Manual and Automatic Transcriptions in Dementia Detection from Speech

Jochen Weiner, Mathis Engelbart, Tanja Schultz

Cognitive Systems Lab, Universität Bremen, Germany

jochen.weiner@uni-bremen.de

Abstract

As the population in developed countries is aging, larger numbers of people are at risk of developing dementia. In the near future there will be a need for time- and cost-efficient screening methods. Speech can be recorded and analyzed in this manner, and as speech and language are affected early on in the course of dementia, automatic speech processing can provide valuable support for such screening methods.

We present two pipelines of feature extraction for dementia detection: the *manual pipeline* uses manual transcriptions while the *fully automatic pipeline* uses transcriptions created by automatic speech recognition (ASR). The acoustic and linguistic features that we extract need no language specific tools other than the ASR system. Using these two different feature extraction pipelines we automatically detect dementia. Our results show that the ASR system's transcription quality is a good single feature and that the features extracted from automatic transcriptions perform similar or slightly better than the features extracted from the manual transcriptions.

Index Terms: computational paralinguistics, dementia, cognitive impairment

1. Introduction

Alzheimer's disease is a degenerative disease which accounts for 70 % of all dementia cases [1]. While there is no known cure, therapy can help delay the course of the disease. The sooner it is diagnosed the more the people affected by it can benefit from therapy [2]. In the optimal case, therapy can start before the disease as wrought any noticeable damage. As many people as possible should be offered therapy at this optimal point in time which means that large numbers of people will have to undergo screening and be diagnosed before the disease has impacted their lives. This requires a cost- and time-effective diagnosis of dementia which the current diagnosing process cannot deliver. To help the diagnosing process meet future demands new automatic methods are being proposed which support clinicians in making their diagnosis [3, 4, 5, 6, 7].

Dementia affects human speech and language from a very early stage and changes in speech and language use are strong indicators for dementia [8, 9]. Automatic speech processing has been shown to be a promising way forward in the diagnosis of dementia. Approaches have used acoustic or prosodic features [3, 4, 5, 10, 11] and linguistic or text-based features [12, 6, 7, 10, 13, 14, 15, 16, 17] in a classification task that aims to distinguish speakers affected by dementia from cognitively health speakers using just their speech.

We can obtain patients' speech and language from written texts and speech recordings. Spontaneous conversational speech is the most natural form of speech and language use. Unlike speech recorded during tests or examinations spontaneous conversational speech provides an unobstructed view on patients' speech capabilities. It can easily be recorded by a person with minimal diagnostic knowledge and can also be recorded remotely (e.g. via telephone). This means that the elderly patients are not required to travel to a clinic in order to be examined for dementia. Once speech is recorded we can use automatic speech recognition (ASR) to create transcriptions automatically. Speech properties such as jitter, shimmer and breathiness change as speakers age [18], which is why the quality of ASR transcripts tends to over time [19]. Nonetheless automatic transcripts of elderly speech have successfully been used to detect dementia in standardized texts [12, 7, 15] although the word error rates for individual subjects reached over 90% [15]. Still, many of the approaches that have used text based features relied on manual transcriptions [10, 13]. If we find dementia detection using features extracted from automatic transcriptions to work as well as dementia detection using features extracted from manual transcripts, then we no longer need to rely on manual transcriptions. This would enable a faster automatic diagnosis. In research we could use more data per patient, since an ASR system will transcribe speech faster and cheaper than a reliable manual transcription process.

In this paper we investigate automatic dementia detection using features extracted from automatic transcriptions of spontaneous conversational speech. We use both acoustic and linguistic features and compare the results to results obtained by features extracted from manual transcriptions. We will show that using our automatic features we achieve dementia detection results that are better than or on par with the result when using manual transcriptions.

This paper has the following structure: In Section 2 we describe our dataset of interview recordings. Section 3 presents the ASR system which we use to extract the features described in Section 4. We then use these features for our experiments in Section 5 and finally discuss our results in Section 6.

2. Database

The speech recordings and cognitive diagnoses that we use in this work are a selection of data [11] from the *Interdisciplinary Longitudinal Study on Adult Development and Aging* (ILSE) [20]. ILSE was created to facilitate research in participants' personality, cognitive functioning, subjective well-being and health. Over the course of more than 20 years participants in Germany were invited to take part in four measurements in which a large corpus of data was collected. From ILSE's wealth of data we use two data sources: recordings of biographic interviews and cognitive diagnoses established by psychiatrists using a range of neuropsychological, anamnestic, clinical, and laboratory tests.

