Kaldi Time; Chronset_NL: automatic Naming Latency obtained with Chronset; SD: Standard Deviation. Negative values indicate an earlier detection; positive values indicate later detection compared to the corresponding manual naming latency.

mentioned above (Table 4). Our comparison of FHNW_NLs with Chronset_NLs for the same data set confirms the superiority of the InitP-based approach. While FHNW_NL stays within the inter-rater variability interval, it varies for Chronset_NL depending on the InitP groups. The same is true for the Kaldi time, which tends towards negative values, i.e., earlier detection in all InitP groups except nasals. This may be due to a low detection threshold and the aim not to miss any syllable for automatic speech recognition. Overall, when examining the precision for our test subset (in terms of absolute differences), FHNW_NL performs twice as well as Kaldi and four times as well as Chronset. Nevertheless, it is important to keep in mind the limitations of the test subset (see section 5.4).

5.3. Initial phoneme-based naming latency detection

FHNW_NL data (Fig. 2 to Fig. 4) showed that more than 50 % of the aNLs remain within our inter-rater variability interval of ± 15 ms for most InitP groups. Comparing individual InitP groups, we improved the algorithm's performance presented by Park et al. [24]. Nasal sounds, which showed the largest difference to mNL, could be improved from $101.7 \text{ ms} \pm 56.0 \text{ ms}$ to $2.6 \pm 40.0 \text{ ms}$. Higher variation of FHNW_NL in test data is observed in the fricative group due to differences in its five InitPs (Fig. 2). Voiced and voiceless InitPs may contribute to this variation, as reported by Kessler et al. [23]. The low energy signal of voiceless fricative InitPs (e.g., f, h, j) can be influenced by following phonemes in generating the speech envelope [24].

Our analysis revealed high variations in FHNW_NL for the InitP /f/ within the fricatives, where the following phoneme—such as in /fa/, /ft/,

and /j/—influences the acoustic characteristics of the /j/ segment. This confirms the observation made by Rastle et al. [15], showing that the following phonemes, e.g., the second and third phonemes, also have an influence on the speech wave and in consequence on the speech envelope. The second phoneme whether a vowel, e.g., /a/, or a consonant, e.g., /t/, /p/, affects the generation of acoustic signals, thresholds and energy with the /j/ InitP. The acoustic energy of /j/ InitP varies depending on the acoustic characteristics of the combination of initial and following phonemes. Consequently, optimal onset detection parameters might differ for the three combinations of /j/ InitP and the following phoneme. Independent optimization would likely improve these results in fricatives.

High variations for /f/ and /h/ can be explained by the fact that only audio of one word (appendix Table A.1) were used per InitP for optimization, making the parameters too specific to generalize to all test data. Additionally, only a small number of audios was available for training, including at least 10 % from data from persons with aphasia. Fig. 3 shows that differences between FHNW_NLs and mNLs from persons with aphasia are higher than in healthy subjects. This may be due to the limited availability of words from persons with aphasia and their lesser integration into the training set. Furthermore, FHNW_NL of persons with aphasia showed a tendency for later detection than mNL. According to the previous study [24], delayed detections of aNL compared to mNL were mainly observed in voiceless fricative InitP containing weak energy. The acoustic feature of persons with aphasia may have less or weaker energy intensity compared to healthy subjects, causing the algorithm to generate its envelope with a delay. Future studies should plan equal amounts of data for both groups and the same words with controlled initial and following phonemes to further investigate their influence.

5.4. Limitations

The present study has the following limitations mainly based on the available data sets: data used were from existing studies and collections with a limited and varying number of words and audios for only some InitPs (Table 3). Despite the small training sets, we improved our aNL detection model and could show the feasibility of the initial phoneme based latency detection. We achieved results closer to mNL than our previous studies [24] and other aNL detection tools as reported above. Comparison with results from other papers with existing tools is limited due to the different data sets, available InitPs, and analysis methods, ranging from percentage of values below a specific time thresholds [17], to differences between signed mean aNL and the signed mean mNL values [26] or the comparison of absolute differences [16]. Nevertheless, by applying a subset of our data to Kaldi and Chronset, a comparison between the different approaches has been made possible. However, the generalizability is limited as only 34 words within six InitP groups and mostly within Plosives and Fricatives (Table 2 and Table 3) were part of this subset.

