Evaluating the Recognition Performance of the RehaLingo Speech Training System with Aphasic Speech

Hans-Günter Hirsch, Yannic Tiggelkamp, Christian Neumann, Tobias Bolten

Institute for Pattern Recognition, Niederrhein University, Krefeld, Germany
Email: {hans-guenter.hirsch}@hs-niederrhein.de

Abstract

RehaLingo is a speech training system designed for individuals with aphasia. Different versions of RehaLingo were tested and employed at a therapy center to support the speech therapy of patients. The system simultaneously utilizes multiple speech recognition frameworks to analyze user input. Speech data were recorded during therapy sessions. This paper focuses on evaluating the performance of the recognition systems with regard to typical aphasic speech artifacts.

1 Introduction

Individuals with aphasia, for example after a stroke, require a long rehabilitation process to regain the ability to associate spoken or written words with their meanings and to use them in everyday communication. We developed a speech training system [1] to support and accelerate this rehabilitation process, providing an additional option alongside the cost-intensive sessions with speech therapists.

The integration of speech processing and recognition technologies into computer-based therapy systems has been explored in previous studies [2, 3]. Furthermore, aspects such as user interface design [4] and usability evaluation [5] are crucial during system development. This paper presents the concept and implementation of our training system, which differs from existing approaches [6, 7, 8, 9] by employing multiple speech recognition systems in parallel. The system runs as a standalone solution on local hardware, eliminating the need for communication with external servers.

In this paper, we describe the setup of the training system and present recognition results obtained from speech data recorded both from individuals without aphasia and from therapy sessions with patients at a therapy center [10].

2 Speech Training System

The hardware used in this project consists of a 2-in-1 laptop and a ReSpeaker microphone array [11]. The laptop operates in tablet mode, enabling patients to interact with the graphical user interface (GUI) via the touchscreen. The ReSpeaker microphone features an array of four microphones and an integrated processing unit, allowing for hands-free speech input while reducing stationary noise signals.

The system requires sufficient computational resources to run three speech recognition modules in parallel, ensuring that results are delivered without significant delay. The internal architecture of the training system is illustrated in Figure 1.

2.1 Training Modes

Aphasia is a neurological condition that impairs an individual's ability to find and produce the correct word corresponding to a given item. The severity and nature of this impairment can vary significantly. In some cases, the primary challenge lies in word retrieval, while in others, additional difficulties with pronunciation may be present.

We developed graphical user interfaces (GUIs) for several training modes. To support the comprehension of spoken words, interfaces were created in which the patient must assign an image or a text label to the auditory representation of a physical item. In this paper, however, we focus on the production task, in which an image – or an image together with the corresponding text – is presented, and the patient is prompted to produce the appropriate word. An example of this GUI is shown in Figure 2, which includes the image representing the target item and a button to initiate speech recording.

The GUIs were designed according to the general concept of incorporating therapist-like behavior. Patients with aphasia are accustomed to direct interaction with a therapist who provides personalized feedback, which is critical for effective rehabilitation.

Therefore, the GUIs developed in this study aim to simulate therapist-patient interaction. For example, short video clips of a therapist are used to introduce tasks and provide individualized feedback. After completing a production task, the patient has the option to listen to the correct word and its pronunciation. A video clip of the therapist is shown, articulating the correct word.

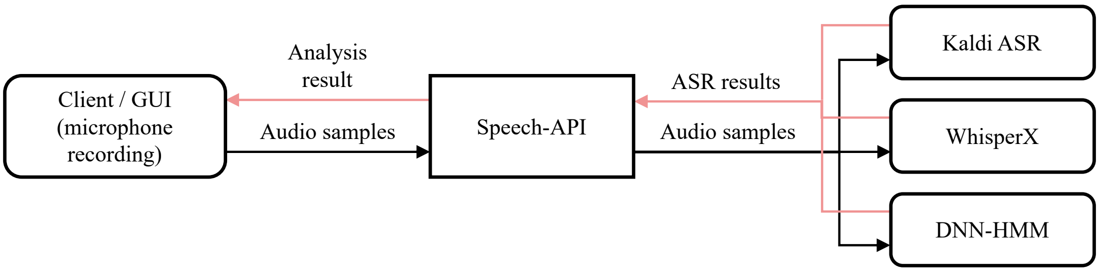


Figure 1: Internal structure of the training system.