The ILSE participants form a group that represents the sampled population (cf. [20, 21]). When the study started, the participants were either 40 or 60 years old. Gerontologists consider people at this age to be young and expect very few cases of cognitive impairment. In accordance with these expectations, most of the ILSE participants had no cognitive impairment when the study began. As the study progressed and the participants grew older, some of them developed cognitive impairments: Our selection of data [11] contains 74 participants in three measurements since data from the fourth measurement is not yet available. The participants are either diagnosed with aging-associated cognitive decline (AACD) or Alzheimers disease (AD), or are members of the control group (see Table 1). The severity of AACD or AD was not recorded in ILSE and is therefore unavailable to us. We do, however, know that some participants dropped out of the study because they felt unable to participate. It's safe to assume that participants with very severe AD are amongst those who dropped out and we expect that we do not have any participants with very severe AD in our dataset. Interviews became shorter as the study progressed but there is no substantial difference in the duration of interviews by speakers with different diagnoses.

Table 1: *Cognitive diagnoses from the three measurements (with times of the data collections) in the data selected from ILSE.*

		Diagnoses		
		control	**AACD**	**AD**
	1 (1993-1996)	51	4	-
Measurement	**2 (1997-2000)**	19	8	-
	3 (2005-2008)	10	1	5
Total		80	13	5

Since the study started with young participants and followed them over time, the cases of cognitive impairment in the dataset developed as the study progressed. The dataset thus represents a distribution of diagnoses which we would expect to see in any representative sample of people in Germany [2, p. 20]. From a machine learning point of view this means that while we have a highly unbalanced dataset, we also have a dataset that represents the real world: if people of a certain age group were consequently screened for cognitive impairment, we would find an unbalanced distribution of diagnoses very similar to the one we find in our dataset.

Using a long audio alignment procedure [20], we segmented the interview recordings and matching manual transcriptions into short segments. Our data selection contains only participant speech and no interviewer speech, a total of 230 hours of audio recordings.

3. Automatic Speech Recognition System

We created automatic transcriptions of the selected interview turns (Section 2) using our in-house speech recognition toolkit BioKIT [22] and our existing automatic speech recognition (ASR) system for the ILSE interviews [23, 24]:

The acoustic model is a deep neural network (DNN) trained on a 256-hour-training set. The input for the DNN is a 440-dimensional feature vector consisting of 11 stacked 40 dimensional LDA transformed stacked MFCC feature vectors. The DNN training in the Kaldi toolkit [25] consists of pre-training, cross-entropy training and finally state-level minimum Bayes risk (sMBR) sequence training. The network has 6 hidden layers with 2,048 nodes each and an output layer with 3,194 nodes. The dictionary contains 76,533 pronunciations for 71,885 distinct words. The language model (LM) is a 3-gram Kneser-Ney language model trained on sentences of the training set. It has a perplexity of 199.38 on the ASR development set.

We ran our error correction system using error signatures [23] on the ASR system and subsequently changed all three components of the system. Table 2 sums up the decoding results for the three cognitive diagnoses in our dataset: For each cognitive diagnosis the middle column shows the WER of all the automatic transcripts of the participants with that diagnosis. The right column shows the standard deviation of the WERs of the automatic transcripts of individual interviews from the mean WER over all interviews. We see that the overall WER increases with cognitive impairment and that the deviation decreases.

Table 2: *Word error rates (WER) for the participant diagnoses.*

participant diagnosis	**Overall WER**	**mean WER**	**std. WER**
control	56.0 %	58.2 %	14.9 %
AACD	60.4 %	60.8 %	12.1 %
AD	70.2 %	70.1 %	7.0 %
all	58.5 %	59.2 %	14.4 %

4. Features for Dementia Detection

We differentiate three categories of features: acoustic features, linguistic features and ASR features. Linguistic features measure *what* participants say while acoustic features measure *how* they speak. ASR features measure ASR performance for individual participants' speech. All features are derived on a per-interview level.

