



Language-independent method for analysis of German stuttering recordings

Tomas Lustyk¹, Petr Bergl¹, Tino Haderlein², Elmar Nöth², Roman Cmejla¹

¹Czech Technical University in Prague, Faculty of Electrical Engineering,
Department of Circuit Theory,
Technická 2, 166 27, Prague, Czech Republic

²Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg,
Martensstraße 3, 91058 Erlangen, Germany

lustytom@fel.cvut.cz, cmejla@fel.cvut.cz, Bergl@seznam.cz, noeth@cs.fau.de,
Tino.Haderlein@cs.fau.de

Abstract

The paper describes experiments where automatic acoustic algorithms initially intended to be used on Czech stuttering speakers were applied on recordings of German stuttering speakers. Four algorithms based on voice activity and abrupt spectral changes detection are introduced. The database consists of 34 speakers. The measure, the number of abrupt spectral changes in speech segments, reached a correlation with fluency rating of 0.85. The other measures have also good agreement with subjective evaluation. Results indicate that it could be basically possible to do language-independent analysis of stuttering, here demonstrated on read recordings of German speakers.

Index Terms: automatic algorithms, stuttering, disfluency, Czech, German, language-independent

1. Introduction

Stuttering belongs to speech fluency disorders and is characterised by symptoms, namely repetitions (of sounds, syllables, word parts, words, or phrases), prolonged sounds, interjections, revisions, incomplete phrases, and broken words [1]. Each patient has her/his individual symptoms. The disorder also affects the psychological and social state of a person [2]. Clinical experts, e.g. stuttering therapist and phoneticians, evaluate the performance of a speaker to get an overall view on the disorder. Correct assessment of the disorder and its follow-up treatment are very difficult tasks. Several stuttering scales have been described, for instance, the Lidcombe Behavioral Language of Stuttering [3] and the Stuttering Severity Index [4], but there has also been a need for an automatic and objective method. Such a method could help with diagnosis, choice of treatment approach, evaluation of treatment progress, or speakers' home evaluation [5, 6].

The symptoms of stuttering may be evaluated in audio recordings to determine the degree of the speech disorder. Acoustical methods have been employed for objective analysis of stuttered speech. Some of researches focused on phoneme duration [7], rate of speech (manually measured) [8, 9], formant frequencies [10], fundamental frequency and voice onset time [11]. For the purpose of evaluation, methods based on speech recognition can be also successfully utilized [12]. An approach to different speech problems can be found as an inspiration for automatic analysis of stuttering, for example, Parkinson's disease [13], children's articulation disorder [14], or second lan-

guage learners' fluency [15, 16].

The following pages give an insight on how the automatic algorithms initially intended to be used on Czech stuttering recordings perform when applying them on German stuttering recordings. One standard measure and four automatic measures of disfluency are introduced and their results are discussed. Questions, we would like to answer, were: Are the measures able to describe the level of speech fluency in different languages? If yes, how good they are? Is it possible to classify fluent and disfluent recordings?

2. Database

The database in this experiment contains 34 recordings – 16 signals of stuttering speakers, 18 recordings of speakers who do not have problems with fluency. All the subjects are males and German native speakers. All participants read the same phonetically rich text "Nordwind und Sonne" (North Wind and Sun). The text is 108 word long and the average duration of reading is 62.5 s (\pm SD, 26.1 s). The number of disfluencies is known for each recording and constitutes the subjective evaluation. The stuttering part of the database was used in the research experiment published in [12].

3. Algorithms

The measurements were designed to take into account symptoms of stuttering and disfluent speech. They were originally proposed to be used for Czech stuttering speakers but they do not take specifics of the Czech language into account. The list of all measures is given in Table 1. The measures use voice activity detection and detection of abrupt spectral changes.

The measures are based on the assumption that the prosody of stutterers' speech is different from fluent speech. For example, subjects suffering from stuttering have a slower speech rate than fluent speakers do [8, 9] and that the speech rate and pauses can discriminate the speech of stutterers from that of nonstutterer [17].

