Björn W. Schuller

The Computational Paralinguistics Challenge

As Friedrich Nietzsche stated more than a century ago, "The most intelligible factor in language is not the word itself, but the tone, strength, modulation, tempo with which a sequence of words is spoken—in brief, the music behind the words, the passions behind the music, the person behind these passions: everything, in other words, that cannot be written" [1]. However, nonlinguistic aspects were broadly conceived as fringe phenomena until the attitude changed at slow pace half a century ago [2]. Paralanguage—literally "alongside" language—is researched more widely only since the term was arguably first mentioned by the linguist Archibald Hill in 1958. Paralinguistics, first named roughly at the same time by George Leonard Trager [3]—can be limited to "vocal factors" according to David Abercrombie and David Crystal roughly a decade later—also all linguists. With the advent of modern computing devices, a new branch of paralinguistics allowed for their automatic processing.

COMPUTATIONAL PARALINGUISTICS

Today's computational paralinguistics, in particular, deals with the computer-based analysis and synthesis of paralinguistic phenomena. In this article, however, we will mostly touch the analysis side. As opposed to many related phenomena, usually states and traits of speakers, the term as such is hardly coined yet, and aims to unite the manifold and diverse corresponding activity in this and related fields such as social signal processing [4] and affective computing [5]. A definition is probably easiest given ex-negativo: It comprises everything that is not dealt with in phonetics or linguistics [2]. Here, we follow the limitation to events primarily encoded in the voice and secondarily in the phonetic content. As for the phenomena, a number of taxonomies can be found to group these, such as measured versus perceived, acted versus spontaneous, felt versus perceived, intentional versus instinctual, consistent versus discrepant, private versus social, universal versus culture-specific, and unimodal versus multimodal [2]. However, the most intuitive grouping may be short term versus long term, and by this, let us exemplify tasks that have been touched in the literature for automatic analysis so far [6]. On the short-term end, one finds the mode such as speaking style and voice quality, emotions including full-blown and prototypical, emotion-related states or affects (e.g., confidence, deception, frustration, interest, intimacy, pain, politeness, pride, sarcasm, shame, stress, and uncertainty). In between short-term states and more permanent traits, there are more or less temporary medium-term states such as friendship and identity, health state, intoxication, mood (e.g., depression), positive/negative attitude, sleepiness, and structural—i.e., behavioral, interactional, and social signals such as entrainment or the role in dyads or groups. Finally, the long-term end contains biological and physical trait primitives such as age, gender, height, weight, group, and ethnicity membership such as race, culture, social class with a weak borderline towards other linguistic concepts (i.e., speech registers such as dialect or nativeness), personality traits such as likability and personality in general and traits that constitute speaker idiosyncrasies, i.e., the identity of a speaker.

TODAY'S COMPUTATIONAL PARALINGUISTICS, IN PARTICULAR, DEALS WITH THE COMPUTER-BASED ANALYSIS AND SYNTHESIS OF PARALINGUISTIC PHENOMENA.

IN DAILY LIFE

Looking at the application of such speaker state and trait information, the following are found among the most promising [4]–[7]. First, it seems obvious that speech recognition and interpretation of speakers' intentions can benefit from paralinguistic information, e.g., when trying to recognize equivocation, irony, or sarcasm. The information can also be exploited to improve recognition of *what* has been said, e.g., by model adaptation. Next, conversation analysis, mediation, and transmission can benefit from paralinguistics, such as in computer-aided analysis of human-human conversations including the investigation of synchrony in the prosody of married couples, specific types of discourse in psychology, or the summarization of meetings. Also, hearing-impaired persons can profit, as cochlear implant processors typically alter the spectral cues which are crucial for the perception of paralinguistic information. Individuals on the autism spectrum may profit from the analysis of

