34. Prosody and spoken-word recognition

James M. McQueen
Radboud University & Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

&

Laura C. Dilley
Michigan State University, East Lansing, USA

Abstract

This chapter outlines a Bayesian model of spoken-word recognition and reviews how prosody is part of that model. The review focusses on the information which 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 which 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

Spoken-word recognition; suprasegmental information; prosodic structure; prosody hierarchy; phonological abstraction; perceptual learning; variability problem; Bayesian model

34.1 Introduction

Each spoken utterance is potentially unique and is one of an infinite range of possible utterances. But 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 she, the listener, achieves spoken-word recognition through Bayesian perceptual inference. Her 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.

34.2 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, focussing on the representations and processes the listener uses as she understands 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, accentuation, grouping, expressive tone of voice, etc., 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 structures (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 below), as well as other kinds of structures (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 among phonological contexts, speakers, speaking styles and listening conditions. We argue that redundancy in encoding multi-leveled 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 cueing of segmental contrasts, including both consonants (Liberman, Cooper, Shankweiler & Studdert-Kennedy 1967; Lisker & Abramson 1964; Wade & 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 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 ones) 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 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 (Beňuš & Šimko 2014; Cho 2016; Cho & Keating 2001; Dilley, Shattuck-Hufnagel & Ostendorf 1996; Fougeron & Keating 1997; Garellek 2014; Krivokapić & Byrd 2012; Tabain 2003; Turk & Shattuck-Hufnagel 2000). 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, McQueen & Cox 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 (Beckman 1986; Bruce 1977; Heldner & Strangert 2001). Even intonation languages (e.g., English, Spanish, German, and Dutch) include lexical contrasts based on stress (e.g., *IMpact* (noun)

vs. *imPACT* (verb)) which may be signalled by a difference in pitch in many structural and communicative contexts, but certainly not all (Fry 1958; Gussenhoven 2004). 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 (Banzina, Dilley & Hewitt 2016; Beckman & Edwards 1994; Mattys 2000; Morrill 2012; Sluijter & van Heuven 1996; Turk & White 1999).

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 (Davis, Johnsrude, Hervais-Adelman, Taylor & McGettigan 2005; Dorman, Loizou & Rainey 1997; Shannon, Zeng, Kamath, Wygonski & Ekelid 1995). Such perceptual recovery of spoken words is possible partly because the listener is able to make contact with her prior experiences of timing and frequency properties of spoken words experienced over her lifetime. 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.

34.3 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 which are perceived are those that ultimately best explain experienced sensory information. Our proposal is that a Bayesian Prosody Recogniser (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 (Gibson, Bergen & Piantadosi 2013; Halle & Stevens 1962; Kleinschmidt & Jaeger 2015; Norris & McQueen 2008; Poeppel, Idsardi & van Wassenhove 2008) which 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' rule, 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 Analyser of Cho et al., 2007). Evidence for predictive and inferential processes in speech recognition is reviewed in multiple sources (Kuperberg & Jaeger 2016; Norris, McQueen & Cutler 2016; Pickering & Garrod 2013; Tavano & Scharinger 2015).

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.

34.3.1 Parallel uptake of information

As we review below, 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 that structures can 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 & Jaeger 2015), words (Norris & 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 construction of different levels of linguistic representation, including the prosodic, phonological, lexical, and higher (syntactic, semantic, pragmatic) levels. In order to accomplish the above, 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 which 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.

34.3.1.1 Influences on processing segmental information

Segments belonging to stressed syllables in sentences are processed more quickly than those belonging to unstressed syllables (Cutler & Foss 1977; Shields, McHugh & Martin 1974). Segmental content in stressed syllables is more accurately perceived than that in unstressed syllables (Bond & Garnes 1980), and mispronounced segments are more easily detected in stressed syllables than in unstressed syllables (Cole & Jakimik 1978). Distortion of normal word stress information also impairs word processing and recognition (Bond & Small 1983; Cutler & Clifton 1984; Slowiaczek 1990, 1991). Recent findings indicate that categorisation of speech segments is modulated by the type of prosodic boundary preceding those segments (Kim & Cho 2013; Mitterer, Cho & Kim 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 & McQueen 2010) can simultaneously help listeners determine which segments they are hearing and the location of word boundaries.