The first measure is the *average length of silence* (ALS). It is based on the voice activity detection (VAD) and assumes that subjects with stuttering have more and/or longer silence and pauses in speech than healthy subjects do, study [17] indicate that pauses are potential perceptual cues for listeners attempting to discriminate the speech of stutterers from that of nonstutterers. When the positions of speech/silence parts is known

Table 1: List of the measures.

measure	description
average length of silence (ALS)	average duration of silent parts identified by voice activity detector
extent of speech fluency (ESF)	number of abrupt spectral changes
spectral changes in short interval (SCSI)	average number of abrupt spectral changes in short time windows
number of spectral changes in speech segments (NSI)	number of abrupt spectral changes included in speech segments

the average duration of silent parts can be computed. The more disfluent a signal is, the more and/or longer pauses appear and the ALS increases. The value of ALS rises with the level of the disorder. To make the difference between fluent and disfluent speech more perceptible, an innovative procedure was added to the algorithm. Short speech parts, e.g. repetitions, superfluous verbal behavior, and parts of incorrectly pronounced words, could be removed by this method and the amount of silence increases in a speech signal after taking them out. The procedure uses successive removal of short segments of speech and silence in regular intervals from 100 up to 1500 ms. The removal threshold is subject of finding the most appropriate setting.

Three remaining measures are based on the detection of abrupt spectral changes (ACs). The significant abrupt spectral changes found in a speech signal should correspond to the phoneme boundaries in the signal. The Bayesian autoregressive changepoint detector (BACD) identifies abrupt spectral changes in this experiment. The detector is based on the analytical solution of the changepoint problem between two autoregressive models [18]. Its detailed description is given in [19]. Firstly, all spectral changes are found in a signal. Followed by selection of those which proved to be significant and correspond to the phoneme boundaries. To distinguish among significant and less significant ACs, a threshold determined as a fraction of the k -th highest maximum in the BACD output is used (the height of the maximum play important role in this selection). The fraction value (multiplication constant) and the rank of maximum are subjects of finding the most appropriate setting. After the identification the procedures of particular measures start to differ.

The *extent of speech fluency* (ESF) counts the number of those significant ACs which are higher than the threshold (correspond to phoneme boundaries). The number of ACs is divided by the duration of the speech signal (analogy with the speech rate). The ESF assumes that the disfluent speakers speak slower (have a lower number of phonemes) than fluent speakers do [8]. In other words, the more disfluent the signal is the less ACs is found in the signal.

The *spectral changes in short interval* (SCSI) assumes that the speech of fluent speakers is more steady and regular, while on the contrary, stuttering speakers interrupt speech with pauses, repetitions, prolongations. In the terms of the time distribution of ACs, there could be differences. If the BACD output of a disfluent signal is sequentially processed by a window (the tested length of window was 1, 2, and 4 s with half overlap), the number of significant ACs in each segment would vary a lot over time and in many cases be zero, when there is no speech there are no ACs, the average number of changes in window would be smaller. Conversely, the number of ACs in the window for healthy speakers could be more stable and regular, and the average value should be higher.

The third measure, the *number of spectral changes in speech segments* (NSI), makes the same assumption as the ESF. It combines the BACD and VAD. The algorithm looks for sig-

nificant spectral changes which are higher than the threshold and within speech segments, then divided by the length of the recording (analogy with the speech rate). A very important part of the NSI algorithm is successive removal of short speech segments (similar to the measure ALS). The time limit used in this experiment were from 100 up to 1500 ms in regular intervals. The measures NSI and ESF should highly coincide with the rate of speech measured as phonemes/time.

A more detailed description of the measures is given in [20] where experiments on read stuttering recording of Czech speakers are carried out.

4. Results

To evaluate the performance of the algorithms on German recordings, Pearson's correlation and statistical analysis ANOVA are utilized. The measures are compared to the number of disfluencies in utterance (subjective evaluation). The total reading time (RT) – the duration of recording in second, a standard measure used to evaluate pathological speech [21, 22], is added and its results are displayed along with others.

First, the correlation between settings of the measures and the subjective evaluation are presented. These are ALS, ESF, SCSI, and NSI in Tables 2, 3, 4, and 5, respectively. The measures reveal good agreement with the subjective evaluation. The ALS (settings 900 and 1000 ms) reached the correlations of 0.81. The ESF has the worst performance of all measures, the highest coefficient was 0.6, achieved by several settings of the algorithm. The moderate correlations are recognized with the SCSI. Many settings have a correlation coefficient of about 0.65, with the best performance at 0.72. The combination of spectral changes and voice activity detection brings the best results. The measure NSI exceeded with several settings a correlation of 0.8, with the highest at 0.85. The standard measure RT reached correlation 0.89, the highest in this study.

