REHALINGO – TOWARDS A SPEECH TRAINING SYSTEM FOR APHASIA

Hans-Günter Hirsch[1], *Christian Neumann*[1], *Yannic Tiggelkamp*[1], *Riccardo Fiorista*[1], *Stefan Knecht*[2], *Alfons Schnitzler*[2], *Katja Biermann-Ruben*[2], *Dietmar Bothe*[3], *Günter Bleimann*[3], *Hendrike Frieg*[4]
[1]*Institute for Pattern Recognition, Niederrhein University of Applied Sciences,*
[2]*University Clinic Düsseldorf, Medical Faculty,* [3]*TEMA AG,*
[4]*University of Applied Sciences and Arts Hildesheim*
hans-guenter.hirsch@hs-niederrhein.de

Abstract: Ideas are presented to develop and design a speech therapy system for people that suffer from aphasia, e.g., after a stroke. We investigate the parallel arrangement of several speech recognition systems to analyze and recognize aphasia speech. Some details of the recognition systems are presented. Furthermore, the design of the user interface is an important aspect. As part of this ongoing project, we want to evaluate the usability of a therapist-like dialogue behavior and the application of a gamification approach.

1 Introduction

After a stroke, people may experience a language disorder, e.g., they cannot produce the words for items in their daily life any longer or they show difficulties in matching a spoken or written word to its meaning. This effect is known as aphasia. They try to recover the associations between words and items during time-consuming and expensive training sessions with a speech and language therapist. The project 'TransLingo' aims at creating a computer-based training system called 'RehaLingo' to support and speed up this recovery process. This way, patients can perform additional training without requiring the direct support by a therapist. Lowering the costs of therapy and offering a chance to deal with shortage of skilled professionals in healthcare is a further benefit. The terms speech therapy and speech training system are used throughout this paper which are synonyms for language therapy and language training system in the field of logopedia. The chances and advantages are described in [1] and [2] for applying speech processing and speech recognition as component of a computer based speech therapy. The design of the user interface [3] and the evaluation of its usability [4] are further important topics to be considered during the development of a therapy system. The concept of our training system is presented in this paper where the parallel usage of several speech recognition systems is a main component and distinguishes our system from comparable approaches [5, 6, 7]. Our goal is an implementation of all recognition components on the autarkic training system itself without the need to communicate with external servers. We want to investigate differerent user interfaces including a therapist-like dialogue behavior and a gamification approach. Furthermore, the training system will contain an operating mode where the difficulty of the training task will be adapted to the users' performance on-the-fly.

2 Speech Training System

A typical training protocol for aphasia contains at least two modes, one for the comprehension of spoken words and one for the word production of shown items, that will be described in the next section. Furthermore, we will present details about the two speech recognition systems that we apply as basis for the recognition of aphasia speech.

Figure 1 – Screenshot of our GUI. The pictures for the six different items are clickable.

2.1 Training Modes

2.1.1 Comprehension Task

The first exercise is usually the comprehension of spoken words or utterances. An example of our graphical user interface (GUI) for the comprehension task is shown in Figure 1.

A word is played back via loudspeaker or headphones. The person in training can listen to it several times by pressing one of two loudspeaker buttons. He has the choice to listen to a female or a male voice. Several images of different items are presented on the screen. The user can touch the image that he assigns to the speech output. He gets the visual feedback, e.g., with appropriate icons, whether he has assigned the correct image. Different variants of this training tool exist with a slightly different learning goal. For example, instead of showing images the orthographic description of several items is presented to train the assignment of the correct orthography. This first version of a user interface contains a simple dialogue between machine and user.

We are developing a concept for two more complex dialogue strategies. The first one tries to follow and imitate the dialogue between therapist and patient in a therapy session. This can be done by presenting short video clips of a therapist or of an avatar representing the therapist. In case of selecting the wrong image, the therapist will make comments and give hints about the correct choice. The second dialogue version is based on a gamification approach. For example, an artificial opponent chooses an image randomly but at a predefined mean accuracy. The user is competing against this opponent. The gaming approach intends to increase the attractiveness of the training program so that users complete more and longer training sessions.

