UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

A machine learning based system for the automatic evaluation of aphasia speech

Christian Kohlschein, Maximilian Schmitt, Björn Schuller, Sabina Jeschke, Cornelius J. Werner

Angaben zur Veröffentlichung / Publication details:

Kohlschein, Christian, Maximilian Schmitt, Björn Schuller, Sabina Jeschke, and Cornelius J. Werner. 2017. "A machine learning based system for the automatic evaluation of aphasia speech." In *IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), 12-15 October 2017, Dalian, China*. Piscataway, NJ: IEEE. https://doi.org/10.1109/HealthCom.2017.8210766.

Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
Deutsches Urheberrecht
Weitere Informationen finden Sie unter: / For more information see:
https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publiz/

A Machine Learning Based System for the Automatic Evaluation of Aphasia Speech

Christian Kohlschein*, Maximilian Schmitt†, Björn Schuller†§, Sabina Jeschke* and Cornelius J. Werner‡
*Institute of Information Management in Mechanical Engineering (IMA), RWTH Aachen University, Germany
†Chair of Complex & Intelligent Systems, University of Passau, Germany
§Department of Computing, Imperial College London, United Kingdom
‡Department of Neurology, Section Interdisciplinary Geriatrics, University Hospital RWTH Aachen, Germany
Email: {kohlschein,jeschke}@ima.rwth-aachen.de, maximilian.schmitt@uni-passau.de, schuller@IEEE.org, cwerner@ukaachen.de

***Abstract*—Aphasia is an acquired language disorder resulting from damage to language related networks of the brain, most often as a result of ischemic stroke or traumatic brain injury. Within the European Union, over 580 000 people are affected each year. Both assessment and treatment of aphasia require the analysis of language, in particular of spontaneous speech. Factoring in therapy and diagnosis sessions, which require the presence of a speech therapist and a physician, aphasia is a resource intensive condition: It has been estimated that in Germany alone, there are 70 000 new cases of stroke-related aphasia every year, 35 000 of which persist over more than six months - all of which should receive formal diagnostic testing at some point. Having an automatic system for the detection and evaluation of aphasic speech would be of great benefit for the medical domain by immensely speeding up diagnostic processes and thus freeing up valuable resources for, e.g., therapy. As a first step towards building such a system, it is necessary to identify the vocal biomarkers which characterize aphasic speech. Furthermore, a database is needed which maps from recordings of aphasic speech to the type and severity of the disorder. In this paper, we present the vocal biomarkers and a description of the existing Aachen Aphasia database containing recordings and transcriptions of therapy sessions. We outline how the biomarkers and the database could be used to construct a recognition system which automatically maps pathological speech to aphasia type and severity.**

***Index Terms*—Aphasia, Assessment Tool, Speech Disorder, Machine Learning, Bag of Audio Words**

I. INTRODUCTION

Stroke is the major cause of acquired disability in adulthood and imposes a severe burden on both the affected individual and society as a whole. This holds particularly true for stroke victims suffering from aphasia, i.e., the acquired full or partial loss of linguistic capabilities. Aphasia rehabilitation is a long and costly process compounded by the fact that 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). Thus, 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 [1]. These are marked by differential loss of putative linguistic modules [2], such as syntax, semantics, phonology and finally motor speech output. Likewise, it is obvious that 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. This traditional way of placing an accurate diagnosis before therapy is usually performed by a highly trained clinician, either a neurologist or a speech and language therapist (SLT). In order to achieve a certain level of objectivity and measurability, clinical tests and scores are employed. In Germany and beyond, the *Aachen Aphasia Test* (AAT) [3] is regarded to be the gold standard in diagnosing and classifying aphasia. This test has the huge advantage that it 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 analysis), and it is at least in part dependent on the experience of the rater. Particularly the former property preclude its widespread use, although it is regarded to be a prerequisite for, e.g., an intensive comprehensive aphasia program. On top of that, the AAT is not very sensitive to intra-individual improvements over the course of rehabilitation, limiting its utility as a feedback and tracking tool. To address this shortcomings, we propose an automated, machine learning based tool which allows for a robust, reliable, and rater-independent syndrome classification and grading of aphasia.