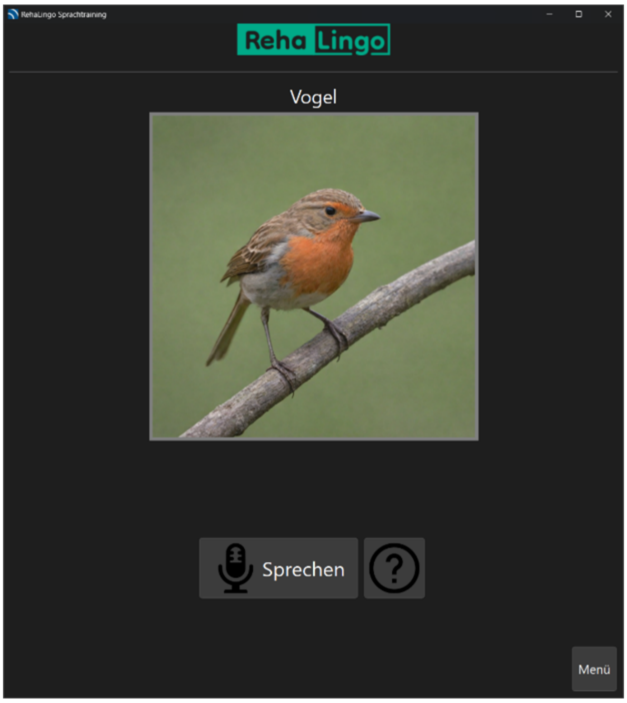


Figure 2: GUI for the speech production task.

2.2 Speech-API

We developed a Speech-API to support the use of multiple speech recognition modules. The recognition results are analyzed and categorized with respect to typical aphasic speech artifacts. During training, an image is shown to individuals with aphasia, prompting them to say the corresponding word. In most cases, the speech input does not consist solely of the correct word or a single utterance. Instead, the input can be categorized into several distinct classes [12]:

- The correct word is uttered along with filler words and hesitations.
- The correct word is produced, but with significant pauses between syllables.
- The correct word is used, but with incorrect grammatical number (singular/plural).
- The speech input is excessively long, although it includes the correct word.
- A semantically similar expression is provided, including:
  - Hyponyms (more specific terms),
  - Hypernyms (more general terms), or
  - Related semantic substitutions.
- A phonetically similar sequence of sounds is uttered, resulting in phonetically similar words or nonwords.
- The response is an unspecific reaction that does not include the target word, such as “I have seen this before.”

Derived from the typical aphasia artifacts, the Speech-API is responsible for mapping the speech input to one of the following nine categories:

- correct
- incorrect
- phonetically similar
- semantically similar
- hypernym
- hyponym
- split word (pauses between subwords or sylabbles)
- wrong numerus (singular/plural)
- excessively long utterance.

The GUI is designed to provide patients with individualized feedback for each case. For example, if the recognized word is phonetically similar to the target, a video of the therapist is shown, explaining that the word sounds similar but is not correct, and offering the patient the opportunity to try again. The categorization of the user’s speech input is the main task of the Speech-API. This is primarly done by majority voting of multiple recognition results. But some special properties of the recognition modules have to be considered as extension of the majority voting. For example, the DNN-HMM modules is less suited to analyze a long speech input containing words unrelated to the target word. Whisper may provide words where the spelling is not correct. The word error rate (WER) is still an indicator for the usefulness of each applied recognition module. But, the correct categorization of the aphasia artifact is more important in this application.

Classification is based on a rule-based approach. Key rules include:

- Words spoken at the end of an utterance are prioritized over those at the beginning.
- In cases where the recognition modules disagree on classification, the result that is closest to the target word is preferred unless it is known that a particular recognizer performs poorly in the relevant category.