Few data sets from persons with aphasia were available and included in this study, making comparison between healthy subjects and persons with aphasia and the applicability to persons with aphasia difficult. Despite this lack of data, the preliminary analysis shows that aNL detection is possible for persons with aphasia, even if differences of aNLs to mNLs vary slightly more for persons with aphasia than for healthy subjects.

With a larger data set covering all InitP groups equally, with more words and data from persons with aphasia, a more detailed analysis on the InitP level and a comparison with the other tools would be possible.

6. Conclusions

In this paper, we presented an algorithm based on a parameter set composed of six features for word onset detection optimized individually for different available InitPs and without the necessity of spectral analysis. We showed that with such an algorithm together with a

reference time and a detection window, the aNL detection quality increases compared to previously presented results and tools. Kaldi time or speech onset times provided by other automatic speech recognition tools could be used as reference time for identifying speech onset in the algorithm. A larger study and data collection with more words per InitP and more data sets per word for all possible InitPs might further improve and extend the obtained results. Further studies would be useful to explore multiple parameter optimizations for certain InitPs, such as /j/, which highly depend on the acoustic characteristics of following phonemes, in order to improve detection performance. Additionally, investigating the influence of PwA on the optimized parameter sets would be valuable. In conclusion, the developed initial phoneme based algorithm has the potential to improve automatic naming latency detection. It could be integrated into a mobile app, providing real-time aNL detection during picture naming tasks in persons with aphasia. This approach may be highly relevant for other fields in research and clinical settings, benefiting audios analysis from both patients and healthy subjects.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank the FHNW University of Applied Sciences and Art Northwestern Switzerland for funding the study E-Inclusion, the study participants for their motivated participation, and the speech and language therapy teams and departments of the practice partners for their dedicated collaboration: REHAB Basel, Reha Rheinfelden, Bad Schinznach (Privat-Klinik Im Park); Cantonal Hospitals Aarau, Baden, Basel-land, Graubünden and Olten (Solothurn Hospitals); Hospitals Schaffhausen and Zofingen; University Hospital Zurich (Clinic for Neurology) and Geriatric Medicine FELIX PLATTER; Stadtsipital Waid and Triemli; speech and language therapy practices Fanny Dittmann-Aubert, Isabelle Facchini-Baumann, LogoTreffpunkt, mund'art, rundum therapie, Petra Dietiker and Unterstrass.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.bspc.2025.108851>.

Data availability

The authors don't have the permission to share the audio data but Matlab code for FHNW_NL detection together with some example audio files (not from study subjects) for testing are available on <https://github.com/IM2Neuroing/Naming-Latency-Detection>. The set of graphic representations used for the naming exercise can be seen and downloaded on the following platform: <https://mediathek.hgk.fhnw.ch/einclusion/>.

References

- [1] M. Cotelli, B. Borroni, R. Manenti, M. Zanetti, A. Arévalo, S.F. Cappa, A. Padovani, Action and object naming in Parkinson's disease without dementia, *Eur. J. Neurol.* 14 (2007) 632–637, <https://doi.org/10.1111/j.1468-1331.2007.01797.x>.
- [2] K. Kohnert, Cognitive and cognate-based treatments for bilingual aphasia: a case study, *Brain Lang.* 91 (2004) 294–302, <https://doi.org/10.1016/j.bandl.2004.04.001>.
- [3] A. Kertesz, The Western Aphasia Battery: a systematic review of research and clinical applications, *Aphasiology* 36 (2022) 21–50, <https://doi.org/10.1080/02687038.2020.1852002>.
- [4] R.H. Brookshire, M.R. McNeil, *Introduction to neurogenic communication disorders*, 8th ed, Mosby, St Louis, MO, US, 2015.