Digital Object Identifier 10.1109/MSP.2012.2192211
Date of publication: 12 June 2012

socioemotional cues as they may have difficulties understanding them. Also, transmitting paralinguistic information along with other message elements can be used to animate avatars, to enrich dictated text messages, or to label calls in voice mailboxes by symbols such as emoticons. In call centers, quality management by monitoring agents and adapting to callers is of commercial interest, including target-group specific advertising. Next, communicative virtual agents and robots are enriched by social competence. Further, there are health- and speech disorder-related applications such as for Parkinson's disease, cancer, cleft lip, dysphonia, and palate. Tutoring and serious gaming systems are another typical field of application where, for example, information on user states such as cognitive load, interest, stress, and uncertainty allows adapting the teaching pace. Similarly, coaching systems for foreign language acquisition and public speeches can advise the speaker based on paralinguistic analysis. In security related situations such as crisis management, piloting, and surgical operations, monitoring of intoxication, sleepiness, and stress level may play a vital role. In addition, counter terrorism or vandalism surveillance may be aided by analyzing aggressiveness of potential aggressors, or fear of potential victims.

In the entertainment sector, less serious applications, such as the Love Detector or Handy Truster (Nemesysco Ltd.) or a console game around deception recognition (*Truth or Lies—Someone Will Get Caught,* THQ Entertainment) already appeared on the market. This holds also for the WhyCry—a device that aims to indicate a newborn's annoyance, boredom, hunger, sleepiness, and stress to less experienced parents. Finally, paralinguistic information allows for various types of multimedia retrieval. However, many ethical issues still need to be elaborated on for most—if not all—of these applications, once we approach the "transparent speaker," concerning how all this information is stored and made use of by technical systems.

CHALLENGES

For a good benchmark overview on current systems' performance, let us next have a look on baselines as were provided in a series of international research challenges and on according best results. These were initiated and coorganized annually by the author at INTERSPEECH conferences since 2009 [2,] [6], [8], [9] and in two workshops titled the "Audio/Visual Emotion Challenge (AVEC)" since 2011 [10] owing to the lack of a well-defined test bed for computational paralinguistics as is given for many related speech processing tasks. Overall, these touched perceived short-term states as were featured in the "INTERSPEECH 2009 Emotion Challenge (IS09EC)" [8] by emotion in two (negative versus idle) and five (anger, emphatic, neutral, positive, and rest) classes, the "INTERSPEECH 2010 Paralinguistic Challenge (IS10PC)" [6] by level of interest on a continuum ($[-1, +1]$), and in AVEC for four-dimensional emotion [activity, expectation, power, and valence—either represented binary above/below mean (2011) [10] or fully continuous in $[-1, +1]$ (2012)] of a speaker. Moving from such perceived short-term states diagonally to measured long-term traits, these were featured in the IS10PC for automatic determination of speakers' age in four groups (child, young, adult, and senior) and gender in three groups (child, female, male). Then, the "INTERSPEECH 2011 Speaker State Challenge (IS11SSC)" [2] provided examples both of measured and perceived medium-term speaker states: intoxication (above or below .5 per ml blood alcohol concentration as measured) and speaker sleepiness (above or below 7.5 on the Karolinska sleepiness scale, i.e., the subject has difficulty in staying awake). Finally, in the "INTERSPEECH 2012 Speaker Trait Challenge (IS12STC)" [9], perceived long-term traits are considered for the first time by personality (openness, conscientiousness, extraversion, agreeableness, and neuroticism), likability, and intelligibility of pathological speakers—all tasks are binarized to above or below average.

PARTITIONING IS STRICTLY SUBJECT INDEPENDENT TO FOSTER REALISM, AS THIS IS A PRECONDITION IN MANY OF THE ABOVE DESCRIBED APPLICATION USE CASES.

