FILLED PAUSE DETECTION BY PROSODIC DISCONTINUITY FEATURES

Uwe D. Reichel[1]*, Benjamin Weiss*[1]*, Thilo Michael*[2]
[1]*audEERING GmbH,* [2]*Technical University Berlin*
ureichel@audeering.com

Abstract: In this study we aim to predict filled pauses (FP) by analyses of energy, f0, and perceived local speaking rate contours. Prosodic feature profiles are introduced for FP and non-FP segments as well as for discontinuities at their transitions. Interpretations of those profiles and their discriminatory power will be given. Based on the extracted prosodic features we trained Random Forest classifiers for FP detection on three different units of classification: manually segmented syllables, automatically detected syllables, and equally spaced time stamps. The advantages and shortcomings of these units are discussed. Based on prosodic features only, well balanced FP recall and precision values between .82 and .86 were achieved.

1 Introduction

Filled pauses (FP) are among the most common types of disfluencies in spontaneous speech. They are realized for example as interjections like *uh* and *uhm* and frequently occur with increased cognitive workload [1, 2], as discourse structure markers [3], and to regulate turn taking [4, 5]. It has been shown e.g. by [6], that FP detection can increase speech recognition performance. Furthermore, it can be utilized amongst others for the automatic assessment of cognitive workload, of speaker states, and of L2 fluency.

So far automatic FP detection has been based on spectral features including spectral stability measures that capture lengthened phone segments and steady central vowels in FPs [7, 6]. Furthermore, adjacent silent intervals [7, 6] and standard pitch and energy features like mean and maximum [8] are employed. [8, 9] predict FPs based on a preceding speech recognition (ASR) step that allows for adding phone and word duration as well as ASR confidence scores [8] to the feature pool. Furthermore, ASR allows for restricting the locations of potential FP on- and offsets [9] and to additionally make use of a language model [9]. [9] account for discontinuities at FP boundaries in terms of F0 differences across these boundaries. As to automatic classification by various machine learning approaches these studies report FP recall values between .36 and .94 – high recall often at the cost of low precision and vice versa [6].

In the current study we focus on prosodic f0, energy, and speaking rate features. We pursue the notion of [9] that FP boundaries are marked by prosodic discontinuities by introducing an extensive feature set capturing such discontinuities in the contours of the three prosodic domains.

2 Data

This study was conducted on the Kiel Corpus of Spontaneous speech [10] (old version), which comprises approximately 3.8 hours of single-turn recordings of German semi-spontaneous speech with hand-segmentations and annotations on the phone, syllable, and word level.

FP segments were identified from the word level annotations according to the corpus documentation. That is, all word tokens labeled as *aeh, aehm, hm*, or *haes* were considered to be FPs. Word tokens labeled by an asterisk were skipped from further analyses.

3 Contour parameterization

3.1 Domains of analysis

For each signal file three contours all with a sample rate of 100 Hz were extracted: fundamental frequency, energy, and perceived local speaking rate [11, 12]. F0 was extracted by autocorrelation (Praat 6.0 [13]; allowed f0 range from 50 to 400 Hz; default settings). Voiceless utterance parts and f0 outliers were bridged by linear interpolation. The contour was then smoothed by Savitzky Golay filtering using third order polynomials in 5 sample windows and transformed to semitones relative to a speaker-dependent base value, that was set to the f0 median below the 5th percentile of the speaker's utterances. Energy in terms of root mean squared deviation was calculated in Hamming windows of 50ms length.

Perceived local speaking rate (PLSR) was introduced by [11, 12]. Based on perceptual speaking rate judgments the PLSR was fitted for German by a linear combination of syllable and phone rate. [14, 15] trained deep neural networks – LSTMs and CNNs – for PLSR prediction. These nets do not require any segmentation into phones and syllables, but are trained on mel frequency cepstral coefficients. For this study PLSR was predicted by the best performing net of [15]. Please see [15] for more details and for a comparative evaluation.

