draft, to appear in: The Oxford Handbook of Language Prosody. Ed. Carlos Gussenhoven and Aoju Chen

Prosody in Automatic Speech Processing

Anton Batliner
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing,
University of Augsburg, Germany
Pattern Recognition Lab, Friedrich-Alexander-Universität
Erlangen-Nürnberg, Germany
anton.batliner@gmail.com

Bernd Möbius
Language Science and Technology,
Saarland University, Saarbrücken, Germany
moebius@coli.uni-saarland.de

April 10, 2018

1 Introduction

We understand *Automatic Speech Processing* (ASP) as covering word recognition (Automatic Speech Recognition, ASR), the processing of higher linguistic components (syntax, semantics, and pragmatics), and the processing of Computational Paralinguistics (CP) which deals with speaker states and traits (Schuller & Batliner 2014). This chapter attempts to track the role of prosody in ASP from the word level up to CP, where the focus initially was on emotion recognition, with the view later broadening to the recognition of health conditions, social signals, and other speaker states (short-term) and traits (long-term).

Automatic processing of prosody means that at least part of the processing is done by the computer; by that, we can save time and money and are able to do things that would be too cumbersome if done fully manually. The automatic part can be small, e.g., pertaining only to pitch extraction, followed by manual correction of the F0 values with subsequent automatic computation of characteristic values such as mean, minimum, or maximum. This is typical for basic, possibly exploratory, research on prosody, or when we want to evaluate models and theories. A fully automatic processing of prosody in ASP, on the other hand, is necessary when we want to employ prosody in a larger context, such as developing a prosody module in a complete dialogue system, or to improve the speech of pathological speakers or foreign language learners by means of screening, monitoring, and feedback with respect to the learning progress, in a stand-alone tool. Nowadays, for many tasks, prosody is used in conjunction with other information, foremost spectral (cepstral) information.

Apart from the *phenomena* we want to investigate, such as prosodic features, emotions and affects, speaker states and traits, or social signals (for details see section 2.2), and from the *speech data* we have to record, the basic ingredients are: the *units of analysis*, suited for both the phenomenon and the type of features we employ; the *features* we have to extract; and *Machine Learning (ML) procedures* that tell us how good we are and, if we are interested, which features computed for which unit are most important.

The units of analysis in the processing of prosody can be given trivially (an entire speech file) or pre-defined (e.g., segments of five seconds or 1/10 of the entire speech file); they can be obtained by voice activity detection (silence as an indicator for major prosodic/syntactic boundaries); or they can be based on ASR that gives us word boundaries, maybe followed by syntactic parsing that gives us phrases and sentences; and we can combine all these strategies.

We will not say much about ML procedures. Many have been employed for prosody in ASP. Roughly speaking, traditional, well established procedures, such as linear classifiers or decision trees, may yield somewhat lower performance but are easier to interpret than more recent ones, especially Deep Neural Networks (DNN) which, however, may yield somewhat higher performance when a sufficient amount of data is used. This observation is not specific for prosody, or ASP in general. The type of speech data – especially, whether tightly controlled data such as read speech or less controlled, 'spontaneous' speech under real life conditions (noise) are used – determines performance: the more controlled the data are, the higher is the performance. This is trivial but worth mentioning, as we cannot simply compare performance across studies that use different types of data. Strictly speaking, this can only be done for the very same data used in the same way (such as identical partitioning into train, development, and test sets).

Two main interests can be seen in the study of prosody in ASP: *performance* and *importance*. Performance can be measured – typically, the result is a numerical value between 0 and 1.0, the higher, the better – or it can be mapped onto such a value. Importance is not as easy to define: it can mean importance for a model or theory, or importance for specific applications, therapies, or treatments. Nowadays, in ASP, performance is the preferred measure. However, an equally important goal, often mentioned in introductory or concluding remarks, is to identify important parameters (pitch, intensity, duration, voice quality) or features characterising these parameters. We will come back to this issue in section 3.

In this contribution, we first present a short history of the field in section 2: a timeline in section 2.1 as well as phenomena addressed in the field and performance obtained in a narrative overview in section 2.2. In section 3, we describe the main aspects of features and feature types used, introducing two concepts: *power features* in section 3.1 and *leverage features* in section 3.2, illustrating them in section 3.3.

2 A short history of prosody in automatic speech processing

2.1 Timeline

The history of prosody in ASP started with pioneering studies on certain prerequisites, such as Lieberman (1960) on 'A simple binary automatic stress recognition program' and Mermelstein (1975) on 'automatic segmentation of speech into syllabic units'. Lieberman (1960) already pointed out the incompleteness of the set of prosodic features used, and that prosody is characterized both by the presence of redundant information and by trading relations between different features. The speech material analysed in these studies consisted of prosodic minimal pairs and elicited, carefully read speech. This was (and quite often still is) the usual procedure to exclude the multifarious intervening factors encountered in real-life situations. This approach typical for basic research was adopted by early attempts at incorporating prosodic knowledge in ASP. Early studies laying the foundations for prosody in ASP in the 1980s were Lea (1980), Vaissière (1988, 1989), Waibel (1988), and Batliner & Nöth (1989).

Table 1 gives an overview of prosody in ASP during the last 40 years. Two prototypical approaches are displayed, one for the early years (left) and one for the later years (right) up until now. The information in the table should be read as follows. Most of the studies conducted in the earlier period can be characterised by the components in the left column and, vice versa, most of the studies from the later period by the components in the right column. Approximately, the year 2000 can be viewed as a doorstep between the classical topics and approaches, culminating in a really working prosody module in an end-to-end system (Batliner, Buckow, Niemann, Nöth & Warnke 2000) and, at the same time, in pointing towards the new focus on paralinguistics, with emotion processing as the first, prototypical topic (Batliner, Huber, Niemann, Nöth, Spilker & Fischer 2000). Approaches from the earlier years of course continued to be pursued further on, but to a lesser extent. The entries under *integration* in Table 1 denote a sliding transition from studies where prosody is processed exclusively, i.e. not in conjunction with other parameters, and *intrinsic*, i.e. prosody being the target of the approach, on the one hand, to studies where prosody is used jointly with other parameters, in an integrated way, towards some *extrinsic* goal, i.e. targeting some application, and thus mostly ceases to be visible.[1] The table can be seen as a box of building bricks: any 'component' in the chain of processing (alone or in combination with some other component) from one of the cells (1–6) can be combined.

