EVALUATING THE USER INTERFACE OF THE REHALINGO SPEECH TRAINING SYSTEM WITH APHASIC PATIENTS

Hans-Günter Hirsch[1], Yannic Tiggelkamp[1], Christian Neumann[1],

Hendrike Frieg[2], Stefan Knecht[3]

[1]Institute for Pattern Recognition, Niederrhein University of Applied Sciences, [2]University of Applied Sciences and Arts Hildesheim, [3]Institute of Clinical Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf
hans-guenter.hirsch@hs-niederrhein.de

Abstract: This study presents the current development of the RehaLingo [1] speech training system, which is designed to help people suffering from aphasia. The system uses three speech recognition frameworks, which operate in parallel to analyze the patients' uttered responses when naming items displayed on a graphical user interface (GUI). Initial experiments were conducted with patients during multiple therapy sessions at a therapy center. This paper discusses the results and insights gained from this first application of RehaLingo in a clinical setting.

1 Introduction

Individuals who have suffered a stroke are often affected by language impairments, commonly referred to as aphasia. Aphasia is characterized by difficulties in associating spoken or written words with their meanings and/or form, and in producing as well as understanding the words necessary for everyday communication. Recovery typically requires extensive and costly training with a speech and language therapist to restore these word-meaning associations.

To address these challenges, the computer-based training system RehaLingo [1] has been developed to support and accelerate the rehabilitation process. At the core of this system is a speech recognition module, enabling the analysis of patients' spoken inputs. Additional training without the physical presence of a therapist reduces the cost of therapy and mitigates the shortage of skilled health professionals. In this paper, the terms *speech therapy* and *speech training system* are used interchangeably with *language therapy* and *language training system*, reflecting terminology commonly employed in the field of speech and language therapy.

The potential of integrating speech processing and recognition technologies into computer-based speech therapy systems has been explored in previous studies [2, 3]. Additionally, the design of user interfaces [4] and the evaluation of their usability [5] are critical aspects to consider during the development of such systems. This paper introduces the concept and implementation of our training system, which is distinguished from existing approaches [6, 7, 8, 9] by its parallel use of multiple speech recognition systems. The system operates as a standalone solution on local hardware, eliminating the need for communication with external servers.

Graphical user interfaces (GUIs) have been developed to support various training modes, allowing customization based on patient needs. The system and its user interfaces were tested and evaluated in collaboration with patients at a therapy center in Lindlar [10]. Insights gained from these evaluations informed the iterative design process, ensuring the system's suitability for individuals with aphasia. Finally, we discuss the implementation of success-dependent transitions between training modes, aimed at enabling patients to independently operate the system without therapist supervision.

2 Aphasic Speech

Patients with aphasia often experience challenges in identifying and correctly pronouncing the appropriate word for a visually presented item. To address this, speech recognition technology is utilized to analyze and classify speech patterns characteristic of aphasia. Speech input from individuals with aphasia typically exhibits a range of artifacts, reflecting the complexity of the disorder. It is uncommon for speech input to consist solely of the correct word or a single utterance. Instead, such input can generally be categorized into several distinct classes [11]:

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

Additionally, many individuals with aphasia experience significant challenges in clearly and correctly producing desired words or phrases. As a result, the automatic recognition of their speech input represents a complex and demanding task. To address this challenge, the proposed system employs the parallel operation of multiple recognition algorithms to analyze and classify the input effectively.

3 Speech Training System

In the following, we describe the hardware components, the Speech-API as interface to the three recognition modules and the software setup for realizing the GUIs and the dialogue with the user.

3.1 Hardware

The hardware employed in this project comprises a 2-in-1 laptop and a Respeaker microphone [12], as illustrated in Figure 1. The laptop operates in tablet mode, allowing patients to interact with the graphical user interface (GUI) via touchscreen. The system requires sufficient computational performance to run three recognition modules concurrently, ensuring that recognition results are delivered without significant delay. Specifically, 16 GB of RAM is adequate to load the models for all recognition servers simultaneously. The laptop is equipped with an Intel Core Ultra 7 165U CPU, featuring 12 cores and a maximum boost clock rate of 4.9 GHz, which provides sufficient processing power to execute the current models in real time.