4.1. Acoustic Features

Acoustic features measure how participants speak. Some of these features like speaking rates require transcripts. We therefore extract these features once using manual transcripts and once using ASR transcripts.

4.1.1. Speech Pause-based Features

Speech pause-based features include speech pause durations, rates, counts and the ratios between speech pauses and words (for a full description see Weiner et al. [11]). These features are calculated from audio, a voice-activity-based speech pause detection and transcriptions.

4.1.2. Speaking Rate Features

Speaking rate is measured in words per second and phones per second [11]. Audio, transcriptions and a pronunciation dictionary are used to calculate these features.

4.2. ASR Features

ASR features make use of the fact that our ASR system performs differently for different participants. We only extract these features when the texts from which the other features are extracted have been created by an ASR system.

Word Error Rate Feature is the word error rate (WER) of the automatic transcription for the recording of an interview. Since we need a textual reference to calculate the WER of an ASR system's output, this feature is not a truly automatic feature. However, we use this feature when we have manual transcriptions available, so that we can learn about the influence of transcription quality on the dementia detection task.

4.3. Linguistic Features

We use linguistic features to measure changes to vocabulary and sentence structure that are caused by dementia. Our linguistic features operate at the word surface level. They require no domain- or language-specific information such as word usage frequencies or tools such as parsers or part-of-speech taggers. This makes them easily portable and less susceptible to errors introduced by external resources or tools. We extract these features once using manual transcripts and once using ASR transcripts.

4.3.1. Lexical Richness Feature

Lexical richness measures the participants' usage of their vocabulary. The lexical richness feature contains these two measures:

Brunet's W index [26, 27] measures the speakers' lexical richness using the number of spoken words N, the vocabulary size V and a constant factor $a = -0.172$ [27]:

$$W = N^{V^{-a}}$$

Honoré's R Statistics [28] makes use of the number of spoken words N, the vocabulary size V and the number of words uttered exactly once V_1:

$$R = \frac{100 \cdot \log N}{1 - (V_1/V)}$$

4.3.2. Perplexity Features

In applications like automatic speech recognition and machine translation, language models are used to model the probability of word sequences. Models are trained on a corpus of texts and their performance is measured by calculating their perplexity on a held-out set of texts. Perplexity measures how well the model fits the data. Intuitively, perplexity can be interpreted as the average branching factor of the texts. The lower the perplexity the better the match between the model (representing its training data) and the held-out texts. We use SRILM [29] to train 1-, 2-, 3-, 4- and 5-gram LMs with Witten-Bell discounting and to evaluate the LMs on texts. As features we extract the results of SRILM's perplexity evaluation: logarithmized text probability (logprob), perplexity (ppl) and perplexity-without-sentence-end (ppl1) for all five n-gram models as well as the number of sentences, the number of words, and the number of out-of-vocabulary words (OOVs). We calculate two different types of perplexity features:

Within-Speaker Perplexity Features are derived by evaluating LMs on texts from the same speaker they were trained on: In a 10-fold cross-validation 90 % of a participant's sentences in one interview are used to train the LMs which are evaluated on the remaining 10 % of the participant's sentences. We then take the arithmetic means between the ten sets of features as final features [14].

Between-Speaker Perplexity Features are obtained by training and evaluating LMs in a leave-one-person-out cross validation: The texts of all but one participant are used to train the LMs. Perplexity features are then calculated for each interview with the left-out participant.

5. Feature Selection and Classification

We extract two different versions of our features from the interview recordings:

(a) the manual version using manual transcriptions

(b) the automatic version using transcriptions created by the ASR system

The features from these pipelines are processed both as individual features or as combined features (early fusion). In both cases we select the best features using mutual information as a selection indicator.

Since we have a small dataset, we train and evaluate our models in a cross-validation. Each participant contributed to the study in more than one measurement, so we need to ensure that we do not include different measurements from one participant in both training and test. We therefore use a leave-one-person-out cross validation[1]: The model is trained on the data from all but one participant and then evaluated on the participant that was not used in training. This cross-validation approach leads to a very robust estimation of the classifier performance.

We trained and evaluated different classifiers: k-nearest neighbors (kNN), linear discriminant analysis (LDA), Gaussian classifier and support vector machines (SVMs). With very few exceptions the Gaussian classifier achieved the best classification results which is why we are only reporting results from the Gaussian classifiers. Our experiments are based on the implementations in scikit-learn [30].