II. RELATED WORK

Previous efforts in automatic aphasia classification were done by Axer et al. [4] and by Hussmann et al. ([5]). While the former one is defunct, the later one is, though highly accurate, not very time-efficient. Our work differs from Hussmann et al. in many aspects, the most important being that the authors used linguistic features, e.g., type-token-ratio of open class words, while our approach will be based solely on acoustic features. The benefits of using a machine learning based approach for the automatic detection and evaluation of speech and language affecting diseases was demonstrated by various research groups. The authors of [6] showed the use of support vector machines and random forests for an high-accuracy classification of Parkinson's disease. Joshi et al. [7] successfully employed the Bag-of-Words (BoW) technique, in which low-level descriptors of different modalities are represented

as a histogram, in depression diagnosis. Similarly, [8], used Gaussian mixture models for the task of modeling depressed speech. The authors of [9] compared GMM-HMM and DNN-HMM approaches for the automatic assessment childhood apraxia of speech. In [10], a proof of concept system based on neural networks was used to yield an assessment of speech development issues in children. More recently, the authors of [11] used a support vector machine to distinguish between normal and pathological voice.

III. Aphasia

Aphasia can be defined as the acquired loss or impairment of language caused by brain damage. As such, aphasia does not pertain to the motor act of producing the sound making up speech or the motor act of writing letters, but to the supra-modal capability of the human brain to produce and comprehend language irrespective of the modality it is presented in or produced in. The remainder of III gives an overview of the various types of Aphasia, the Aachener Aphasia Test, the vocal biomarkers for aphasia detection and the Aachen aphasia database.

A. Aphasia types

In clinical routine and in scientific endeavours, there is a crude distinction into four large syndrome groups, namely global aphasia, Broca, Wernicke and anomic/amnestic aphasia ([3]). Very rare syndromes are conduction aphasia and transcortical aphasia, which by some are not regarded as central aphasias but as peripheral or dysexecutive aphasias [1]. Remarkably, up to 50% of all cases cannot be classified at all and are usually labeled as unclassified aphasia. The three largest groups (global, Broca and Wernicke) can be explained by the respective vascular territories that are affected by the causative ischemic stroke, with Brocas aphasia resulting from occlusion of anterior branches / superior division of the middle cerebral artery (MCA), Wernickes aphasia resulting from occlusion of posterior branches / inferior division of the MCA, and global aphasia being the result of occlusion of the common trunk (M1 segment) of the MCA. However, by employing statistical symptom-lesion mapping techniques, specific symptoms can be mapped to brain areas that are not defined by their vascularization ([12]). Generally speaking, the congruence between anatomical lesion patterns and neurolinguistic deficits is notoriously low ([13]), thus emphasizing again the need for tools based on clinical features. Also, deficit-specific therapy focuses more on individual deficits across the aforementioned linguistic modules than on syndromes alone.

B. Vocal biomarkers for aphasia detection

Markers of aphasia in speech differ according to the specific aphasia syndrome, with the main distinction being made between fluent and non-fluent syndromes. Global aphasia is the most severe form of non-fluent aphasia, with utterances consisting of one to two words only (mainly nouns) or even no verbal communication at all. In some cases, only automatisms or stereotypical utterances (“tan-tan-tan”) can be elicited. Neologisms are an important feature of global aphasia. Broca’s aphasia is a non-fluent aphasia, too, and is characterized by so-called agrammatism, i.e., a marked reduction in syntactic complexity sometimes leading to very brief sentences made up from subject and verb only. Additional word-finding difficulties will lead to many interjections (such as “eh” or “hm”) and to a reduction in lexicalic diversity. Finally, prosody is often lost in Broca’s aphasia. Wernicke’s aphasia on the other hand is the prototype of a fluent aphasia with preserved prosody. However, syntax is often overly complex with extremely long and twisted sentences, which is called paragrammatism. Overuse of function words leads to a relative reduction in open class words. Erroneous phonematic substitutions or paraphasias (“crain” instead of “train”) can occur as well as semantic paraphasias (“mother” instead of “wife”). Sometimes, unstoppable utterances occur consisting of meaningless semantic paraphasias or neologisms, but with fully preserved prosody, a phenomenon called semantic jargon. Amnestic or anomic aphasia is characterized by a pronounced deficit in word-finding capabilities. It therefore can be either fluent (if many circumlocutions are used) or non-fluent (if interjections and breaks occur). While lexicalic diversity can be reduced, prosody is often preserved. It should be remembered, though, that up to 50% of cases cannot be classified with a high degree of confidence using traditional algorithms such as ALLOC [10].