2.3 Recognition Modules

In addition to the previously mentioned artifacts, many individuals with aphasia face considerable difficulties in clearly and accurately producing the intended words or phrases. As a result, the automatic recognition of their speech input is a complex and challenging task.

To address this, the proposed system employs the parallel operation of multiple speech recognition algorithms to analyze and classify the input more effectively. Each recognition module offers specific features and strengths. The underlying idea is to combine these diverse approaches in a meaningful and synergistic way in order to achieve high recognition performance.

2.3.1 DNN-HMM

We employ our own implementation of a DNN-HMM-based recognition system, which combines a deep neural network (DNN) with a Hidden Markov Model (HMM), as described in [13]. The DNN estimates the probabilities of assigning speech segments to triphones, while the HMM uses these probabilities to compute the likelihood that an utterance corresponds to a word sequence from a predefined vocabulary. Recognition is guided by an item-specific grammar tailored to the expected input, which includes phonetically or semantically similar words related to the target item.

Each grammar is represented as a graph of nodes (words) and transitions (connections). Grammar generation begins by identifying words relevant to the task. To account for aphasia-related artifacts, the grammar incorporates phonetically and semantically similar words, paraphrases, neologisms, and unspecific responses in addition to the correct target word. Techniques for selecting these words are context-sensitive and allow for single or multi-word responses. For example, when shown an image of a bun, the response "bread" may also be considered valid. Detailed methods for word selection are described in [13].

The DNN-HMM algorithm is focused on recognizing the target item and related input patterns associated with known aphasic artifacts. However, it is less suitable for recognizing speech that does not refer to the target item at all. Therefore, two additional recognition systems are employed.

2.3.2 Kaldi Recognition System

The Kaldi recognition system [14] is used to process German speech with a large vocabulary, enabling the analysis of arbitrary spoken content. We employ a German Kaldi model developed and optimized by the Language Technology Group at the University of Hamburg [15]. For real-time inference, the Kaldi GStreamer server is utilized.

Although Kaldi performs well in recognizing unrestricted German speech, it has difficulties detecting out-of-vocabulary words, which are frequently encountered in aphasic speech. This limitation stems from its pre-trained language model. Potential future improvements include training a custom language model tailored to the specific characteristics of aphasic speech.

2.3.3 Whisper

Whisper, developed by OpenAI [16], is a robust speech recognition framework. In our system, we use the WhisperX implementation, which performs efficiently on CPUs. Unlike traditional vocabulary-based systems, Whisper recognizes word tokens that are subsequently combined into coherent text. This approach allows it to process typical aphasic speech, including mispronunciations that do not correspond to standard words.

However, Whisper is computationally demanding. Several model variants are available, each with significantly different requirements in terms of processing power and memory [16]. These range from the "tiny" model, based on a network with 39 million parameters, to the "large" model with 1,550 million parameters. High-accuracy models require powerful GPUs to achieve real-time performance. Fine-tuning existing models with task-specific data is also possible and has been explored in related work [17].

3 Data Acquisition

At the beginning of our investigations, we had no access to individuals with aphasia. With the assistance of speech therapists, we defined several lists of phrases containing typical aphasic speech artifacts for recording. We recorded approximately 1,330 utterances from 20 individuals without aphasia. Later, we had the opportunity to test our training system with patients at the therapy center in Lindlar [10]. Throughout 2024, we conducted tests with a total of sixteen patients over four sessions, each reflecting a specific stage of the system's development. Each patient participated in a one-hour session, supervised by a member of our team, on an individual basis. The primary focus of these tests was observing word production tasks. Insights and experiences gathered from each session, which typically involved four patients, were immediately incorporated into the system's ongoing development.

Whisper model	Base	Small	Medium	Large	Small-finetuned
WER/%	74.8	51.4	44.6	37.5	16.6

Table 1: WER for different Whisper models.