3.2 Units of analysis

We defined three types of FP onset candidates: manually annotated syllable boundaries, automatically extracted syllable boundaries, and equally spaced time stamps. Separately for each of these units we carried out contour stylization and FP detection, the latter comparatively evaluated in section 5.3.

Manually segmented syllables are provided by the corpus annotations, and automatically extracted boundaries were derived as introduced in [16]. In short, first, interpausal units (IPUs) were extracted as the inverse result of an energy-based silent speech pause detector. The required minimum silent pause length was set to 0.15s. Syllable nuclei were then located within each IPU in the bandpass-filtered signal at local energy maxima that fulfill three constraints: to exceed a locally and a globally defined energy threshold, and to result in syllable durations above a temporal threshold. Syllable boundaries are then simply inserted at the energy minimum between two extracted syllable nuclei.

For the time stamp based approach, in each IPU we shifted an analysis window of 1s length by 10 ms steps to decide for each frame whether or not its center is the onset of an FP.

3.3 Discontinuity

For each of the domains: f0, energy, and PLSR we parameterized each boundary defined by the respective unit analogously to [17] the following way:

We centered an analysis window of 1s length on the syllable boundary or time stamp for which to decide whether or not it is the onset of an FP. At the onset of IPUs the initial half of the analysis window was shifted left in front of the silent pause. Since the data comprises single turn recordings only, these silent pauses don't separate turns but partly indicate hesitation. Within the analysis window centered on the potential FP onset boundary, we fitted three regression lines on the respective domain's contour as shown in Figure 1: one line to the pre-boundary segment (the first half of the analysis window), one line to the post-boundary segment (the second half of the window), and one line to the entire window spanning both segments. Discontinuity is represented in the degree of deviation of the pre- and post-boundary lines as well as in the their deviation from a common fate represented by the joint segment's line. [17] have found

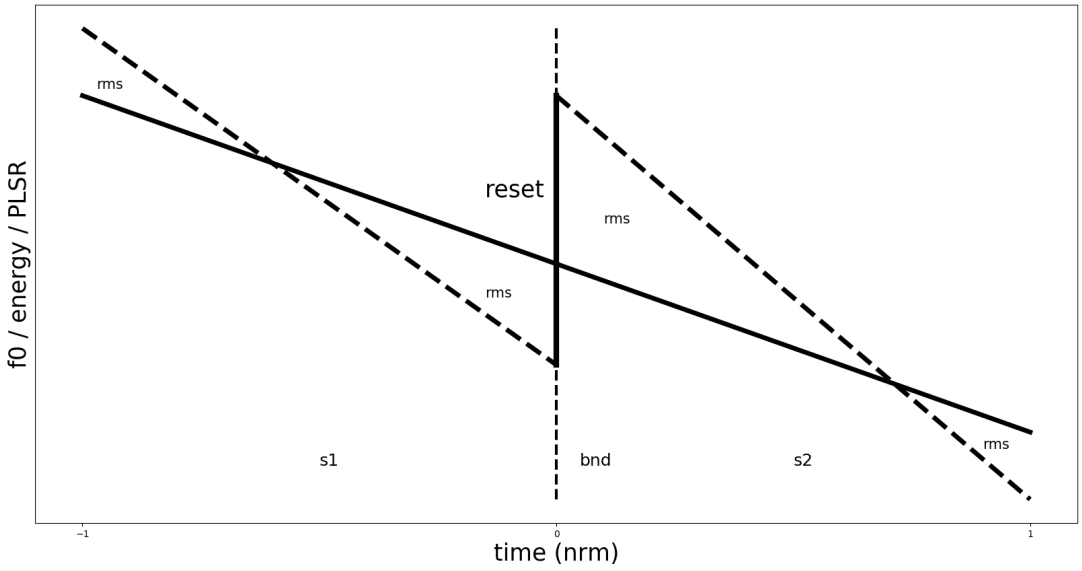

Figure 1 – Discontinuity stylization at a given boundary *bnd* in terms of 3 linear fits through the pre-boundary (*s1*), post-boundary (*s2*), and the joint segment. Two discontinuity features, the line reset and the RMSD between the fitted lines are shown. See Table 1 for a description of all features.