The Respeaker microphone includes an array of four microphones, enabling hands-free speech input. It also features a built-in processing unit that supports various signal preprocessing techniques. For this implementation, the microphone is configured to perform stationary noise suppression, effectively reducing persistent background noise in the recording environment. The noise-suppressed signal from a single channel is then utilized as input for the recognition systems.

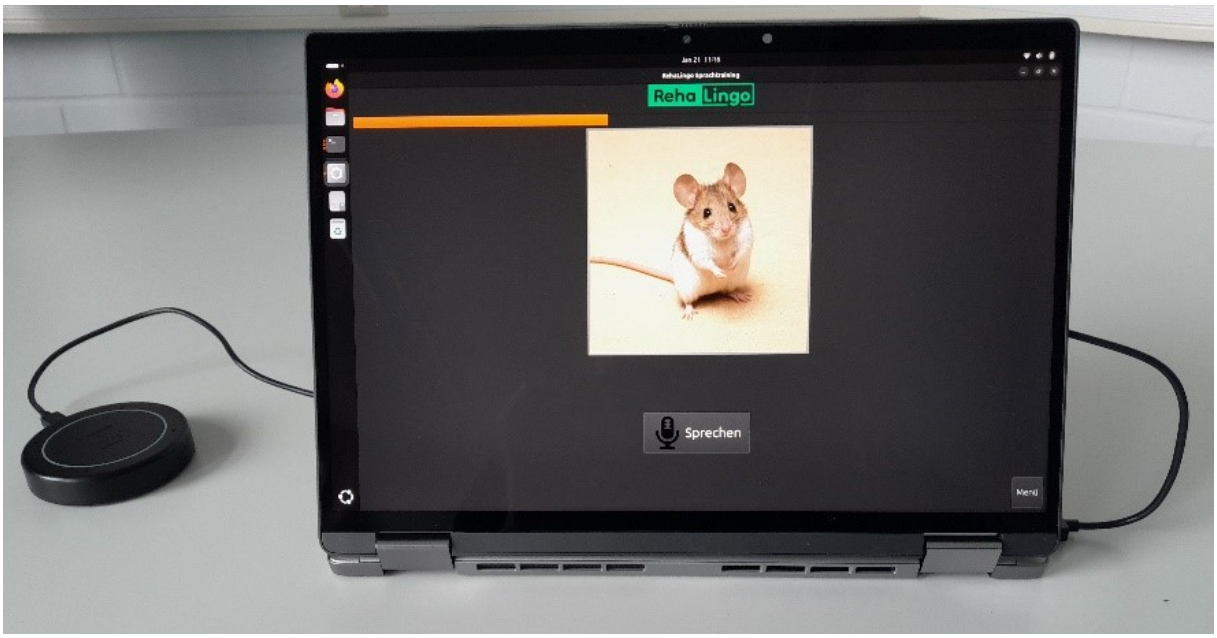

Figure 1: The hardware setup for the RehaLingo speech training

3.2 Speech-API

We developed a Speech-API server to process recorded speech samples by interfacing with multiple recognition systems. The API transmits speech data to the enabled recognition systems and consolidates their outputs into a unified result, as illustrated in Figure 2. The results are then interpreted and returned to the training system. The following sections provide a detailed description of the recognition modules and the Speech-API.