2.2 Phenomena and performance

In this section we offer a short and rough account of phenomena addressed and performance obtained for them. This is intended to be a compact narrative overview and not a systematic meta-review.

In the 1990s, we see speech processing in a narrower sense, viz. as focusing on word and phrase prosody (accents and boundaries), intonation models[2], and, based on that, syntax (parsing), semantics (salience), and dialogue acts, with respect to both segmentation of such units (chunking) and classification. This trend went together with the general development of automatic speech and language systems, moving from read speech to less controlled speech in more natural situations and leading to conversational speech and dialogue act modelling. In this first phase, most of the time, only prosodic features – sometimes enriched with features from higher linguistic levels – were used. Of course, this line of inquiry continued to be pursued after the turn of the century but was complemented and essentially substituted by a strong focus on paralinguistics, starting with emotion recognition (Daellert et al. 1996) and eventually including all kinds of speaker states and traits, such as the long-term traits age, gender, group membership, and personality; the medium-term traits sleepiness and health state; the short-term states emotion and affects (stress, uncertainty, frustration); and based on all that, interactional/social signals, to mention just a few. In this second phase, normally, prosodic features were used together with other features, especially spectral (cepstral) ones. Thus, we have to keep in mind that performance measures are usually not obtained by using prosodic features alone.

Status reports in terms of overviews and accounts of the state of the art regarding the first period of prosody in ASP are offered in Shriberg & Stolcke (2001) for work done at SRI and in Batliner et al. (2001) for work done at Erlangen University, providing – together with studies referenced therein – a fairly generic account of pertinent research

[1] Prosody 'survives' in stand-alone applications, especially when described for (semi-)commercial products: often, some prosodic parameters are highlighted as most important, because the reader has some idea why, for instance, pitch or loudness should be important. However, such advertisement needs to be distinguished from serious research: notwithstanding the fact that such prosodic parameters can indeed be important, statements about the performance of applications are most of the time counterfactual and cannot be verified.

[2] We use 'intonation' in a narrower sense, comprising only pitch plus delimiters of pitch configurations (boundaries), and 'prosody' in a wider sense, comprising pitch and duration (rhythm), loudness, and voice quality, too.

Table 1: Development of prosody in ASP during the last 40 years. Prototypical approaches for earlier (left) and later years (right) are characterised along several aspects of the chain of processing with different components, with the year 2000 representing a kind of doorstep between traditional topics and the new focus on paralinguistics.

<table>
<tr><th colspan="2">1980 1990 2000 2010 2020</th></tr>
<tr><td colspan="2">(1) motivation</td></tr>
<tr><td>getting wiser; basic knowledge; deciding between theoretical constructs; models / theories</td><td>getting better; successful performance / intervention; applications</td></tr>
<tr><td colspan="2">(2) phenomena</td></tr>
<tr><td>phonetics/linguistics (speech): accents, boundaries, dialogue acts; parsing, dialogue systems; speaker adaptation / verification / identification; ... ± intermediate levels such as tone representation</td><td>paralinguistics (speaker): states (emotion, ...) and traits (personality, ...); diagnostics / teaching / therapy; towards ‘direct’ representation (raw audio in – classes out)</td></tr>
<tr><td colspan="2">(3) data</td></tr>
<tr><td>controlled, constructed; ‘interesting’ phenomena; prompted/acted; lab recordings; one (a few) speaker(s); small segments (units of analysis trivially given)</td><td>less restricted data (more speakers, noisy environment); more spontaneous; from lab to real life; big data; segmentation / chunking into units of analysis necessary</td></tr>
<tr><td colspan="2">(4) features</td></tr>
<tr><td>a few theoretically and/or empirically motivated; only intonational (tunes, pitch patterns, e.g., ToBI); only prosodic (pitch/loudness/duration plus/minus voice quality); syntactic features; speech only (uni-modal)</td><td>many (brute forcing) low-level descriptors (LLD) and functionals; together with other types (spectral (cepstral), ...); all kind of linguistic features; multi-modal (together with facial and body gestures)</td></tr>
<tr><td colspan="2">(5) procedures</td></tr>
<tr><td>‘traditional’ (k-Nearest-Neighbor, Linear Classifiers, Decision Trees, Artificial Neural Networks); feature selection/reduction</td><td>‘modern’ ones (Support Vector Machines, ensemble classifiers (Random Forests)); all varieties of Deep Neural Networks; feature selection/reduction not necessary</td></tr>
<tr><td colspan="2">(6) aim; cf. above motivation</td></tr>
<tr><td>within theory: interpretability, deciding between alternatives, explicit modelling; within applications: employed for syntactic/semantic ‘pre-processing’</td><td>performance; applications: e.g., semantic salience, states and traits; big data, data mining; (towards) implicit modelling of prosody</td></tr>
<tr><td colspan="2">(7) integration</td></tr>
<tr><td colspan="2">stand-alone/intrinsic → intrinsic+extrinsic/±integrated → extrinsic/integrated/± visible</td></tr>
</table>

topics.[3]

The successful incorporation of a prosody module into the end-to-end translation system VERBMOBIL (Batliner, Buckow, Niemann, Nöth & Warnke 2000, Nöth et al. 2000) at the turn of the century can be seen as highlighting the possible impact that prosody can have for ASP.[4] However, such an integration comes at a cost, as described in Spilker et al. (2001) for speech repairs and in Streit et al. (2006) for modelling emotion. The interaction of the prosody module with other modules is highly complex and to some extent instable. In general, the modular and partly knowledge-based design of such systems gave way to integrated ML processing only, which proved to be successful in subsequent years. In a state-of-the-art paper (Xiong et al. 2016) on conversational speech recognition, prosody is not even mentioned. This might be the main reason why the focus of prosody research in ASP, and concomitantly the visibility of prosody in ASP, shifted to the domain of paralinguistics, whereas ASP (and especially ASR) systems today employ prosodic information, if at all, in a rather implicit way, for instance by using prosodic features in combination with all kinds of other features in a large, brute-force feature vector. Yet, there are many studies dealing with the assessment of non-nativeness or specific speech pathologies that address the impact of prosodic features, aiming at identifying the (most) important features (see section 3). Shriberg (2007) gives an overview of higher level (amongst them, prosodic) features in speaker recognition. A survey of studies on computational paralinguistics, not only but also those using prosody, is given in chapters 4 and 5 of Schuller & Batliner (2014).