6. Experimental Results

We use *unweighted average recall (UAR)* to evaluate our experiments. This metric gives equal weight to all three classes (*control*, *AACD* and *AD*) and is therefore more suitable to this unbalanced dataset than a weighted metric such as accuracy [31]. The classification chance level for a three-class classification is at UAR = ⅓. In the result graph, the chance level is shown by a dotted line.

Figure 1 shows the results of the dementia detection experiments. The automatic feature versions achieve better results than the manual versions for all but three features. Furthermore, the automatic version of the majority of the features outperforms the manual version by a good margin. The best performing automatic version of a feature is the automatic version of the within-speaker perplexity with UAR $= 0.623$. While this is the best result of all the automatic versions of the features, it is also the overall best result in our experiments. The automatic versions of the features do not only outperform their manual counterparts for most features, but the best overall result is also achieved by an automatic feature version.

There are only small differences between the results from the two versions of the pause-based features, speaking rate features, lexical richness features and the combination of pause-based features and speaking rate features. The manual versions of three of these features obtain results better than their automatic counterparts. The manual version of the lexical richness feature achieves the best result of all manual feature versions with UAR $= 0.606$. Overall this is the second best result. The next ranks in the result comparison all go to automatic versions of features.

The word error rate (WER) by itself is a strong feature (UAR $= 0.520$). This shows that the ASR system provides different transcription qualities for the different diagnoses even

[1] In previous work [11] we used a different cross-validation method.

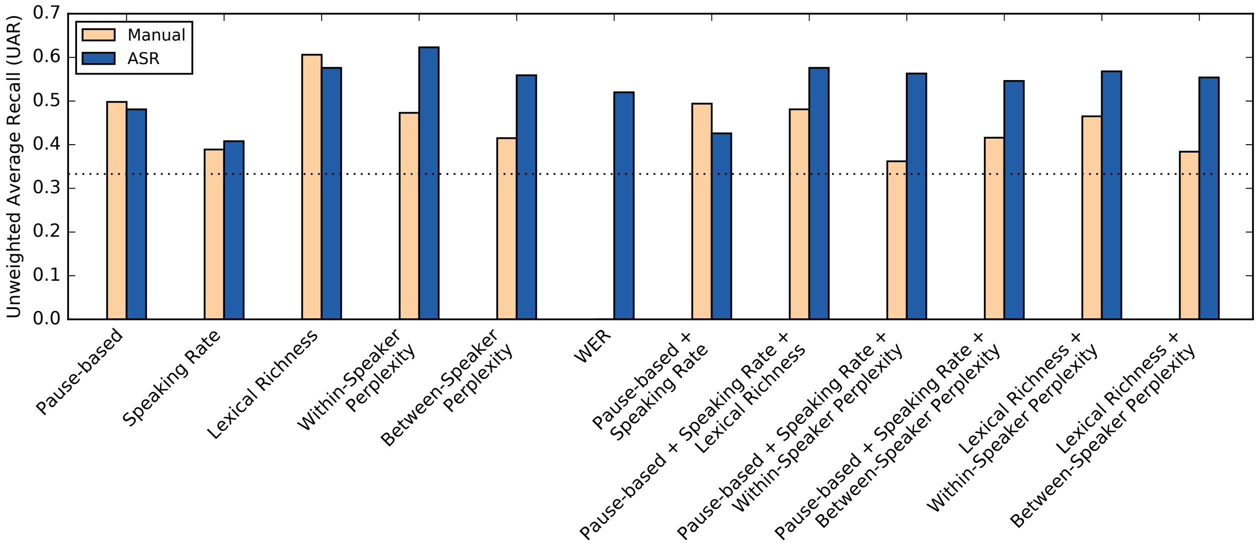

Figure 1: *Comparison of dementia detection results between features extracted from manual and automatic transcriptions. The six results on the left are produced by individual features and the six results on the right by combinations of features. The dotted line represents the chance level (UAR = ⅓).*

at the high overall WER level of the ASR system and this information is encoded in the texts from which all the automatic variations of the features are extracted.