a positive correlation for f0 between the proposed discontinuity representation and perceived word boundary strength. From this representation we calculated a set of features which is summarized in Table 1. Most features were extracted by the CoPaSul toolkit [18] written in Python3, which is available on GitHub.[1]

4 Discontinuity profiles

For the time stamps derived from the manually segmented syllables we obtained profiles for transitions into filled pauses (*FP*) and transitions into words (*–FP*) by plotting the features' medians for both conditions. The features are given on the y-axis, their centered and scaled values on the x-axis. The solid line is derived from all syllable boundaries, for which the post-boundary segment is an FP. The dashed line is derived from all boundaries, for which the post-boundary segment is part of a non-FP word. To ease readability the features are sorted in order to minimize the number of profile crossings. The domains f0, energy, and PLSR are plotted in Figures 2, 3, and 4, respectively. The left plots contain the features obtained separately for the pre- and post-boundary segments, and the right plots contain the discontinuity features between these two segments. These profiles serve to visualize feature value tendencies at the onsets of FPs as opposed to fluent speech.

5 Filled pause detection

5.1 Units

For each of the chosen units of analysis, manually segmented syllables (*ManSyl*), automatically detected syllables (*AutSyl*) and equally spaced time stamps (*TS*), we trained classifiers to predict for each analysis window, whether or not the right window half contains an FP. *AutSyl* and *TS* can be obtained directly from the signal without any manual preprocessing.

[1] https://github.com/reichelu/copasul/

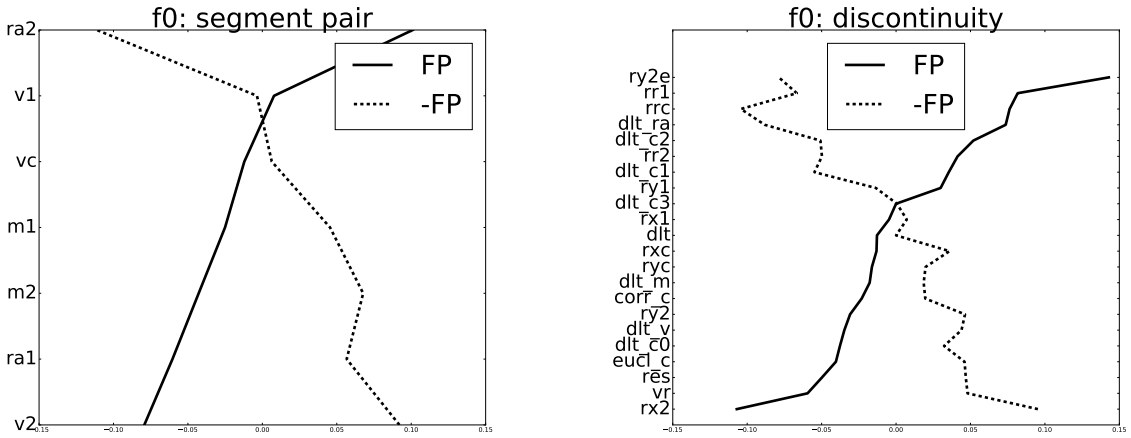

Figure 2 – Feature profiles for f0 for the filled pause FP and fluent speech –FP condition. **Left:** calculated separately for the pre- and post-boundary segment. **Right:** Discontinuity features.

