Prosody in Automatic Speech Processing

Anton Batliner and Bernd Möbius

The Oxford Handbook of Language Prosody
Edited by Carlos Gussenhoven and Aoju Chen

Print Publication Date: Dec 2020 Subject: Linguistics, Phonetics and Phonology
Online Publication Date: Feb 2021 DOI: 10.1093/oxfordhb/9780198832232.013.42

Abstract and Keywords

Automatic speech processing (ASP) is understood as covering word recognition, the processing of higher linguistic components (syntax, semantics, and pragmatics), and the processing of computational paralinguistics (CP), which deals with speaker states and traits. This chapter attempts to track the role of prosody in ASP from the word level up to CP. A short history of the field from 1980 to 2020 distinguishes the early years (until 2000)—when the prosodic contribution to the modelling of linguistic phenomena, such as accents, boundaries, syntax, semantics, and dialogue acts, was the focus—from the later years, when the focus shifted to paralinguistics; prosody ceased to be visible. Different types of predictor variables are addressed, among them high-performance power features as well as leverage features, which can also be employed in teaching and therapy.

Keywords: automatic speech processing, word recognition, computational paralinguistics, word level, prosody, predictor variables, leverage features

46.1 Introduction

WE understand ‘automatic speech processing’ (ASP) to mean word recognition (automatic speech recognition (ASR)), processing of higher linguistic components (syntax, semantics, and pragmatics), and processing of computational paralinguistics (CP). This chapter attempts to describe the role of prosody in ASP from the word level up to the level of CP, where the focus was initially on emotion recognition and later expanded to the recognition of health conditions, social signals such as back-channelling, and speaker states and traits (Schuller and Batliner 2014).

‘Automatic processing’ of prosody means that at least part of the processing is done by the computer. The automatic part can be small, for example pertaining only to pitch extraction, followed by manual correction of the fundamental frequency (f0) values with subsequent automatic computation of characteristic values such as mean, minimum, or maximum. This is typically done in basic, possibly exploratory, research on prosody and in studies aiming to evaluate certain models and theories. A fully automatic processing of

prosody, on the other hand, is necessary when we employ prosody in conjunction with other information in a larger context, such as developing a prosody module in a complete speech-to-speech dialogue system, or improving the speech of pathological speakers or foreign language learners via screening, monitoring, and feedback on the learning progress in a stand-alone tool.

Apart from the *phenomena* to be investigated—such as prosodic parameters, emotions and affects, speaker states and traits, and social signals (for details see §46.2.2)—and the speech *data* to be recorded, the basic ingredients of automatic processing of prosody are (i) the units of analysis, suited to both the phenomenon and the type of features we employ; (ii) the features to be extracted; and (iii) machine learning (ML) procedures that tell us how good we are (i.e. which classification performance we obtain) and, if relevant, which features are most important, and for which units.

The units of analysis in the processing of prosody may be implicit (e.g. an entire speech file), be temporally defined (e.g. segments of five seconds or one tenth of the entire speech (p. 634) file), or be obtained via pre-processing, such as voice activity detection (e.g. using silence as an indicator for major prosodic/syntactic boundaries), ASR yielding word boundaries, syntactic parsing that generates phrase and sentence boundaries, or a combination of these strategies.

Regarding ML procedures, many have been employed for the processing of prosody in ASP. Generally speaking, traditional, well-established procedures, such as linear classifiers and decision trees, tend to yield somewhat lower but more interpretable performance than the more recently developed procedures, such as deep neural networks, which tend to yield better results on larger data sets. Additionally, more controlled data, such as read speech, is likely to yield a better performance than spontaneous speech. This point may seem trivial but is worth stressing, since comparisons across different types of speech data are not uncommon. Strictly speaking, a comparison of performance obtained by, for example, different ML procedures can only be done for the very same data used in the same way, including, for instance, identical partitioning into train, development, and test sets.