34.3.1.2 Influences on lexical segmentation

Consistent with the BPR, the metrical properties of a given syllable affect the likelihood with which listeners infer the syllable to be word-initial (Cutler, Dahan & van Donselaar 1997; Cutler & Norris 1988). For instance, strong syllables are more likely heard as word-initial in errors in perception (Cutler & 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, McQueen, Cutler & Butterfield 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, White & Melhorn 2005).

34.3.1.3 Influences on lexical selection

Suprasegmental pronunciation modifications modulate which words the listener considers and which words she eventually recognizes. For example, subtle differences in segment durations or whole syllables can help her determine the location of syllable boundaries (Tabossi, Collina, Mazzetti & Zoppello 2000), word boundaries (Gow & 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* (Blazej & Cohen-Goldberg 2015; Davis, Marslen-Wilson & Gaskell 2002; Salverda, Dahan & McQueen 2003). Additional kinds of suprasegmental acoustic-phonetic information, including pitch and intensity, also modulate perception of syllable boundaries (Garellek 2014; Heffner, Dilley, McAuley & Pitt 2013; Hillenbrand & Houde 1996). The rapidity with which these kinds of lexical disambiguation take place (as measured, e.g., 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 intonation phrase boundary or not) has also been shown to modulate lexical selection (Cho et al. 2007; Christophe, Peperkamp, Pallier, Block & Mehler 2004; see also Tremblay, Broersma & Coughlin 2018; Tremblay, Broersma, Coughlin & Choi 2016 for similar non-native language effects).

Some earlier studies (Cutler 1986; Cutler & Clifton 1984) 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, 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 (Cooper, Cutler & Wales 2002; Cutler & van Donselaar 2001; Soto-Faraco, Sebastian-Galles & Cutler 2001). 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 (Brown, Salverda, Dilley & Tanenhaus 2015; Reinisch, Jesse & McQueen 2010). 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, Zhang & Peng 2018).

34.3.1.4 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 of syntactic structure are interdependent (Buxó-Lugo & Watson 2016; Carlson, Clifton & Frazier 2001) and prosody influences semantic and pragmatic inference (Ito & Speer 2008; Rohde & Kurumada 2018).

34.3.2 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 which 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.

34.3.2.1 Left-context effects

Under the BPR account, regularities in context which are statistically predictive of properties of upcoming words will be used to infer lexical properties of upcoming words, giving rise 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 & 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 & Schaffer 1978). Further, a metrically regular speech context has also been shown to promote spoken-word recognition (Quené & 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 engenders predictability of structure and timing of upcoming sentential elements (Jones 1976; Martin 1972) at multiple, hierarchical levels and points (Liberman & Prince 1977). Statistical regularities in stress alternation and timing are attested in speech production experiments, corpus studies and theoretical linguistics (Farmer, Christiansen & Monaghan 2006; Hayes 1995; Kelly & Bock 1988; Selkirk 1984). 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., VOT) are well known, particularly for timing (Kidd 1989; Miller & Liberman 1979; Repp 1982) but also for pitch (Dilley 2010; Dilley & Brown 2007; Holt 2006; 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 (Kidd 1989; Newman & Sawusch 1996; Sawusch & Newman 2000; Summerfield 1981; but see Wade & 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 & McAuley 2008; Pitt, Szostak & Dilley 2016). Statistical distributions of distal contextual speech rates influence listeners' word perception over the course of ~1 hour (Baese-Berk et al. 2014). Further, the patterns of pitch and timing on prominent and nonprominent syllables in the left context influences where listeners hear word boundaries in lexically-ambiguous sequences such as *crisis turnip* vs. *cry sister nip* (Dilley, Mattys & Vinke 2010; Dilley & McAuley 2008; Morrill, Dilley & McAuley

2014). These patterns also influence the extent to which listeners hear reduced words or syllables (Baese-Berk, Dilley, Henry, Vinke & Banzina 2019; Morrill, Dilley, McAuley & Pitt 2014). 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 (Breen, Dilley, McAuley & Sanders 2014; Brown, Salverda, Dilley & Tanenhaus 2011; Brown et al. 2015). Consistent with 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).