The history of the *Tones and Break Indices* (ToBI) model (Silverman et al. 1992) regarding its relationship to ASP nicely illustrates how a genuinely phonological-prosodic approach has been harnessed but eventually abandoned by ASP. One of the aims of ToBI was a close collaboration between prosody researchers and engineers (Silverman et al. 1992). Especially during the 1990s, researchers tried to employ ToBI categories for mainstream ASP; see Rosenberg & Hasegawa-Johnson (2018) (this volume) for an assessment of ToBI and its automatic processing ('labelling') for intonation research and typology. We previously elaborated on the drawbacks of this approach in Batliner & Möbius (2005). In short, when tones are used as features in ML procedures, ToBI introduces a quantisation error by reducing detailed information entailed in prosodic features to only a few parameters. A reduced set of ToBI labels, i.e. a 'light' version proposed by Wightman (2002), which was based on results from perception experiments and would recognize classes of tones and breaks rather than the full set of ToBI labels, actually corresponded closely to the labels used in the Verbmobil project (Batliner et al. 1998). In other words, a *functional* model based on the annotation and classification of *perceived* accents and

[3] Needless to say, automatic processing of prosody has been pursued at other sites as well, cf., for example, Price et al. (1991), Wang & Hirschberg (1992), Ostendorf et al. (1993).

[4] Syntactic-prosodic boundary detection reduced the search space for parsing considerably, yielding tolerable response times. This was a limited yet pivotal contribution.

boundaries, is preferred to a *formal* model relying on the annotation and classification of intonational forms, i.e., pitch configurations with delimiters (break indices as quantised pauses), without a clear-cut relationship of these forms to functions.

Table 2 lists representative phenomena that have been addressed, ordered from linguistics to paralinguistics, and from more basic to more complex ones; this vertical order is mirrored in the horizontal timeline in Table 1. Where we report, in a generic way, performance obtained, we refrain from giving exact figures – this would be misleading because performance depends on a plethora of factors such as type of data and features employed. Moreover, it makes a big difference whether Weighted Average Recognition (WAR) or Unweighted Average Recognition (UAR) is used.[5] Instead, we map exact figures for a 2-class UAR problem with a 50% chance level onto ranges in the following way, with the corresponding values for Pearson's correlation, following Coe (2002); UAR is given in percent, followed by Pearson's r in parentheses: *excellent:* $>$ 90% ($>$ .80); *good:* 80 – 90% (0.63 – 0.80); *medium:* 70 – 80% (0.46 – 0.63); *low:* 60 – 70% (0.24 – 0.46); *very low:* $<$ 60% ($<$ .24).

Note again that for all phenomena, controlled (read) speech yields (much) better performance than uncontrolled, spontaneous speech. The performance given in Table 2 can be seen as a blurred snapshot, but it gives an impression of the range of 'goodness of fit' that we can expect. The role of prosody is more salient in the first phase; as already mentioned, it was complemented by other types of features, especially in the second phase. In this overview, we cannot disentangle the contribution of parameters and feature types. Databases employed in CP so far are much smaller than those used for ASR. Larger databases will surely improve performance. Yet, a conservative stance towards improvements to come with more and even 'big' data is advisable: often, in CP, we do not have any ground truth but some gold standard, based on human assessment or labelling and moderate inter-rater agreement.

3 Features and their importance

There are different types of (prosodic) *features* – as they are called in ASP – that are used as independent (predictor) variables and are, e.g., declared as attributes in the standard Attribute-Relation File Format ARFF (Hall et al. 2009). Features can be (1) *low-level descriptors* (LLDs), such as frame-wise fundamental frequency (F0); (2) *functionals*, such

[5] For WAR, chance level is the frequency in percent of the most frequent class. UAR was introduced in the VERBMOBIL project as 'average of the class-wise recognition rates' (Batliner et al. 1998), to facilitate a comparison of performance across results with different numbers of syntactic-prosodic boundary classes (skewed class distributions, up to 25 classes); it has been used as a standard measure in the Paralinguistic Challenges at Interspeech since 2009 (Schuller et al. 2009, Rosenberg 2012). It is fair towards sparse classes (often, these are more interesting than the most frequent class); moreover, the chance level is known when the number of classes is known (50% for 2 classes, 33.3 % for 3 classes, and so on).

Table 2: Phenomena and performance: a rough overview. Qualitative performance terms are typeset in italics.

word recognition: prosody contributes little (*low* performance)
lexicon (word accent, stress): roughly the same performance as for accents
accents: phrase (primary, sentence) accent: *medium* to *good*; secondary accents markedly worse
boundaries: major and minor boundaries, purely prosodic and/or syntactic; major boundaries *good*, sometimes *excellent*; minor boundaries worse; boundaries can be better classified than accents – they display a more categorical distribution
syntactic parsing: based on accent and boundary detection; successful
sentence mood: mainly statement vs. question but others as well (imperative, ...); depends on type of sentence mood: questions vs. statements *medium* to *good*
semantic salience (topic spotting): cf. accents above: islands of reliability, salient topics; closely related to phrase accent processing
dialogue acts: cf. above, sentence mood; sometimes good if pronounced, e.g. back-channelling with duration (here, duration is not really a prosodic feature but simply reflects the fact that back-channellings normally consist of very short words)
agrammatical phenomena: filled/unfilled pauses, false starts, hesitations: *low* to *good*
biological and cultural traits: sex/gender (pitch register): *good* to *very good*
personality traits: big five or single traits; depends on the trait to be modelled: *good* for those that display clear acoustic correlates such as loudness (extraversion), *low* for others
emotional/affective states: same as for personality; arousal *good*, valence *low* (especially if only acoustic features are used); emotions that display pronounced acoustic characteristics can be classified better, cf. anger vs. sadness; yet, anger with high arousal can be be confused with happiness with high arousal
typical vs. atypical speech: pathological speech, non-native speech, temporary deviant speech (duration (non-natives), rhythm, loudness (Parkinson)); *good*, almost on par with single human expert annotator for assessment of intelligibility/naturalness
discrepant speech: irony/sarcasm, deceptive speech (lying): *medium* for controlled speech, rather *very low* for un-controlled speech; off-talk (speaking aside): *medium* to *good*
entrainment / (phonetic) convergence: mutual adaptation of speakers in conversational settings, employing many of the above mentioned phenomena
social/behavioural signals: modelling of speakers in interactional/conversational settings, employing many of the above mentioned phenomena

