Automatic Processing of Clinical Aphasia Data collected during Diagnosis Sessions: Challenges and Prospects

Christian Kohlschein*, **Daniel Klischies***, **Tobias Meisen***, **Björn W. Schuller**†, **Cornelius J. Werner**+
*Institute of Information Management in Mechanical Engineering (IMA), RWTH Aachen University, Germany
†GLAM – Group on Language Audio & Music, Imperial College London, United Kingdom
+Department of Neurology, Section Interdisciplinary Geriatrics, University Hospital RWTH Aachen, Germany
Corresponding authors: christian.kohlschein@ima.rwth-aachen.de and cwerner@ukaachen.de

Abstract

Aphasia is an acquired language disorder, often resulting from a stroke, affecting nearly 580,000 people Europe alone each year (Huber et al., 2013) . Depending on the type and severity, people with aphasia suffer, in varying degrees, from the impairment of one or several of the four communication modalities. To choose an appropriate therapy for a patient, the extent of the aphasia at hand has to be diagnosed. In Germany and other countries this is done using the Aachen Aphasia Test (AAT). The AAT consists of a series of tests, requiring the patient to talk, read and write over the course of up to two hours. The AAT results then have to be evaluated by a speech and language therapist, which takes around 6 hours. In order to further objectify the manual diagnosis and speed up the process, a digital support system would be highly valuable for the clinical field. To facilitate such a system, we have collected, cleaned and processed real-life clinical aphasia data, coming from AAT diagnosis sessions. Each dataset consists of speech data, a transcript and rich linguistic AAT annotations. In this paper, we report on both challenges and early results in working with the (raw) clinical aphasia data.

Keywords: Clinical Aphasia Data, Multimodal Language Data, Rich Metadata

1. Introduction

Aphasia, i. e., the full or partial loss of linguistic capabilities in adults, is usually an acquired condition, mostly due to damage inflicted to the brain by ischemic or hemorrhagic stroke, but also due to head injury, tumours or neurodegeneration. The loss of linguistic capabilities neither pertains to the motor acts of speaking or writing nor the sensory capabilities of hearing or seeing, but rather to damage to the human brain's 'supra-modal' capability of producing and comprehending language. The consequences of aphasia for the patient are immense: as language, both spoken and written, is our main tool of communication, affected persons are largely cut off from basic social interaction, leading to severe disability, social isolation, loss of health-related quality of life and depression. The socioeconomic impact also is enormous, as persons suffering from aphasia are less likely to return to their jobs (Wozniak and Kittner, 2002). Thus, every effort has to be made to keep this percentage of people dropping out of their jobs as small as possible, necessitating the need for intensive rehabilitation. However, as language is an extremely complex function of the human brain supported by a widespread network of neurons throughout the human brain (albeit with a left-hemispheric predominance), different patterns of damage to the human brain, e. g., by occlusion of different vessels or by trauma to different brain locations, will result in different aphasic syndromes (Ardila, 2010). These are marked by differential loss of putative linguistic modules (Heilman, 2006), such as syntax, semantics, phonology and finally motor speech output. Thus, it is obvious that aphasia rehabilitation is a non-trivial task, and any success in rehabilitation can only occur if and when the prominently hit modules are identified correctly, resulting in a syndromal diagnosis also encompassing the severity of the damage, as there is no general 'aphasia' rehabilitation. In order to achieve a certain level of objectivity and measurability in diagnosing and grading aphasia syndromes, clinical tests and scores are employed. In Germany and beyond, the *Aachen Aphasia Test* (AAT) (Huber et al., 2013) is regarded to be the gold standard in diagnosing and classifying aphasia. This test allows to assess different language modalities at all linguistic levels. Beyond that, it also yields information of probabilistic syndrome classification and syndrome severity. Its disadvantages are that the AAT is immensely time-consuming (up to 8 hours for one patient including data acquisition and evaluation), it does not encompass all linguistic symptoms a patient can exhibit, and it is at least in part dependent on the experience of the rater. Particularly the requirements on human resources preclude its widespread use, although it is regarded to be a prerequisite for, e. g., an intensive comprehensive aphasia program. Besides, the AAT is not very sensitive to changes over time, limiting its utility as a feedback and tracking tool.

Therefore, an automatic aphasia diagnosis system based on the AAT would be highly valuable for patients and clinicians alike. Clinicians would profit from an increased objectivity of the AAT. Having an objective system in place across different hospitals would also enable aphasia rehab units to offer individualized rehabilitation strategies to their (prospective) patients, because they could correlate their language profiles with outcomes of therapeutic success within a specific facility. Patients, e. g. mobility impaired stroke victims, would also benefit from an automatic AAT diagnosis system within their home, making it a non necessity to go the hospital every time for follow-up aphasia examinations. In order to facilitate such a system, a high-quality data and, ideally, large collection of speech and language data along with diagnosis annotations is a prerequisite. During aphasia diagnosis sessions over the course of roughly 20 years at the University Hospital Aachen, clinician-patient speech was recorded, transcribed and, along with the corresponding tests results, digitally archived. The data is in a variety of formats, not available in one homogeneous database but rather spread over multiple systems and the speech data is a mix between clinician and patient speech. Nevertheless, to the best of our knowledge, this data is one of the richest collections of aphasia data in Germany. We therefore strive to utilize this data to built an automatic AAT system. This paper will not focus on the architecture of the system, but rather present and discuss the challenges we encountered in dealing with the clinical speech and language data itself.