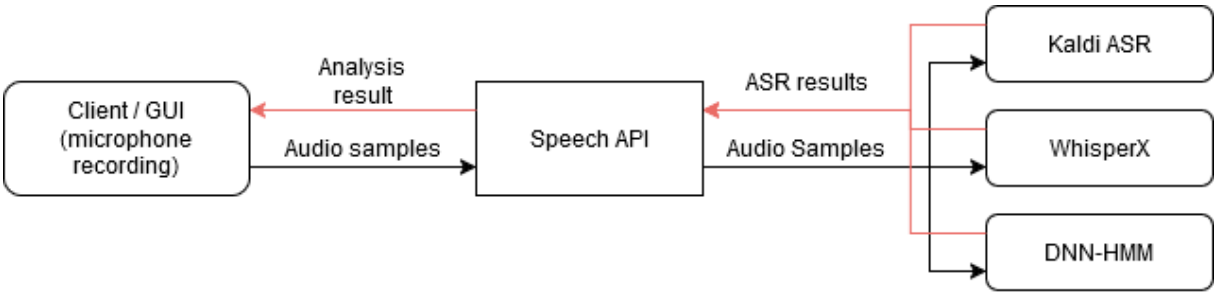

Figure 2: Speech API workflow

3.2.1 Speech-API Architecture and Functionality

The recording of speech data from the ReSpeaker microphone is triggered by the graphical user interface (GUI). Optionally, a voice activity detector (VAD) can be employed to detect the beginning and/or end of the speech utterance. During recording, audio samples are streamed in real-time to the Speech-API, which acts as a proxy to forward data to all available recognition servers. These servers initiate their respective recognition pipelines as soon as sufficient data is available, thereby minimizing latency.

Upon completion of the recording, the Speech-API signals all recognition servers to finalize their processing pipelines. It then transitions to a waiting state until all recognition results are received. For each recognition server, the Speech-API generates a JSON-formatted result object containing the predicted transcription, word-level timestamps, confidence scores, and any system-specific metadata.

Once all results are gathered, the API's analysis module consolidates them into a single JSON response for the client. The primary purpose of this analysis is to classify the patient's speech into predefined categories (as described in Chapter 2), enabling the client to generate an appropriate response. Classification is achieved using a rule-based approach, key rules include:

- Words spoken at the end of an utterance are prioritized over those at the beginning.
- In cases of classification disagreement between recognizers, preference is given to the result closest to the target word, unless a specific recognizer is known to perform poorly for the given category.

3.2.2 Kaldi Recognition System

The Kaldi recognition system [13] is employed to process German speech with a large vocabulary, enabling the analysis of arbitrary speech content. We use a German Kaldi model developed and optimized by the Language Technology Group at Hamburg University [14]. For real-time inference, the Kaldi GStreamer server is utilized.

While Kaldi excels in recognizing arbitrary German speech, it struggles with detecting out-of-vocabulary words, which are common in aphasic speech. This limitation arises from its pre-trained language model. Future enhancements, such as training a custom language model, could address this issue. For now, preference is given to other systems when handling such cases.

3.2.3 Whisper

Whisper, developed by OpenAI [15], is a robust speech recognition framework. We use the WhisperX implementation, which performs efficiently on CPUs. Unlike traditional vocabulary-based systems, Whisper recognizes word tokens, which are later combined into text. This approach enables it to handle typical aphasic speech, including mispronunciations that do not produce standard words.

However, Whisper's computational demands are significant. High-accuracy models require powerful GPUs to achieve real-time performance. The best model compatible with our hardware, "whisper-small," initially exhibited a high word error rate (WER) above 100% on our aphasic speech dataset due to a high number of insertions. To mitigate this, we fine-tuned the model using aphasic speech recordings from the ReSpeaker microphone array, achieving an improved WER of 4%. Nevertheless, the limited availability of training data constrains the effectiveness of fine-tuning and the reliability of evaluations, highlighting an avenue for future research.

3.2.4 DNN-HMM

The DNN-HMM recognition system combines a deep neural network (DNN) with a Hidden Markov Model (HMM), as described in [1]. The DNN estimates probabilities of assigning speech segments to triphones, while the HMM uses these probabilities to determine 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, accounting for artifacts observed in individuals with aphasia.

Each grammar is represented as a graph of nodes (words) and transitions (connections). Grammar generation starts by identifying words relevant to the task. To account for aphasia-related artifacts, the grammar includes phonetically or semantically similar words, paraphrases, neologisms, and unspecific responses besides the correct target-word. Techniques for selecting these words are context-sensitive, with single or multiple-word responses possible. For example, when presented with an image of a bun, "bread" may also be valid. Detailed methods for word selection are provided in [1].