as the first and second derivatives (delta and delta-delta) or maximum, minimum, skewness, and other values characterizing a distribution; or (3) *structured features*, i.e. LLDs and/or functionals, computed for units like syllables, words, sentences, or paragraphs. Employing (some of) these three types of features, we can obtain (4) *categorical* features, e.g., ToBI tones and breaks (Silverman et al. 1992).[6] The feature set consisting of prosodic and other types of features contains sometimes just a few and sometimes several thousand features. The phenomena to be modelled, such as accents and boundaries; focal structure in syntax; paralinguistic categories (emotions such as anger or happiness) or dimensions (such as arousal or valence in emotion modelling), which traditionally had to be established and annotated manually, are learned from annotated data in a first step and subsequently detected, classified, or evaluated with regression and correlation procedures. In the future, the effort for time-consuming annotations may be reduced, by means of automatic and semi-supervised or unsupervised learning and by end-to-end processing that takes a speech signal (sample values) as input and outputs, e.g., conversational speech in an automatic dialogue system.

Maybe the central question to be asked in prosody research is: 'What is (are) the most important feature(s) for which phenomenon?' To address this question, automatic processing has some advantages: it can handle larger data and feature sets and is therefore more objective than an approach in which the relevance of features is assumed *a priori*. However, the advantages of automatic processing come at a price: it is more cumbersome to handle – because of the sheer number of features – and results are often less clear. We can circumvent this issue by simply relying on a large, brute-force feature vector (Schuller & Batliner 2014, 232ff); for example, the ComParE feature vector used since the Interspeech COMputational PARalinguistics challengE (ComParE) 2014 consisted of 6373 acoustic features – mostly spectral (cepstral) and prosodic ones. This means that most likely, the most important features are indeed captured, although implicitly and along with many other features.[7]

To establish the optimal procedure, we should model all potentially relevant (types of) features, deal with a representative data set, and employ the best feature selection or reduction procedure. This, however, is the Holy Grail and impossible to obtain – we always have to aim at some approximation. Therefore, we should assume that we have a fairly complete feature vector at hand, such as those provided by toolkits like

[6] Phonological, categorical features such as ToBI tones and breaks are, in fact, when used in automatic processing and created by tools like AUToBI (Rosenberg 2009), simply two-step features: LLDs and functionals are used in a first step as features to create phonological categories, and these are then employed in the same way as the other feature types in the second step.

[7] In comparison, the main drawback of the traditional approach to feature relevance is expressed by the rule *What You Are Looking For is What You Get* (WYALFIWYG, Batliner (1989)). In intonation models such as ToBI, just a few (accent and boundary) tones are modelled explicitly. Only when other types of features were eventually modelled explicitly together with pitch, it was revealed that duration is indeed more important for phrase accent in German and English (Batliner et al. 1999, Kochanski et al. 2005); cf. similar results on the word level (e.g. Dogil 1999).

the generic set openSMILE (Eyben et al. 2013), and then employ some state of the art classification and selection or reduction procedure, such as the tried and trusted combination of support vector machines (SVM) and wrapper[8] (e.g., Batliner et al. 2008). Note, however, that such generic feature vectors are not always competitive and have to be complemented by (types of) features especially suited for the given task; cf. Hönig et al. (2012) where structured prosodic, especially rhythmic features, outperformed openSMILE features by a large margin, for the assessment of non-native speech.[9] We also have to decide on a stopping criterion, i.e. on how many *most important features* we want to obtain. Ideally, finding a clear break between important features and those that contribute little, employing the so called 'elbow method' (Thordike 1953), would be helpful, but in practice the curve displaying the improvement of incrementally adding another feature is often rather flat. For convenience, some arbitrary but round number such as 10, 50, 100, or 400 can be chosen for the number of features to be handled and interpreted. Basic functionals such as minimum, maximum or range of values of some parameter are easy to interpret. However, a brute-force vector often results in some derivative of some functional of some low-level descriptor, which are difficult to interpret and explain, and without a corroboration by means of replications or meta-studies it is not possible to assess how reliable and credible a result is in the long run.

Aiming at the relative importance of feature groups instead of single most important features is a feasible alternative (cf. Batliner et al. 2011), but it does not tell us which single features are really important. Yet, types of functionals such as higher variability can as well be employed as most important features. Candidates for such feature types are variability (expressed in terms of several parameters) or expanded range (lowered minima, raised maxima, or both).

3.1 Power features

In this section we sketch what kind of performance we can expect from feature selection, for different constellations of feature vector length and classifier adequacy.

For a large but not well-suited feature vector and/or suboptimally adequate classifiers, the curve is (slightly) rising towards a convex plateau, then somewhat (slightly) falling. Of course, we may observe unexpected irregularities in the curve shape as well.

For a large, well-suited feature vector and adequate classifiers, the curve is (slightly)

[8] Wrappers are computationally costly because for each subset of features, a model is tested, but they normally yield highly competitive performance. Other methods are, e.g., based on correlation or information gain, cf. (Schuller & Batliner 2014, 235ff).