2.1.2 Word Production Task

Besides the comprehension task with an assignment to the acoustic or written presentation, the second training step is the correct production of a visually shown item. At this point, speech recognition is needed to analyze and classify the speech produced by a patient. A single training cycle starts by showing the image of an item. The user is asked to produce the corresponding word. The speech input of people with aphasia can show a combination of many different artifacts. It will be the exception that only the correct or a single word is uttered. In general, the speech input in case of aphasia can be assigned to several categories [8]:

- The correct word is uttered in combination with filler words and hesitations.
- The user's utterance is an unspecific reaction that does not contain the desired word like 'I have seen this before'.
- A semantical paraphrase or a sementically similar word is uttered.
- A similar sequence of phonemes is spoken which can be a nonsensical articulation or a phonetically similar word.
- The user's utterance is a morphological paraphrase, e.g., a corresponding verb is spoken instead of the subject.

Furthermore, a lot of people with aphasia have problems in general to clearly and correctly articulate words or phrases. Thus, the automatic recognition of the speech input is a complex and challenging task. The idea is the parallel usage of several recognition systems to analyze and classify the input. Figure 2 contains the parallel arrangement of three recognition systems.

The system in the middle branch can recognize a large vocabulary of several ten thousand of German words so that speech phrases with arbitrary content can be analyzed. We will apply a pretrained Kaldi system [9] here.

The system in the upper branch is a DNN-HMM recognition system [10] for small vocabularies of up to several hundred words. An individual grammar is defined to perform a specific recognition task. In this project, the idea is the usage of an item-specific grammar that will cover all the known artifacts of aphasia mentioned before. When, for example, an image for 'mouse' is shown, the grammar contains the restriction of the recognition to

- the correct word in combination with filler words and/or hesitations,
- similar phoneme sequences or phonetically similar words like 'house',
- semantical paraphrases or semantically similar words like 'rat',
- unspecific reactions, and
- generic terms like 'small animal'.

The idea is the setup of a recognizer that focuses on the expected input for the specific item under consideration of artifacts known to appear frequently with aphasia patients.

The third recognition system in the lower branch recognizes the sequence of phonemes. This can either be done on the whole speech input or only on the segments where the other recognition systems have detected the correct item. The DNN-HMM system will be also used to perform the recognition of phoneme sequences by including an appropriate grammar. The idea is a further detailed check whether the pronunciation of the item is correct. A further processing block might be needed to analyze and combine the results of all recognition systems, e.g., by an approach from the domain of natural language processing. We will investigate different approaches in a later phase of the project.

As for the comprehension task, two dialogue approaches will be investigated. The first one tries to follow and imitate the dialogue between therapist and patient. This will be more complex in comparison to the comprehension task given the broad range of possible speech input. The second realization will contain a gaming approach. We present some details about the Kaldi and the DNN-HMM recognition systems and the creation of the item-specific grammar in the following sections.

2.2 Kaldi

The recognition system for the German language with a large vocabulary is implemented using Kaldi [9]. Right now, we use a Kaldi model for German that was developed and optimized by the Language Technology Group at Hamburg University [11]. We are also active in training our

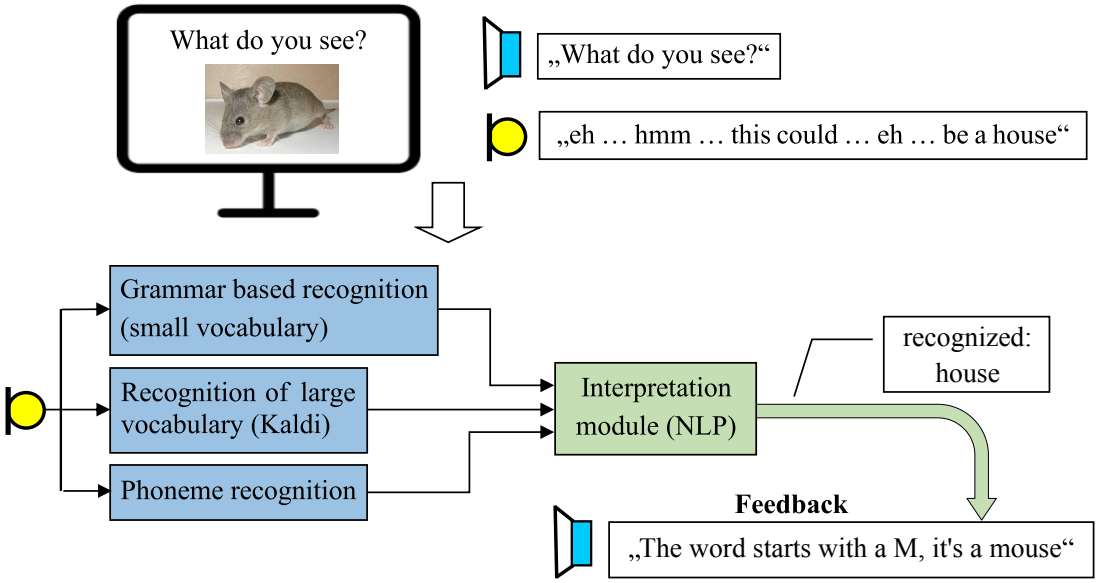

Figure 2 – Parallel setup of several recognition systems.