3.3 GUI software

The current graphical user interface (GUI) implementation is developed in Python using the PySide6 framework [16]. PySide6 is a Python-based implementation of the widely used Qt library [17], which facilitates the development of cross-platform user interfaces. PySide6 was selected for this project due to its modern architecture, ease of use, and sufficient feature set to support the requirements of a simple GUI. Python, as the programming language, offers similar advantages, including simplicity and versatility. Furthermore, since the SpeechAPI is also implemented in Python, the use of PySide6 allowed seamless integration without the need to rewrite client software.

4 Graphical User Interfaces (GUIs)

This section provides an overview of the conceptual design of the graphical user interfaces (GUIs) developed for aphasia training. We describe the general principles guiding the development of the GUIs and detail specific interfaces designed for speech comprehension and production tasks.

4.1 Training Modes

Aphasia is a neurological condition, which impairs an individual's ability to find and produce the correct word corresponding to an item. The severity and nature of this impairment can vary widely. In some cases, the primary challenge is finding the correct word, while in others, additional difficulties with pronunciation may arise. In general, an item can be represented through three primary modalities:

- Visual representation: an image of the item
- Orthographic representation: the written name of the item
- Auditory representation: the spoken word

This study focuses exclusively on physical items and does not address abstract concepts, as their visual representation can be challenging. Based on these three modalities, aphasia training typically involves presenting an item in one or two modalities and requiring the patient to identify or produce the corresponding representation in the missing modality. Additionally, training may involve contextual presentations where the item itself is absent, and the patient must infer the target term, such as in a cloze task with a missing word.

Numerous training modes can be devised by combining these modalities, and therapists tailor the choice of modes based on the patient's specific impairments and progress during rehabilitation. This study focuses on commonly employed training modes rather than attempting to exhaustively cover all possibilities.

4.2 General Concept

Patients with aphasia often engage in therapy sessions involving direct interaction with a therapist, who provides personalized feedback. Such feedback is critical for effective rehabilitation. Consequently, the GUIs developed in this study aim to simulate the therapist-patient interaction. For example, short video clips of a therapist are used to introduce tasks and deliver individualized feedback. Future iterations are anticipated to include an avatar-based therapist utilizing text-to-speech and AI-trained dialogue systems. By analyzing recorded therapist-patient interactions, it may be possible to generate automated feedback and reactions through the avatar.

Given that strokes, a leading cause of aphasia, may also result in physical disabilities, the user interfaces are designed to minimize interaction complexity for the patient. The GUIs

employ touchscreens for input, with minimal use of buttons. Whenever possible, textual labels on buttons are replaced by intuitive icons or images to facilitate usability.