[9] To speculate about the reasons why: Generic feature vectors may be better at modelling global characteristics such as high/low arousal modulated onto speech than at modelling time-dependent, structured relationships such as consonant–vowel transitions or rhythm, which can be characteristic in non-native or pathological speech.

rising and then flat or slightly, asymptomatically rising towards a ceiling.[10] Given this constellation, we may see a steeper rise, singling out a small number of features or just one individual feature already contributing the lion's share of performance. We illustrate these two constellations in section 3.3.

An individual most important feature can be called a *power feature.* If there is a small number of most important features, we can speak of a *group of power features.* Speech tempo and silent pause duration, i.e. grammatical and ungrammatical (hesitation) pauses, have been found to be good predictors of fluency – the faster, the more fluent – and therefore also of language proficiency, for instance for the assessment of non-native speech (Hönig 2016). In the same vein, Black et al. (2015) established knowledge-inspired, competitive features for the same task. Other examples of a power feature are maximum or range of pitch and intensity for emotion (arousal) or, to a lesser extent, for (focal) accent. Bone et al. (2014) describe three power features, namely median pitch, median vocal intensity, and HF500 – the ratio of high-frequency to low-frequency energy with a 500 Hz cutoff – for the rating of emotional arousal.

For the purpose of speeding up processing, in practice, such power features might suffice. However, processing speed is increasingly becoming less of a concern and even large, brute-force feature vectors can now be processed in less than real-time. It is therefore not necessary anymore to reduce the number of features,[11] although speed might still be a criterion for certain time-critical applications (cf. Batliner & Huber 2007).

Another nice example of a power feature can be found in (Rosenberg 2009, 131): for the Boston Direction Corpus – a well-designed corpus with a few speakers, thus, performance can be high – silence ('empty pause') as the only feature for predicting intonational phrase boundaries yields an accuracy of 95.4% for read and 91.4% for spontaneous data. When duration and pitch features are used additionally, only a small gain can be observed, to 95.6% for read and 93.1% for spontaneous data. All features combined yield the best performance, but one single power feature is almost as good. Thus it depends on our intentions whether we employ all features or only the one most important feature.

In Hönig et al. (2014), 27 features were selected manually as acoustic correlates for sleepiness according to the pertinent literature, out of a large vector encompassing 3705 features. Of course, employing all features yielded the best results, but the manually selected features turned out to be on par with the same number of automatically selected features which, however, are often not easy to interpret; for instance, the 75% quantile of the tenth MFCC coefficient on consonantal frames was the second most important

[10] This might look like a 'post hoc ergo propter hoc' explanation: vector and classifier are adequate because they happen to produce the desired result. Of course, we need replication and detailed comparison of feature vectors and classifiers employed.

[11] Note that the *curse of dimensionality*, i.e., the problem of employing too many features in situations of data sparsity (only a few cases), is not relevant if classifiers such as SVM or Random Forests (RF) are used: SVMs are robust towards this problem, and RFs circumvent it by fusing many decision trees each of them having only a small number of features.

feature obtained in the data-driven feature selection.

3.2 Leverage features

Power features may not always be ideal in the context of human-machine interaction. For instance, instructing non-native speakers to speed up is not sufficient to reduce the degree of non-nativeness; in fact, it might be better to advise them to use more pauses, i.e., to slow down, in order to improve intelligibility. Thus, we also need a different type of features – which we call *leverage features* – that can be conveyed easily in teaching or therapy to learners or patients and at the same time contribute to making their speech more natural or typical. In the clinical context, loudness (energy) seems to be a leverage feature for patients with Parkinson's condition (Villa-Cañas et al. 2015), and variability seems to be a leverage feature for patients with depression or children presenting with Autism Spectrum Condition. These features are good for classification and also good for teaching. Chances are that they are highly correlated with other features: loudness is often correlated with F0 maximum and range and with longer duration, and variability of one specific parameter will be correlated with the variability of other parameters, too. Two of the power features used by Bone et al. (2014) – median pitch and median intensity – are good candidates for leverage features.

An interesting case of both a power and a possible leverage feature, but with cross-cultural constraints, is speaker overlap (Hilton 2016). On its own, it is very good at predicting conflict: in Grèzes et al. (2013), speaker overlap as a single feature exceeded the baseline for conflict obtained with 6,373 features by 3% absolute. Such a feature can be used for detection and for teaching and coaching. However, sociocultural conventions prevent this 'Anglo' style from being a universally applicable leverage feature. For instance, in the 'Latin' conversational style, overlap is commonplace and indicates interest rather than conflict, whereas in some Asian cultures ('Oriental' style), overlap is associated with impoliteness and therefore generally avoided, which leads to rather long pauses, irrespective of a possible conflict (Trompenaars & Hampden-Turner 1998, Fitzgerald 2003).

In order to find prosodic features that can be easily conveyed to children on the Autism Spectrum Continuum (ASC), Marchi et al. (2012) compared 15 prosodic features – three prosodic low-level descriptors (energy, pitch and duration) with basic functionals (such as mean, standard deviation, 1st percentile and 99th percentile), manually preselected from a large feature vector, with 15 features obtained from automatic selection procedures out of the same large feature vector. These prosodic features were, for the arousal dimension, superior to the same number of features automatically selected based on information gain.

In a similar approach, Corrales-Astorgano et al. (2018) address the role prosodic features play in the speech of people with Down Syndrome. In Schuller et al. (2018), for classifying arousal, the third quartile of the 25 % spectral roll-off point was the best single feature; it relates to a large portion of higher frequencies but is easier to compute and more robust than F_0. It is therefore a power feature for classification but can be substituted by a – related – pitch feature when we need a leverage feature.

Another option is to consider parameters used in teaching or treatment and then try to identify those that yield a satisfactory performance and can be conveyed to learners or patients, too. Yet, we do not know of any study where this has been done in an integrated, systematic way, comparing brute-force feature vectors, automatically selected subsets, and features derived from therapy or teaching, employing the same group of subjects.