The remainder of the paper is structured as follows: In Section 2. we present related work. Section 3. discusses aphasia, its diagnosis in general and introduces the Aachen Aphasia Test in its current form. Following that, Section 4. discusses our work regarding the assembly of the database and the dataset itself, including a description of its modalities. In Section 5., preliminary results will be presented and discussed. Afterwards, in Section 6., we conclude the paper. Finally, in Section 7., we outline future work.

2. Related Work

Computer programs designed to help diagnose and treat aphasia can be categorized into three different groups (Katz, 2010): Tools for 'alternative and augmentative communication (AAC)', which offer additional ways for aphasia patients to communicate, 'Computer-only treatment (COT)' such as smartphone apps designed to be used by aphasia patients to practice speaking without a therapist, and 'Computer-assisted treatment (CAT)' systems, which help therapists during the therapy. Our system is initially designed as a CAT system: While conducting a conversational speech test, the system analyses the patients speech and returns an aphasia score, as outlined in (Kohlschein et al., 2017). This contrasts many existing projects, which are designed as COT systems.

A COT system which allows patients to build sentences out of predefined clauses via a touchscreen interface, and then requests that the patient reads out the sentence was presented by (Le et al., 2016). The system aims to provide feedback to the patient, such that the patient can practice correct speech. For all predefined clauses, they recorded healthy speech during development of the application. Furthermore, this procedure provides, by design, a transcript of the sentence the patient attempted to say. Additionally, the audio file is transcribed after recording. Possession of a transcript currently leads to better detection of aphasic and especially paraphasic speech (Le et al., 2017). The transcript allows to compare healthy speech to aphasic speech on a per-word basis, and therefore to determine the fraction of correct words compared to the total number of words. Additionally, transcripts based on the recordings can be used as training data for automatic speech recognition (ASR) systems, while knowledge about which sentence the patient attempts to say constrains the search space for ASR (Le et al., 2016). Since our goal is to perform a rating on completely spontaneous clinical speech in the context of CAT systems, we do not have predefined sentences or clauses. However, we have aphasia syndrome and severeness ratings for all recordings, which were made by speech therapists or neurologists. This contrasts the ratings used by Le et al. which were made by trained students, and led to the requirement of a reduced number of severeness categories because the agreement on ratings of the same utterance between different evaluators was low. In 2013, (Fraser et al., 2013) compared different approaches to automatically identify subtypes of primary progressive aphasia. They compared two different techniques for feature detection. The first approach they tried is to perform a Welch t-test on features extracted from audio and transcript files of aphasic speech, compared to healthy speech. Then, they ranked the results based on the p-values obtained from the t-test results and selected only the most significant features. Their second approach is based on the minimum-redundancy-maximum-relevance (mRMR) technique proposed by (Peng et al., 2005). Subsequently, Fraser et al. compared a probabilistic Naive Bayes classifier to Support Vector Machines (SVMs) and Random Forests (RFs). Their results showed that, aphasia subtype detection is more accurate when combining acoustic and transcript data, compared to acoustic data alone. However, even if only acoustic data is available, classification of primary progressive aphasia patients and control group members had an average accuracy of 74.05 %, with Random Forests applied on a feature set chosen by an mRMR algorithm performed best at close to 90 % accuracy. Interestingly, the mRMR selection performed worse than the p-value feature selector when applied to a decision problem between the two aphasia subtypes.

The available aphasia speech data in the University Hospital Aachen consists of spontaneous speech interviews between a clinician and a patient. As an alternative to segmenting all the data manually, we investigated automatic systems as well, i. e., using speaker diarization. Speaker diarization can be classified into bottom-up and top-down approaches. These are based on splitting the audio sample into segments using an heuristic identifying changes in loudness, bandwidth and frequency, which implicate speaker changes. In the next step, these segments are clustered and segments in the same cluster are recombined (Tranter and Reynolds, 2006). The goal of the clustering is to form one cluster per speaker, requiring a clustering based on a method that distinguishes between speakers, but does not discriminate intra class. The top-down approach is based on starting with one cluster and iteratively differentiating it into an ideal amount of clusters, while the bottom-up approach starts with a high number of clusters and iteratively merges similar clusters (Bozonnet et al., 2010). Different approaches for clustering have been proposed. These include using Gaussian Mixture Models to model speakers (Castaldo et al., 2008) based on a sliding window and using eigenvoices as features. Eigenvoices are feature vectors in a vector space whose basis was determined using principle component analysis on the extracted features, causing a model that is based on dimensions which had a high variance in the original feature set (Kuhn et al., 2000). Another method, introduced in (Sell and Garcia-Romero, 2014), is to apply agglomerative hierarchical clustering based on scores retrieved by computing the pairwise similarity of all i-vectors using probabilistic linear discriminant analysis, merging those that are most similar. There also have been approaches based on identifying speakers by training deep neural networks to identify speakers and subsequently extracting their hidden layer feature activations, under the assumption that similar activation patterns imply that two speakers are the same (Rouvier et al., 2015). The authors of (Isik et al., 2016) also presented an approach based on deep clustering capable of single-channel multi-speaker separation. Finally, (Zhang et al., 2017) presented a diarization approach based on paralingustic cues, e. g., age and gender.