In total, we generated images and videos for approximately 180 items. The items were categorized into 10 thematic groups, such as animals and vehicles, from which one or more groups could be selected for a training cycle. During the initial patient sessions, the system was in the early stages of development and relied exclusively on Kaldi recognition. This occasionally resulted in recognition failures for spoken terms. The evaluation of whether a term was named correctly depended on subjective judgment. Errors occurred more frequently when participants produced only a single word, which can be attributed to Kaldi's design for analyzing and recognizing entire sentences in context. By the final session, all three recognizers were employed in parallel, leading to a substantial improvement in recognition accuracy. Errors in identifying correctly named terms in utterances were reduced to nearly zero.

The speech input during the production tasks was recorded with the patients' consent. Consequently, we created a small database containing 1,466 utterances to conduct offline recognition experiments. To facilitate annotation, we developed a graphical user interface for semi-automatically labeling the recorded utterances. Based on the outputs of the different recognition systems, we manually assigned the correct label information to each utterance.

In total, approximately 2,800 utterances are available, comprising 1,330 from speakers without aphasia and 1,466 from patients with aphasia, which were recorded during test sessions with our system. All data were recorded using the ReSpeaker microphone, which serves as the input device for our system.

4 Evaluation of Recognition

At the beginning, we started with our own DNN-HMM and Kaldi recognition systems. Our basic concept is to enable the incorporation of additional recognition modules. The goal is to allow patients the use of the training system wherever and whenever they want. Therefore, it must be able to operate locally in order to avoid communicating with external servers.

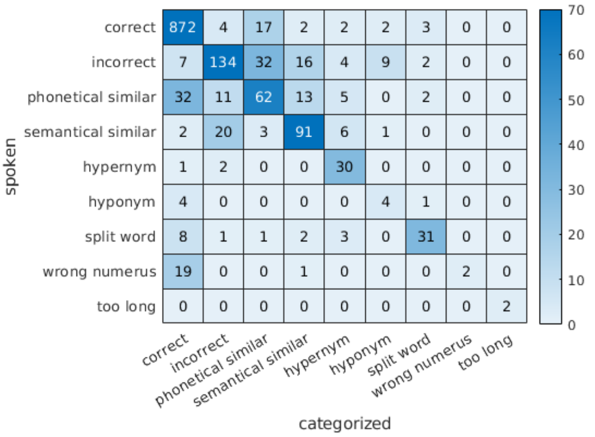
4.1 Recognition with Whisper

After the Whisper recognition system became available, we integrated it into our setup. We conducted recognition experiments with various Whisper models. For testing, we randomly selected 10% of the 2,800 recordings and used the remaining 90% to adaptively train the "small" Whisper model. This model was chosen because it performs recognition on the currently used hardware without noticeable delay. Word error rates, including insertions, are listed in Table 1 for the different Whisper models and the fine-tuned model.

As expected, the WER increases for less complex models with fewer parameters. Fine-tuning the "small" model significantly reduces the WER. Several factors

Recogn.	Whisper	Speech	Whisper	Speech
system	(without)	API	(with fine.)	API
Correct	42%	66%	84%	82%

Table 2: Percentages of correctly categorized utterances.



	correct	incorrect	phonetical similar	semantical similar	hypernym	hyponym	split word	wrong numerus	too long
correct	872	4	17	2	2	2	3	0	0
incorrect	7	134	32	16	4	9	2	0	0
phonetical similar	32	11	62	13	5	0	2	0	0
semantical similar	2	20	3	91	6	1	0	0	0
hypernym	1	2	0	0	30	0	0	0	0
hyponym	4	0	0	0	0	4	1	0	0
split word	8	1	1	2	3	0	31	0	0
wrong numerus	19	0	0	1	0	0	0	2	0
too long	0	0	0	0	0	0	0	0	2

Figure 3: Confusion matrix of category detection.

affect WER estimation. Whisper outputs a sequence of characters with spaces delimiting words. Without applying a specific language model, as Kaldi does, word errors may arise from orthographic differences, e.g., a single “n” instead of a double “nn”. Additionally, the final annotation was performed manually, so annotations of artifacts, such as hesitations, may differ from Whisper’s character feedback. Postprocessing would be necessary when applying the usual tools to estimate the WER. Due to these limitations, the character error rate (CER) can serve as an alternative metric. CER, which includes insertion errors, improves from 103% for the small model to 6.3% for the fine-tuned small model. Through adaptive training, the Whisper model adapts to the characteristics of the ReSpeaker microphone and some specific annotation conventions. The WER here primarily serves to compare Whisper models.