Evaluating the role of prosody in ASP has focused on two issues: *performance* and *importance*. Performance can be measured: typically, the result is a numerical value between 0 and 1.0 (the higher, the better) or can be mapped onto such a value (Schuller and Batliner 2014). 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, performance is the preferred measure in ASP. However, an equally important issue, often mentioned in introductory or concluding remarks, is to identify salient parameters (pitch, intensity, duration, voice quality) or features characterizing these parameters (see more on this in §46.3).

In this chapter, we first present a short history of the field (§46.2), including a timeline in §46.2.1 and an overview of the phenomena addressed in the field and performance obtained in §46.2.2. We then describe the main aspects of prosodic features and feature

types used in ASP in §46.3., introducing two concepts: ‘power features’ in §46.3.1 and ‘leverage features’ in §46.3.2. We then illustrate these concepts in §46.3.3, which is followed by concluding remarks in §46.4.

46.2 A short history of prosody in automatic speech processing

46.2.1 Timeline

The history of prosody in ASP started with pioneering studies on the prerequisites for automatic processing of prosody, such as Lieberman (1960: 451) on ‘a simple binary automatic stress recognition program’[1] and Mermelstein (1975) on ‘automatic segmentation of speech into syllabic units’. 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 (p. 635) procedure used to exclude the multifarious confounding factors encountered in real-life situations. This approach, typical of basic research, was adopted by early attempts at incorporating prosodic knowledge in ASP.

Table 46.1 gives an overview of research on prosody in ASP over the past 40 years. Most of the studies conducted in the earlier period can be characterized by the components in the left column and most of the studies from the later period by the components in the right column. The entries under ‘integration’ in Table 46.1 denote a sliding transition from studies where prosody is processed alone (stand-alone) and as sole topic (intrinsic), by that being visible, to studies where prosody is used jointly with other parameters in an integrated way, towards some extrinsic goal (i.e. targeting some application), and leading to prosody becoming invisible as a contributing factor. Early studies that laid the foundations for prosody in ASP in the 1980s include Lea (1980), Vaissière (1988, 1989), and Batliner and Nöth (1989). The year 2000 can be viewed as a turning point away from these classical approaches, culminating in a functional prosody module in an end-to-end system (Batliner et al. 2000a) and moving towards new approaches with a focus on the processing of paralinguistics, starting with emotion recognition (Batliner et al. 2000b). Approaches from the earlier years nevertheless continued to be pursued after 2000, but to a lesser extent.

Table 46.1 can be seen as a set of building blocks: 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. Normally, only cells from the left or cells from the right are combined with each other unless a comparison of methodologies is aimed at (see, for instance, Batliner et al. 2000c).

Table 46.1 Prototypical approaches in research on prosody in automatic speech processing over the past 40 years (1980–2020), with the year 2000 as a turning point from traditional topics to a new focus on paralinguistics

1980	1990	2000	2010	2020
		1. Motivation		
Getting wiser; basic knowledge; deciding between theoretical constructs; models/theories		Getting better; successful performance/intervention; applications		
		2. Phenomena		
Phonetics/linguistics (speech): accents, boundaries, dialogue acts; parsing, dialogue systems; speaker adaptation/verification/identification; … ± intermediate levels such as tone representation		Paralinguistics (speaker): states (emotion, pain, etc.) and traits (personality, ethnicity, etc.); diagnostics/teaching/therapy; towards 'direct' representation (raw audio in–classes out)		
		3. Data		
Controlled, constructed; 'interesting' phenomena; prompted/acted; lab recordings; one (a few) speaker(s); small segments (units of analysis trivially given)		Less restricted data (more speakers, noisy environment); more spontaneous; from lab to real life; big data; segmentation/chunking into units of analysis necessary		
		4. Features		