Few collections of aphasic data are publicly available, the most prominent being the AphasiaBank (MacWhinney et al., 2011), which is mostly for the English language domain. More recently, a Greek data set (GREECAD) was made available by (Varlokosta et al., 2016). Both data sets contrast our data collection in several ways. GREECAD was assembled with scientific purposes in mind and subsequently annotated and transcribed by humans in a predefined way, thereby maximizing the agreement between evaluators to get uniform and coherent annotations. Additionally, machine readability and processability was taken into account when choosing the data format and recording the patients. In contrast, the data set of the University Hospital Aachen was solely collected for clinical diagnosis purposes during assessment sessions over a couple of years. Therefore, machine readability was not taken into account while assembling and recording the data, which in turn poses challenges for the automatic processing of it. These challenges include, but are not limited to missing or incorrect meta data, such as therapist attribution, and mono-channel recordings with low cost microphones, requiring a speaker diarization procedure capable of handling open speaker groups, with high noise tolerance and which does not rely on language models, as these do not apply to aphasic speech.

Transcripts and annotations were made by clinical speech and language therapists for the aphasia domain, whereas the Greek data set was transcribed by linguists (graduate or post graduate students). Our data currently contains transcripts roughly four times the amount of aphasic utterances in GREECAD, but does not contain a control group (due to the origin of the data). The AphasiaBank data set has similar properties as the Greek data set, albeit being larger. Additionally, the AphasiaBank contains video recordings of patients (which are not available for both GREECAD and Aachen data sets).

3. Aphasia Syndromes and Diagnosis

Due to the fact that linguistic modules usually are located in distinct neuroanatomical regions of the brain, and that the vascular supply also encompasses distinct areas, occlusion of the trunk or a particular branch of the middle cerebral artery (MCA) leads to typical combinations of linguistic symptoms, called aphasic syndromes. Testing the different linguistic domains thus allows classification of the aphasic syndrome and prediction of the location of the lesion. However, anatomical variations, incomplete or pre-existing lesions or non-vascular lesions can lead to non-standard syndromes, which are then called unclassified aphasia. Additionally, some symptoms can be mapped to anatomical areas that are not solely defined by their vascular supply (Henseler et al., 2014). Typically, however, the following syndromes will occur after an ischemic stroke: occlusion of the main trunk of the MCA (M1 segment) leads to destruction of almost all perisylvic areas concerned with speech and language and subsequent *Global aphasia*. The resulting speech is characterized by a profound loss of syntax and severe disturbances in word retrieval and semantics, sometimes leaving the patient with recurring utterances or automatisms only. Full mutism can occur and language comprehension is severely affected. Occlusion of the anterior branches usually leads to so-called *Broca's aphasia*, marked by non-fluent spontaneous speech (which is monotonous and lacking prosody) and agrammatism. Language comprehension is relatively spared. Lesions in areas supplied by posterior branches of the MCA can lead to *Wernicke's aphasia* which is characterized by fluent spontaneous speech, which however is accompanied by severe disturbances in language comprehension and the use of overshooting, long and tortuous sentences filled with neologisms and paraphasias – a symptom that is called paragrammatism. Prosody usually is preserved. *Amnestic aphasia* is caused by a prominent deficit in word-finding capabilities, while language comprehension and prosody are usually preserved.

Thus, a diagnosis of aphasia is made by testing the presence and severity of the different linguistic symptoms. For this purpose, many validated tests are available in addition to the clinician's expertise that probe variable aspects of the patient's linguistic capabilities. As outlined above, the gold standard in Germany for aphasia diagnosis is the Aachen Aphasia Test (AAT) (Huber et al., 2013). Its purpose is to assess different language modalities (i. e., understanding, writing, reading, speaking) at all linguistic levels. Beyond that, it also yields information of probabilistic aphasia syndrome classification and syndrome severity. The AAT consists of six parts in total, testing different speech and language modality impairments and differentiations. First, and most-important for our current research, an approximately 10 minutes long semi-structured interview is conducted by a clinician. The purpose of the interview is to assess the spontaneous speech capabilities of the patient. Usually, the patient gets to tell about the circumstances the aphasia syndromes first appeared (e. g., when and where a stroke happened and what they where doing), about treatment, family and job etc. The interview is followed by a series of five tests where the patients gets to read, write and has to identify certain tokens. During the AAT, the clinician records the answers on an protocol sheet and takes notes. The interview of the spontaneous speech part is recorded using a basic microphone setup and later transcribed by the clinician, typically a speech and language therapist (SLT). Both the evaluation sheet, the recording and the transcription then constitute the basis for the subsequent diagnosis, which takes up to 6 hours.