The range of values (mean and standard deviation) of each algorithm is given in Table 6 for two groups (0 - without any sign of stuttering or fluent, 1 - subjects with stuttering). The typical values are followed by the results of the ANOVA analysis.

The ALS and comparative RT increase with the level of fluency disorder, on the other hand all the BACD-based measures decrease with the level of speech fluency disorder.

According to the best performance (highest correlation) the algorithm settings were selected to be presented in ANOVA analysis.

- ALS, time threshold for successive removal 900 ms;
- ESF, multiplication constant 0.1, $k = 2$;
- SCSI, multiplication constant 0.1, $k = 2$, window length 2 s;
- NSI, multiplication constant 0.1, $k = 2$, threshold for successive removal 900 ms.

Table 2: Pearson’s correlation for all settings of the ALS in comparison to the number of disfluencies (subjective evaluation), setting of the measure: the time threshold for successive removal.

setting [ms]	correlation
100	0.51
150	0.58
200	0.65
300	0.69
400	0.71
500	0.75
700	0.78
800	0.80
900	0.81
1000	0.81
1100	0.75
1200	0.74
1300	0.76
1400	0.77
1500	0.76

Table 3: Pearson’s correlation for all settings of the ESF in comparison to the number of disfluencies (subjective evaluation), setting of the measure: k -th highest maximum and multiplication constant.

multiplication constant, correlation					
k	0.10	0.15	0.20	0.25	0.30
1	-0.60	-0.59	-0.54	-0.50	-0.50
2	-0.65	-0.60	-0.58	-0.55	-0.53
3	-0.60	-0.54	-0.52	-0.50	-0.46
4	-0.60	-0.54	-0.53	-0.51	-0.47
5	-0.60	-0.54	-0.53	-0.52	-0.47
6	-0.59	-0.52	-0.52	-0.50	-0.47
7	-0.59	-0.54	-0.52	-0.51	-0.48
8	-0.60	-0.55	-0.54	-0.53	-0.49
9	-0.59	-0.54	-0.53	-0.51	-0.50

The measures ALS, ESF, and SCSI are able to find statistically significant difference between fluent and disfluent at the level 0.01, while the NSI at the level 0.001, and the comparative measure RT at the level 0.05.

Another view on results show Figure 1, where the NSI is compared to the subjective evaluation. The range and dependency of the measure on the level of the disorder can be seen.

The classification using linear discriminant analysis and leave-one-out cross-validation was done to validate the results of correlations and ANOVA analysis. The NSI algorithm classified 26 (77%) of 34 recordings correctly, incorrectly 8 (23%, 5 disfluent, 3 fluent). The ALS, ESF, and SCSI placed 24 (71%) recordings correctly, 10 (29%) incorrectly. The comparative measure assigned 24 (71%) speakers to the correct class.

5. Discussion

The paper describes experiment where four automatic and objective measures initially intended for Czech stuttering recordings are applied on German stuttering recordings of read speech. The symptoms of stuttering were considered in design of the measures. Due to the method which are used they do not take

Table 4: Pearson’s correlation of the SCSI in comparison to the number of disfluencies (subjective evaluation), setting of the measure: k -th highest maximum, multiplication constant, and window length 2 s.

multiplication constant, correlation					
k	0.10	0.15	0.20	0.25	0.30
1	-0.68	-0.66	-0.63	-0.59	-0.58
2	-0.72	-0.68	-0.57	-0.64	-0.61
3	-0.66	-0.61	-0.61	-0.58	-0.54
4	-0.67	-0.62	-0.61	-0.60	-0.55
5	-0.66	-0.62	-0.60	-0.60	-0.54
6	-0.65	-0.60	-0.60	-0.58	-0.55
7	-0.65	-0.59	-0.60	-0.59	-0.56
8	-0.67	-0.61	-0.62	-0.61	-0.58
9	-0.66	-0.61	-0.60	-0.59	-0.57

Table 5: Pearson’s correlation of the NSI in comparison to the number of disfluencies (subjective evaluation), setting of the measure: k -th highest maximum, multiplication constant, and time threshold 900 ms.