4.3 Comprehension

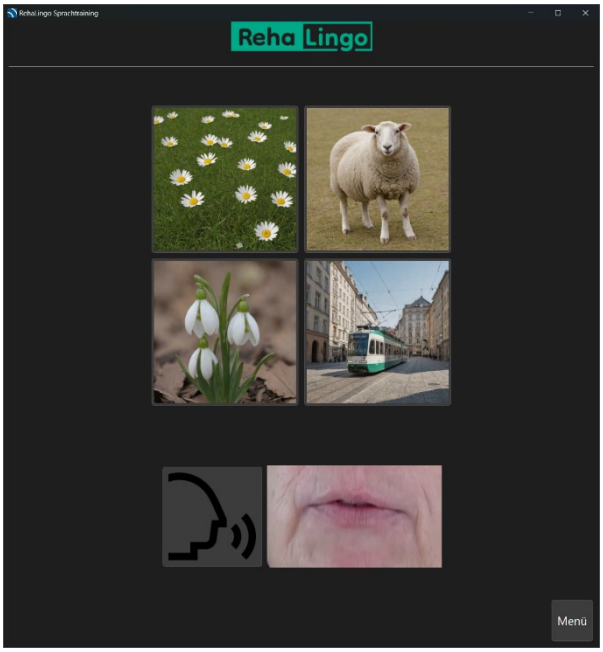

Figure 3: RehaLingo comprehension task

To facilitate the relearning of the mapping between items and their pronunciation, two GUIs were developed for speech comprehension tasks: one for matching a spoken word to an image and another for matching a spoken word to its orthographic representation. An example of the former is shown in Figure 3. The interface displays several images alongside a video of a therapist pronouncing the corresponding word, enabling the patient to observe lip movements.

The task requires the patient to select the image that corresponds to the spoken word by touching it. Feedback is provided immediately after the selection through visual signal colors and auditory cues, indicating whether the choice was correct. The next item is then presented with a new set of images. The complexity of the task can be adjusted by varying the number of images displayed.

4.4 Word Production

To support the development of word production skills, a GUI was created where patients are presented with the image of an item or both its image and text. The patient is tasked with pronouncing the corresponding word. This task is inherently more challenging than comprehension, as it requires both word retrieval and accurate production. An example of this GUI is shown in Figure 4, which includes a central image representing the target item and a button to initiate speech recording.

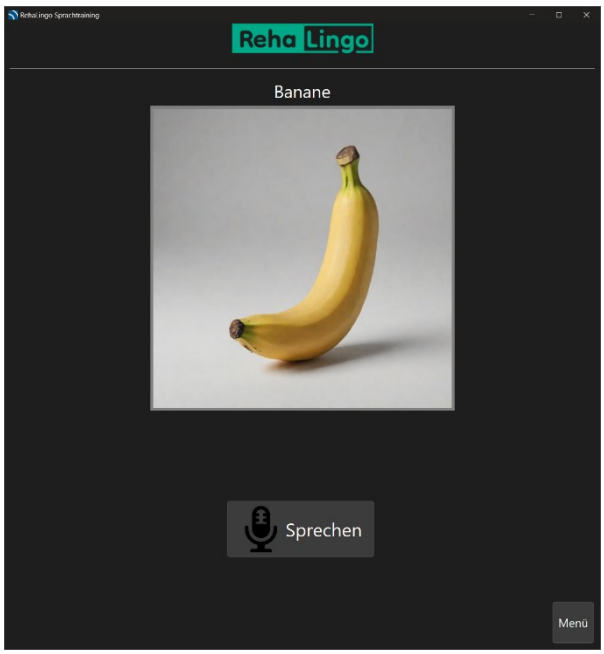

Figure 4: RehaLingo speaking task

When the patient activates the recording button, their speech is recorded and analyzed, as described in Section 3.2.1. Feedback is then provided based on the analysis of the patient's input. This feedback consists of visual signal colors and a short video clip of the therapist providing corrective guidance. For instance, if the patient's pronunciation closely resembles the target word, the therapist might highlight the phonetic similarity. Additionally, the patient can listen to the correct pronunciation and retry the task if necessary. If the patient successfully completes the task or chooses not to retry, they can proceed to the next task by pressing the corresponding button.

5 Tests and Evaluation

We are deeply grateful for the opportunity to test our training system with patients at the therapy center in Lindlar [10].

5.1 Setup

Throughout 2024, we conducted tests of the system with a total of 16 patients over 4 sessions, each reflecting a specific stage of the system's development. Each patient participated individually in a one-hour session supervised by a member of our team. The primary focus of these tests was observing the word production tasks. Insights and experiences gathered from each session, involving 4 patients at a time, were immediately incorporated into the system's further development.

Thus far, we have examined the various training modes independently, without implementing automatic transitions between modes. To facilitate this, an additional user interface was used to select the desired training mode. Additionally, we categorized the 180 terms for which we generated images and videos into 10 thematic groups (e.g., animals, vehicles), from which one or more could be chosen for a session.