While the concrete answers of the patients for each of the five non-interview tests are not directly accessible by us, we only have their final AAT evaluation results, we have access to the raw speech recordings, transcripts and diagnosis results of the (spontaneous speech) interview section. This data forms the basis for our research and the topics discussed in this paper. Each spontaneous speech sample together with its corresponding transcript is evaluated on six different speech impairment levels and on a six point scale (with 0 being the most severe and 5 meaning no impairment) by a clinician. The levels are (Huber, 1983):

1. Communication behavior: Describes the ability of the patient to conduct a dialog, i. e., to understand questions from the clinician and respond to them, to utter speech-based information.

2. Articulation and prosody: Impairments of the speech are described in this level, in particular fluidity, vocalization, preciseness, speed, rhythm.

3. Automatic speech: Features of the speech which are produced automatically by the patient during the dialog are accounted for in this level, e. g., recurring utterances or echophrasias (e. g., repeating phrases of what the clinician said).

4. Semantic structure: This level evaluates the ability of the patient to pick words and to differentiate between their meaning. Furthermore, it evaluates if the patient picks meaningless set phrases.

5. Phonemic structure: Evaluates the order of phonemes in uttered words, e. g., if they are added, dropped, repeated or shuffled.

6. Syntactic structure: This level accounts for the completeness and complexity of sentence parts, their order and amount, and for inflections.

During diagnosis, items 1. and 2. are mostly evaluated on a qualitative level, e. g., is the patient able to communicate daily matters, while 3. – 6. are evaluated on a quantitative level, e. g., the amount of automatisms in the transcript is counted manually.

4. Clinical Aphasia Data Collection and Preprocessing

The available aphasia data in the University hospital consists of several hundred AAT sessions over the course of nearly 20 years. This data was spread over multiples systems within the aphasia ward and was not available in one homogeneous file format (i. e., a mix of txt, doc, docx and PDF documents). To make the data usable for research, we first had to consolidate this data and integrate it into one database. Furthermore, not all datasets were usable for the goal of developing an automatic AAT and had to preprocessed. Some patients had no transcripts, some had no diagnosis sheet, while others where lacking the speech recordings. After a mixture of automatic and manual consolidation, we arrived at a database of 442 complete AAT diagnosis results from 343 patients (some patients took the AAT several times, i. e., for follow-up exams). Each AAT result has a corresponding speech sample in audio format and 388 of them are transcribed. The speech sample stems from the recording of the spontaneous speech evaluation, i. e., the interview, conducted with the clinician. The following sections describes each modality in detail.

4.1. Ratings

Each patient's spontaneous speech performance is rated according to the six categories listed above (see section 3.). The corresponding rating distributions are shown in Figure 2.

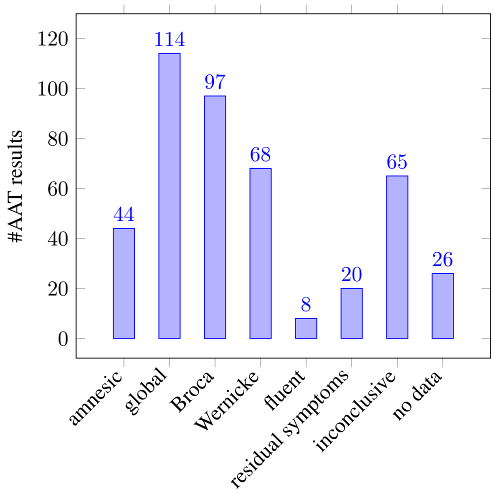

Figure 1: Distribution of aphasia syndromes in the UKA AAT database

Aphasia Type	#Patient	#Utt. Patient (Avg.)
Amnestic	40	491 (12.28)
Broca	53	1225 (23.11)
Global	61	1562 (25.61)
Wernicke	40	612 (15.30)

Table 1: Amount of transcribed utterances available for each of the four most prevalent aphasia syndromes in the UKA AAT database

Notably, there is no test result with a communication impairment rating of zero, as this would be equal to not showing any reactions at all during a conversation, including any non-verbal reactions such as gestures. Additionally, most of the of the samples contain an aphasia severeness rating and an aphasia syndrome diagnosis (e. g., a mild Broca Aphasia). The severeness is rated in five severeness levels, but apparently only mild, moderate to severe and severe are used by most therapists. The aphasia syndrome is classified in six categories, with the most prevalent syndromes being global aphasia, Broca's aphasia, and Wernicke's aphasia. There is an additional category for inconclusive syndromes, i. e., syndromes that are not clearly distinguishable between multiple categories or which do not fit into any category at all (see Figure 1). Furthermore, each AAT sheet also contains the ratings of the 5 other tests, such as the token test. About half of the available AATs also contain information on which therapist conducted the test. There are 104 different therapist names. The most involved therapist conducted 75 tests, while the overwhelming majority of therapist names occurs only once (however this information is not normalized as it was manually entered by therapists). It is entirely possible that the same therapist is referred to under different names such as initials and surname. Due to privacy concerns, information about the patients was anonymized, i. e., neither name, age or gender is given in the data.