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)	Many (brute forcing) low-level descriptors and functionals; together with other types (spectral (cepstral)); all kind of linguistic features; multi-modal (together with facial and body gestures)
	5. Procedures
'Traditional' (k-nearest-neighbour, linear classifiers, decision trees, artificial neural networks); feature selection/reduction	'Modern' ones (support vector machines, ensemble classifiers (random forests)); all varieties of deep neural networks; feature selection/reduction not necessary
	6. Utilization
Within theory: interpretability, deciding between alternatives, explicit modelling; within applications: employed for syntactic/semantic 'pre-processing'	Performance; applications: e.g. semantic salience, states and traits; big data, data mining; (towards) implicit modelling of prosody
	7. Integration
Stand-alone, intrinsic, visible	→ Integrated, extrinsic, not visible

46.2.2 Phenomena and performance

In this section we take a closer look at the phenomena addressed in past studies on prosody in ASP (Table 46.1) and performance obtained for them in ASP. This is intended as a compact narrative overview instead of a systematic meta-review.

In the second phase (after the year 2000), prosodic features were mainly used together with other features, especially spectral (cepstral) ones. It is therefore important to keep in mind that performance measures are usually not obtained by using prosodic features alone. In the 1990s, speech processing focused narrowly on the role of word and phrase prosody (accents and boundaries), intonation models,[2] syntax (parsing) based on prosodic models, semantics (salience), and segmentation and classification of dialogue acts. This trend went in tandem with the general development of automatic speech and language processing systems, moving from read speech to less controlled speech in more natural situations and leading to conversational speech and dialogue act modelling. In the first phase (before 2000), most of the time, only prosodic features—sometimes enriched with features from higher linguistic levels—were used; see reviews of state-of-the-art systems in Shriberg and Stolcke (2001) and Batliner et al. (2001), as well as Price et al. (1991), Wang and Hirschberg (1992), and Ostendorf et al. (1993). This line of inquiry continued to be pursued after the turn of the century but was complemented and essentially replaced by a strong focus on paralinguistics, starting with emotion recognition (Daellert et al. 1996) and eventually extending to all kinds of speaker states and traits, including long-term traits, such as age, gender, group membership, and personality; medium-term traits, such as sleepiness and health state; short-term states, such as emotion and affect (e.g. stress, uncertainty, frustration); and interactional/social signals.

The successful incorporation of a prosody module into the end-to-end translation system VERBMOBIL (Batliner et al. 2000a; Nöth et al. 2000) has highlighted the impact that (p. 637) prosody can have for ASP.[3] 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 unstable. In general, the modular and partly knowledge-based design of such systems gave way to an integrated ML approach, which proved to be successful in subsequent years: in a state-of-the-art paper (Xiong et al. 2017) 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, has 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 concerned 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 §46.3.[4]

The implementation of the Tones and Break Indices (ToBI) model (Silverman et al. 1992) in ASP nicely illustrates how a genuinely phonological-prosodic approach was harnessed

but eventually abandoned by ASP. One of the aims of ToBI was to foster a close collaboration between prosody researchers and engineers (Silverman et al. 1992). Especially during the 1990s, researchers tried to employ ToBI categories in mainstream ASP. However, using tonal categories as features in ML procedures introduces a quantization error by reducing detailed prosodic information to only a few parameters (Batliner and Möbius 2005). A reduced set of ToBI labels—that is, a light version proposed by Wightman (2002), which was based on results from perception experiments and would recognize classes of tones and breaks instead of 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 syntactic-prosodic boundaries should be preferred to a *formal* model relying on the annotation and classification of intonational forms—that is, pitch configurations with delimiters (break indices as quantized pauses), without a clear-cut relationship of these forms to functions.

In Table 46.2, we report performance obtained for a selection of representative phenomena that have been addressed, ordered vertically from linguistic features to paralinguistic features, and from the more basic ones to the more complex ones, largely corresponding to the entries listed under ‘phenomena’ in Table 46.1. 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 of presenting exact figures, we map the figures onto ranges of performance, (p. 638) following Coe (2002); UAR for a two-class problem with 50% chance level is given in per cent, 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*: 0.63–0.80 (0.24–0.46); *very low*: <60% (<.24).

