Prosody and Spoken-Word Recognition

James M. McQueen and Laura Dilley

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/9780198832232.013.33

Abstract and Keywords

This chapter outlines a Bayesian model of spoken-word recognition and reviews how prosody is part of that model. The review focuses on the information that assists the listener in recognizing the prosodic structure of an utterance and on how spoken-word recognition is also constrained by prior knowledge about prosodic structure. Recognition is argued to be a process of perceptual inference that ensures that listening is robust to variability in the speech signal. In essence, the listener makes inferences about the segmental content of each utterance, about its prosodic structure (simultaneously at different levels in the prosodic hierarchy), and about the words it contains, and uses these inferences to form an utterance interpretation. Four characteristics of the proposed prosody-enriched recognition model are discussed: parallel uptake of different information types, high contextual dependency, adaptive processing, and phonological abstraction. The next steps that should be taken to develop the model are also discussed.

Keywords: Bayesian model, spoken-word recognition, prosodic structure, perceptual inference, prosodic hierarchy, parallel uptake, high contextual dependency, adaptive processing, phonological abstraction

36.1 Introduction

EACH spoken utterance is potentially unique and is one of an infinite range of possible utterances. However, each is made from words that usually have been heard before, sampled from the finite set of words the speaker/listener knows. To understand the speaker's intended message in any utterance, therefore, the listener must recognize the utterance's words. We argue here that listeners achieve spoken-word recognition through Bayesian perceptual inference. Their task, over and over again for each word, is to infer the identity of the current word and build an interpretation, integrating current acoustic information with prior knowledge. In this chapter, we consider the role of 'prosody' in this process of perceptual recovery of spoken words.

36.2 Defining prosody in spoken-word recognition

We begin with a definition of ‘prosody’. This is not only because it can mean different things to different people, but also because one of our goals is to highlight the utility of an abstract definition of prosody that has to do with structures built in the mind of the perceiver. Critically, this definition is tied to the cognition in question: the process of spoken-word recognition. Our definition therefore does not start from linguistic material (words, sentences) or from the acoustic properties of speech (e.g. spectral and durational features) but instead from a psychological perspective, focusing on the representations and processes listeners use as they understand speech.

The basis of our definition is that, during word recognition, two types of structure are built in the listener’s mind. The former structures are ‘segmental’ in that they are based on abstractions about segments—the traditional combinatorial ‘building blocks’ of words. The latter structures are ‘suprasegmental’ and relate to abstractions about the prominence, (p. 510) accentuation, grouping, expressive tone of voice, and so on of syllables relative to each other and also of words relative to each other. The latter structures are prosodic, and hence to understand the role of prosody in word recognition is to have an adequate account of how these structures are built, but also how the segmental structures are built, and how these two types of structure jointly support speech understanding.

This definition thus highlights the interdependency, during processing, of signal characteristics often classified as ‘segmental’ and ‘suprasegmental’. For example, pitch characteristics (i.e. perceptual indices of fundamental frequency variations)—often considered to be suprasegmental in the spoken-word recognition literature—may frequently contribute simultaneously to extracting both segmental and suprasegmental structures, as well as other kinds of structure (e.g. syntactic). In the same vein, acoustic characteristics relating to distributions of periodic (i.e. vocal fold vibration) or aperiodic energy—often considered to be segmental in the spoken-word recognition literature—contribute to extracting both segmental structures (e.g. words) and suprasegmental structures (e.g. prosodic phrase-level structures through domain-initial strengthening of segments, see later in this section), as well as other kinds of structure (e.g. syntactic). Again, this happens in an interdependent fashion across levels of structure. That such interdependences among different levels of structure exist in spoken-word recognition is consistent with the observation that lexical entries are defined in part by the constructs of ‘syllable’ and ‘stress’—each of which has both a ‘segmental’ and a ‘suprasegmental’ interpretation. That a given acoustic attribute (e.g. fundamental frequency in speech, which gives rise to a harmonic spectrum) contributes simultaneously to perception of both segmental and suprasegmental structures has long been recognized (e.g. Lehiste 1970).

Consideration of this interdependence across different levels of the linguistic hierarchy during structure extraction is also motivated by our perspective on speech recognition. In

our view, a core challenge to be explained is how words are extracted from the speech stream in spite of considerable variability. That is, a spoken-word recognizer needs to be robust in the face of acoustic variability of various kinds—for example, differences between phonological contexts, speakers, speaking styles, and listening conditions. We argue that redundancy in encoding multi-levelled tiers of structure across different kinds of acoustic information means that the system is more robust to any one kind of acoustic degradation. That is, listeners build interlocking segmental and suprasegmental phonological structures as a means to solving the variability problem.