C. Aachen Aphasia Test (AAT)

The Aachen Aphasia Test is considered the gold standard for diagnosing and grading aphasia syndromes [14]. It has been translated to several other languages, such as English, Italian, French, Dutch, Portuguese and Thai, some of which have been validated in aphasic patients and show similar psychometric properties ([15], [16], [17], [18] and [19]). The AAT consists of mainly two parts: examination of spontaneous language, five subtests plus a token test. The interview examining spontaneous language is performed as a semi-structured interview of about 10 minutes, which is recorded during therapy sessions and transcribed afterwards (hence our data). The transcript is rated later according to a standardized manual in six domains, namely communicative behavior, articulation/prosody, formulaic language, semantics, phonology and syntax. The five additional subtests aim at examining naming, comprehension, repetition, reading and writing by respectively testing several sets of 10 items each. The grading of severity for each (sub)test ranges from 0 (unable to perform or massive deficits) to 5 (normal performance). The token test consists of 20 tokens that the patient has to arrange in front of himself according to increasingly complex criteria. This test is sensitive to even mild forms of aphasia and serves as a measure of overall aphasia severity. The resulting test profile can then be entered into classification algorithms such as ALLOC [20] to obtain a syndrome classification. According to Murray and Coppens, the AAT is used throughout Western Europe in addition to being mandatory for aphasia diagnosis in Germany.

IV. Aachen aphasia database

In this section, we describe the structure of the Aachen aphasia database and the pre-processing necessary to yield a structure suitable for the task at hand.

A. General structure of the database

The aphasia database was assembled in the Department of Neurology, University Hospital Aachen, during assessment and therapy sessions of aphasic patients over the course of roughly 20 years (1996 – 2016). The database consists of two main directory parts: patient records, e.g., notes and assessments, and speech samples, recorded during therapy sessions. Both the patient record directory and the speech sample directory contain year-based subdirectories. For each patient, an individual folder is created in the year-based subdirectory. It is named according to the first name, last name and day of birth, e.g., "Doe, John – born 08/11/1991". There can be several directories for the same patient, if the patient showed up in several years, e.g., 2011 and 2013. The speech sample part of the main directories consists of spontaneous speech samples in the mp3 file format. It further includes speech samples of the five AAT sub tests, also in mp3 file format. The spontaneous speech samples were conducted in form of interviews by speech therapists. Their voices were recorded as well and are included in each spontaneous speech sample (though in varying degrees). Each mp3 file is named according to the convention above, but contains further labeling like "Doe, John – born 08/11/1991 – spontaneous speech sample session 04/07/2011.mp3". The patient record part of the main directories contains, among other files, e.g., notes taken by the therapist, the assessment file of the spontaneous speech sample potion of the database. The assessment files are given in the Microsoft Word binary document format (.DOC). Here, the labeling of the assessment files is also done according to above convention, e.g., "Doe, John – born 08/11/1991 – spontaneous speech sample session assessment 04/07/2011.mp3". The assessment files contain the rating of each speech sample according to the six AAT rating scales: "Communicative Behavior", "Articulation and Prosody", "Formulaic Language", "Semantic Structure", "Phonological Structure" and "Syntactic Structure". For each of the rating scales the interviewer, e.g., the patients Speech and Language Therapist (SLT), had to assign a value from 0 (most severely impaired) to 5 (unimpaired) within the document.

B. Cleansing and consolidation

In order to yield a copy of the database suitable for using it in an automated way (i.e., as opposed to a doctor working manually with the database) , preprocessing of the database structure (IV-A) was necessary. As a first task, the patient record part and the speech sample part were consolidated into one unique directory for each patient. This novel directory then contained all of the patient records and the speech samples as compiled in several years. This task was done using regular expressions and yield some caveats. Foremost, the naming of files, which was performed manually by therapists, was sometimes not consistent or contained minor typos, e.g., within the patient name. Hence, the automatic identification of individual patients within the database was sometimes not possible, thus yielding two directories for the same patient. Furthermore, for some patients there was neither a speech sample nor a rating file available. These directories were omitted by our script. The second part was the mandatory anonymisation of the patient names throughout the database. Each patient name was replaced with a random UUID, e.g, "d98ab219f1f2". The replacing of the names had to be done in the file names and in the Microsoft Word DOC files. In order to achieve the latter one, all (binary) DOC files had to be converted into the XML based open DOCX file format, thus making it possible to search for all occurrences of the patient name within the file and replace them with an UUID. After cleansing and consolidation the database contains 705 spontaneous speech samples, some of them with transcriptions in the records, of 600 aphasia patients (though some of them can be the same person, see above). The speech samples total to roughly 107 hours. Statistics of the database are summarized in I.