multiplication constant, correlation					
k	0.10	0.15	0.20	0.25	0.30
1	-0.84	-0.81	-0.77	-0.71	-0.68
2	-0.85	-0.81	-0.78	-0.74	-0.70
3	-0.83	-0.79	-0.76	-0.72	-0.67
4	-0.84	-0.80	-0.77	-0.73	-0.67
5	-0.84	-0.80	-0.78	-0.75	-0.69
6	-0.83	-0.80	-0.78	-0.74	-0.70
7	-0.83	-0.80	-0.78	-0.75	-0.71
8	-0.84	-0.81	-0.78	-0.76	-0.72
9	-0.84	-0.81	-0.78	-0.76	-0.72

specifics of any language into account. Therefore, the main goal is to find out their possible use for different languages, in other words their language-independence.

The experiment analyses read recordings of 16 stuttering and 18 fluent native German speakers. The number of disfluencies was counted in each recording and it constitutes the subjective evaluation. These recordings were used in the study where speech recognition technology was applied to look for stuttering events in recordings of read text [12].

The main finding of the experiment is that the measures are able to describe the level of the speech fluency disorder for German read speech. The finding is supported by the results where two of four measures exceeded correlation with the reference evaluation of 0.8, the highest is 0.85. The standard measure, the total reading time, achieved a correlation of 0.89. The ANOVA analysis confirms the results, three of the measures found statistically significant difference between fluent and stuttering recordings at the level 0.01, for the number of spectral changes in speech segments even 0.001, while for the standard measure, the total reading time, at the level 0.05.

To validate the results of correlation and ANOVA analysis the classification using linear discriminant analysis and leave-one-out cross-validation was carried out. The NSI measure classified 26 (77%) of 34 recordings correctly, incorrectly 8 (23%). The other measures as well as the total reading time classified 71% into correct class. The overlapping of the groups of normal fluent speakers and speaker at lower levels of disfluency

Table 6: The mean \bar{x} and standard deviation SD of fluency measures and statistical significance by means of the ANOVA analysis with comparison between classes by the post hoc Bonferroni adjustment.

	ALS		ESF		SCSI		NSI		RT	
	\bar{x}	SD	\bar{x}	SD	\bar{x}	SD	\bar{x}	SD	\bar{x}	SD
Fluent (0)	0.21	0.05	7.89	0.98	1.23	0.05	0.94	0.05	51.8	8.1
Stuttering (1)	0.36	0.17	6.32	1.45	1.13	0.1	0.78	0.14	74.1	34.1
Comparison between the classes										
ANOVA F(1, 33)	12.92*		11.29*		12.1*		17.41*		7.05*	
0 vs. 1	$p < 0.01$		$p < 0.01$		$p < 0.01$		$p < 0.001$		$p < 0.05$	
NS = not significant										
* $p < 0.001$										

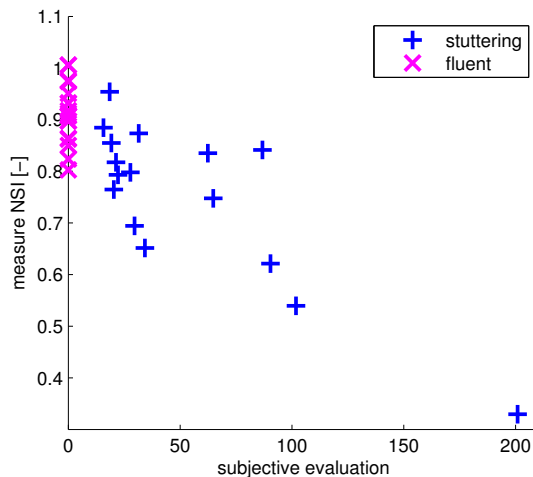


Figure 1: The comparison of the NSI with the number of all disfluencies (subjective evaluation).

(see Figure 1) should not be interpreted as a failing of the measures; it is a common thing that speech of normal fluent speaker contains disfluencies [23], also [8] found a high overlap in the speech rate of fluent speakers and participants who were classified as stutterers.

The standard measure, total reading time, was found distinctive for evaluation of pathology in read speech [21, 22]. Also experiments with Czech stuttering speakers acknowledged this statement [20]. The measure was added to the experiment to have a comparison with the introduced measures. The results of the total reading time turned out to be very good and in some cases better than those of the described methods (correlation of 0.89). However, there is an advantage of the automatic methods over the total reading time. The standard tasks in evaluation of stuttering (as clinical evaluation of monologue or picture description) are usually limited by time [8], therefore the total duration of recordings is not applicable but the automatic measures would be able to do the assessment.