We believe that our cognitive definition of prosody allows us to avoid several problems. In particular, we do not need to define particular types of acoustic cue as strictly either 'segmental' or 'suprasegmental'. Such attempts come with the implication that whatever phonetic properties are taken to define 'suprasegmental'—usually timing and pitch—are via logical opposition 'not segmental', and thus that these do not cue segmental contrasts. Indeed, such a view is highly problematic, as has been noted by many researchers (e.g. Lehiste 1970). Much work has documented the role of timing in the cueing of segmental contrasts, including both consonants (Lisker and Abramson 1964; Liberman et al. 1967; Wade and Holt 2005) and vowels (cf. vowel length or tenseness; Ainsworth 1972; Miller 1981).

Under our proposal, acoustic information can nevertheless still be categorized as that which assists the listener in recognizing either the segments of an utterance ('segmental information') or its prosodic structure ('suprasegmental information'). Our definition is in (p. 511) service of the view that spoken-word recognition involves simultaneously recognizing the words being said, the prosodic (e.g. grouping, prominence) structures associated with those words, and the larger structures (e.g. syntactic) in which the words are embedded. On this view, it becomes easier to see how diverse acoustic cues—ranging from pitch to timing to allophonic phonetic variation—could be employed to help extract structure (lexical and otherwise) at various hierarchical levels.

The same acoustic information can therefore help the listener to simultaneously identify segmental and prosodic structures. Take the case of domain-initial strengthening, in which acoustic cues for consonants and vowels tend to be strengthened (e.g. become longer or louder, or add glottal stops or other fortification) at the beginnings of prosodic domains (Dilley et al. 1996; Fougeron and Keating 1997; Turk and Shattuck-Hufnagel 2000; Cho and Keating 2001; Tabain 2003; Krivokapić and Byrd 2012; Beňuš and Šimko 2014; Garellek 2014; Cho 2016). Domain-initial strengthening affects pitch, timing, and spectral details, but also concerns systematic variation at the lexical level, such that it can help with lexical disambiguation (Cho et al. 2007) and at the utterance level (such that it helps the listener with sentential parsing and interpretation building). That is, domain-initial strengthening concerns variation simultaneously at (at least) two levels of structure.

Domain-initial strengthening is an example of cross-talk between segmental and suprasegmental domains. Another example relates to the widespread usage of pitch in

the world's languages to convey lexical contrast. Not only is pitch used throughout the lexicon to convey lexical contrasts in lexical tone languages (e.g. Mandarin, Thai, Igbo) but pitch also plays a role in distinguishing words in languages such as Japanese and Swedish (Bruce 1977; Beckman 1986; Heldner and Strangert 2001). Even intonation languages (e.g. English, Spanish, German, and Dutch) include lexical contrasts based on stress (e.g. *IMpact* (N) vs. *imPACT* (V)) that may be signalled by a difference in pitch in many structural and communicative contexts, but certainly not all (Fry 1958; see also chapter 5). Indeed, the acoustic cues that signal lexical stress contrasts are many and varied and include not only segmental vowel quality differences but also differences in timing, amplitude, and/or spectral balance as well as pitch (Beckman and Edwards 1994; Sluijter and van Heuven 1996a; Turk and White 1999; Mattys 2000; Morrill 2012; Banzina et al. 2016).

Our definition also highlights how prosody can assist in the perceptual recovery of spoken words when the speech signal is degraded. For example, fine spectral details in signals usually associated with segmental information can be replaced with a few frequency bands of noise, producing noise-vocoded speech, or the dynamic formants can be replaced with sine waves, producing sinewave speech. Such degraded speech is often highly intelligible, especially with practice (Shannon et al. 1995; Dorman et al. 1997; Davis et al. 2005). Such perceptual recovery of spoken words is possible partly because listeners are able to make contact with their prior experiences of timing and frequency properties of spoken words experienced over their lifetimes. That is, this ability indicates that stored knowledge about word forms may include timing, pitch, and amplitude information.

A critical feature of our fundamentally cognitive definition is thus that it refers not only to relevant acoustic information but also to relevant prior knowledge. To explore prosody in spoken-word recognition is thus to ask how suprasegmental information and prior knowledge about prosodic structures, together with segmental information and prior knowledge about segments, jointly support speech comprehension. We propose that the answer to this question is that speech recognition involves Bayesian inference.