TABLE I
Properties of the cleaned and consolidated database

Property	**Value**
Number of patients	600
Total hours of spontaneous speech	106:39 hrs
Number of spontaneous speech samples	705
Average length of speech sample	09:04 minutes

V. Automatic detection and evaluation of aphasia speech

In this section we describe our proposed approach for utilizing the biomarkers mentioned in section III-B and the preprocessed database described in section IV-B to build a system for automatic aphasia classification. The core idea behind the proposed system is to utilize the spontaneous speech samples of the database and their corresponding assessment according to the 6 dimensions mentioned in section III-C, e.g., articulation/prosody and phonology, to train a machine learning system. Once the system is trained, it shall, when presented with a novel speech sample of an Aphasia patient, extract relevant features of the sample, and subsequently rate it along the six dimension. Figure 1 depicts the proposed processing pipeline. The core of the system will make be based on the Bag-of-Audio-Words (BoAW) approach, as described in [7]. The remainder of this section is organized as follows: We first describe the necessary preprocessing of the spontaneous speech data in the database, namely through speaker speaker diarization and overlap detection. Next, we describe how we intend to use the vocal biomarkers to extract features of aphasia speech from the patients speech. Following the description of the feature extraction, we will describe the BoAW approach. The section concludes with a description of the classifiers and the medical classes we intend to use for the classifying of aphasic speech. Within the classifier

section, we also describe an alternative approach to aphasia classification, namely Long Short-Term Memory Recurrent Neural Networks.

A. Preprocessing of the database

As stated in section IV-A, the spontaneous speech samples contain not only patient speech, but also speech samples of the therapist. Since we are only interested in training the system on Aphasia speech, we have to extract the speech portion of the patient from the data, while omitting the speech of the therapist. Doing so, the total amount of hours of aphasic speech will be below 107 hours.For the extraction task, we will perform speaker diarization. Depending on whether therapist and patient speak in parallel or in sequence, the task will obviously differ in difficulty. Since we cannot be certain for every recording in the database which case occurs, we will employ overlap detection for every speech sample during speaker diarization, e.g., as outlined in [21]. According to the authors of [22], there are currently two main clustering approaches to speaker diarization: the bottom-up and the top-down approach. While usually only one cluster is used in the top-down approach, many clusters are used in the bottom-up approach. Which approach will work best in the aphasia domain at hand, will be evaluated during our experiments. Nevertheless, following the findings of [22], we will start with the seemingly better performing bottom-up approaches. Independently of the findings we will have to solve the task of automatically evaluating who of the two speakers represents the patient and who the interviewer. As a first step to tackle this challenge we propose to use simple heuristics derived from listing to a representative set of the speech samples. For instance, if it could be derived that the majority of the conversation is done by the patient, we could automatically assign the smaller portion of the diarizations output to the therapist. Since this approach, including the speaker diarization itself, is highly error prone, we propose to post-process its output manually before the actual feature extraction task, in order to ensure a high data quality.

B. Feature extraction

Generally, feature extraction in the domain of machine learning is the task of representing a set of data, e.g., a speech sample, by a fixed set of properties, e.g., length of the sample and average amplitude. For the feature extraction task of our idea we will explore two alternative approaches. For the first approach, we will evaluate the suitability of the open-source feature extractor openSMILE [23], while for the second approach we will test the vocal biomarkers as identified by the medical experts. So far, openSMILE has been successfully employed in language disease classification. For instance, the authors of [24] report how they used a total of 1 582 features for severity classification of Parkinsons Disease. Cummins et al. used openSMILE to extract a 39-dimensional feature vector for the classification of depressed speech [25] and in [26], it was used for autism spectrum disorder recognition. Given the the usage of openSMILE in these language disease classification settings, we will evaluate its performance on the given aphasia database. What and how many features we will use, will be evaluated within this work. Regarding the vocal biomarkers, we will have to evaluate which of them is suitable for the use in our intended system. As a first step, we will model the acoustic characteristics of word retrieval difficulties, e.g., stuttering.