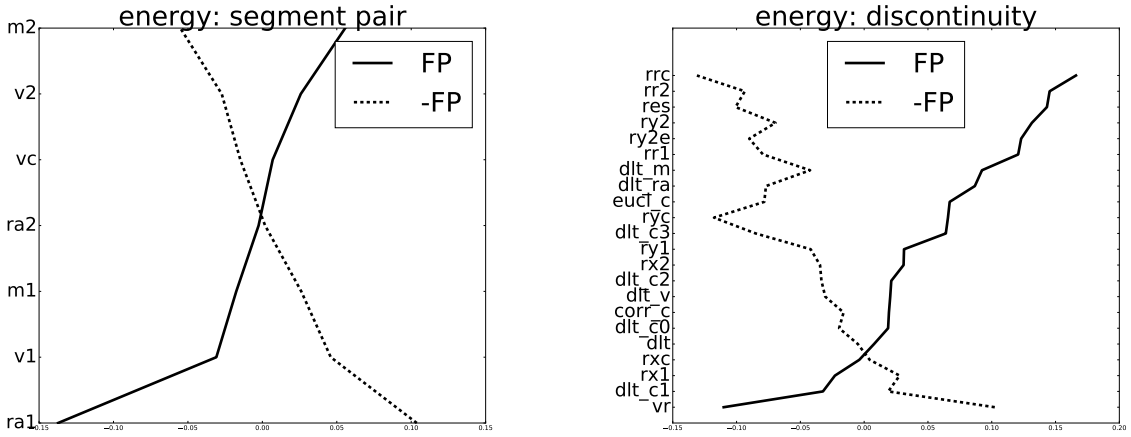

Figure 3 – Feature profiles for energy for the filled pause FP and fluent speech –FP condition. **Left:** calculated separately for the pre- and post-boundary segment. **Right:** Discontinuity features.

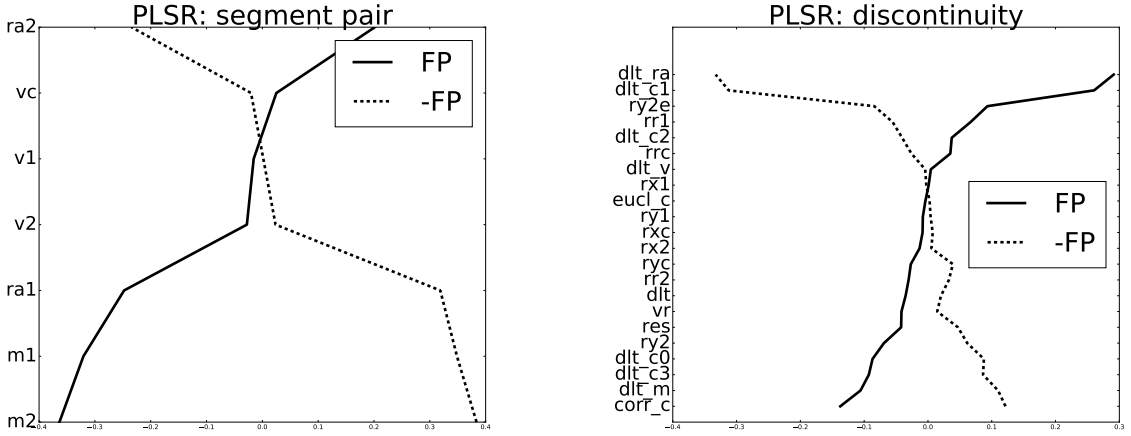

Figure 4 – Feature profiles for PLSR for the filled pause FP and fluent speech –FP condition. **Left:** calculated separately for the pre- and post-boundary segment. **Right:** Discontinuity features.

Table 1 – Feature description. All features are extracted for each f0, energy, and PLSR contour at the boundaries given by syllable segmentation or equally spaced time stamps. Three different analysis windows are used: the pre-boundary segment (*s1*), the post-boundary segment (including the boundary; *s2*), and the joint segment (*sc*). The respective segment within which the feature is extracted, is given in the *Segment* column. In this column *bnd* indicates, that the feature is calculated locally at the boundary time point relative to the left-adjacent time point (delta values).