(p. 512) 36.3 The bayesian prosody recognizer: robustness under variability

A growing body of evidence supports a Bayesian account of spoken-word recognition in which simultaneous multiple interdependent hypotheses are considered about the words being said, their component segments, and aspects of expressiveness that are heard to accompany those words. According to this view, the linguistic structures that are perceived are those that ultimately best explain experienced sensory information. Our proposal is that a Bayesian Prosody Recognizer (BPR) supports this inferential process by extracting prosodic structures (syllables, phrases) and words while deriving utterance interpretations. The BPR draws inspiration from other Bayesian models of speech recognition and understanding and analysis-by-synthesis approaches (Halle and Stevens 1962; Norris and McQueen 2008; Poeppel et al. 2008; Gibson et al. 2013; Kleinschmidt and Jaeger

2015) that envision the inferential, predictive process of spoken language understanding as involving simultaneous determination of multiple levels of linguistic structures, including hierarchical prosodic structures. In essence, as guaranteed by Bayes's theorem, the listener combines prior knowledge with signal-driven likelihoods to obtain an optimal interpretation of current input. The BPR also draws inspiration from previous accounts arguing that speech recognition requires parallel evaluation of segmental and suprasegmental interpretations (in particular the Prosody Analyzer of Cho et al. 2007). Evidence for predictive and inferential processes in speech recognition is reviewed in multiple sources (Pickering and Garrod 2013; Tavano and Scharinger 2015; Kuperberg and Jaeger 2016; Norris et al. 2016).

A central motivation for the BPR is the variability problem, as already introduced: structure extraction needs to be robust in spite of variability in speech. Bayesian inference is a response to this challenge because it ensures optimal interpretation of the current input. The BPR instantiates four key characteristics about prosodic processing in spoken-word recognition. All are further specifications of how the BPR offers ways to ensure robustness of recognition under acoustic variability.

36.3.1 Parallel uptake of information

As we review in the following subsections, considerable evidence from studies examining the temporal dynamics of the recognition process supports our contention that timing and pitch characteristics constrain word identification, and that they do so at the same time as segmental information. In our view, parallel uptake of information has at least two important consequences. First, it makes it possible for structures to be extracted at different representational levels simultaneously. This can readily be instantiated in the BPR. Just like there can be, in a Bayesian framework, a hierarchy of segments (Kleinschmidt and Jaeger 2015), words (Norris and McQueen 2008), and sentences (Gibson et al. 2013), there can also be a Bayesian prosodic hierarchy, potentially from syllables up to intonational phrases. Second, it means that the same acoustic information can contribute simultaneously to the construction of different levels of linguistic representation, including the prosodic, phonological, lexical, and higher (syntactic, semantic, pragmatic) levels. In order to accomplish the above, (p. 513) the BPR must analyse information across windows of varying sizes simultaneously (some quite long, such as recognizing a tune or determining turn-taking structures in discourse). As an example of both of the above, consider that as durational information for a prosodic word (i.e. a single lexical item) accumulates, it can also provide the basis of evidence for a phrase that contains that word. Evidence about that word influences the interpretation of syntactic information, and so forth. Suprasegmental information (as acoustically defined) has been shown to influence recognition in at least four different ways.

36.3.1.1 Influences on processing segmental information

Segments belonging to stressed syllables in sentences are processed more quickly than those belonging to unstressed syllables (Shields et al. 1974; Cutler and Foss 1977). Segmental content in stressed syllables is more accurately perceived than that in unstressed

syllables (Bond and Garnes 1980), and mispronounced segments are more easily detected in stressed syllables than in unstressed syllables (Cole and Jakimik 1978). Distortion of normal word stress information also impairs word processing and recognition (Bond and Small 1983; Cutler and Clifton 1984; Slowiaczek 1990, 1991). Recent findings indicate that categorization of speech segments is modulated by the type of prosodic boundary preceding those segments (Kim and Cho 2013; Mitterer et al. 2016). All of the above evidence supports the view that suprasegmental and segmental sources of acoustic information in words are the basis of parallel inference processes at multiple levels of linguistic structure. In keeping with this view, it has been shown that the same information (durational cues; Tagliapietra and McQueen 2010) can simultaneously help listeners to determine which segments they are hearing and the locations of word boundaries.

36.3.1.2 Influences on lexical segmentation

Consistent with the BPR, the metrical properties of a given syllable affect the likelihood of listeners inferring the syllable to be word-initial (Cutler and Norris 1988; Cutler et al. 1997). For instance, strong syllables are more likely heard as word-initial in errors in perception (Cutler and Butterfield 1992). There is evidence that listeners use multiple cues (some lexical and some signal-driven, based on segmental and suprasegmental acoustic properties) to segment continuous speech into words (Norris et al. 1997). Suprasegmental cues appear to play a more important role under more difficult listening conditions. Thus, for example, the tendency to assume that strong syllables are word-initial is stronger when stimuli are presented in background noise than when there is no noise (Mattys 2004; Mattys et al. 2005).