C. Bag-of-audio-words

In the original BoW method in *natural language processing*, word histograms are used for classification of text documents, but the idea has been adapted for video and audio recognition tasks, introducing a preceding step of vector quantization (VQ) of numeric feature vectors. In the audio domain, the method is called Bag-of-Audio-Words (BoAW). The VQ is done based on a codebook of audio words, learned from the acoustic *low-level descriptors (LLDs)* of the audio data. Those LLDs are, e.g., *Mel-frequency cepstral coefficients*, spectral, or prosodic (such as the pitch of the human voice) features [23]. The BoAW method has already been thoroughly studied in the domain of *multimedia event detection* [27] and *acoustic event classification* [28]. In the field of medical diagnosis, besides *depression monitoring* [7], it has also been exploited for classification of *snore sounds* and recognition of a cold [29], [30]. Recently, the toolkit openXBOW has been published, which facilitates the generation of BoW in arbitrary domains [31]. Our goal is now to apply it for the task of aphasia classification. BoAW representations might be especially beneficial in unconstrained recording conditions as they are known to be robust. If the histograms are normalized, BoAW have the capability of capturing the nature of the signal independent from its length [32].

D. Classifier

In order to classify the feature representations which have been generated in an unsupervised manner so far, models are learned employing methods of machine learning (ML). One commonly used ML scheme is *support vector machine* (SVM), where a discriminative classifier is trained targeting a maximum margin between the defined classes in the feature space. In case of continuously valued target labels, e.g., describing the level of Aphasia speech, *support vector regression* (SVR) is used correspondingly. SVM – and respectively SVR – have the potential to cope with problems, where the underlying feature space is not linearly separable w.r.t. the target labels, by using a nonlinear *kernel*. A kernel which is especially suitable for histogram-based features, such as BoAW, is the so called *histogram intersection kernel* [27]. Besides, novel methods of deep learning will be explored for classification. In contrast to conventional feed-forward or recurrent neural networks, Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) are well-suited for classification or detection tasks in sequential data, such as speech signals [33], [34]. This is because they internally store information for an arbitrary period of time and thus are aware of the long-term context of the input signal without any need to represent the whole

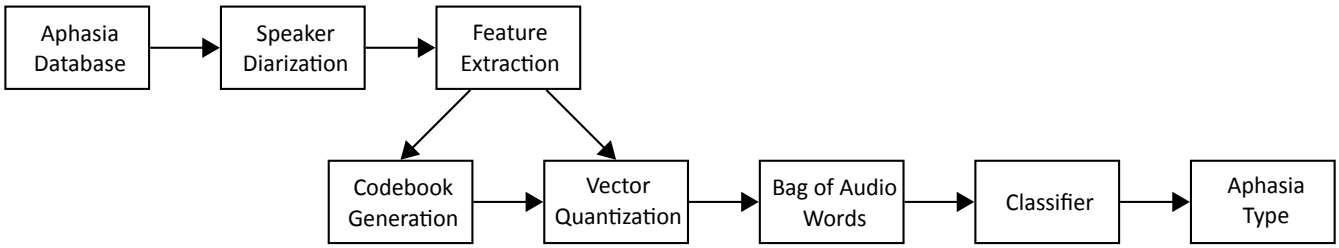

Fig. 1. *Processing pipeline of the aphasia system. Based on [7].*

segment of interest in one single BoAW feature vector. In case of using bidirectional LSTM-RNN, also future information is taken into account.

VI. Impact for patients, therapist and the research community

Although the inter-rater reliability of the AAT is quite good, it is not perfect. It gets worse with less-than-optimally-trained test persons. An automated algorithm employed at different rehab units across the country would make inter-site comparisons (e.g., in the frame of multi-center studies) much more valid. It would open new and efficient ways of conducting clinical trials for scientists. For service providers, it would enable them to monitor outcomes at multiple clinics, thus facilitating quality management across sites. For patients, it would potentially enable them to monitor their progress during home training programs, as stroke victims are often not very mobile and thus cannot attend SLP outpatient clinics easily. Finally, for highly specialized rehab units offering intensive comprehensive aphasia programs (such as Aachen Aphasia Ward of the University hospital), they could correlate language profiles sent to them from across the country with probabilities of therapeutic success (calculated ,e.g., from data also aggregated at our facility) to select suitable patients from those that might benefit more from other measures. This could pave the way to individualized rehabilitation strategies, potentially enabling insurance companies to cover a course of intensive language therapy for the right patients, while saving money on others.