own models for Kaldi in which we want to apply German databases, e.g., [12] and [13], as well as application-specific speech data which we are going to record ourselves.

For inference, we use the Kaldi GStreamer server which allows real-time speech recognition using a pretrained Kaldi model [14]. Server communication is done via websockets that accept a speech signal and return a JSON object with corresponding recognition results. This allows for an easy and platform-independent integration with our two other recognition systems.

The two key components of the result, we are most interested in, are the sequence of predicted words and their temporal alignment to the original speech signal. The latter is needed for aligning the recognition result to those of the other two systems.

2.3 DNN-HMM Recognition System

The DNN-HMM recognition is based on the substitution of the Gaussian Mixture Model (GMM) as a component of the well known GMM-HMM approach by a DNN to estimate the emission probabilities of all HMM states. We use a phoneme-based HMM recognition scheme with a total of 4715 tied states for modeling all triphones. The DNN is applied to sequences of feature vectors in order to calculate probabilities for the 4715 tied states. A DFT of length 512 is applied every 10 ms to speech segments of length 25 ms containing 400 samples at a sampling frequency of 16 kHz. The magnitude DFT components in the range from 250 to 7000 Hz are merged to 138 spectral features [10]. A logarithmic mapping and a mean and variance normalization are applied to the spectral features.

60 vectors with 138 normalized features are taken as input to the neural network. The DNN consists of, in order, one convolutional layer, three LSTM layers and three dense layers (see Figure 3). The input is first convolved by a 11×9 kernel. After the convolution, the tensor has the size of 50 vectors. The tensor is then flattened and fed into our LSTM stack. All LSTM layer have 512 cells and are in many-to-many configuration, so that input and output dimensions are equal. Furthermore, the LSTM layer stack includes residual connections helping to increase convergence speed during training. Now, a dropout layer is inserted to lower the tendency to overfit. Three dense layers, two of which with 2048 hidden units, are applied over the temporal dimension. The last dense layer maps to the 4715 tied states. The output is activated by softmax to get a probability vector for every speech segment.

For training the DNN, we split the sequence of feature vectors of each speech utterance

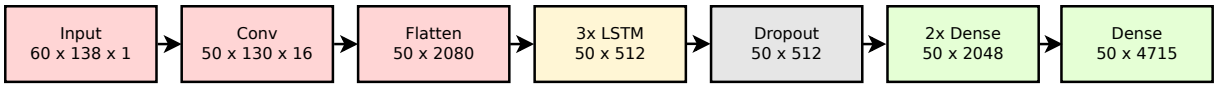

Figure 3 – Network architecture for the DNN. In each block the first line specifies the layer and the second line is the output shape. Residual connections are not shown.

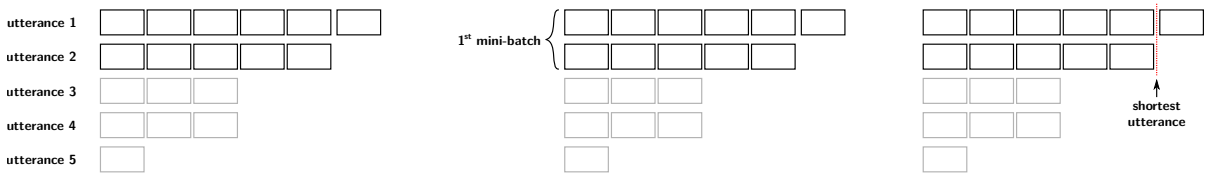

Figure 4 – Example for one sequence being truncated ($B = 2$, left: after sorting, center: after grouping, right: required truncation point).

in subsequences containing 50 successive vectors that represent a speech segment of 500 ms. These subsequences are taken as input for the stateful training of the DNN. In order to provide enough data for the first convolution in the temporal dimension, the subsequences are extended by $\frac{11-1}{2} = 5$ earlier and later feature vectors. The size of the overlapping subsequences is 60 vectors.