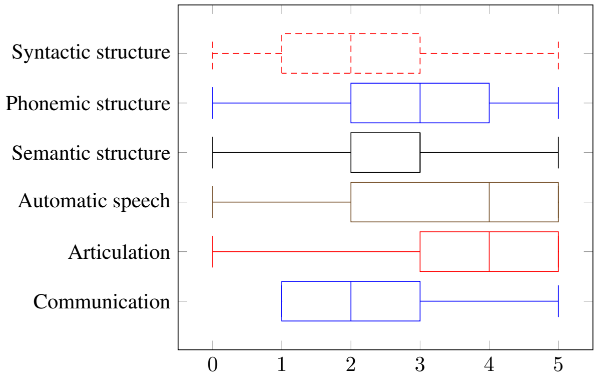

Figure 2: Distribution of AAT speech impairment ratings in the UKA AAT database

4.2. Speech

Each spontaneous speech sample is available as an MP3; most of them are mono recordings. Since the data stems from the course of 20 years, we cannot state the exact type of audio setup for each recording session. As of 2017, the audio setup consists of one microphone positioned between the patient and the clinician. The recording is started manually by the clinician once the spontaneous speech test starts and stopped afterwards. The total duration of all recordings combined is around 63.7 hours. This includes both patient and clinician speech. In order to be able to evaluate aphasic speech, we needed to extract the patient portion of the interview. This can either be done manually or using an automatic speaker diarization system. A completely manual source separation is a very time consuming matter. We found that it took at least 5 – 7 minutes on average to split 1 minute of interview speech (currently, the segmentation is ongoing). Depending on the aphasia syndrome, especially in global aphasia, patients talk only briefly, sometimes uttering just an interjection, before the clinician talks again. That contributes to the necessary time invest, because one has to constantly start and pause the recordings to do the tagging. On the contrary, patients with Wernicke’s aphasia tend to talk much longer, but from time to time the clinician makes a comment, leading to an overlap between patient and clinician speech. Again, this segments have to be identified by hand. For a comparison of two different aphasia speech sections see Figures 3 and 4.

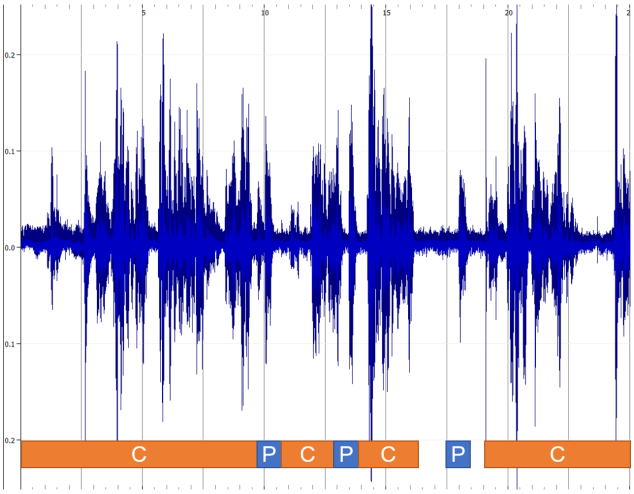

Figure 3: 25 seconds snippet of an global aphasia speech interview. To each question (e. g., “how did the disease start?”) the patient (P) responds with a short “um” utterance.

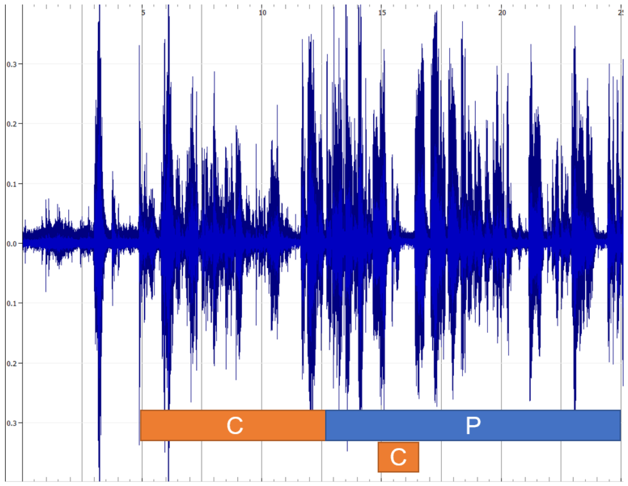

Figure 4: 25 seconds snippet of an Wernicke’s aphasia speech interview. The patient (P) answers fluently, but the clinician (C) makes interjections.