We have found similarities in experiments on German recordings and previous experiments on Czech recordings [20]. Firstly, the range of all measures (mean and standard deviation) are very similar for both languages. Secondly, similar settings of the algorithms performed well on both databases. The measures ESF and SCSI have got little worse correlation for Ger-

man recordings (0.65, 0.72 vs. 0.76, 0.80 on Czech recordings). The ALS and NSI achieved higher correlations 0.81, 0.85 on German vs. 0.68, 0.82 on Czech recordings. We can also conclude that the measures based on Bayesian change point detection can be considered as robust because there are several settings which reached consistent results. The same was observed for Czech recordings.

The stuttering part of the database in this experiment was used in the paper [12]. One can see that the results of the current experiment are not as good as in that research. The highest correlation here is 0.85, on the contrary, [12] reported correlation of 0.99. The reasons for it could be, that introduced measures do not take specifics of German or Czech language into account, therefore they could be considered to be language-independent to some extent. Moreover, algorithms based on hidden Markov models were specially trained for German language, using the German-stuttering adapted grammar. The language-independence is an advantage of introduced measures although their performance on different languages could be either better or worse in comparison to the language-dependent. The language-independence of the measures could be limited by additional rules when segmenting to phonemes [24] or specific prosody features [25] of different languages, more specifically, for Czech and German [26].

A weak point of this experiment could be regarded the database which consists of only 34 speakers. In spite of the entire spectrum of speech fluency disorder is represented it would be preferable to have more subjects on different levels.

6. Conclusions

The experiment was carried out in order to find out possible use of automatic and objective measures of disfluency on different languages. The measures were originally examined on evaluation of Czech stuttering recordings and now applied on recordings of German speakers. The following conclusion can be drawn. The measures are able to describe the level of the speech fluency disorder in German and Czech language in read speech recordings and distinguish between fluent and stuttering speakers. In other words, it could be basically possible to do language-independent analysis of stuttering.

7. Acknowledgements

We would like to thank everyone who helped. This research was supported by the project GACR P102/12/2230 and by the Grant Agency of the Czech Technical University in Prague, grants SGS12/185/OHK4/3T/13 and SGS13/138/OHK3/2T/13.