CORPORA

High realism was emphasized throughout any of these challenges. This is also well reflected in the choice of all challenge data: The speaker states and traits were realistic throughout, i.e., for example, genuine intoxication and sleep deprivation were given and speech is spontaneous when it comes to perceived states. Partitioning is strictly subject independent to foster realism, as this is a precondition in many of the above-described application use cases. Partly, the data also stems from broadcast or telephone transmission with according low sample rate of 8 kHz. Only the first challenge did not feature a development partition. In the subsequent challenges, the partitioning roughly followed a 40:30:30 partitioning as for training, development, and test sets, whereby training and development were united for baselines. Test data was available to the participants, but of course without the target labels—only five trials of result submissions were usually allowed per competing site, except for the first challenge with 25, and the second with only two. Table 1 summarizes the corpora by giving insight into the total speech time (TT), number of instances (INST), subjects (SUB), and labelers (LAB), the type (TY) of speech: spontaneous (S) or prompted (P) if at least part of the corpus contained prompted speech, the spoken language (LAN) by country code, and the audio quality by broadcast speech (FM), lab recording (LAB), or telephone transmission (TEL), and sample rate in kHz. Each challenge further featured multiple subchallenges, usually by target task. Only the IS09EC and AVEC feature

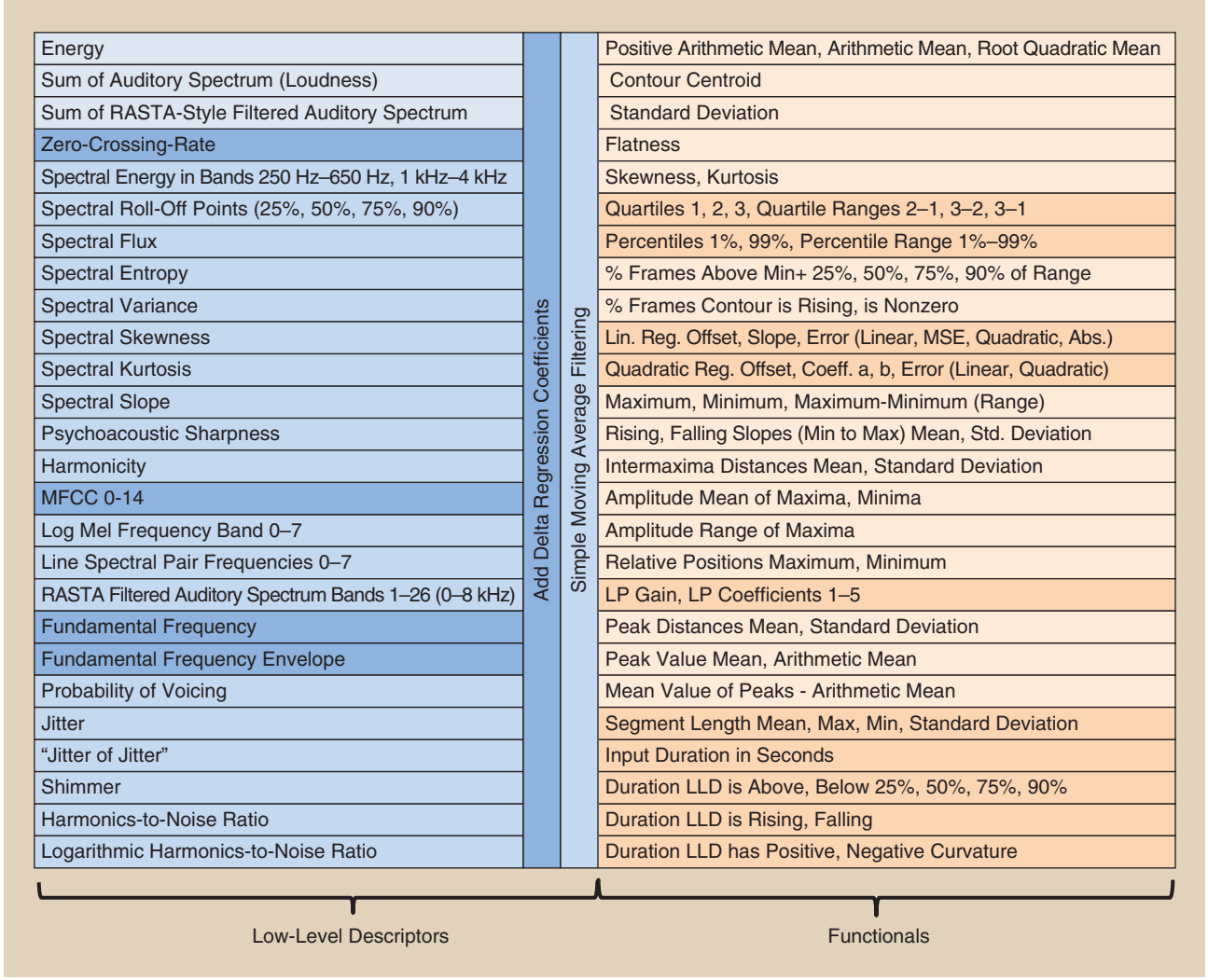
subchallenges on the same task in principle—emotion—but focusing on other aspects such as features, classifiers, or modality. As can be seen, the unit of analysis shows quite some deviation in terms of length per instance—in between single words, linguistically and syntactically motivated "chunks," and complete turns. AVEC 2012 is the first of the challenges that features frame-level audio analysis. Languages cover mostly the Germanic family with French as one exception.