As an alternative to a completely manual split of the speech data, we also tested a commercial tool and the open-source framework pyAudioAnalysis (Giannakopoulos, 2015) for speaker diarization. Neither automatic tool could provide the quality of segmentation needed for our research. We attribute this to the difficulty of speaker diarization itself and the complexity of our disease related data. Sometimes, the segmentation contained alternating patient and clinician speech, sometimes both parties were talking, sometimes a mono person segment was labeled as patient when it was in fact the clinician talking and vice versa. We experimented with counteracting the later case by building a binary classifier able to distinguish between aphasia and non-aphasia speech. For this, we extracted 45,912 utterances from the English AphasiaBank corpus ((MacWhinney et al., 2011)) and split these into a train (70 %) and test (30 %) group, based on which sub data set they belong to. Basing the split on the sub data set affiliation prevents us from training and validating based on the same therapists. This results in 25,414 utterances in the training set and 20,498 utterances in the test set (The discrepancy to our 70:30 quota is caused by different sizes of the sub data sets, and the test data set containing larger sub datasets). We subsequently extracted a feature vector for each utterance, using openSMILE (Eyben et al., 2013) with the `IS13_ComParE` feature set (Schuller et al., 2013). These feature vectors, along with the speaker labels extracted from the transcripts, have been used to train a Gradient Boosting classifier to discriminate between clinician and patient. The Gradient Boosting was implemented using scikit-learn 0.19.1 (Pedregosa et al., 2011). The resulting model was evaluated by calculating the mean accuracy of its predictions on the test set, resulting in a mean unweighted

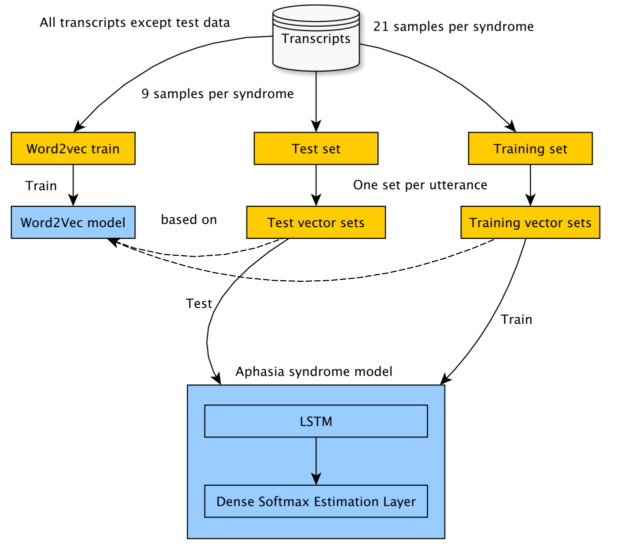

Figure 5: Transcript based aphasia syndrome classification pipeline

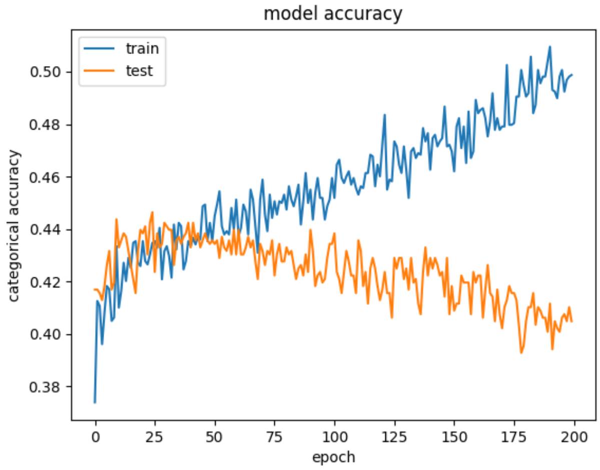

Figure 6: Categorical accuracy of an LSTM estimating the aphasia syndrome. Best performance from epoch 25 to 60, with peak accuracy of 44.3% ($\mu = 25\%$).

accuracy of 83.27 % ($\mu = 50\%$). This is too inaccurate for usage in our system. Additionally, this does not provide any segmentation, but requires a segmentation beforehand, possibly lowering its accuracy even further if the provided segmentation (using an automatic diarization system for preprocessing) is not as accurate as the segmentation of the AphasiaBank.

4.3. Transcripts

After the therapy session is completed, the clinician starts to transcribe the recording of the spontaneous speech session. The speech is transcribed as it is, including interjections like "hmm", or speech and articulation errors. Furthermore, the clinician might also include remarks like "patient is laughing" or "patient is thinking" in curly brackets within the patient portion of the transcript. The clinician also transcribes her own speech. In our data, each transcript is then an alternating list of texts, tagged with either patient or clinician. In Table 1, the amount of utterances available for each of the four standard aphasia syndromes is stated.