36.3.1.3 Influences on lexical selection

Suprasegmental pronunciation modifications modulate which words listeners consider and which words they eventually recognize. For example, subtle differences in segment durations or whole syllables can help them to determine the location of syllable boundaries (Tabossi et al. 2000), word boundaries (Gow and Gordon 1995), and prosodic boundaries (e.g. in making the distinction between a monosyllabic word such as *cap* and the initial syllable of a longer word such as *captain*; Davis et al. 2002; Salverda et al. 2003; Blazej and Cohen-Goldberg 2015). Additional kinds of suprasegmental acoustic-phonetic information, including pitch and intensity, also modulate perception of syllable boundaries (Hillenbrand and Houde 1996; Heffner et al. 2013; Garellek 2014). The rapidity with which (p. 514) these kinds of lexical disambiguation take place (e.g. as measured with eye tracking; Salverda et al. 2003) indicates that suprasegmental processing is not delayed relative to segmental processing. Variation in pronunciation associated with distinct positions of words in prosodic phrases (e.g. whether the two words in the phrase ‘bus tickets’ span an intonational phrase boundary or not) has also been shown to modulate lexical selection (Christophe et al. 2004; Cho et al. 2007; see also Tremblay et al. 2016, 2018 for similar non-native language effects).

Some earlier studies (Cutler and Clifton 1984; Cutler 1986) suggested that stress differences cued by suprasegmental information (e.g. the distinction between the ‘ancestor’

and ‘tolerate’ senses of ‘forbear’, which is not due to a difference in the segments of the words; Cutler 1986) did not constrain lexical access substantially. Subsequent experiments, however, have indicated that stress does constrain lexical access, albeit to different extents in different languages, as a function of the informational value of suprasegmental stress cues in the language in question (Cutler and van Donselaar 2001; Soto-Faraco et al. 2001; Cooper et al. 2002). For example, the influence of suprasegmental stress cues on word recognition is stronger in Dutch, where such cues tell listeners more about which words have been spoken, than in English, where segmental differences are more informative (Cooper et al. 2002). Eye-tracking studies indicate that suprasegmental cues to stress are taken up without delay and can thus support lexical disambiguation before any segmental cues could disambiguate the input (Reinisch et al. 2010; Brown et al. 2015). Relatedly, work on word recognition in tone languages has shown how pitch characteristics of the input constrain word identification in parallel with the uptake of segmental information (Lee 2009; Sjerps et al. 2018).

36.3.1.4 Influences on inferences about other structures

Consistent with the BPR, there is considerable evidence that suprasegmental information influences the listener’s inferences about various levels of structure beyond the word level, simultaneously, in real time. The focus of this chapter is on spoken-word recognition, but since perception of lexical forms influences higher levels of linguistic structure and inference, it is important to note that there is evidence that prosody and other higher levels of linguistic knowledge are extracted in parallel. That is, perception of prosodic information and perception of syntactic structure are interdependent (Carlson et al. 2001; Buxó-Lugo and Watson 2016) and prosody influences semantic and pragmatic inference (Ito and Speer 2008; Rohde and Kurumada 2018).

36.3.2 High contextual dependency

Another characteristic of prosodic processing in spoken-word recognition is its high contextual dependency. That is, the interpretation of the current prosodic event depends on the context that occurs before and/or after that event. Context can be imagined as a timeline, where ‘left context’ temporally precedes an event and ‘right context’ follows it.

36.3.2.1 Left-context effects

Under the BPR account, regularities in context that are statistically predictive of properties of upcoming words will be used to infer lexical properties of upcoming words, giving rise (p. 515) to left-context effects. It is well attested that suprasegmental aspects of sentential context affect the speed of processing of elements. For example, suprasegmental cues in a sequence of words preceding a given word affect processing speed on that word (Cutler 1976; Pitt and Samuel 1990) and accuracy of word identification (Slowiaczek 1991). The rhythm of stressed and unstressed syllables is an important cue for word segmentation in continuous speech (Nakatani and Schaffer 1978). Further, a metrically regular speech context has also been shown to promote spoken-word recognition (Quené and Port 2005). Our BPR proposal accounts for these findings in terms of statistical inference

on the basis of regularities in the speech signal. Structures in utterances formed by prosodic (e.g. rhythmic) patterning in production engender predictability of structure and timing in perception of upcoming sentential elements (Jones 1976; Martin 1972) at multiple hierarchical levels and points (Liberman and Prince 1977). Statistical regularities in stress alternation and timing are attested in speech production experiments, corpus studies, and theoretical linguistics (Selkirk 1984; Kelly and Bock 1988; Hayes 1995; Farmer et al. 2006). Changes in the priors in a Bayesian model can account easily for the effects of left prosodic context (and other types of preceding context) on recognition of the current word.