2.3.1 Stateful Training

In stateful training mode, the memory cells of the LSTM layers are not reset after every subsequence of feature vectors. Instead, they only need to be reset at the end of a speech utterance. During training, multiple examples per training step are processed, i.e. we use mini-batching. The parameter that determines the number of examples processed in one training step is the batch size B. The reset is applied to all entries in a mini-batch at the same time. Now, the challenge is to reset the states at the correct time. Our solution is to reorganize the input data. If we can make sure that all speech sequences contributing to a mini-batch end at the same time then a single reset is sufficient across all batch entries. Thus, the reset times need to be synchronized within groups of B sequences and a flag indicating the end of the sequences needs to be stored.

When we don't have multiples of B sequences of equal duration, which is naturally very likely, we have to adjust the duration of some sequences. This can either be done by zero-padding or truncating the end of some sequences. Considering any zero-padding will be masked out during inference, learning a mapping from zeroes to a silence state is unnecessary. Also, losing a small percentage of the possible training data is irrelevant because the number of available sequences as well as their duration is sufficiently high. Thus, we opted for truncation of the end. In the following, we describe the algorithm to rearrange the speech data for minimal truncation loss.

We can minimize the number of feature vectors lost due to truncation. In this case, a greedy algorithm is optimal (see Figure 4). First, we sort all sequences by duration in descending order. Then we organize them into groups of B elements. For each group we truncate the first $B-1$ sequences to the duration of the shortest sequence, i.e. the last one.

2.3.2 Dataset Balancing

Right now, our DNN for the German language is trained on audio from the MLS [12] and RVG [13] corpus. In the MLS corpus the duration of all utterances widely spreads for all speakers (see Figure 5). Especially one of the 236 speakers dominates by contributing 3% ($\approx$ 10 hours, $\approx$ 71 times the mean duration per speaker) of all audio. We improved our recognition results by extracting training data with respect to the cumulated duration per speaker. Using a Round-

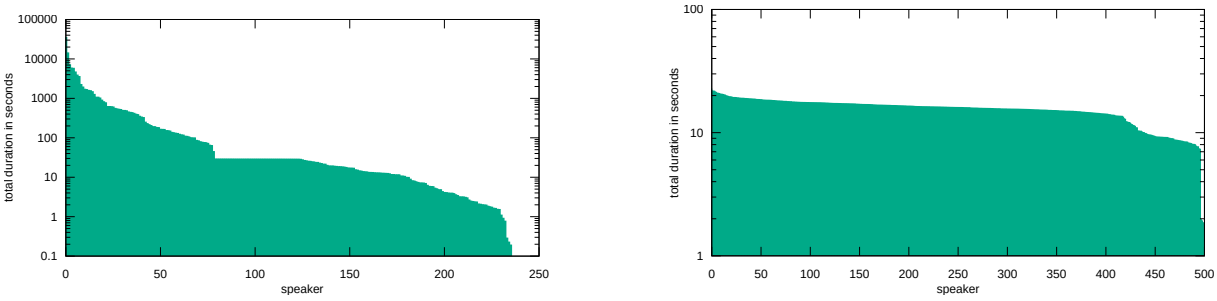

Figure 5 – Total durations of utterances per speaker in our MLS (left) and RVG (right) corpora.

Robin sampling results in less examples taken from MLS in general and, as long as enough audio is available, a nearly uniform speaker distribution.

Additionally, we applied data augmentations. This includes additive noises chosen from a commercial noise database [15], with a random SNR from 10 dB to 20 dB, and reverberation by convolution with room impulse responses [16].

2.4 Item-specific Grammar

As described before, the DNN-HMM Recognition system will make use of item-specific grammars that predefine possible paths during Viterbi alignment. We intend to include about 100 different items in the first version of the training system. Therefore, we are investigating how we can create the item specific grammar. The challenge is to do this automatically as far as possible with inclusion of all mentioned aphasia artifacts.

Fundamentally, each grammar is a graph with a set of nodes (words) and transitions between those. Therefore, the first part of grammar generation is identifying which words are relevant to the current recognition task. Besides including the correct words representing the item, the aphasia artifacts imply the inclusion of phonetically, syntactically, or semantically similar words, paraphrases, mock words, and general phrases for unspecific reactions. Each of the above categories is populated with words by different techniques.