4.2 Recognition of Aphasia Artifacts

For patient communication within the user interface, it is crucial to correctly map the recognition results to the typical aphasic artifacts. To evaluate this, we manually annotated the 1,466 aphasic patient recordings with labels for the nine artifact classes as mentioned before.

We analyzed the mapping of 152 patients’ utterances that were not used for adaptive training of the Whisper small model. Kaldi correctly mapped 57% of these, DNN-HMM 44%. Percentages of correct mappings for Whisper without and with fine-tuning, as well as for the Speech-API (which combines DNN-HMM, Kaldi, and Whisper results), are presented in Table 2.

The “small” Whisper model without fine-tuning showed the worst performance. In this case, combining the three recognition results via the Speech-API improved performance relative to individual modules. Using the fine-tuned Whisper model increased the percentage of correctly categorized utterances substantially, from 42% to 84%. The combined analysis by the Speech-API performed slightly worse than the fine-tuned Whisper model alone. This is attributed to the large performance gap between Kaldi and DNN-HMM on the one hand, and the fine-tuned Whisper on the other. Further adaptation of the Kaldi model and improvements to the DNN-HMM system should reduce this disparity and potentially enhance combined recognition results.

The confusion matrix for the categorization into the nine aphasic artifact classes is shown in Figure 3. We examined misclassifications with respect to patients’ personal ratings and perceptions. Seventeen correct utterances were classified as phonetically similar, suggesting doubts about pronunciation accuracy. Thirty-two incorrect inputs were mapped as phonetically similar and sixteen as semantically similar, potentially causing patients to misjudge their responses as nearly correct. Twenty semantically similar utterances were categorized as incorrect, mainly due to incomplete semantic similarity lists. Thirty-two phonetically similar utterances were classified as correct, highlighting the challenges of fine-tuning recognition interpretations. These may depend on individual patients and therapy stages. Early in therapy, accepting phonetically similar utterances and providing positive feedback may be beneficial, while later stages should demand stricter pronunciation accuracy. The tolerance threshold should therefore be adjustable.

5 Outlook

Feedback from patients using our speech training system was predominantly positive, largely due to the system’s effective recognition of utterances containing the correct target word. Repeated system testing provided valuable insights for enhancing both recognition accuracy and the dialogue-based training process.

Thus far, the speech training has focused on word production tasks, in which patients are prompted to name individual items shown as images. These images were mostly generated using AI-based tools. Next, we plan to expand the training to include describing short scenes verbally, such as shopping or visiting a café. This will support patients’ reintegration into everyday social contexts.

To achieve this, we have created short, AI-generated video clips depicting simple scenes, such as two people greeting each other with a handshake or a hug. Instead of static item images, these clips are shown in the GUI. Then, the virtual therapist prompts the patient to verbally describe the scene.

This new training mode was recently piloted during a session at the therapy center, and speech data from patients were recorded for further analysis. As the task requires a higher level of linguistic competence, it is intended for patients in advanced stages of therapy. Depending on the type and severity of aphasia, the verbal response may range from a short sentence to a sequence of semantically related words. To process such inputs, speech recognition must be extended with natural language processing (NLP) techniques to extract meaning and assess intent. We plan to evaluate suitable NLP tools for this purpose (e.g. [18, 19]).

Acknowledgements

We acknowledge support by the Federal Ministry of Education and Research (BMBF) under grant no. 13GW0481D.

References

[1] H.G. Hirsch, Y. Tiggelkamp, C. Neumann, H. Frieg, S. Knecht, "Evaluating the User Interface of the RehaLingo Speech Training System with Aphasic Patients", in *Proc. Elektronische Sprachsignalverarbeitung*, Halle, Germany, Mar. 2025, pp. 61–68.