Figure 1 also shows the results of experiments conducted with combinations of linguistic features and acoustic features. The automatic versions of the linguistic features and any feature combinations that contain linguistic features outperform both automatic and manual versions of purely acoustic features (pause-based features, speaking rate features and their combination). This shows that our linguistic features are robust against transcription errors.

We see, however, that the mutual information criterion does not find the optimal feature selection: No combination of manual feature versions achieves a result close to or better than the result with the individual lexical richness feature. Looking at the automatic versions of features, no feature combination achieves a result comparable to that of only the within-speaker perplexity. An optimal feature selection would find a selection of features that yields at least have the performance of that feature since it could select just that feature. We therefore tried feature selection based on the ANOVA f-value, but this method does not find the best selection of features either. The best result in our experiment are therefore achieved by individual features extracted from ASR transcripts.

7. Conclusions and Future Work

The increasing numbers of elderly people who need to be screened for dementia require automatic approaches to support clinicians in making dementia diagnoses. We have investigated a fully automatic approach to dementia detection using speech. This approach uses an ASR system to transcribe speech and then extracts acoustic and linguistic features from the audio and the automatic transcriptions. Our linguistic features have the advantage that they do not depend on language-specific tools or resources.

In our experiments the automatic versions of our features outperform their manual counterparts. The best overall result (UAR $= 0.623$) is achieved by the automatic version of a linguistic feature. Together with the high word error rates of our automatic transcriptions this result shows that our choice of features is robust against transcription quality. Furthermore we found the transcription quality itself (represented by the word error rate) to be a good feature for dementia detection. Since manual transcripts are required to measure transcription quality in terms of word error rate, the word error rate feature is not a fully automatic feature. In the future we will therefore investigate replacing the word error rate feature with a measure of transcription quality that does not require manual input.

In the experiments we presented in this paper feature combination by early fusion did not yield any improvements in the classification result. Our next step will therefore involve investigating whether late fusion (result fusion) will lead to a more rewarding feature combination.

Our findings let us conclude that we do not need a manual transcription of our participants' speech in order to detect dementia on the ILSE corpus using our features. Instead we can rely on transcriptions created fully automatically by ASR systems which gives us access to a lot more data than is currently available with a manual transcriptions. In ILSE, for example, there are 8,000 hours of untranscribed interviews which we will now access with automatic speech recognition. Since we are already using spontaneous conversational speech we have shown that a recording of a conversation that keeps the patient comfortable may be all the input needed to fully automatic detect dementia using acoustic and linguistic features.

8. Acknowledgements

We thank Johannes Schröder at the University of Heidelberg and the University Hospital Heidelberg, Germany for providing us with the ILSE data and for sharing with us his expertise on speech and language use of people with dementia.

9. References

[1] World Health Organization and Alzheimers Disease International, *Dementia: a public health priority*. World Health Organization, 2012.

[2] M. Prince, A. Wimo, M. Guerchet, G.-C. Ali, Y.-T. Wu, and M. Prina, *World Alzheimer Report 2015. The Global Impact of Dementia: an Analysis of Prevalence, Incidence, Cost and Trends*. London: Alzheimer's Disease International, 2015.

[3] A. Satt, R. Hoory, A. König, P. Aalten, and P. H. Robert, "Speech-based automatic and robust detection of very early dementia." in *INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association*, 2014, pp. 2538–2542.

[4] F. Espinoza-Cuadros, M. A. Garcia-Zamora, D. Torres-Boza, C. A. Ferrer-Riesgo, A. Montero-Benavides, E. Gonzalez-Moreira, and L. A. Hernandez-Gómez, "A spoken language database for research on moderate cognitive impairment: design and preliminary analysis," in *Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2014, pp. 219–228.

[5] L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, and G. Szatlóczki, "Automatic detection of mild cognitive impairment from spontaneous speech using asr," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, 2015, pp. 2694–2698.

[6] E. T. Prud'hommeaux and B. Roark, "Extraction of narrative recall patterns for neuropsychological assessment." in *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association*, 2011, pp. 3021–3024.

[7] M. Lehr, E. T. Prud'hommeaux, I. Shafran, and B. Roark, "Fully automated neuropsychological assessment for detecting mild cognitive impairment." in *INTERSPEECH 2012 – 13th Annual Conference of the International Speech Communication Association*, 2012, pp. 1039–1042.