As far as we can see, leverage features are often power features; they will most likely – when used alone – result in some lower performance. On the other hand, they may be highly effective, for instance, in therapy and teaching, provided that the feature is easy to explain and imitate. If this condition is met, the client will (1) understand what to do, and (2) to some extent co-vary other features that contribute to the desired outcome; for instance, higher pitch range will co-vary with longer duration. When we analyse the contribution of features for classification and regression, we should find this co-variation in a higher correlation between these features and their functionals.

3.3 An illustration

Figure 1 illustrates the idea behind power (and therefore leverage) features. The dashed line displays a 'typical' curve: (slightly) rising, without a clear-cut elbow that could serve as a criterion to distinguish most important from less important features. Note that the goal is not just to identify the best features for a specific problem and database but, of course, we want to find a small, generic feature set that will work for similar problems as well. Thus, it may be advisable to include a larger number of features, even if performance gain is low.

By intention, the y-axis in Figure 1 has no concrete values: it depends on the phenomenon whether the ceiling is at, say, 70% or 90%. The x-axis values are just examples: the number of features may range from a few to several hundred. The continuous line shows a sharp rise caused by one or a few power features that contribute the bulk of performance and can easily be distinguished from the remaining features. This performance pattern can be obtained (1) simply from a large feature vector, in which the power features can be more or (usually) less interpretable; (2) from additional features that are based on expert knowledge; or (3) from a knowledge-based selection from the large vector. If we are lucky, (1) and (3) have much in common. Normally, however, we actively have to search the literature and define possible candidate features that are found in our large

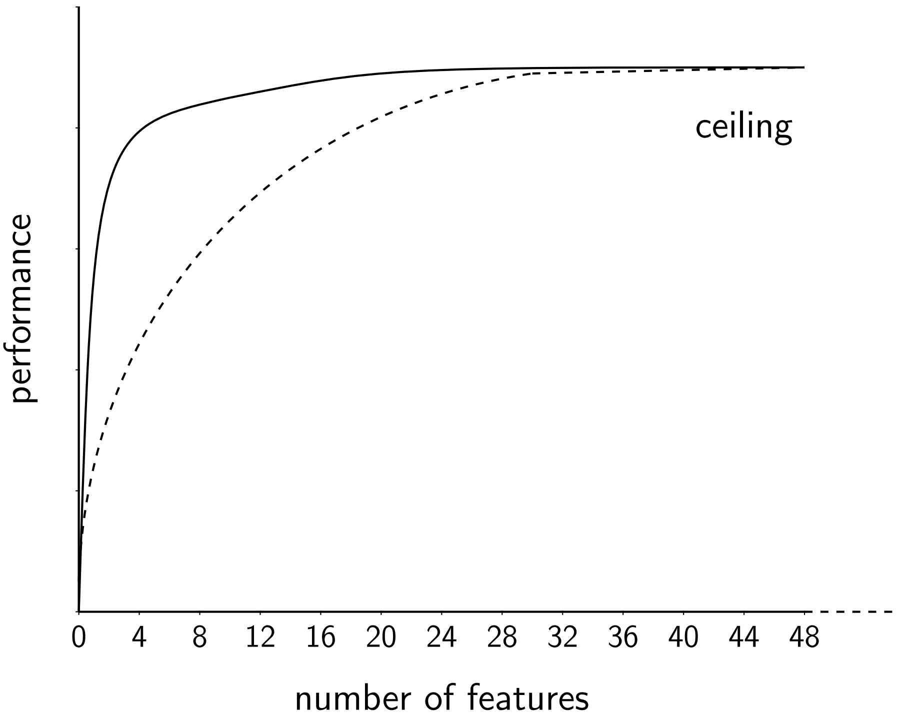

Figure 1: Effect of power features on performance: a few power features contribute strongly to performance (continuous line), whereas often, there is no clear indication of which features contribute most (dashed line).

vector, and if things go well, this manually selected small vector yields a performance that matches the one obtained with the same number of automatically selected features, yielding a high or at least acceptable performance.

In sections 3.1 and 3.2, we referred to a few exemplary studies where these strategies have been applied. There are not many of them because the two sub-cultures – (applied) phonetics and linguistics on the one hand, and engineering approaches on the other hand – are not yet integrated as much as we would wish.

4 Concluding remarks

In this article, we have not addressed in detail the phenomena and algorithmic procedures that have been dealt with in the field of automatic processing of prosody, and we refrained from presenting exact performance measures across studies – this is often done in survey articles on paralinguistics but of doubtful value because a strict comparability is almost never met. Moreover, we have not given a full account of the history and the state-of-the-art. Instead, we have tried to sort out the most important methodological trends along the timeline from the 1980s up until today, from a bird's eye perspective. We have seen that in the first phase, prosody was very visible. In the second phase, with the advent of heavy ML in integrated approaches, prosody has ceased to be visible.

This might change again if we explicitly address power and leverage features and their relationship to linguistic structure, to wit not only in basic research but in applications as well. Another interesting research avenue is the combination of acoustic-prosodic features and text-based features in applications of natural language processing such as question answering, sentiment analysis, and the analysis of referring expressions in discourse and dialogue.

Traditional linguistic treatments of prosody and ASP have an important aspect in common. They are both eschatological to some degree. In linguistic theory, newly invented models are assumed to be, and presented as, definitive and necessarily superior to the older ones. In ASP, new methodological frameworks such as, at this time, deep learning are assumed to present the solution for every problem. History tells us that none of this is very likely to be the case in the longer run. Although scientific paradigms (Kuhn 1970) are persistent, it is difficult to predict which theories and methods will prevail in the medium-term future. But it is not risky to predict that there will be no big, unified approach embracing both theories and machine learning, and we may not see much convergence and collaboration between theory and models on the one hand, and engineering approaches on the other hand. Yet, we can hope for some practical convergence – and by that, a higher visibility of prosody – due to the possibility of ubiquitous applications, which will make it necessary to find links between automatic processing and analysis, synthesis, and learning and therapy, eventually yielding further insight in the intricate relationship between power and leverage features.

5 Acknowledgements

This work was supported by the European Union's Horizon 2020 Programme under grant agreement No. 688835 (RIA DE-ENIGMA). We want to thank two anonymous reviewers for their comments. The responsibility lies with the authors.