VII. Conclusion

In this work, we outlined an approach to automatic aphasia detection and classification, given an already available large and annotated medical database of aphasia patient-interviewer dialogs. After the introduction in section I and the discussion of related work (section II), we gave a general overview over aphasia in section III. Next, we discussed the Aphasia Database of the Aachen University hospital IV and described how we consolidated and cleaned the data. We then outlined how we will use the data to build a system for the automatic evaluation of aphasia speech. To use the database for the given task at hand, we first pre-process its recordings to extract the patient portion of the dialog using speaker diarization and overlap detection. Next, we described our approach towards the feature extraction task. Here, we aim at two methods: on the one hand, the popular open-source toolkit openSMILE, which has be successfully employed for language disease classification, e.g., Parkinsons Disease; on the other hand, a set of defined vocal aphasia biomarkers, to extract aphasia features from the patients speech recordings. Using the BoAW approach, we represent each recording as as a bag of words. Given the aphasia classification of each recording within the database we then train a classifier to distinguish between severity levels of aphasia (e.g., "1" on the word retrieval scale). As an alternative approach, exploration of Long Short-Term Memory Recurrent Neural Networks appears promising given their suitability for aphasic speech classification. While large portions of the work remain to be implemented, it has strong foundations in previous findings. Furthermore, with a running system established, the data recorded offers the potential of tests under real life conditions using therapy sessions held in the hospital. After pre-processing the data in the database and speaker diarization, we will make the preprocessed database and its annotations available to the research community aiming at establishing a well-defined test-bed such as in the framework of the Interspeech Computational Paralinguistics Challenge [30] annually organised by part of these authors. Furthermore, as a parallel approach to the acoustic signal analysis and since a large portion of the database is already transcribed, we aim to test the additional usage of linguistic features for aphasia classification, e.g., cohesion, coherence and grammar. The bag-of-word principle provides an elegant fusion of the audio and linguistic word entities. A further future approach would be to also include the other, non-spontaneous, speech samples, into the training material and assess their value for building and evaluating the proposed system. Further options include the usage of Convolutional Neural Networks - potentially pre-trained on image data to ensure large data availability. The basis of analysis is then formed by spectograms in a "deep spectogram" [35].

Acknowledgements

The authors would like to thank Mr. Daniel Klischies for providing support regarding the cleaning and consolidation tasks of the database.

References

[1] A. Ardila, "A proposed reinterpretation and reclassification of aphasic syndromes," *Aphasiology*, vol. 24, no. 3, pp. 363–394, 2010.

[2] K. M. Heilman, "Aphasia and the diagram makers revisited: an update of information processing models," *Journal of Clinical Neurology*, vol. 2, no. 3, pp. 149–162, 2006.

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

[4] H. Axer, J. Jantzen, and D. G. von Keyserlingk, "An aphasia database on the internet: a model for computer-assisted analysis in aphasiology," *Brain and language*, vol. 75, no. 3, pp. 390–398, 2000.

[5] K. Hussmann, M. Grande, E. Meffert, S. Christoph, M. Piefke, K. Willmes, and W. Huber, "Computer-assisted analysis of spontaneous speech: quantification of basic parameters in aphasic and unimpaired language," *Clinical linguistics & phonetics*, vol. 26, no. 8, pp. 661–680, 2012.

[6] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinsons disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, May 2012.

[7] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Multimodal assistive technologies for depression diagnosis and monitoring," *Journal on MultiModal User Interfaces*, vol. 7, no. 3, pp. 217–228, 2013.

[8] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization." in *Proc. INTERSPEECH*, 2011, pp. 2997–3000.

[9] M. A. Shahin, B. Ahmed, J. McKechnie, K. J. Ballard, and R. Gutierrez-Osuna, "A comparison of gmm-hmm and dnn-hmm based pronunciation verification techniques for use in the assessment of childhood apraxia of speech," in *INTERSPEECH*, 2014.

[10] L. Ward, A. Stefani, D. Smith, A. Duenser, J. Freyne, B. Dodd, and A. Morgan, "Automated screening of speech development issues in children by identifying phonological error patterns," in *INTERSPEECH*, 2016.

[11] R. Amami and A. Smiti, "An incremental method combining density clustering and support vector machines for voice pathology detection," *Computers & Electrical Engineering*, vol. 57, pp. 257–265, 2017.