[TABLE 1] CHALLENGE DATA INTERSPEECH 2009–2012 AND AVEC 2011–2012.

CORPUS	SUBCHALLENGE	TT[H]	INST #	SUB #	LAB #	TY	LAN	AUDIO KHZ
SPC	PERSONALITY	1.7	640	322	11	S	FR	FM 8
SLD	LIKABILITY	0.7	800	800	32	P	DE	TEL 8
NCSC	PATHOLOGY	2.0	2386	55	13	P	NL	LAB 16
ALC	INTOXICATION	43.8	12360	162	–	P	DE	LAB 16
SLC	SLEEPINESS	21.3	9089	99	4	P	DE	LAB 16
AGENDER	AGE, GENDER	50.6	65364	945	–	P	DE	TEL 8
TUM AVIC	AFFECT	2.3	3880	21	4	S	UK	LAB 44
FAU AIBO	EMOTION	8.9	18216	51	5	S	DE	LAB 16
AVEC	EMOTION	7.5	50350	24	≥ 2	S	UK	LAB 48

FEATURES

In each challenge, a baseline feature set was given for baseline calculation and participants' optional usage by Technische Universität München's (TUM's) open source openSMILE feature extractor [11]. Figure 1 and Table 2 give an overview on the challenges' feature sets and their size. As can be seen by multiplying the number of low-level descriptors (LLDs) times two for addition of delta regression coefficients with the number of functionals, a strict brute-forcing was followed only in 2009.



[FIG1] Acoustic features used in the INTERSPEECH 2009–2012 and AVEC 2011–2012 Challenges. From left to right, brute forcing takes place by adding delta LLDs, low-pass filtering, and subsequent functional application for projection onto a single value per LLD. Different shading borders groups.

[TABLE 2] FEATURES INTERSPEECH 2009–12 AND AVEC 2011–12.

# FEATURES	IS09EC	IS10PC	IS11SSC	IS12STC	AVEC11/12
LLDS	16	38	59	64	31
FUNCTIONALS	12	21	39	61	42
TOTAL	384	1582	4368	6125	1941/1841

[TABLE 3] RESULTS OF THE INTERSPEECH 2009–2012 CHALLENGES.