Table 46.2 Phenomena and performance: a rough overview (qualitative performance terms appear 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, subjunctive, etc.); 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 rather *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 confused with happiness with high arousal
Typical vs. atypical speech: pathological speech, non-native speech, temporary deviant speech (duration (non-natives), rhythm, loudness (Parkinson's condition)); *good*, almost on par with single human expert annotators for assessment of intelligibility/naturalness
Discrepant speech: irony/sarcasm, deceptive speech (lying): *medium* for controlled speech, but *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

Note that all of the phenomena in Table 46.2 fare better in read speech than in spontaneous speech. The qualitative performance terms refer to the range of performance levels that we may expect.[6] Because the role of prosody was more easily identifiable in the first (p. 639) phase than in the second phase, when non-prosodic features were included, the contribution of prosody to performance cannot easily be disentangled from the contribution of those other features. Additionally, the databases employed in CP today are much smaller than those used for ASR, and, unsurprisingly, larger databases will yield better performance. Nevertheless, it is wise to adopt a conservative stance when it comes to expectations from 'big' data, since so far we usually only have a *gold standard* to measure performance, as established by human assessment or labelling with moderate inter-rater agreement, rather than a *ground truth* (i.e. an objective set of crsiteria).[7] Moreover, it is as yet unclear whether modern approaches towards enlarging databases (such as crowdsourcing, transfer-learning, and zero-shot learning[8]) will really result in big data whose size can be compared to the corpora available for ASR.

46.3 Features and their importance

Various types of prosodic features are used as independent (predictor) variables; in ASP, predictor variables are simply referred to as 'features', and a set of features constitutes a 'feature vector' in ML processing. Features can be (i) *low-level descriptors* (LLDs), such

as frame-wise f0; (ii) *functionals*, such as the first and second derivatives (delta and delta-delta) or maximum, minimum, skewness, and other values characterizing a distribution of LLDs; or (iii) *structured features*, which are LLDs and/or functionals, computed for units such as syllables, words, sentences, or paragraphs (Schuller and Batliner 2014). Employing (some of) these three types of features, we can obtain (iv) *categorical* features, such as ToBI tones and breaks (Silverman et al. 1992).[9] A feature set (feature vector) consisting of prosodic and other types of feature can contain a few to several thousand features. The phenomena to be modelled—such as accents and boundaries, focal structure in syntax, and paralinguistic categories (emotions such as anger or happiness) or dimensions (such as arousal or valence in emotion modelling)—which traditionally need to be established and annotated manually, are learned initially from annotated data and subsequently detected, classified, or evaluated via regression and correlation procedures. In the future, the effort of 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 output (e.g. conversational speech in an automatic dialogue system).

The central question to be asked in prosody research may be this: What is (are) the most important feature(s) for which phenomenon? To address this question, automatic processing (p. 640) 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 than clear-cut. We can circumvent this issue by simply relying on a large, brute-force feature vector (Schuller and Batliner 2014: 232–234). For example, the ComParE feature vector used since the Interspeech Computational Paralinguistics Challenge (ComParE) 2013 consists of 6,373 acoustic features—mostly spectral (cepstral) and prosodic ones (Schuller et al. 2013). This means that, most likely, the most important features are indeed captured, although implicitly and along with many other features.[10]

To establish the optimal procedure resulting in a feature vector that can be interpreted and yields good performance, 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: impossible to obtain but well worth trying to approximate. Therefore, we should make sure that a fairly complete feature vector is available, such as the one provided by the generic toolkit 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 (SVMs) and wrappers[11] (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 to the given task; see Hönig et al. (2012), where structured prosodic, especially rhythmic, features outperformed openSMILE features by a large margin in the assessment of non-native speech.[12] We also have to decide on a limiting (*stopping*) criterion for the number of *most important features* we aim to obtain. Ideally, to find a clear break between important features and those that contribute little, it would be helpful to employ the ‘elbow

method' (Thordike 1953), but in practice the curve showing the improvement of incrementally adding another feature is often rather flat. For convenience, an arbitrary but round number (e.g. 10, 50, 100, or 400) can be chosen for the number of features to be handled and interpreted. Furthermore, basic functionals such as the minimum, maximum, or range of values of some parameter are easy to interpret. By contrast, a brute-force vector often results in some derivatives of some functional or some LLD, which are difficult to interpret and explain; moreover, without replications or meta-studies, it is not possible to assess how reliable and credible a result will be in the long run.