Contextual influences of suprasegmental cues on perception of segmental information (e.g. voice onset time) are well known, particularly for timing (Miller and Liberman 1979; Repp 1982; Kidd 1989) but also for pitch (Holt 2006; Dilley and Brown 2007; Dilley 2010; Sjerps et al. 2018). However, such effects have by and large been found to involve fairly proximal speech context within about 300 ms of a target segment (Summerfield 1981; Kidd 1989; Newman and Sawusch 1996; Sawusch and Newman 2000; but see Wade and Holt 2005).

More recent work has shown that suprasegmental information from the more distant ('distal') left context can also influence which words are heard—including how syllables are grouped into words, and even whether certain words (and hence certain phonemes) are heard at all. For example, the rate of distal context speech influences whether listeners hear reduced words such as *are* spoken as 'err' (Dilley and McAuley 2008; Pitt et al. 2016). Statistical distributions of distal contextual speech rates influence listeners' word perception over the course of around one hour (Baese-Berk et al. 2014). Further, the patterns of pitch and timing on prominent and non-prominent syllables in the left context influence where listeners hear word boundaries in lexically ambiguous sequences such as *crisis turnip* vs. *cry sister nip* (Dilley and McAuley 2008; Dilley et al. 2010; Morrill et al. 2014a). These patterns also influence the extent to which listeners hear reduced words or syllables (Morrill et al. 2014b; Baese-Berk et al. 2019). Distal rate and rhythm influence lexical processing early in perception and modulate the extent to which lexically stressed syllables are heard to be word-initial (Brown et al. 2011b, 2015; Breen et al. 2014). Consistent with the BPR, whether a listener hears a word depends in gradient, probabilistic fashion on the joint influence of distal rate cues and proximal information signalling a word boundary (Heffner et al. 2013).

36.3.2.2 Right-context effects

Information that follows can be informative about lexical content that may have already elapsed. A growing body of evidence indicates that listeners often commit to an interpretation of lexical content only *after* the temporal offset of that content (Bard et al. 1988; Connine et al. 1991; Grossberg and Myers 2000; McMurray 2007). In segmental perception, (p. 516) temporal information to the right of a given segment can influence listeners' judgements of segmental perception (e.g. Miller and Liberman 1979). Eye-tracking studies show that later-occurring distal temporal information (e.g. relative duration of a subsequent phoneme sequence that includes the morpheme /s/) can influence whether listen-

ers hear a prior reduced function word (Brown et al. 2014). All of these findings indicate that acoustic information must be held in some kind of memory buffer and hence that perceptual decisions can be delayed until after the acoustic offset of that information. The extent to which listeners hold alternative parses in mind after a given portion of signal consistent with a given word has elapsed, as opposed to abandoning them, is an active area of research and debate (Christiansen and Chater 2016).

While the effects of right context might at first glance appear to be more problematic, they too can be explained in a Bayesian framework. The key notion here is that different hierarchical levels of structure and constituency (e.g. segments, syllables, words, prosodic phrases) entail different time windows over which relevant evidence is collected and applied to generate inferences about representations at that level. This implies that acoustic evidence at a given moment might be taken as highly informative for structure at one level, while simultaneously being taken as only weakly informative (or indeed uninformative) about structure at another level. Depending on the imputed reliability of evidence as it appertains to each level, inferences about structure at different levels may be made at different rates (i.e. are staggered in time). Because evidence bearing on the structure of a larger constituent (e.g. a prosodic phrase) typically will appear in the signal over a longer time span than evidence bearing on the structure of a smaller one (e.g. a syllable), completion of the inferences about the larger constituent may often entail consideration of evidence from some amount of subsequent ‘right-context’ material. This apparent delay with respect to inferences about the structure of the larger constituent does not imply that the BPR does not always attempt to use all information simultaneously or that it does not attempt to draw inferences at different levels simultaneously. Rather, it implies only that in some cases the current information is insufficient for inferences at a given level of structure to be made with confidence, and hence that the BPR may wait for further information in the upcoming context before committing to an interpretation of structure at that level. This view also entails that later-occurring information might provide evidence that an earlier assumption about structure was not well supported and hence the possibility of revision of inferences drawn earlier.

36.3.2.3 Syntagmatic representation of pitch

Phonological interpretation of pitch cues in spoken language comprehension requires consideration of both left and right pitch context (Francis et al. 2006; Sjerps et al. 2018). Left and right context is also important in listeners drawing abstractions about the tonal properties of a given syllable, including that which is relevant to perceiving distinct lexical items (Wong and Diehl 2003; Dilley and Brown 2007; Dilley and McAuley 2008). Such findings support a view in which the representation of linguistically relevant pitch information is fundamentally syntagmatic (i.e. relational) and in which paradigmatic aspects of tonal information involve inferences driven by abstract knowledge about a typical speaker’s pitch range in relation to incoming pitch information (Dilley 2005, 2008; Lai 2018; Dilley and Breen, in press). This view is adopted in the BPR.