- [5] J. Brumbi, A. Winkelmann, I. Ablinger, Der Einsatz der LetMeTalk-App bei globaler Aphasie, 17. Jahrestagung der Gesellschaft für Aphasieforschung und -behandlung (GAB), Berlin.
- [6] J.P. Johnson, E.L. Meier, Y. Pan, S. Kiran, Pre-treatment graph measures of a functional semantic network are associated with naming therapy outcomes in chronic aphasia, *Brain Lang.* 207 (2020) 104809, <https://doi.org/10.1016/j.bandl.2020.104809>.
- [7] W. Huber, K. Poeck, K. Willmes, *The Aachen Aphasia Test*, *Adv. Neurol.* 42 (1984) 291–303.
- [8] P. Bonin, M. Chalard, A. Méot, M. Fayol, The determinants of spoken and written picture naming latencies, *Br. J. Psychol.* 93 (2002) 89–114, <https://doi.org/10.1348/000712602162463>.
- [9] S. LaGrone, D.H. Spieler, Lexical competition and phonological encoding in young and older speakers, *Psychol. Aging* 21 (2006) 804–809, <https://doi.org/10.1037/0882-7974.21.4.804>.
- [10] M. Vitkovitch, L. Tyrrell, Sources of Disagreement in Object Naming, *The Quarterly Journal of Experimental Psychology Section A* 48 (1995) 822–848, <https://doi.org/10.1080/14640749508401419>.
- [11] M. Calabria, N. Grunden, M. Serra, C. García-Sánchez, A. Costa, Semantic Processing in Bilingual Aphasia: evidence of Language Dependency, accessed May 13, 2022, *Front. Hum. Neurosci.* 13 (2019), <https://www.frontiersin.org/article/10.3389/fnhum.2019.00205>.
- [12] S. Kargel, A. Rosenkranz, M. Merz, U. Domahs, F. Domahs, Latenzmessungen in der Diagnostik diskreter Benennstörungen - Eine praktikable Ergänzung zum Aachener Aphasie Test [Latency measurements for the assessment of discrete naming disorders - Supplementary normative data for the Aachener Aphasia Test], *Logos* 23 (2015) 92–99, <https://doi.org/10.7345/prolog-1502092>.
- [13] S. Mätzig, J. Druks, J. Masterson, G. Vigliocco, Noun and verb differences in picture naming: past studies and new evidence, *Cortex* 45 (2009) 738–758, <https://doi.org/10.1016/j.cortex.2008.10.003>.
- [14] C. Reymond, S. Widmer Beierlein, C. Müller, R. Reutimann, K.P. Kuntner, N. Falcon Garcia, I. Grumbinaite, S. Hemm Ode, M. Degen, F. Parrillo, S. Karlin, S. Park, M. Renner, A. Blechschmidt, Naming Images in Aphasia, in: *Effects of Graphic Representations and Photographs on Naming Performance in Persons with and without Aphasia*, *Aphasiology* 0, 2022, pp. 1–23, <https://doi.org/10.1080/02687038.2022.2064421>.
- [15] K. Rastle, M.H. Davis, On the complexities of measuring naming, *J. Exp. Psychol. Hum. Percept. Perform.* 28 (2002) 307–314, <https://doi.org/10.1037/0096-1523.28.2.307>.
- [16] F. Roux, B.C. Armstrong, M. Carreiras, Chronset: an automated tool for detecting speech onset, *Behav. Res.* 49 (2017) 1864–1881, <https://doi.org/10.3758/s13428-016-0830-1>.
- [17] P.A. Jansen, S. Watter, SayWhen: an automated method for high-accuracy speech onset detection, *Behav. Res. Methods* 40 (2008) 744–751, <https://doi.org/10.3758/BRM.40.3.744>.
- [18] C. Breitenstein, T. Grewe, A. Flöel, W. Ziegler, L. Springer, P. Martus, W. Huber, K. Willmes, E.B. Ringelstein, K.G. Haeusler, S. Abel, R. Glindemann, F. Domahs, F. Regenbrecht, K.-J. Schlenck, M. Thomas, H. Obrig, E. de Langen, R. Rocker, F. Wigbers, C. Rühmkorf, I. Hempen, J. List, A. Baumgaertner, A. Villringer, M. Bley, M. Jöbges, K. Halm, J. Schulz, C. Werner, G. Goldenberg, G. Klingenberg, E. König, F. Müller, B. Gröne, S. Knecht, R. Baake, J. Knauss, S. Miethe, U. Steller, R. Sudhoff, E. Schillkowski, G. Pfeiffer, K. Billo, H. Hoffmann, F.-J. Ferneding, S. Runge, T. Keck, V. Middeldorf, S. Krüger, B. Wilde, K. Krakow, C. Berghoff, F. Reinhuber, I. Maser, W. Hofmann, C. Sous-Kulke, W. Schupp, A. Oertel, D. Bätz, F. Hamzei, K. Schulz, A. Meyer, A. Kartmann, O. Som, S.-B. Schipke, S. Bamborschke, Intensive speech and language therapy in patients with chronic aphasia after stroke: a randomised, open-label, blinded-endpoint, controlled trial in a health-care setting, *Lancet* 389 (2017) 1528–1538, [https://doi.org/10.1016/S0140-6736\(17\)30067-3](https://doi.org/10.1016/S0140-6736(17)30067-3).