(p. 641) Aiming at the relative importance of feature groups instead of the single most important features is a feasible alternative (Batliner et al. 2011), but it does not tell us which single features are really important. Yet, (derivatives of) functionals such as higher variability (expressed in terms of several parameters) and expanded range (lowered minima, raised maxima, or both) can be employed as the most important features, as the underlying features are positively correlated with each other (§46.3.2).

46.3.1 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 (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) rising and then flat or slightly, asymptotically rising towards a ceiling.[13] Given this constellation, we may see a steeper rise, singling out a small number of features or just one individual feature that is already contributing the lion's share of performance. We illustrate these two constellations in §46.3.3.

A single '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'. For instance, speech tempo and silent pauses (i.e. grammatical and ungrammatical (hesitation) pauses) have been found to be good predictors of fluency—the faster and the fewer pauses, the more fluent—and therefore also of language proficiency, for the assessment of non-native speech (Hönig 2016). In the same vein, Black et al. (2015) established a group of knowledge-inspired, competitive features modelling speaking rate and pauses 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) described three power features for the rating of emotional arousal, namely median pitch, median vocal intensity, and HF500 (i.e. the ratio of high-frequency to low-frequency energy with a 500 Hz cutoff).

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, which means that performance can be high—using 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 speech 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 single most important feature.

In Hönig et al. (2014a), 27 features were selected manually as acoustic correlates of sleepiness according to the pertinent literature, from a large vector encompassing 3,705 features. Although using all features yielded the best results, the performance of the 27 manually (p. 642) selected features turned out to be on par with that of the same number of automatically selected features, which are often not easy to interpret: for instance, the 75% quantile of the tenth MFCC (mel frequency cepstral coefficient) on consonantal frames was the second most important feature obtained in the data-driven feature selection. This approach was also adopted for the modelling of depressed speech in Hönig et al. (2014b).

Using such smaller sets of power features may well serve to speed up processing. However, processing speed is increasingly becoming less of a concern and even large, brute-force feature vectors can now be processed in a very short time, even in less than real time, making reduction of the number of features unnecessary,[14] although speed might still be an issue for certain time-critical, not server-based but embedded, applications.

46.3.2 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. For instance, a foreign language teacher or a speech therapist can elicit higher variability (corresponding, e.g., to more extreme f0/energy maxima and minima) in their students’ speech by telling them ‘Please, do not speak that monotonously, speak in a more lively manner’ and by demonstrating these two different styles. In this section, we will list possible candidate features and refer to pertinent studies.

An interesting case of both a power and a possible leverage feature, but with cross-cultural constraints, is speaker overlap (Hilton 2016).[15] 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 change. Such a feature can be used for detection and for teaching and coaching. However, sociocultural conventions prevent this ‘Anglo’ conversation style from being a universally applicable leverage feature. For instance, in the ‘Latin’ conversation 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 and Hampden-Turner 1998; Fitzgerald 2003).

In the clinical context, loudness (energy) would appear to be a leverage feature for patients with Parkinson's condition (Villa-Cañas et al. 2015), and variability would appear to be a leverage feature for patients with depression or children diagnosed with autism spectrum condition (ASC). These features are good for classification and also good for teaching. (p. 643) 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. In order to find leverage features for children with ASC, Marchi et al. (2012) compared 15 prosodic features—three prosodic LLDs (energy, pitch, and duration) with basic functionals (such as mean, standard deviation, 1st percentile, and 99th percentile), manually pre-selected from a large feature vector—with 15 features automatically selected from the same large feature vector, based on information gain. The manually selected prosodic features were, for the arousal dimension, superior to the same number of automatically selected features. In a similar approach, Corrales-Astorgano et al. (2018) examined the role of prosodic features in the speech of patients with Down syndrome.