(p. 517) 36.3.3 Adaptive processing

The perceptual apparatus must dynamically adapt to variation in order to remain robust in understanding intended messages. The available evidence suggests that prosodic processing is indeed very flexible. For instance, listeners adapt rapidly to the rate of compressed speech (Dupoux and Green 1997). The evidence just reviewed on context effects shows that listeners track characteristics of the current speech (e.g. distributional properties of speaking rate variation and the metrical properties of utterances) and flexibly adjust to that context (Dilley and McAuley 2008; Dilley and Pitt 2010; Baese-Berk et al. 2014; Morrill et al. 2015).

Another way in which prosodic processing has been shown to be adaptive is that it involves perceptual learning. It has been established that listeners can adapt to variation in the realization of segments (Norris et al. 2003; Samuel and Kraljic 2009): they tune in, as it were, to the segmental characteristics of the speech of the current talker. It is thus plausible to expect that there are similar adjustments with respect to suprasegmental characteristics. There is indirect evidence that this may be the case. Listeners adapt to the characteristics of accented as well as distorted speech (Bradlow and Bent 2008; Mitterer and McQueen 2009; Borrie et al. 2012; Baese-Berk et al. 2013), which presumably includes adjustments to suprasegmental features. But there is also more direct evidence. Dutch listeners in a perceptual-learning paradigm can adjust the way they interpret the reduced syllables of a particular Dutch speaker (Poellmann et al. 2014), and Mandarin listeners adjust the way they interpret the tonal characteristics of syllables through exposure to stimuli with ambiguous pitch contours in contexts that encourage a particular tonal interpretation (Mitterer et al. 2011).

The BPR therefore needs to be flexible. Detailed computational work on perceptual learning in a Bayesian model with respect to speech segments has already been performed (Kleinschmidt and Jaeger 2015). The argument, in a nutshell, is that learning is required for the listener to be able to recognize speech optimally, in the context of an input that is noisy and highly variable due, for instance, to differences between talkers (Norris et al. 2003; Kleinschmidt and Jaeger 2015). That is, the ideal observer needs to be an ideal adapter. Exactly the same arguments apply to prosodic variability. Learning processes, for example based on changes in the probability density function of a given prosodic constituent for a given idiosyncratic talker, should be instantiated in the BPR in a similar way to those already implemented for segments.

36.3.4 Phonological abstraction

The final characteristic of prosodic processing in spoken-word recognition is that it is based on phonological abstraction. The listener must be able to form abstractions so as to remain optimally robust and capable of handling not-yet-encountered variation. Phonological abstraction is thus also a feature of the BPR. As in the previous Bayesian accounts focusing on segmental recognition (Norris and McQueen 2008; Kleinschmidt and Jaeger 2015), the representations that inferences are drawn about are abstract categories so

that (as the adaptability of the BPR also guarantees) the recognition process is robust to variation due to differences across talkers and listening situations. Evidence suggests that the abstractions (p. 518) about categories entail generalizations about segmental structures and allophonic variation (Mitterer et al. 2018); lexical stress and tone (Sulpizio and McQueen 2012; Ramachers 2018; Sjerps et al. 2018); pitch accent, pitch range, and boundary tone types (Cutler and Otake 1999; Dilley and Brown 2007; Dilley and Heffner 2013); and relationships between phonological elements and other aspects of the linguistic structure of information, such as grammatical categories (Kelly 1992; Farmer et al. 2006; Söderström et al. 2017).

Prosodic processing in speech recognition appears to involve phonological abstraction. One line of evidence for this comes from the learning studies just reviewed. If perceptual learning generalizes to the recognition of words that have not been heard during the exposure phase, then some type of abstraction must have taken place—the listener must know which entities to apply the learning to (cf. McQueen et al. 2006). The studies on learning about syllables (Poellmann et al. 2014) and tones (Mitterer et al. 2011) both show generalization of learning to the recognition of previously unheard words.

Experiments on learning novel words also provide evidence that listeners have abstract knowledge about prosody. In these experiments (on prosodic words in Dutch: Shatzman and McQueen 2006; on lexical stress in Italian: Sulpizio and McQueen 2012), listeners learned new minimal pairs of words, and the new words were then acoustically altered to remove suprasegmental cues that distinguished between the pairs. In the final test phase, the listeners heard the altered (training) words and their unaltered (original) variants. Eye-tracking measures revealed that the listeners had knowledge about the suprasegmental cues that they could apply to the online recognition of the novel words, even though they had never heard those words with those cues (for the Dutch listeners, durational cues distinguishing monosyllabic words from the initial syllables of disyllabic words; for the Italian listeners, durational and amplitude cues to antepenultimate stress in trisyllabic words). These findings suggest that processing of prosody in spoken-word recognition involves not only the uptake of fine-grained acoustic-phonetic cues to prosodic structure but also the storage of abstract knowledge about those cues. That is, while the fine phonetic details about the prosody in the current utterance are key determinants of word recognition and speech comprehension, the listener abstracts over those details in order to be able to understand future utterances.