YEAR	UA [%]/*CC	CLASSES	BASE	BEST	VOTE
2012	OPENNESS	2	59.0	–	–
	CONSCIENTIOUSNESS	2	79.1	–	–
	EXTRAVERSION	2	75.3	–	–
	AGREEABLENESS	2	64.2	–	–
	NEUROTICISM	2	64.0	–	–
	LIKABILITY	2	59.0	–	–
	INTELLIGIBILITY	2	68.9	–	–
2011	INTOXICATION	2	65.9	70.5	72.2 (3)
	SLEEPINESS	2	70.3	71.7	72.5 (3)
2010	AGE	4	48.9	52.4	53.6 (4)
	GENDER	3	81.2	84.3	85.7 (5)
	INTEREST	[−1,1]	.421*	.428*	–
2009	EMOTION	5	38.2	41.7	44.0 (5)
	NEGATIVITY	2	67.7	70.3	71.2 (7)

Later, for example, in contrast to AVEC 2011, the AVEC 2012 feature set was reduced by 100 features, which were found to carry very little information, as they were (close to) zero most of the time. Also, not all functionals are applied to all LLDs.

RESULTS

Great effort was laid on reproducibility also for the learning algorithms—available standard toolkits were used throughout—namely HTK and WEKA 3 [12]. For the calculation of baselines, support vector machines were used in all of the challenges. In some, also random forests were used (IS10PC, IS11STC). Hidden Markov models (HMMs) were considered as an alternative in the first challenge operating on LLD level, but observed as less suited owing to the fact that they do not well model suprasegmental information. There is a typical high imbalance of instances across classes encountered in many computational paralinguistics tasks—for example, there is usually a high neutral speech presence in emotion analysis, but comparably limited negative speech. This is often dealt with in learning, where synthetic minority over-sampling (SMOTE) was repeatedly used for baseline calculation [12]. Table 3 depicts the challenge results by year and task including the number of classes or the value range in case of continuous task representation by baselines (BASE), if already available the results of the challenge winners (BEST), and by majority voting over the optimal number (shown in parentheses in the table) of best results (VOTE). The competition measure throughout is unweighted accuracy (UA) in the sense of unweighted average of the recalls per class. This is well suited for many computational paralinguistic tasks, as it takes the mentioned typical class-distribution imbalance well into account. As a benefit, it intuitively contains information on chance levels: For two, three, four, and five classes, it is 50%, 33%, 25%, and 20%, respectively. Thus, all results are significantly above chance level, and all differences between BASE, WIN, and VOTE are as well (highly) significant—in fact also the winners mostly were significantly better than the concurring results of other participants. In case of continuous level of interest determination, the correlation coefficient (CC) between the learner's prediction and the gold standard was used as measure, and no fusion of participants was carried out.

LOOKING AT METHODS EMPLOYED BY THE PARTICIPANTS, SUPRASEGMENTAL MODELING OF PARALINGUISTIC INFORMATION PREVAILS BY FAR.

LESSONS LEARNED

The number of participants shows the increasing popularity of these events and computational paralinguistics: for INTERSPEECH, the number of registered sites was 33, 32, 34, and 52 (2009–2012); for AVEC, it increased from 12 to 24 (2011–2012). A couple of reoccurring patterns can be observed from these challenges. Most of all, it was shown that fusion of best participant results exceeds individual winners throughout, and in fact, these "multiple-site" results are so far not reached by individual attempts. Further, looking at methods employed by the participants, suprasegmental modeling of paralinguistic information prevails by far. Only sparsely dynamic algorithms such as HMMs or dynamic Bayesian models were encountered. Another interesting trend is that stimulation across disciplines seems to be successful: approaches have been repeatedly shown as highly beneficial that were originally developed for similar related tasks, but have been applied for the first time to the challenge tasks at hand.

LITTLE WHITE LIES

Looking in more general at the field of computational paralinguistics, a number of considerations and promising avenues shall conclude this article. A key characteristic of the field is that there is often no solid ground truth—in

fact, only gender, age, and intoxication from the above exemplified challenge tasks are based on a more or less reliable one. To worsen things, task representation itself can be highly ambiguous and difficult, such as in the case of emotion or personality. Thus, one has to bear on mind that the recognition systems often aim to best mirror a few raters' opinion. Recently, the evaluator weighted estimator was applied increasingly to "denoise" human ratings by weighting raters. However, their number is typically very limited, and labeling is laborsome. Crowd sourcing, semisupervised, unsupervised, and active learning may help to ease this fact, and certainly interrater deviation can also be better exploited in the learning process. Further, whereas we claim data to be "realistic" in the challenges and for many of today's data, obviously conditions can often be more demanding in a real-life system: Crosstalk, nonstationary noises, reverberation, microphone type and position variation, and coding artifacts are just a few examples that are often ignored. In addition, variation of linguistic content, language, and cultural background of subjects will usually be considerably higher, and tasks will likely not vary in isolation.