Two of the power features used by Bone et al. (2014)—median pitch and median intensity—are good candidates for leverage features. In Schuller et al. (2019), for classifying arousal, the third quartile of the 25% spectral roll-off point was the best single feature; it relates to a large proportion of higher frequencies but is easier to compute and more robust than f0. 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 and their shape used in teaching or treatment and then identify those features that yield a satisfactory performance while also being easily conveyed to learners or patients. Yet, we do not know of any study that systematically compares brute-force feature vectors, automatically selected subsets, and features derived from therapy or teaching, employing the same group of subjects.

It seems to be plausible that leverage features are also power features; in the same way, they will most likely—when used alone—result in some lower performance compared to a full feature vector. Yet, they can be more generic across databases, languages, and cultures. Moreover, they can 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 (i) understand what to do and (ii) to some extent co-vary other features that contribute to the desired outcome; for instance, a wider 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.

46.3.3 An illustration

Figure 46.1 illustrates the idea behind power (and leverage) features. By intention, the *y*-axis in Figure 46.1 has no concrete values: it depends on the phenomenon whether the ceiling is at, say, 70% or 90%. The values on the *x*-axis stand for the number of features, which may range from a few to several hundred. The dashed line shows a 'typical' curve: (slightly) rising, without a clear-cut elbow that could serve as a criterion to distinguish the 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 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. The solid 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 (i) simply from a (p. 644) large feature vector, in which the power features can be more or (usually) less interpretable; (ii) from a knowledge-based selection from the large vector; and (iii) from additional features that are based on expert knowledge or (iv) maybe even from parameters used in teaching and treatment. In the ideal scenario, the features stemming from (i) and (ii) have much in common with and can be mapped onto features stemming from (iii) and (iv). However, we often need to define possible candidate features that are found in our large vector on the basis of thorough literature research. It can only be hoped that this manually selected small number of features will yield a performance that matches the one obtained with the same number of automatically selected features, yielding a high or at least acceptable performance.

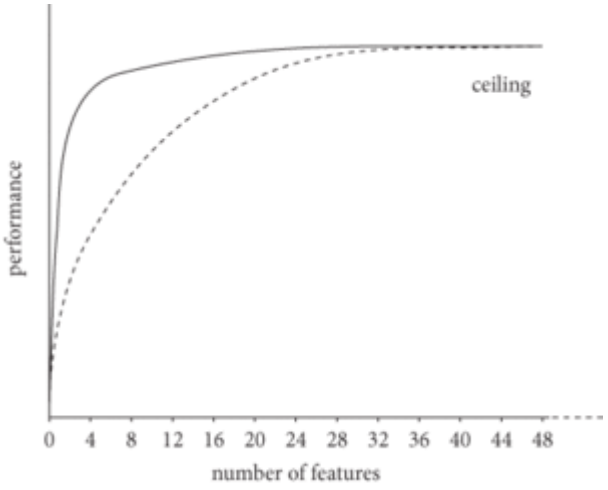

Figure 46.1 Effect of power features on performance: a few power features contribute strongly to performance (continuous line), whereas often there is no clear indication of which features contribute most (dashed line).

In §46.3.1 and §46.3.2, we referred to a few exemplary studies where these strategies have been applied. Such studies are still sparse; interdisciplinary collaboration between

(applied) phonetics and linguistics on the one hand and engineering approaches on the other hand are called for in future research.