References

Batliner, A. (1989), Eine Frage ist eine Frage ist keine Frage. Perzeptionsexperimente zum Fragemodus im Deutschen, *in* H. Altmann, A. Batliner & W. Oppenrieder, eds, 'Zur Intonation von Modus und Fokus im Deutschen', Niemeyer, Tübingen, pp. 87–109.

Batliner, A., Buckow, A., Niemann, H., Nöth, E. & Warnke, V. (2000), The Prosody Module, *in* W. Wahlster, ed., 'Verbmobil: Foundations of Speech-to-Speech Translations', Springer, Berlin, pp. 106–121.

Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E. & Niemann, H. (1999), Prosodic Feature Evaluation: Brute Force or Well Designed?, *in* 'Proc. of ICPhS', San Francisco, pp. 2315–2318.

Batliner, A. & Huber, R. (2007), Speaker Characteristics and Emotion Classification, *in* C. Müller, ed., 'Speaker Classification I: Fundamentals, Features, and Methods', LNAI, Springer, Berlin-Heidelberg, pp. 138–151.

Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J. & Fischer, K. (2000), The Recognition of Emotion, *in* W. Wahlster, ed., 'Verbmobil: Foundations of Speech-to-Speech Translations', Springer, Berlin, pp. 122–130.

Batliner, A., Kompe, R., Kießling, A., Mast, M., Niemann, H. & Nöth, E. (1998), 'M = Syntax + Prosody: A syntactic–prosodic labelling scheme for large spontaneous speech databases', *Speech Communication* **25**, 193–222.

Batliner, A. & Möbius, B. (2005), Prosodic Models, Automatic Speech Understanding, and Speech Synthesis: Towards the Common Ground?, *in* W. Barry & W. Dommelen, eds, 'The Integration of Phonetic Knowledge in Speech Technology', Springer, Dordrecht, pp. 21–44.

Batliner, A. & Nöth, E. (1989), The Prediction of Focus, *in* 'Proc. of EUROSPEECH', Paris, pp. 210–213.

Batliner, A., Nöth, E., Buckow, J., Huber, R., Warnke, V. & Niemann, H. (2001), Whence and Whither Prosody in Automatic Speech Understanding: A Case Study, *in* M. Bacchiani, J. Hirschberg, D. Litman & M. Ostendorf, eds, 'Proc. of the Workshop on Prosody and Speech Recognition 2001', Red Bank, NJ, pp. 3–12.

Batliner, A., Schuller, B., Schaeffler, S. & Steidl, S. (2008), Mothers, Adults, Children, Pets — Towards the Acoustics of Intimacy, *in* 'Proc. of ICASSP', Las Vegas, NV, pp. 4497–4500.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V. & Amir, N. (2011), 'Whodunnit – Searching for the Most Important Feature Types Signalling Emotional User States in Speech', *Computer Speech and Language* **25**, 4–28.

Black, M. P., Bone, D., Skordilis, Z. I., Gupta, R., Xia, W., Papadopoulos, P., Chakravarthula, S. N., Xiao, B., Segbroeck, M. V., Kim, J., Georgiou, P. G. & Narayanan, S. S. (2015), Automated evaluation of non-native English pronunciation quality: combining knowledge- and data-driven features at multiple time scales, *in* 'Proc. of INTERSPEECH', Dresden, Germany, pp. 493–497.

Bone, D., Lee, C.-C. & Narayanan, S. (2014), 'Robust Unsupervised Arousal Rating: A Rule-Based Framework with Knowledge-Inspired Vocal Features', *IEEE Transactions on Affective Computing* **5**, 201–213.

Coe, R. (2002), 'It's the effect size, stupid: What effect size is and why it is important', Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, 12-14 September, accessed from www.leeds.ac.uk/educol/documents/00002182.htm on 13 March 2018.

Corrales-Astorgano, M., Escudero-Mancebo, D. & Gonzalez-Ferreras, C. (2018), 'Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome', *Speech Communication* . to appear; doi: 10.1016/j.specom.2018.03.006.

Daellert, F., Polzin, T. & Waibel, A. (1996), Recognizing emotion in speech, *in* 'ICSLP', Vol. 3, Philadelphia.

Dogil, G. (1999), The phonetic manifestation of word stress in Lithuanian, German, Polish and Spanish, *in* H. van der Hulst, ed., 'Word Prosodic Systems in the Languages of Europe', De Gruyter, Berlin, pp. 273–311.

Eyben, F., Weninger, F., Gross, F. & Schuller, B. (2013), Recent developments in opensmile, the munich open-source multimedia feature extractor, *in* 'Proceedings of the 21st ACM International Conference on Multimedia', Barcelona, Spain, pp. 835–838.

Fitzgerald, H. (2003), *How Different are We? Spoken Discourse in Intercultural Communication*, Multilingual Matters, Clevendon, UK.

Grèzes, F., Richards, J. & Rosenberg, A. (2013), Let me finish: automatic conflict detection using speaker overlap, *in* 'Proc. of INTERSPEECH', Lyon, pp. 200–204.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. (2009), 'The WEKA Data Mining Software: An Update', *SIGKDD Explorations* **11**.

Hilton, K. (2016), The Perception of Overlapping Speech: Effects of Speaker Prosody and Listener Attitudes, *in* 'Proc. of INTERSPEECH', San Francisco, CA, pp. 1260–1264.

Hönig, F. (2016), Automatic Assessment of Prosody in Second Language Learning, PhD thesis, Friedrich-Alexander Universität Erlangen-Nürnberg, Berlin.

Hönig, F., Batliner, A., Nöth, E., Schnieder, S. & Krajewski, J. (2014), Acoustic-Prosodic Characteristics of Sleepy Speech – between Performance and Interpretation, *in* 'Proc. of Speech Prosody 2014', Dublin, pp. 864–868.

Hönig, F., Bocklet, T., Riedhammer, K., Batliner, A. & Nöth, E. (2012), The Automatic Assessment of Non-native Prosody: Combining Classical Prosodic Analysis with Acoustic Modelling, *in* 'Proc. of Interspeech', Portland Oregon, USA.