Correct words are not trivial to determine due to context-sensitivity. Depending on the tasks difficulty this class may be a single word or multiple words. For example, when presenting a bun the word ‘bread’ can also be seen as a correct answer.
Phonetically similar words are automatically retrieved from a phonetic lexicon. For each word in question, similar words are found by calculating a token-based Levenshtein distance to every other word in the dictionary. For the German language, we use the PHONOLEX [17] dictionary. The same technique can be used on a regular lexicon to retrieve syntactically similar words.
Semantically similar words are found by traversing a lexical semantic network. We use GermaNet [18, 19] for this. This network sets all kinds of nouns, verbs, and adjectives in relation to each other. By traversing the graphs edges for a set distance, we can find words with a certain degree of semantic similarity.
Paraphrases and mock words are not easy to generate automatically. Our current solution consists of a manual compilation of words and sentences. Silence, noise, hesitations, and generic phrases like ‘I am not sure’ are also manually selected but apply to every task, independent of the currently wanted keyword.

The second part of the grammar generation is to include the graphs’ edges which represent possible transitions between words. Because the speech of people with aphasia can abruptly switch topic or can be affected by artifacts, we decided to start with an open approach and simply allow transitions from any node to any other node. This way, we loose some of the

power a grammar normally provides in detecting grammatically correct sentences but we can not always expect such production from aphasia patients. The effectiveness of this approach is currently being tested.

3 Results

First versions of a graphical user interfaces for the comprehension task have been implemented in Matlab. So far, no experimental studies have been carried out with people that are affected by aphasia. A Kaldi recognition system has been trained with several thousand hours of speech from different German databases. The interface for the parallel access to several recognition systems from Matlab has been implemented. We are developing tools for the automatic generation of the item-specific grammar including the access to phonetic and semantic dictionaries.

4 Conclusions

This paper describes work in progress. We presented our concept for a speech training system to support the therapy of patients with aphasia. The key features for the setup of the training system are

- the parallel usage of several recognition systems to analyze the speech of people with aphasia,
- the evaluation of a user interface with a therapist like dialog behavior,
- the investigation of a gamification-based user interface to increase the patient's load utilization period, and
- the adaptation of task complexity dependent on the user's performance.

The evaluation of our approaches is subject of further research.

Acknowledgements

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

References

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

[2] 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.

[3] 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.

[4] 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.

[5] TEMA TECHNOLOGIE MARKETING AG: *aphavox*. 2018. URL `https://aphavox.de`.

[6] 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.

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

[8] 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.

[9] POVEY, D., 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 *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[10] HIRSCH, H.-G.: *Speech Assistant System With Local Client and Server Devices to Guarantee Data Privacy*. *Frontiers in Computer Science*, 4, 2022. doi:10.3389/fcomp.2022.778367.

[11] 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.

[12] PRATAP, V., Q. XU, A. SRIRAM, G. SYNNAEVE, and R. COLLOBERT: *MLS: A Large-Scale Multilingual Dataset for Speech Research*. *ArXiv*, abs/2012.03411, 2020. doi:10.48550/arXiv.2012.03411.

[13] BURGER, S. and F. SCHIEL: *RVG 1 - A Database for Regional Variants of Contemporary German*. In *Proceedings of the First International Conference on Language Resources and Evaluation*, pp. 1083–1087. 1998.

[14] ALUMÄE, T.: *Full-duplex Speech-to-text System for Estonian*. In *Baltic HLT 2014*, pp. 3–10. 2014.

[15] AVOSOUND: *Digiffects Sound Library*. 2022. URL `https://www.avosound.com/en/sound-libraries/digiffects/`.

[16] JEUB, M., M. SCHAFER, and P. VARY: *A Binaural Room Impulse Response Database for the Evaluation of Dereverberation Algorithms*. In *2009 16th International Conference on Digital Signal Processing*, pp. 1–5. IEEE, 2009. doi:10.1109/ICDSP.2009.5201259.

[17] SCHIEL, F., C. DRAXLER, and H. G. TILLMANN: *The Bavarian Archive for Speech Signals: Resources for the Speech Community*. In *Fifth European Conference on Speech Communication and Technology*. 1997.

[18] HAMP, B. and H. FELDWEG: *GermaNet - a Lexical-Semantic Net for German*. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. 1997.

[19] HENRICH, V. and E. HINRICHS: *GernEdiT - the GermaNet Editing Tool*. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. 2010.