5. Early Results and Discussion

For an initial analysis of the data and due to the challenges with speaker diarization we described in 4.2., we started with the goal of predicting the aphasia syndrome type based on the transcripts by configuring a baseline setup. Therefore, a subset of the data has been partitioned into four groups of 30 AAT tests each, such that each group contains patients of one of the four most prevalent aphasia syndroms: global aphasia, Broca's aphasia, Wernicke's aphasia and amnesic aphasia. From each of the four groups representing syndroms, we used 70% for training and 30% for testing purposes. In order to classify the aphasia syndrome based on transcripts, we converted each patient utterance into a list of words and trained a word2vec (Mikolov et al., 2013) model. We chose a window size of three and required each word in the word2vec space to occur at least two times in our utterances. To train the word2vec model, we use our training data set described above, along with phrases from patients which we did not include in the training and test sets before, for instance because they had an inconclusive aphasia syndrome diagnosis. In order to train our aphasia syndrome classifier, we subsequently transform all training utterances into lists of 20-dimensional word vectors, padding them to a length of 30 vectors per utterance. Each of these lists has an assigned aphasia syndrome label and is used to train a pipeline of an long short-term memory (LSTM) layer, followed by a densely connected layer featuring a softmax activation function. This is implemented using Keras (Chollet and others, 2015). The LSTM has been configured to use a 30 % chance of unit dropout and 40 % chance of unit dropout in the recurrent state, while using 80 memory units. We only use a single layer LSTM configuration, as the goal is to provide a baseline for further developments. The model uses an categorical cross entropy loss function and estimates a four dimensional normalized tensor, with each dimension representing one aphasia syndrome. The result is evaluated based on categorical accuracy, which is the percentage of correctly predicted classes, with the "predicted class" being the greatest element of the softmax output tensor. The evaluation has been performed on the test set described above. Plots of accuracy and loss attributes over 200 epochs are depicted in Figures 6 and 7, while the classification pipeline is depicted in Figure 5. The increasing loss function indicates that the model overfits around 100 epochs. Further increasing loss values did not show any meaningful improvements, indicating that more training samples might be the better way to cope with this issue. In summary, the baseline setup shows both the potential and the challenges with clinical aphasia data. While it was possible to perform an initial classification, the usage in clinical scenarios depends on higher accuracies and further improvements (see Section 7.).

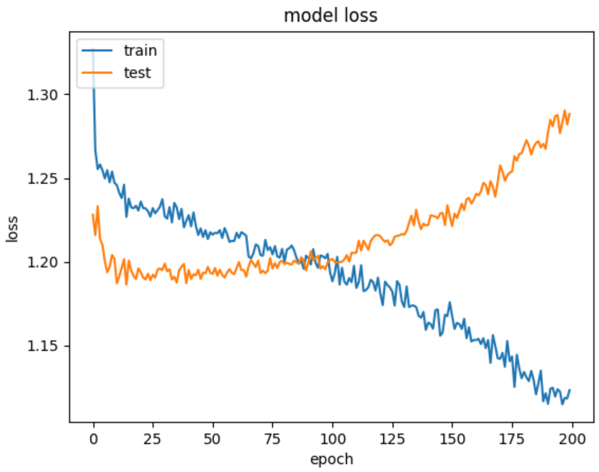

Figure 7: Loss of an LSTM estimating the aphasia syndrome.

6. Conclusion

In this paper, we presented challenges and early results in the automatic processing of real-world clinical aphasia data. We described our data collection of aphasia spanning many years of diagnosis sessions in the university hospital. Each data point in our collection consists of speech recording data, transcripts and rich meta data. The speech data consists of patient clinician interviews and has to be segmented before it can be utilized. We therefore reported on challenges with speaker diarization. The meta data was extracted from diagnosis sheets and contains aphasia syndrome and severeness classification, as well as scores and evaluations of the spontaneous speech section. The scores contain six different categories, which, among others grade the prosody, syntax and phonematic structure of the patient speech. We aim to use this data collection to build and automatic aphasia test, based on the German AAT. Such a system would both benefit clinicians and patients. E. g., patients, many of them mobility impaired stroke victims, could have a continuous spontaneous speech evaluation system at home without the need to go to the hospital every time. In our work, we started with building a baseline syndrome classifier based on an LSTM using the transcript portion of the dataset.