Feature	Description	Segment
dlt	delta value at boundary	bnd
res	reset (delta of onset – offset of the fitted lines in s2 and s1)	bnd
ry1	RMSD between joint vs pre-bnd linear fit	s1
ry2	RMSD between joint vs post-bnd linear fit	s2
ryc	RMSD between joint vs pre+post-bnd linear fit	sc
ry2e	RMSD between extrapolated pre-bnd fit and post-bnd fit	s2
rx1	RMSD between joint fit and pre-bnd contour	s1
rx2	RMSD between joint fit and post-bnd contour	s2
rxc	RMSD between joint fit and pre+post-bnd contour	sc2
rr1	ratio of RMSD(joint fit) and RMSD(pre-bnd fit)	s1
rr2	ratio of RMSD(joint fit) and RMSD(post-bnd fit)	s2
rrc	ratio of RMSD(joint fit) and RMSD(pre+post-bnd fit)	sc
ra1	rate of fitted line	s1
ra2	rate of fitted line	s2
dlt_ra	ra2-ra1	s1, s2
dlt_c3	cubic coef delta of cubic contour fits in s1 and s2	s1, s2
dlt_c2	quadratic coef delta	s1, s2
dlt_c1	linear coef delta	s1, s2
dlt_c0	intercept coef delta	s1, s2
eucl_c	euclidean distance of cubic polynomial coef vectors	s1, s2
corr_c	correlation of cubic polynomial coef vectors	s1, s2
v1	variance	s1
v2	variance	s2
vc	variance	sc
vr	variance ratio: mean(v1,v2)/vc	s1, s2, sc
dlt_v	v2-v1	s1, s2
m1	mean	s1
m2	mean	s2
dlt_m	m2-m1	s1, s2
p	pause length	bnd

5.2 Classification

To balance classes we undersampled the majority class (*–FP*) to the size of class *FP*. We robustly centered and scaled the features on their medians and interquartile ranges, on which we then trained and evaluated Random Forest classifiers (500 trees; split quality criterion: Gini impurity; min. leaf samples: 2; min. node splitting samples: 6). The entire feature processing and classification pipeline was carried out with the Python scikit-learn package version 0.20.0.

5.3 Results

Since variant *AutSyl* depends on the quality of the automatic syllable detection we also assessed the FP recall of the syllable extractor, that is the proportion of FP intervals within which one or more syllable nuclei were detected: recall amounts .92, that is 8% of the FPs are ignored by this

method.

Table 2 shows the mean classification performance after 10-fold cross-validation. Due to the class balancing and the nicely balanced precision and recall values, accuracy and F1 scores are by and large the same so that we just report the latter. F1 values from .83 to .84 were obtained with equally high precision and recall values between .82 and .86 for both classes. The automated variants *AutSyl* and *TS* yield as high results as the variant *ManSyl* requiring manual annotations.

Table 2 – Mean performance of the filled pause (*FP*) vs fluent speech (*–FP*) detection assessed by 10-fold cross validation. Units *ManSyl* – manually annotated syllable boundaries, *AutSyl* – automatically detected syllable boundaries, *TS* – equally spaced time stamps. *N*: number of items. $F1$: F1 macro mean, which due to class balancing and close precision and recall values amounts to the same values as $F1$ micro mean and accuracy.

Unit	**N**	**precision**		**recall**		**F1**
		FP	*–FP*	*FP*	*–FP*	
ManSyl	1316	.82	.84	.84	.82	.83
AutSyl	1580	.83	.85	.84	.83	.83
TS	4110	.83	.85	.86	.82	.84

6 Discussion

6.1 Profiles

In section 4 we plotted feature profiles extracted from the manually annotated syllable condition for the pre- and post-boundary segments and for the discontinuity at their transition, for which we will provide interpretations in the following. Due to the large amount of features making it difficult to correct for type 1 errors we do not back up these profiles by statistical tests. However, we still consider them to be an appropriate instrument to visualize tendencies.