BIG BLACK HOLES

Given these deficits, a couple of milestones need to be reached before we will likely see computational paralinguistics applied by and large in our everyday lives. Given the recent huge success in application of server-based automatic speech recognition as empowering factor for genuine broad user application, distributed processing of computational paralinguistics seems a promising step to be taken. By that, the ever-present and dominating bottleneck of data-sparseness may be partly overcome, though ethical considerations will need to be evaluated carefully. Given the shown state-of-the-art performances of current implementations, it becomes obvious that confidence measurements should be provided by recognition engines—this is at present a surprisingly untouched topic in this field's literature. On a higher representation level, new standards will then be needed—Emotion Markup Language is a good example, but for other tasks these are still missing. Further, closing the gap between analysis and synthesis seems promising. As an example, it was shown that synthesizing of learning material can be beneficial in the recognition of emotional speech [3]. On the other hand, synthesis of speech can be based on experience from the analysis side. In fact, however, today's methods and features differ considerably when, for example, analyzing or synthesizing emotional voices. Filling these gaps, and providing highly efficient separation of the speakers in real-life audio streams together with massive semi- or unsupervised multitask paralinguistic learning of feature spaces and models, we may soon reach machines' holistic computational analysis of paralanguage—imagine for an illustration a machine listening and stating: "I hear a mother—guess around mid-forties—talk to a young, yet taller boy in a friendly tone. Seems not be her child, though. He appears to be a rather open nature, yet maybe a bit tired."

WHILE MANY CHALLENGES REMAIN, THEY ARE BEST SOLVED WITH COMBINED COMMUNITY EFFORTS AND IN CLOSE CONNECTION TO RELATED FIELDS.

While many challenges remain, they are best solved with combined community efforts and in close connection to related fields.

AUTHOR

Björn W. Schuller (Schuller@tum.de) is a senior lecturer at TUM in Germany. He is a Member of the IEEE.

REFERENCES

[1] F. Nietzsche, S. Werke. *Kritische Studienausgabe in 15 Baenden*, v. Giorgio Colli and Mazzino Montinari (Eds.). 2nd edition, Berlin/München: DTV/de Gruyter, vol. 10, 1988, p. 89.

[2] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, to be published.

[3] G. L. Trager, "Paralanguage: A first approximation," *Stud. Linguist.*, vol. 13, pp. 1–12, 1958.

[4] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 69–87, 2012.

[5] R. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997.

[6] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language—State-of-the-art and the challenge," *Comput. Speech Lang.*, to be published. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2012.02.005.

[7] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human–computer interaction," *IEEE Signal Processing Mag.*, vol. 18, pp. 32–80, Jan. 2001.

[8] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun. (Special Issue on Sensing Emotion and Affect—Facing Realism in Speech Processing)*, vol. 53, pp. 1062–1087, Nov./Dec. 2011.

[9] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 speaker trait challenge," in *Proc. INTERSPEECH,* Portland, OR, ISCA, Sept. 2012.

[10] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011 – The first international audio/visual emotion challenge," in *Proc. 4th Int. Conf. Affective Computing and Intelligent Interaction, ACII,* Springer-Verlag, Oct. 2011, vol. 2, pp. 415–424.

[11] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE—The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia,* Florence, Italy, ACM, Oct. 2010, pp. 1459–1462.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explor.*, vol. 11, 2no. 1, 2009.

[SP]