[8] J. Appell, A. Kertesz, and M. Fisman, "A study of language functioning in Alzheimer patients," *Brain and language*, vol. 17, no. 1, pp. 73–91, 1982.

[9] R. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance," *Aphasiology*, vol. 14, no. 1, pp. 71–91, 2000.

[10] A. Khodabakhsh, F. Yesil, E. Guner, and C. Demiroglu, "Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–15, 2015.

[11] J. Weiner, C. Herff, and T. Schultz, "Speech-Based Detection of Alzheimer's Disease in Conversational German," in *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association*, 2016.

[12] D. Hakkani-Tür, D. Vergyri, and G. Tür, "Speech-based automated cognitive status assessment." in *INTERSPEECH 2010 – 11th Annual Conference of the International Speech Communication Association*, 2010, pp. 258–261.

[13] L. Hernández-Domínguez, E. García-Cano, S. Ratté, and G. Sierra-Martínez, "Detection of Alzheimer's disease based on automatic analysis of common objects descriptions," in *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, 2016.

[14] S. Wankerl, E. Nth, and S. Evert, "An Analysis of Perplexity to Reveal the Effects of Alzheimer's Disease on Language," in *12th ITG Conference on Speech Communication*, 2016.

[15] L. Zhou, K. C. Fraser, and F. Rudzicz, "Speech recognition in alzheimers disease and in its assessment," in *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association*, 2016, pp. 1948–1952.

[16] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp, "Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech," in *IEEE International Conference Mechatronics and Automation, 2005*, vol. 3, 2005, pp. 1569–1574 Vol. 3.

[17] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*, 2014, pp. 27–36.

[18] V. Young and A. Mihailidis, "Difficulties in Automatic Speech Recognition of Dysarthric Speakers and Implications for Speech-Based Applications Used by the Elderly: A Literature Review," *Assistive Technology*, vol. 22, no. 2, pp. 99–112, 2010.

[19] R. Vipperla, S. Renals, and J. Frankel, "Longitudinal study of asr performance on ageing voices," in *INTERSPEECH 2008 – 9th Annual Conference of the International Speech Communication Association*, 2008, pp. 2550–2553.

[20] J. Weiner, C. Frankenberg, D. Telaar, B. Wendelstein, J. Schröder, and T. Schultz, "Towards Automatic Transcription of ILSE – an Interdisciplinary Longitudinal Study of Adult Development and Aging," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016.

[21] P. Martin and M. Martin, "Design und Methodik der Interdisziplinären Längsschnittstudie des Erwachsenenalters," in *Aspekte der Entwicklung im mittleren und höheren Lebensalter: Ergebnisse der Interdisziplinären Längsschnittstudie des Erwachsenenalters (ILSE)*, P. Martin, K. U. Ettrich, U. Lehr, D. Roether, M. Martin, and A. Fischer-Cyrulies, Eds. Steinkopff, 2000, pp. 17–27.

[22] D. Telaar, M. Wand, D. Gehrig, F. Putze, C. Amma, D. Heger, N. T. Vu, M. Erhardt, T. Schlippe, M. Janke, C. Herff, and T. Schultz, "BioKIT - Real-time decoder for biosignal processing," in *INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association*, 2014, pp. 2650–2654.

[23] D. Telaar, J. Weiner, and T. Schultz, "Error Signatures to identify Errors in ASR in an unsupervised fashion," in *Proceedings of the Errare Workshop (ERRARE 2015)*, 2015.

[24] D. Telaar, "Error Correction based on Error Signatures applied to automatic speech recognition," Ph.D. dissertation, Karlsruhe Institute of Technology, 2015.

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

[26] E. Brunet, *Le vocabulaire de Jean Giraudoux, structure et évolution*. Geneva: Slatkine, 1978.

[27] F. J. Tweedie and R. H. Baayen, "How Variable May a Constant be? Measures of Lexical Richness in Perspective," *Computers and the Humanities*, vol. 32, no. 5, pp. 323–352, 1998.

[28] A. Honoré, "Some Simple Measures of Richness of Vocabulary," *Association for Literary and Linguistic Computing Bulletin*, vol. 7, no. 2, pp. 172–177, 1979.

[29] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *International Confonference on Spoken Language Processing*, 2002.

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

[31] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association*, 2016.