6.1.1 Pre- and post-boundary segments

The left plots in Figures 2, 3, and 4 show the medians of those features extracted for the segments preceding and following the syllable boundary, marked by the indices 1 and 2, respectively. As an example, *v1* denotes the variance of the contour values in the pre-boundary segment, and *v2* the variance in the post-boundary segment, which either is a filled pause *FP* or fluent speech *–FP*.

PLSR Lower mean PLSR values *m1* and *m2* can be observed both in the FP as well as in the preceding segment. Thus, not surprisingly, FPs go along with a reduced speaking rate both reflecting a higher amount of cognitive workload [2]. Due to the size of the PLSR smoothing window of 625 ms [15], it cannot be decided at the current stage whether a low $m1$ is only a smoothing artifact or also indicates that FPs can to some extent be anticipated from a left context deceleration. Regardless of its explanation the large *m1* difference suggests this features to be valuable for FP detection. The potential deceleration as well as the smoothing artifact is further reflected in lower *ra1* values, which stands for falling PLSR contours in front of FPs.

F0 *m2* and *v2* show lower f0 means and variance in FPs. This indicates that FPs are not likely to bear pitch accents to encode information status, and in the examined single-turn data are neither

used as a high pitch turn holding signal.

Energy Opposite to F0, FPs show higher energy means and variances that are a challenge for automatic syllable detection which will be addressed in section 6.2.

6.1.2 Discontinuities

The right plots in Figures 2, 3, and 4 show delta, correlation, and RMSD features measuring in how far the pre- and post-boundary segment differ in the respective domain.

Among the most stable discontinuity indicators is *ry2e* measuring the fitting error when applying the pre-boundary linear trend to the post-boundary segment. *ry2e* is higher in the FP condition for all, the f0, energy, and PLSR domain. *corr_c* is the correlation of coefficients between two 3rd-order polynomial contour stylizations for the pre- and post-boundary segment, respectively. For f0 and PLSR it shows lower values in the FP condition indicating that in these conditions the contours differ more strongly, which again is an aspect of discontinuity.

6.2 Detection

The syllable qualifies as a unit for FP prediction, since FPs generally correspond to one-syllable segments. However, the drawback is, that errors in automatic syllable detection add on to the FP detection errors. As reported in section 5.3, 8% of the FPs are already missed due to not detected syllables. Unfortunately, especially for FPs, syllable detection is challenging, since some FPs as *hm*, glottalizations or fricative noises, all found in our data, lack a high energy vocalic part. Less crucial for the FP prediction, but also challenging are the partly very long durations of FPs, in our data up to 1.57s. In these long segments more than one local energy peak is likely which leads to the erroneous assignment of several syllable nuclei. This multiple assignment occurred 161 times in our data and is responsible for the diverging *N* values of the conditions *ManSyl* and *AutSyl* in Table 2.

Thus we consider the agnostic condition *TS*, i.e. the FP prediction for equally spaced time stamps to be a more robust approach. Table 2 shows comparably high performances for all three conditions, which again qualifies equally spaced time stamps to be appropriate analysis units for FP detection.

In contrast to previous studies balanced precision and recall values between .82 and .86 have been obtained based on prosodic features only. The next step is to merge analyses from this prosodic and the segmental level for the FP prediction task.

References

[1] REYNOLDS, A. and A. PAIVIO: *Cognitive and emotional determinants of speech. Canadian Journal of Psychology*, 22, pp. 164–175, 1968.

[2] ROCHESTER, S.: *The significance of pauses in spontaneous speech. Psycholinguistic Research*, 2, pp. 51–81, 1973.