[2] L. Tuschen, "Einsatz von Sprachverarbeitungstechnologien in der Logopädie und Sprachtherapie", *Sprache· Stimme· Gehör*, 46(01), pp. 33–39, 2022.

[3] H. Frieg, J. Muehlhaus, U. Ritterfeld, and K. Bilda: ISi-Speech, "A Digital Training System for Acquired Dysarthria", in *Studies in Health Technology and Informatics*, vol. 242, pp. 330–334, 2017, doi:10.3233/978-1-61499-798-6-330.

[4] P. Cuperus, D. De Kok, V. De Aguiar and L. Nickels, "Understanding User Needs for Digital Aphasia Therapy: Experiences and Preferences of Speech and Language Therapists", *Aphasiology*, pp. 1–23, 2022. doi:10.1080/02687038.2022.2066622.

[5] H. Jakob, J. Pfab, A. Prams, W. Ziegler and M. Späth, "Digitales Eigentraining bei Aphasie: Real-World-Data-Analyse von 797 Nutzern*innen der App »neolexon Aphasie«", *Neurologie & Rehabilitation*, 28(2), pp. 61–67, 2022, doi:10.14624/NR2202002.

[6] M. Späth, E. Haas and H. Jakob, "neolexon-Therapiesystem", *Forum Logopädie*, 31(3), pp. 20–24, 2017, doi:10.2443/skv-s-2017-53020170304.

[7] J. Netzebandt, D. Schmitz-Antonischki and J. Heide, "Hochfrequente Wortabruftherapie mit LingoTalk: Eine Einzelfallstudie zum Eigentraining mit automatischer Spracherkennung", *Forum Logopädie*, 36(3), 2022, doi:10.2443/skv-s-202253020220303.

[8] E. Rykova, M. Walther, "AphaDigital – Digital Speech Therapy Solution for Aphasia Patients with Automatic Feedback Provided by a Virtual Assistant", in *Proc. Hawaii International Conference on System Sciences*, Honolulu, USA, 2024, pp.3385–3394

[9] TEMA Technologie Marketing AG: aphavox, 2018, `https://aphavox.de`.

[10] Logopädisch-interdisziplinäres Therapiezentrum Lindlar: `https://www.logozentrumlindlar.de/`

[11] ReSpeaker Microphone Array: `https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/`

[12] N. Stadie, S. Hanne, A. Lorenz, N. Lauer and D. Schrey-Dern, "Lexikalische und semantische Störungen bei Aphasie", Georg Thieme Verlag KG, 2019, doi:10.1055/b006-149440.

[13] H.G. Hirsch, C. Neumann, Y. Tiggelkamp, R. Fiorista, S. Knecht, A. Schnitzler, K. Biermann-Ruben, D. Bothe, G. Bleimann, H. Frieg, "Rehalingo – Towards s Speech Training System for Aphasia", in *Proc. Elektronische Sprachsignalverarbeitung*, Magdeburg, Germany, 2023

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit", in *Proc. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society, 2011.

[15] R. Geislinger, B. Milde, and C. Biemann, "Improved Open Source Automatic Subtitling for Lecture Videos", in *Proc. Conference on Natural Language Processing*, Potsdam, Germany, 2022, pp. 98–103.

[16] A. Radford, J.W. Kim et al., "Robust Speech Recognition via Large-Scale Weak Supervision", OpenAI, arXiv: 2212.04356v1, 2022.

[17] S. Gandhi, "Fine-Tune Whisper For Multilingual ASR with Transformer", `https://huggingface.co/blog/fine-tune-whisper`, 2022

[18] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python", 2020, doi:10.5281/zenodo.1212303, `https://github.com/explosion/spaCy`

[19] T. Bocklisch, J. Faulkner, N. Pawlowski, A. Nichol, "Rasa: Open Source Language Understanding and Dialogue Management", arXiv: 1712.05181, 2017.