Speakers also form phonological abstractions based on long-term knowledge of phonetic properties of talker attributes, such as gender (Johnson et al. 1999; Lai 2018), that contribute to Bayesian inferences about spoken words and other aspects of linguistic meaning. Phonological abstractions are also formed based on simultaneous or sequential statistical correspondences between phonetic properties, such as pitch and non-modal voice quality, which are phonetic properties that co-vary in many lexical tone languages (Gordon and Ladefoged 2001; Gerfen and Baker 2005; Garellek and Keating 2011; Garellek et al. 2013). Such phonological abstraction—formed from long-term statistical knowledge of correspondences—is essential for drawing correct inferences based on otherwise highly

ambiguous suprasegmental cues (including those for pitch and duration) about, for example, intended words, meaning, and structure (Gerfen and Baker 2005; Bishop and Keating 2012; Lai 2018). For instance, knowledge about co-occurrences of pitch and spectral (e.g. formant frequency) information for male versus female voices can be used to infer a typical or mean pitch of a talker's voice and/or pitch span, from which Bayesian inferences can be drawn about phonological structures (such as those for pitch accents and lexical tones) and associated meanings (Dilley 2005; Dilley and Breen, in press). The BPR assumes that such long-term (p. 519) abstracted statistical knowledge about talkers and the simultaneous and sequential distributional properties of the phonetic cues they produce is, along with talker-independent abstract phonological knowledge, the basis of the Bayesian probabilistic inferences that enable optimal decoding of spoken signals.

36.4 Conclusions and future directions

We have argued that spoken-word recognition is robust under speech variability because it is based on Bayesian perceptual inference and that a vital component of this process is the BPR. As a spoken utterance unfolds over time, the BPR, based on prior knowledge about correspondences between acoustic variables, on the one hand, and meanings and structures, on the other, makes Bayesian inferences about the prosodic organization, lexical content, and semantic and pragmatic information in the utterance, among other inferences. These inferences are both signal and knowledge driven, and concern abstract structures at different levels in the prosodic hierarchy that are computed in parallel, informed by statistical distributions of relationships between acoustic cues often considered segmental or suprasegmental. Inferences about a given stretch of input are influenced by earlier input and by inferences about it, and can be revised based on later input. Importantly, the BPR adapts to current input to optimize its inferences.

We have suggested that the goal of the BPR is to derive the metrical and grouping structures in each utterance at different levels in the prosodic hierarchy. Especially for utterance-level inferences, the representation must include a sparse set of tones, including pitch accents, boundary tones, and/or lexical tones, which are autosegmentally associated with particular positions in metrical and grouping structures indexed to the lexicon (Gussenhoven 2004; Ladd 2008b; Dilley and Breen, in press). Establishing how listeners recover this prosodic hierarchy, and the number of levels that need to be built, are important challenges for future research.

The BPR will need to be implemented as part of a full Bayesian model of speech recognition, which includes, but is not limited to, prosodic inferences. Our view is that segmental and suprasegmental structures are built in parallel, based on information that may inform inferences about either or both types of structure. Over time, inferences about prosodic structure feed into (and are in turn influenced by) inferences made about segments and words of the unfolding utterance and its current interpretation. The model will need to specify how interacting processes determine spoken-word recognition and how infer-

ences drawn about the speech signal change over time. It will also need to be tested, through simulations and experimentation.

One way to evaluate and develop the BPR would be to compare it to other models on the role of prosody in spoken-word recognition. Unfortunately, no such alternative models currently exist. Shuai and Malins (2017) have recently proposed TRACE-T, an implementation of TRACE (McClelland and Elman 1986) that seeks to account for the processing of tonal information in Mandarin monosyllabic words. While this is a very welcome addition to the literature, TRACE-T is much more limited in scope than the BPR. Comparisons could potentially also be made to the Prosody Analyzer (Cho et al. 2007; but the BPR can be seen as a development of that model) and to Shortlist B (Norris and McQueen 2008; but Shortlist B (p. 520) is limited, with respect to prosody, to the role of metrical structure in lexical segmentation, and again the BPR is largely inspired by the earlier model). Detailed comparisons of the BPR to other models (e.g. Kurumada et al. 2018) will have to wait for the implementation of the BPR and for the development of competitor models of equivalent scope.