7. Future Work

Our initial implementation of an automatic aphasia syndrome categorizer shows the challenge of the task of usage in a real world scenario. As higher accuracies will be needed before such systems can be used in everyday clinical settings, in the future, we aim to increase its performance in several ways, such as performing a majority vote based on the categorization of all utterances of a patient or additional layers within the classification model. These layers might use information like word histograms and utterance length distributions. Additionally, it might be possible to constrain the decision space for certain combinations of meta information. The latter could be an especially valuable approach when estimating speech impairment factors like automatic speech, as the AAT limits the possible ratings by measurable factors like misplaced words. This would help to cope with the lack of training data, since a first attempt in using an LSTM to do this expressed signs of underfitting and thus yielded a low accuracy. Regarding the segmentation of speech data, we plan to further investigate the possibility of using an automatic speaker diarization system, or at least applying a semi-automatic approach. We think that it might be helpful to include clues about one speaker having impaired speech in the process, i. e., analogous to the paralinguistic approach presented by (Zhang et al., 2017). Finally, we plan to include the speech section as well in order to build a model able to draw from both speech and transcript data. Furthermore, we plan to use the UKA AAT DB (including speech, transcript and rating data) for a challenge, e.g. ComParE at Interspeech, and release it to the research community afterwards. The DB will then include distinct portions for training, development and testing.

8. Acknowledgements

The authors would like to thank Rena Overbeck, M.Sc., for taking the time for patiently answering all questions regarding the concrete execution and evaluation of the AAT.

9. Bibliographical References

Ardila, A. (2010). A proposed reinterpretation and reclassification of aphasic syndromes. *Aphasiology*, 24(3):363–394.

Bozonnet, S., Evans, N., Fredouille, C., Wang, D., and Troncy, R. (2010). An integrated top-down/bottom-up approach to speaker diarization. In *Interspeech 2010, September 26-30, Makuhari, Japan*, pages Interspeech–2010.

Castaldo, F., Colibro, D., Dalmasso, E., Laface, P., and Vair, C. (2008). Stream-based speaker segmentation using speaker factors and eigenvoices. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4133–4136. IEEE.

Chollet, F. et al. (2015). Keras. `https://github.com/keras-team/keras`.

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM.

Fraser, K. C., Rudzicz, F., and Rochon, E. (2013). Using text and acoustic features to diagnose progressive aphasia and its subtypes. In *INTERSPEECH*, pages 2177–2181.

Giannakopoulos, T. (2015). pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12):e0144610.

Heilman, K. M. (2006). Aphasia and the diagram makers revisited: an update of information processing models. *Journal of Clinical Neurology*, 2(3):149–162.

Henseler, I., Regenbrecht, F., and Obrig, H. (2014). Lesion correlates of patholinguistic profiles in chronic aphasia: comparisons of syndrome-, modality-and symptom-level assessment. *Brain*, page awt374.

Huber, W., Poeck, K., and Springer, L. (2013). *Klinik und Rehabilitation der Aphasie: eine Einführung für Therapeuten, Angehörige und Betroffene*. Georg Thieme Verlag.

Huber, W. (1983). *Aachener aphasie test (AAT)*. Verlag für Psychologie Dr. CJ Hogrefe.

Isik, Y., Roux, J. L., Chen, Z., Watanabe, S., and Hershey, J. R. (2016). Single-channel multi-speaker separation using deep clustering. *arXiv preprint arXiv:1607.02173*.

Katz, R. C. (2010). Computers in the treatment of chronic aphasia. In *Seminars in speech and language*, volume 31, pages 034–041. Published by Thieme Medical Publishers.

Kohlschein, C., Schmitt, M., Schuller, B., Jeschke, S., and Werner, C. J. (2017). A machine learning based system for the automatic evaluation of aphasia speech. In *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*.

Kuhn, R., Junqua, J.-C., Nguyen, P., and Niedzielski, N. (2000). Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8(6):695–707.

Le, D., Licata, K., Persad, C., and Provost, E. M. (2016). Automatic assessment of speech intelligibility for individuals with aphasia. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2187–2199.

Le, D., Licata, K., and Provost, E. M. (2017). Automatic paraphasia detection from aphasic speech: A preliminary study. *Proc. Interspeech 2017*, pages 294–298.

MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238.

Rouvier, M., Bousquet, P.-M., and Favre, B. (2015). Speaker diarization through speaker embeddings. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 2082–2086. IEEE.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., et al. (2013). The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Sell, G. and Garcia-Romero, D. (2014). Speaker diarization with plda i-vector scoring and unsupervised calibration. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 413–417. IEEE.

Tranter, S. E. and Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. *IEEE Transactions on audio, speech, and language processing*, 14(5):1557–1565.

Varlokosta, S., Stamouli, S., Karasimos, A., Markopoulos, G., Kakavoulia, M., Nerantzini, M., Pantoula, A., Fyndanis, V., Economou, A., and Protopapas, A. (2016). A greek corpus of aphasic discourse: Collection, transcription, and annotation specifications. In *Proceedings of LREC 2016 Workshop. Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016), Monday 23rd of May 2016*, number 128. Linköping University Electronic Press.

Wozniak, M. A. and Kittner, S. J. (2002). Return to work after ischemic stroke: a methodological review. *Neuroepidemiology*, 21(4):159–166.

Zhang, Y., Weninger, F., Liu, B., Schmitt, M., Eyben, F., and Schuller, B. (2017). A paralinguistic approach to speaker diarisation: Using age, gender, voice likability and personality traits. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 387–392. ACM.