[12] I. Henseler, F. Regenbrecht, and H. Obrig, "Lesion correlates of patholinguistic profiles in chronic aphasia: comparisons of syndrome-, modality- and symptom-level assessment," *Brain*, p. awt374, 2014.

[13] K. Willmes and K. Poeck, "To what extent can aphasic syndromes be localized?" *Brain*, vol. 116, no. 6, pp. 1527–1540, 1993.

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

[15] N. Miller, K. Willmes, and R. D. Bleser, "The psychometric properties of the english language version of the aachen aphasia test (eaat)," *Aphasiology*, vol. 14, no. 7, pp. 683–722, 2000.

[16] R. De Bleser, G. Denes, C. Luzzati, A. Mazzucchi *et al.*, "L'aachener aphasie test (aat): I. problemi e soluzioni per una versione italiana del test e per uno studio crosslinguistico dei disturbi afasici." *Archivio di psicologia, neurologia e psichiatria*, 1986.

[17] P. Graetz, R. De Bleser, K. Willmes *et al.*, *Akense afasie test*. Swets & Zeitlinger Lisse, 1992.

[18] M. Lauterbach, I. P. Martins, P. Garcia, J. Cabeca, A. C. Ferreira, and K. Willmes, "Cross linguistic aphasia testing: the portuguese version of the aachen aphasia test (aat)," *Journal of the International Neuropsychological Society*, vol. 14, no. 06, pp. 1046–1056, 2008.

[19] N. Pracharitpukdee, K. Phanthumchinda, W. Huber, and K. Willmes, "The thai version of the german aachen aphasia test (aat): description of the test and performance in normal subjects." *Journal of the Medical Association of Thailand= Chotmaihet thangphaet*, vol. 81, no. 6, pp. 402–412, 1998.

[20] C. Lang, M. Spambalg, W. Sauerbrei, and T. Treig, *Computergestützte Klassifikation von Sprachstörungen bei Alzheimer-Demenz*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1989, pp. 129–131. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-83771-5_19

[21] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2007, pp. 683–686.

[22] X. A. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 356–370, 2012.

[23] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of the 21st ACM International Conference on Multimedia*, ser. MM '13, 2013, pp. 835–838.

[24] A. Bayestehtashk, M. Asgari, I. Shafran, and J. McNames, "Fully automated assessment of the severity of parkinson's disease from speech," *Computer speech & language*, vol. 29, no. 1, pp. 172–185, 2015.

[25] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke, "Modeling spectral variability for the classification of depressed speech," in *Proc. INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 857–861.

[26] H.-Y. Lee, T.-Y. Hu, H. Jing, Y.-F. Chang, Y. Tsao, Y.-C. Kao, and T.-L. Pao, "Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition," in *Proc. INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 215–219.

[27] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Proc. INTERSPEECH*, Portland, USA, 2012, pp. 2105–2108.

[28] H. Lim, M. J. Kim, and H. Kim, "Robust sound event classification using lbp-hog based bag-of-audio-words feature representation," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 3325–3329.

[29] M. Schmitt, C. Janott, V. Pandit, K. Qian, C. Heiser, W. Hemmert, and B. Schuller, "A bag-of-audio-words approach for snore sounds' excitation localisation," in *Proc. ITG Speech Communication*, VDE. Paderborn, Germany: IEEE, 2016, pp. 230–234.

[30] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. S. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Proc. INTERSPEECH*. Stockholm, Sweden: ISCA, 2017, to appear.

[31] M. Schmitt and B. W. Schuller, "openXBOW-introducing the passau open-source crossmodal bag-of-words toolkit," *preprint arXiv:1605.06778*, 2016.

[32] M. Riley, E. Heinen, and J. Ghosh, "A text retrieval approach to content-based audio hashing," in *ISMIR 2008, 9th International Conference on Music Information Retrieval*, Philadelphia, USA, 2008.

[33] R. Brückner and B. Schuller, "Social signal classification using deep blstm recurrent neural networks," in *Proc. of the 39th IEEE International Conference on Acoustics, Speech, and Signal Processing*. Florence, Italy: IEEE, 2014, pp. 4856–4860.

[34] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, November 2015.

[35] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller, "Snore Sound Classification Using Image-based Deep Spectrum Features," in *Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*, ISCA. Stockholm, Sweden: ISCA, August 2017, 5 pages.