Another important aspect of future work will be cross-linguistic comparison. Most work on prosody in spoken-word recognition has been done on English or a small set of related European languages. There are some exceptions to this Eurocentric bias (Cutler and Otake 1999; Ye and Connine 1999; Lee 2007), and there has been an upsurge of work on, for example, pitch cues in conveying lexical and other meanings in typologically diverse languages (Kula and Braun 2015; Ramachers 2018; Sjerps et al. 2018; Wang et al. 2018; Yamamoto and Haryu 2018; Genzel and Kügler, in press). Much research nevertheless still is needed to explore how the full set of prosodic phenomena in the world's languages modulates the recognition process. We do not expect that experiments on non-European languages will lead to falsification of the Bayesian model. For example, pitch conveys different kinds of structure simultaneously in a given language, and is used to convey lexical information to different degrees in different languages. Pitch is simply less informative about lexical structure in a Bayesian statistical sense in intonation languages than in tone languages and thus will be relied on less in discriminating between and recognizing words in intonation languages. While such cross-linguistic differences can thus readily be captured in a Bayesian model, it will be important to explore how pitch information can simultaneously inform inferences about words and inferences about intonational structures in a tone language, and how this weighting changes in intonation versus lexical tone languages.

The Bayesian model will need to be developed in the direction of neurobiological implementation. As in psycholinguistic research (including computational modelling), much work in cognitive neuroscience focuses on how segments (e.g. individual consonants or vowels) are recognized, and how that contributes to word recognition. Prosody had tended to be ignored. There are some interesting new approaches—for example, evidence of neural entrainment to the 4 Hz oscillations at which speech tends to be spoken (i.e. the 'syllable rate') (Giraud and Poeppel 2012; Ding et al. 2017). Nevertheless, much work still needs to be done to specify the brain mechanisms that support spoken-word recognition

as a process that depends on parallel inferences about segmental content and prosodic structures (e.g. whether entrainment is modulated by information arriving in the speech signal at faster or slower rates than 4 Hz).

It will also be necessary to specify how the proposed model relates to other aspects of language processing, speech production in particular. Knowledge that is needed to support recognition (e.g. the acoustic characteristics of words with penultimate stress) may not be relevant in speech production. It remains to be determined whether and to what extent the processes and representations involved in recognition are shared with those involved in production. It is already clear, however, that there is an intimate relationship between input and output operations. For example, the Bayesian recognition process depends on the ability of the recognizer to track production statistics. There are undoubtedly constraints on which statistics are tracked (e.g. with respect to the size of the structures that are tracked), but future work will need to establish what those constraints are and why certain statistics are tracked and not others.

There is also a need to evaluate the model not only relative to other domains of cognitive psychology (such as speech production, language acquisition, and second language (p. 521) processing) but also relative to other domains of linguistics. The representations of prosodic structure that a listener needs for efficient speech recognition may or may not have a one-to-one correspondence with those that are most relevant (for example) to language typology. It is theoretically possible, for example, that a structure such as the prosodic word may have an essential role in typological work and yet have no role in processes relating to the cognitive construction of prosodic structures during spoken-word recognition. It is another important challenge for future research to establish the extent to which representations of prosody indeed vary across different domains of linguistic enquiry.

We have here reviewed the state of the art of research on prosody in spoken-word recognition. Rather than being theoretically neutral, we have advocated a specific model. We look forward to future research testing our central claim that prosody influences speech recognition through Bayesian perceptual inference.

James M. McQueen

James M. McQueen is Professor of Speech and Learning at Radboud University. He studied experimental psychology at the University of Oxford and obtained his PhD from the University of Cambridge. He is a principal investigator at the Donders Institute for Brain, Cognition and Behaviour (Centre for Cognition) and is an affiliated researcher at the Max Planck Institute for Psycholinguistics. His research focuses on learning and processing in spoken language: How do listeners learn the sounds and words of their native and non-native languages, and how do they recognize them? His research on speech learning concerns initial acquisition processes and ongoing processes of perceptual adaptation. His research on speech processing addresses core computational problems (such as the variability and segmentation problems).

He has a multi-disciplinary perspective on psycholinguistics, combining insights from cognitive psychology, phonetics, linguistics, and neuroscience.

Laura Dilley

Laura Dilley is Associate Professor in the Department of Communicative Sciences and Disorders at Michigan State University. She received her BS in brain and cognitive sciences with a minor in linguistics in 1997 from MIT and obtained her PhD in the Harvard–MIT Program in Speech and Hearing Biosciences and Technology in 2005. She is the author of over 60 publications on prosody, word recognition, and other topics.