Kochanski, G., Grabe, E., Coleman, J. & Rosner, B. (2005), 'Loudness predicts Prominence; Fundamental Frequency lends little', *Journal of the Acoustical Society of America* **11**, 1038–1054.

Kuhn, T. S. (1970), *The Structure of Scientific Revolutions*, 2 edn, University of Chicago Press, Chicago. International Encyclopedia of Unified Science, Vol. II,2.

Lea, W. (1980), Prosodic Aids to Speech Recognition, *in* W. Lea, ed., 'Trends in Speech Recognition', Prentice–Hall Inc., Englewood Cliffs, New Jersey, pp. 166–205.

Lieberman, P. (1960), 'Some Acoustic Correlates of Word Stress in American English', *Journal of the Acoustical Society of America* **32**, 451–454.

Marchi, E., Schuller, B., Batliner, A., Fridenzon, S., Tal, S. & Golan, O. (2012), Emotion in the Speech of Children with Autism Spectrum Conditions: Prosody and Everything Else, *in* 'Proceedings of the 3rd Workshop on Child, Computer and Interaction', Portland, OR.

Mermelstein, P. (1975), 'Automatic segmentation of speech into syllabic units', *Journal of the Acoustical Society of America* **58**, 880–883.

Nöth, E., Batliner, A., Kießling, A., Kompe, R. & Niemann, H. (2000), 'Verbmobil: The Use of Prosody in the Linguistic Components of a Speech Understanding System', *IEEE Transactions on Speech and Audio Processing* **8**, 519–532.

Ostendorf, M., Wightman, C. & Veilleux, N. (1993), 'Parse Scoring with Prosodic Information: an Analysis/Synthesis approach', *Computer Speech and Language* **7**, 193–210.

Price, P., Ostendorf, M. & Shattuck-Hufnagel, S. (1991), Disambiguating Sentences using Prosody, *in* 'Proc. of ICSLP', Aix–en–Provence, pp. 418–421.

Rosenberg, A. (2009), Automatic detection and classification of prosodic events, PhD thesis, Dissertation,Columbia University.

Rosenberg, A. (2012), Classifying skewed data: Importance weighting to optimize average recall, *in* 'Proc. of INTERSPEECH', Portland, OR, pp. 2242–2245.

Rosenberg, A. & Hasegawa-Johnson, M. (2018), 'Chapter 47. Automatic labelling and assessment ', in this volume.

Schuller, B. & Batliner, A. (2014), *Computational Paralinguistics – Emotion, Affect, and Personality in Speech and Language Processing*, Wiley, Chichester, UK.

Schuller, B., Steidl, S. & Batliner, A. (2009), The INTERSPEECH 2009 Emotion Challenge, *in* 'Proc. of INTERSPEECH', Brighton, pp. 312–315.

Schuller, B., Weninger, F., Zhang, Y., Ringeval, F., Batliner, A., Steidl, S., Eyben, F., Marchi, E., Vinciarelli, A., Scherer, K., Chetouani, M. & Mortillaro, M. (2018), 'Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge', *CSL* . to appear.

Shriberg, E. (2007), Higher-level features in speaker recognition, *in* C. Müller, ed., 'Speaker Classification I: Fundamentals, Features, and Methods', Springer, Berlin, Heidelberg, pp. 241–259.

Shriberg, E. & Stolcke, A. (2001), Prosody Modeling for Automatic Speech Understanding: An Overview of Recent Research at SRI, *in* M. Bacchiani, J. Hirschberg, D. Litman & M. Ostendorf, eds, 'Proc. of the Workshop on Prosody and Speech Recognition 2001'.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992), ToBI: a standard for labeling English prosody, *in* 'Proc. of ICSLP', Banff, Canada, pp. 867–870.

Spilker, J., Batliner, A. & Nöth, E. (2001), How to Repair Speech Repairs in an End-to-End System, *in* 'Proc. of the ISCA Workshop on Disfluency in Spontaneous Speech', Edinburgh, pp. 73–76.

Streit, M., Batliner, A. & Portele, T. (2006), Emotions Analysis and Emotion-Handling Subdialogues, *in* W. Wahlster, ed., 'SmartKom: Foundations of Multimodal Dialogue Systems', Springer, Berlin, pp. 317–332.

Thordike, R. L. (1953), 'Who Belongs in the Family?', *Psychometrika* **18**, 267–276.

Trompenaars, F. & Hampden-Turner, C. (1998), *Riding the Waves of Culture: Understanding Diversity in Global Business*, 2nd edn, McGraw-Hill Companies, Incorporated.

Vaissière, J. (1988), The Use of Prosodic Parameters in Automatic Speech Recognition, *in* H. Niemann, M. Lang & G. Sagerer, eds, 'Recent Advances in Speech Understanding and Dialog Systems', Vol. 46 of *NATO ASI Series F*, Springer–Verlag, Berlin, pp. 71–99.

Vaissière, J. (1989), On Automatic Extraction of Prosodic Information for Automatic Speech Recognition Systems, *in* 'Proc. of EUROSPEECH', Paris, pp. 202–205.

Villa-Cañas, T., Arias-Londoño, J. D., Orozco-Arroyave, J. R., Vargas-Bonilla, J. F. & Nöth, E. (2015), Low-frequency components analysis in running speech for the automatic detection of Parkinson's disease, *in* 'Proc. of INTERSPEECH', pp. 100–104.

Waibel, A. (1988), *Prosody and Speech Recognition*, Morgan Kaufmann Publishers Inc., San Mateo.

Wang, M. & Hirschberg, J. (1992), 'Automatic Classification of Intonational Phrase Boundaries', *Computer Speech and Language* **6**, 175–196.

Wightman, C. (2002), ToBI or not ToBI, *in* 'Proc. of Speech Prosody', Aix en Provence, pp. 25–29.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D. & Zweig, G. (2016), 'Achieving Human Parity in Conversational Speech Recognition', http://arxiv.org/abs/1610.05256.