- [19] R. Palmer, M. Dimairo, C. Cooper, P. Enderby, M. Brady, A. Bowen, N. Latimer, S. Julious, E. Cross, A. Alshreef, M. Harrison, E. Bradley, H. Witts, T. Chater, Self-managed, computerised speech and language therapy for patients with chronic aphasia post-stroke compared with usual care or attention control (big CACTUS): a multicentre, single-blinded, randomised controlled trial, *Lancet Neurol.* 18 (2019) 821–833, [https://doi.org/10.1016/S1474-4422\(19\)30192-9](https://doi.org/10.1016/S1474-4422(19)30192-9).
- [20] S. Widmer Beierlein, C. Reymond, K.P. Kuntner, A. Blechschmidt, M. Degen, C. Müller, N. Falcón García, S. Altermatt, C. Elsener, S. Karlin, S.H. Park, R. Reutimann, F. Parrillo, S. Bucheli, I. Grumbinaite, A. Jochmann, M. Harvey, J. Lee, J. Loew, L. Meier, L. Poffet, M. Renner, S. Schiltknecht, A. Shah, G. Schneider, A. Strub, K. Trachsel, M. Winkler, S. Wyss, S. Hemm-Ode, E-Inclusion – Ein interdisziplinäres, Schweizerisches Aphasie-Forschungsprojekt (2021). <https://doi.org/10.26041/fhnw-4213>.
- [21] S. Wyss, E. Rickert, S. Altermatt, K.P. Kuntner, M. Degen, C.M. Reymond, S. Widmer Beierlein, A. Blechschmidt, S. Hemm, Patient-friendly speech recognition feedback for aphasia patients, in: *Conf Proc EMBEC* (2020) 283, <https://doi.org/10.26041/fhnw-4206>.
- [22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, The Kaldi speech recognition toolkit, in: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society, 2011.
- [23] B. Kessler, R. Treiman, J. Mullennix, Phonetic Biases in Voice Key Response Time Measurements, *J. Mem. Lang.* 47 (2002) 145–171, <https://doi.org/10.1006/jmla.2001.2835>.
- [24] S. Park, S. Altermatt, S.W. Beierlein, A. Blechschmidt, C. Reymond, M. Degen, E. Rickert, S. Wyss, K.P. Kuntner, S. Hemm, Evaluation of the potential of Automatic Naming Latency Detection for different initial Phonemes during Picture Naming Task, in: *In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2021, pp. 945–950, <https://doi.org/10.1109/EMBC46164.2021.9630770>.
- [25] E. Bates, S. D'Amico, T. Jacobsen, A. Székely, E. Andonova, A. Devescovi, D. Herron, C. Ching Lu, T. Pechmann, C. Pléh, N. Wicha, K. Federmeier, I. Gerdjikova, G. Gutierrez, D. Hung, J. Hsu, G. Iyer, K. Kohnert, T. Mehotcheva, A. Orozco-Figueroa, A. Tzeng, O. Tzeng, Timed picture naming in seven languages, *Psychon. Bull. Rev.* 10 (2003) 344–380, <https://doi.org/10.3758/BF03196494>.
- [26] L. Schillingmann, J. Ernst, V. Keite, B. Wrede, A.S. Meyer, E. Belke, AlignTool: the automatic temporal alignment of spoken utterances in German, Dutch, and British English for Psycholinguistic Purposes, *Behav. Res.* 50 (2018) 466–489, <https://doi.org/10.3758/s13428-017-1002-7>.
- [27] F. Schiel, Automatic Phonetic Transcription of Non-Prompted Speech, in: *Proceedings of the XIVth International Congress of Phonetic Sciences : ICPHS 99*, San Francisco, 1999: pp. 607–610. <http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-13682-6>.
- [28] F. Schiel, A statistical model for predicting pronunciation, in: *Proc. of the International Congress of Phonetic Sciences*, Glasgow, 2015.
- [29] A. Fairs, K. Strijkers, Can we use the internet to study speech production? yes we can!, Evidence Contrasting Online versus Laboratory Naming Latencies and Errors, *PLOS ONE* 16 (2021) e0258908, <https://doi.org/10.1371/journal.pone.0258908>.
- [30] A. Vogt, R. Hauber, A.K. Kuhlen, R.A. Rahman, Internet-based language production research with overt articulation: Proof of concept, challenges, and practical advice, *Behav. Res.* 54 (2022) 1954–1975, <https://doi.org/10.3758/s13428-021-01686-3>.