8. References

- [1] O. Bloodstein and N. Bernstein Ratner, *A Handbook on Stuttering*, 6th ed. Clifton Park, NY: Delmar, Cengage Learning, 2008, 1–552.
- [2] J. Kalinowski, “Self-reported efficacy of an all in-the-ear-canal prosthetic device to inhibit stuttering during one hundred hours of university teaching: an autobiographical clinical commentary,” *Disability and Rehabilitation*, vol. 25(2), pp. 107–111, 2003.
- [3] K. Teesson, A. Packman, and M. Onslow, “The Lidcombe behavioral data language of stuttering,” *J. Speech Lang. Hear. Res.*, vol. 46, pp. 1009–1015, 2003.
- [4] G. D. Riley, “A stuttering severity instrument for children and adults,” *J. Speech Hear. Disord.*, vol. 37, pp. 314–322, 1972.
- [5] D. E. Metz, V. J. Samar, and P. R. Sacco, “Acoustic analysis of stutterers’ fluent speech before and after therapy,” *J. Speech Hear. Res.*, vol. 26, pp. 531–536, 1983.
- [6] J. Van Borsel, G. Reunes, and N. Van den Bergh, “Delayed auditory feedback in the treatment of stuttering: clients as consumers,” *Int. J. Lang. Commun. Disord.*, vol. 38(2), pp. 119–129, 2003.
- [7] F. G. Di Simony, “Some preliminary observations on temporal compensation in the speech of children,” *J. Acoust. Soc. Am.*, vol. 56, pp. 697–699, 1974.
- [8] W. Johnson, “Measurements of oral reading and speaking rate and disfluency of adult male and female stutterers and nonstutterers,” *J. Speech Hear. Disord.*, vol. 7, pp. 1–20, 1961.
- [9] C. R. F. de Andrade, L. M. Cervone, and F. C. Sassi, “Relationship between the stuttering severity index and speech rate,” *Sao Paulo Med. J.*, vol. 121 (2), pp. 81–84, 2003.
- [10] J. S. Yaruss and E. G. Conture, “F2 transitions during sound/syllable repetitions of children who stutter and predictions of stuttering chronicity,” *J. Speech Hear. Res.*, vol. 36, pp. 883–896, 1993.
- [11] E. C. Healey and B. Gutkin, “Analysis of stutterers’ voice onset times and fundamental frequency contours during fluency,” *J. Speech Hear. Res.*, vol. 27, pp. 219–225, 1984.
- [12] E. Nöth, H. Niemann, T. Haderlein, M. Decher, U. Eysholdt, F. Rosanowski, and T. Wittenberg, “Automatic stuttering recognition using hidden Markov models,” in *Sixth International Conference on Spoken Language Processing*, vol. 4, Beijing, China, 2000, pp. 65–68.
- [13] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, “Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson’s disease,” *J. Acoust. Soc. Am.*, vol. 129, pp. 350–367, 2011.
- [14] A. Maier, F. Hönig, T. Bocklet, E. Nöth, F. Stelzle, E. Nkenke, and M. Schuster, “Automatic detection of articulation disorders in children with cleft lip and palate,” *J. Acoust. Soc. Am.*, vol. 126(5), pp. 2589–2602, november 2009.
- [15] C. Cucchiari, H. Strik, and L. Boves, “Quantitative assessment of second language learners’ fluency by means of automatic speech recognition technology,” *J. Acoust. Soc. Am.*, vol. 107, pp. 989–999, 2000.
- [16] A. Maier, F. Hönig, V. Zeissler, A. Batliner, E. Körner, N. Yamanaka, P. Ackermann, D. Peter, and E. Nöth, “A language-independent feature set for the automatic evaluation of prosody,” in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*. Brighton, England: 10th Annual conference of the International Speech Communication Association (Interspeech 2009), 2009, pp. 600–603.
- [17] R. A. Prosek and C. M. Runyan, “Temporal characteristics related to the discrimination of stutterers’ and nonstutterers’ speech samples,” *J. Speech Lang. Hear. Res.*, vol. 25 (1), pp. 29–33, 1982.
- [18] J. Ruanaidh and W. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*. New York, NY: Springer-Verlag, 1996, chap. 5, pp. 96–101.
- [19] R. Cmejla, J. Rusz, P. Bergl, and J. Vokral, “Bayesian changepoint detection for the automatic assessment of fluency and articulatory disorders,” *Speech Commun.*, vol. 55, pp. 178–189, 2013.
- [20] T. Lustyk, P. Bergl, and R. Cmejla, “Evaluation of disfluent speech by means of automatic acoustic measurements,” *J. Acoust. Soc. Am.*, vol. 135(3), pp. 1457–1468, March 2014.
- [21] A. L. Foundas, A. M. Bollich, J. Feldman, D. M. Corey, M. Hurley, L. C. Lemen, and K. M. Heilman, “Aberrant auditory processing and atypical planum temporale in developmental stuttering,” *Neurology*, vol. 64 (9), pp. 1640–1646, 2004.
- [22] A. Maier, F. Hönig, S. Steidl, E. Nöth, S. Horndach, E. Saurerhöfer, O. Kratz, and G. Moll, “An automatic version of a reading disorder test,” *ACM Trans. Speech Lang. Process. (TSLP)*, vol. 7(4), pp. 17:1–17:15, 2011.
- [23] P. M. Roberts, A. Meltzer, and J. Wilding, “Disfluencies in non-stuttering adults across sample lengths and topics,” *J. Commun. Dis.*, vol. 42 (6), pp. 414–427, 2009.
- [24] O. Scharenborg, V. Wan, and M. Ernestus, “Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries,” *J. Acoust. Soc. Am.*, vol. 127 (2), pp. 1084–1095, 2010.
- [25] J. Vaissiere, “Language-independent prosodic features,” in *Prosody: Models and Measurements*, A. Cutler and R. Ladd, Eds. Springer Berlin Heidelberg, 1983.
- [26] J. Matousek, D. Tihelka, J. Psutka, and J. Hesova, “German and Czech speech synthesis using HMM-based speech segment database,” in *Text, Speech and Dialogue*. Springer Berlin Heidelberg, 2002, pp. 173–180.