5.2 Experiences

Each session began with comprehension training. Since most participants were already at an advanced stage of therapy, they quickly progressed to naming training. During the initial sessions, the system was still in a developmental phase and relied exclusively on the Kaldi recognizer. This approach occasionally resulted in recognition failures for spoken terms. The evaluation of whether a term has been correctly named depends on subjective judgment. Errors are more frequent when participants produce 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 described recognizers were utilized in parallel, resulting in a substantial improvement in recognition accuracy. Errors in identifying correctly named terms in utterances have been reduced to nearly zero, aligning with results obtained from offline tests conducted using our aphasia test dataset. This dataset was compiled from recordings collected during earlier sessions. Incorporating WhisperX and the DNN-HMM recognizer into the analysis improved absolute classification accuracy from 66%, achieved with Kaldi alone, to 87%.

The remaining errors can primarily be attributed to two factors. First, the DNN-HMM recognizer is still under development and frequently misclassifies phonetically similar words. Second, the current configuration of our analysis module is intentionally lenient, leading to false positives in the classification of some utterances.

During the initial sessions, the graphical user interface (GUI) for the naming task displayed an image, and the therapist prompted the user to describe what they saw. The system then automatically initiated recording and used a voice activity detector (VAD) to determine when to stop recording, requiring no manual input from the patient. However, this approach did not allow patients time to reflect on the target term. In the latest version, a microphone button was introduced, enabling patients to initiate the recording themselves. This adjustment proved effective, as most patients had no difficulty operating the button and benefited from the additional time to think.

5.3 Future Perspective

We are currently developing a version of the training system that incorporates automatic transitions between different training modes. The goal is to create a system that allows patients to train and learn independently. This system will also provide therapists with configuration

options, such as predefining specific categories for term selection, setting the number of terms, and determining the starting training mode and applicable training modes based on the patient's therapy stage.

Our current approach involves presenting all terms in a training mode three times. If a patient can correctly identify or name each term at least once, the system transitions to the next, more challenging training mode. If the patient struggles with certain terms, the system shifts to a simpler mode, focusing on the problematic terms.

6 Acknowledgements

We acknowledge support by the Federal Ministry of Education and Research (BMBF) under grant no. 13GW0481D. We would like to thank the therapy center [10] and especially Ms. Tina Keck for enabling and supporting the tests of our training system in Lindlar.

7 References

[1] HIRSCH, H.G.. et al.: Rehalingo – Towards s Speech Training System for Aphasia, In *Proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, Magdeburg, 2023.

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

[3] FRIEG , H., 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] CUPERUS , P., 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] JAKOB , H., 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:https://doi.org/10.14624/NR2202002.

[6] SPÄTH , M., E. HAAS , and H. JAKOB : neolexon-Therapiesystem. Forum Logopädie, 31(3), pp. 20–24, 2017. doi:10.2443/skv-s-2017-53020170304.

[7] NETZEBANDT , J., 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-2022-53020220303.

[8] aphaDigital: https://aphadigital.sprechwiss.uni-halle.de/

[9] TEMA TECHNOLOGIE MARKETING AG: aphavox. 2018. URL https://aphavox.de

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

[11] STADIE , N., S. HANNE , A. LORENZ , N. LAUER , and D. SCHREY -DERN : Lexikalische und semantische Störungen bei Aphasie. Georg Thieme Verlag KG, 2019. doi:10.1055/b-006-149440.

[12] Respeaker microphone array: https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/

[13] POVEY , et al.: The Kaldi Speech Recognition Toolkit. In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, 2011.

[14] GEISLINGER , R., B. MILDE , and C. BIEMANN : Improved Open Source Automatic Subtitling for Lecture Videos. In Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022), pp. 98–103. KONVENS 2022 Organizers, 2022.

[15] RADFORD, A., J.W. KIM et al.: Robust Speech Recognition via Large-Scale Weak Supervision, OpenAI, arXiv: 2212.04356v1, 2022

[16] Python module PySide6: https://pypi.org/project/PySide6/

[17] Library Qt for GUI design: https://wiki.qt.io/About_Qt