[3] SWERTS, M.: *Filled pauses as markers of discourse structure. J. Pragmatics*, 30, pp. 485–496, 1998.

[4] CLARK, H. and J. TREE: *Using uh and um in spontaneous speaking. Cognition*, 84, pp. 73–111, 2002.

[5] NAVARRETTA, C.: *The functions of fillers, filled pauses and co-occurring gestures in Danish dyadic conversations*. In *Proc. 3rd European Symposium on Multimodal Communication*, pp. 55–61. Dublin, Ireland, 2015.

[6] STOUTEN, F. and J.-P. MARTENS: *A feature-based filled pause detection system for Dutch*. In *Proc. ASRU*, pp. 309–314. St. Thomas, U.S., 2003.

[7] GABREA, M. and D. O'SHAUGHNESSY: *Detection of filled pauses in spontaneous conversational speech*. In *Proc ICSLP*. Beijing, China, 2000.

[8] MEDEIROS, H., H. MONIZ, F. BATISTA, I. TRANCOSO, and H. MEINEDO: *Experiments on automatic detection of filled pauses using prosodic features*. In *Actas de Inforum*, pp. 335–345. 2013.

[9] E., S., R. BATES, and A. STOLCKE: *A prosody-only decision-tree model for disfluency detection*. In *Proc. Eurospeech*, pp. 2383–2386. Rhodes, Greece, 1997.

[10] KOHLER, K.: *Phonetisch-akustische Datenbasis des Hochdeutschen. Kieler Arbeiten zu den PHONDAT-Projekten 1989 – 1992 (aipuk)*. Tech. Rep. 26, Institut für Phonetik und digitale Sprachverarbeitung der Christian-Albrechts-Universität Kiel, 1992.

[11] PFITZINGER, H.: *Local speech rate as a combination of syllable and phone rate*. In *Proc. ICSLP*, vol. 3, pp. 1087–1090. Sydney, Australia, 1998.

[12] PFITZINGER, H.: *Local speech rate perception in German speech*. In *Proc. 14th ICPhS*, vol. 2, pp. 893–896. San Francisco, USA, 1999.

[13] BOERSMA, P. and D. WEENINK: *PRAAT, a system for doing phonetics by computer*. Tech. Rep., Institute of Phonetic Sciences of the University of Amsterdam, 1999. 132–182.

[14] WEISS, B., T. MICHAEL, and S. HILLMANN: *Kontinuierliche schätzung von sprechgeschwindigkeit mit einem rekurrenten neuronalen netzwerk*. In A. BERTON, U. HAIBER, and W. MINKER (eds.), *Elektronische Sprachverarbeitung 2018*, vol. 90 of *Studientexte zur Sprachkommunikation*, pp. 186–191. TUDpress, Ulm, Germany, 2018.

[15] WEISS, B., T. MICHAEL, U. REICHEL, and O. PAULY: *Vergleich verschiedener Machine-Learning Ansätze zur kontinuierlichen Schätzung von perzeptivem Sprechtempo*. In P. BIRKHOLZ and S. STONE (eds.), *Elektronische Sprachverarbeitung 2019*, vol. this volume. TUDpress, Dresden, Germany, 2019.

[16] REICHEL, U.: *Unsupervised extraction of prosodic structure*. In J. TROUVAIN, I. STEINER, and B. MÖBIUS (eds.), *Elektronische Sprachverarbeitung 2017*, vol. 86 of *Studientexte zur Sprachkommunikation*, pp. 262–269. TUDpress, Dresden, Germany, 2017.

[17] REICHEL, U. and K. MÁDY: *Comparing parameterizations of pitch register and its discontinuities at prosodic boundaries for Hungarian*. In *Proc. Interspeech 2014*, pp. 111–115. Singapore, 2014.

[18] REICHEL, U.: *CoPaSul Manual – Contour-based parametric and superpositional intonation stylization*. RIL, MTA, Budapest, Hungary, 2016.