46.4 Concluding remarks

A few words of caution and notes about limitations of the overview given in this chapter seem appropriate. We have not addressed in detail the phenomena and algorithmic procedures that have been dealt with in the field of automatic processing of prosody. We have also refrained from presenting exact performance measures across studies, which is often done in survey articles on paralinguistics but is of doubtful value because strict comparability is almost never given. Moreover, we have not given a full account of the history and state of the art. Instead, we have tried to present the most important methodological trends in the period from the 1980s up to the present day. We have seen that in the first phase prosody (p. 645) was very visible, and 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 relationships to linguistic structure, to wit not only in basic research but also in applications. 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 to every problem. History tells us that none of this is very likely to be the case in the long run. Although scientific paradigms are persistent (Kuhn 2012), it is difficult to predict which theories and methods will prevail in the medium-term future. But it is safe to predict that there will be no big, unified approach embracing both linguistic/prosodic theories and ML. We may not see much convergence and collaboration between the two sides or a higher visibility of prosody. Yet, 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, we might eventually develop a better understanding of the intricate relationship between power and leverage features.

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.

Notes:

([1]) Lieberman 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.

([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.

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

([4]) Shriberg (2007) gives an overview of higher-level (including prosodic) features in the field of automatic speaker recognition. Schuller and Batliner (2014: chs. 4, 5) survey studies on CP, again including prosodic ones.

([5]) For WAR, chance level is the frequency in per cent of the most frequent class. UAR reports the mean of the diagonal in a confusion matrix in per cent; chance level is always 50% for two classes, 33.3% for three classes, and so on. UAR was introduced in the VERBMOBIL project as the ‘average of the class-wise recognition rates’ (Batliner et al. 1998: 216), 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 Interspeech Computational Paralinguistics Challenge since 2009 (Schuller et al. 2009; Rosenberg 2012a).

([6]) Phonetic convergence and social signals are complex (‘bags of’) phenomena and related to each other: when speakers converge, this can be seen as a social signal, indicated by one or more of the parameters listed in Table 46.2.

([7]) More challenges for the integration of prosody into speech technologies are discussed in Rosenberg (2018).

([8]) In crowdsourcing, annotation is done by a large, paid group of internet users; in transfer-learning, knowledge is transferred from one domain to another domain; and in zero-shot learning, no labelled data are needed.

([9]) Phonological, categorical features such as ToBI tones and breaks are, in fact, simply two-step features when used in automatic processing and created by tools such as AuToBI (Rosenberg 2009): 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. The first step reduces variability, which is unfavourable for ML modelling (cf. Parada-Cabaleiro et al. 2019).

([10]) 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 feature were eventually modelled explicitly together with pitch was it 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 1999b).

([11]) Wrappers are computationally costly because a model is tested for each subset of features, but they normally yield highly competitive performance. Other methods are, for example, based on correlation or information gain (Schuller and Batliner 2014: 235–238).

([12]) 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 of non-native or pathological speech.

([13]) 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 a detailed comparison of the feature vectors and classifiers employed.

([14]) Note that the ‘curse of dimensionality’, i.e. the problem of employing too many features in conditions of data sparsity (only a few cases), is not relevant if classifiers such as SVMs or random forests (RFs) are used: SVMs are robust regarding this problem and RFs circumvent it by fusing many decision trees, with each of them having only a small number of features.

([15]) On the conversation level, speaker overlap can be seen as ‘negative pause’ and thus as a genuine prosodic phenomenon.

Anton Batliner

Anton Batliner is Senior Research Fellow affiliated with the chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg. He obtained his PhD at LMU Munich in 1978. He has published widely on prosody and paralinguistics and coauthored Computational Paralinguistics (Wiley, 2014, with Björn Schuller), besides being an active editor and conference organizer. His earlier affiliations were with the Pattern Recognition Lab at the University of Erlangen-Nuremberg and the institutes for Nordic Languages and German Philology (both LMU Munich).

Bernd Möbius

Bernd Möbius is Professor of Phonetics and Phonology at Saarland University and was editor in chief of Speech Communication (2013–2018). He was a board member of the International Speech Communication Association (ISCA) from 2007 to 2015, a founding member and chair (2002–2005) of ISCA’s special interest group on speech synthesis, and has served on ISCA’s Advisory and Technical committees. A central theme of his research concerns the integration of phonetic knowledge in speech

technology. Recent work has focused on experimental methods and computational simulations to study aspects of speech production, perception, and acquisition.