34.3.2.2 Right-context effects

Information which follows can be informative about lexical content that may have already elapsed. A growing body of evidence that listeners often commit to an interpretation of lexical content only *after* the temporal offset of that content (Bard, Shillcock & Altmann 1988; Connine, Blasko & Hall 1991; Grossberg & Myers 2000; McMurray 2007). In segmental perception, temporal information to the right of a given segment can influence listeners' judgments of segmental perception (e.g., Miller and Liberman 1979). Eyetracking 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 listeners hear a prior reduced function word (Brown, Dilley & Tanenhaus 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 & Chater 2016).

While 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 generation of 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 nor 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.

34.3.2.3 Syntagmatic representation of pitch

Phonological interpretation of pitch cues in spoken language comprehension requires consideration of both left and right pitch context (Francis, Ciocca, Wong, Leung & Chu 2006; Sjerps et al. 2018). Left and right context is also important for listeners to draw abstractions about the tonal properties of a given syllable, including that which is relevant to perceiving distinct lexical items (Dilley & Brown 2007; Dilley & McAuley 2008; Wong & Diehl 2003). Such findings support a view in which the representation of linguistically-relevant pitch information is fundamentally syntagmatic (i.e., relational), and that 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; Dilley & Breen to appear; Lai 2018). This view is adopted in the BPR.

34.3.3 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 & 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 (Baese-Berk et al. 2014; Dilley & McAuley 2008; Dilley & Pitt 2010; Morrill, Baese-Berk, Heffner & Dilley 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, McQueen & Cutler 2003; Samuel & 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 (Baese-Berk, Bradlow & Wright 2013; Borrie et al. 2012; Bradlow & Bent 2008; Mitterer & McQueen 2009), 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, Bosker, McQueen & Mitterer 2014), and Mandarin listeners adjust the way they interpret the tonal characteristics of syllables through exposure to stimuli with ambiguous pitch contours in contexts which encouraged a particular tonal interpretation (Mitterer, Chen & Zhou 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 & Jaeger 2015). The argument, in a nutshell, is that learning is required for the listener to be able to recognise speech optimally, in the context of an input that is noisy and highly variable due, for instance, to differences among talkers (Kleinschmidt & Jaeger 2015; Norris et al. 2003). 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.

34.3.4 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 focussing on segmental recognition (Kleinschmidt & Jaeger 2015; Norris & McQueen 2008), 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 about categories entail generalizations about segmental structures and allophonic variation (Mitterer, Reinisch & McQueen 2018), lexical stress and tone (Ramachers 2018; Sjerps et al. 2018; Sulpizio and McQueen 2012), pitch accent, pitch range and boundary tone types (Cutler & Otake 1999; Dilley & Brown 2007; Dilley & Heffner 2013), and relationships between phonological elements and other aspects of the linguistic structure of information, such as grammatical categories (Farmer et al. 2006; Kelly 1992; Söderström, Horne, Mannfolk, van Westen & Roll 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, Cutler & Norris 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 & McQueen 2006; , and on lexical stress in Italian, Sulpizio & McQueen 2012) listeners learned new minimal pairs of words; the new words were acoustically altered to remove suprasegmental cues that distinguished between the pairs. In a 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 on-line 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, Strand & D'Imperio 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 among phonetic properties, such as pitch and non-modal voice quality, which are phonetic properties that co-vary in many lexical tone languages (Garellek & Keating 2011; Garellek, Keating, Esposito & Kreiman 2013; Gerfen & Baker 2005; Gordon & Ladefoged 2001). 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 (Bishop & Keating 2012; Gerfen & Baker 2005; Lai 2018). For instance, knowledge about co-occurrences of pitch and spectral (e.g., formant frequency) information for male vs. 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 & Breen to appear). The BPR assumes that such long-term 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 which enable optimal decoding of spoken signals.

34.4 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 and meanings and structures, 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 which are computed in parallel, informed by statistical distributions of relationships among 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 (Dilley & Breen to appear; Gussenhoven 2004; Ladd 2008). 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 inferences 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, and implementation of TRACE (McClelland & Elman 1986) which 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 & McQueen 2008; but Shortlist B 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, Brown & Tanenhaus 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 & Otake 1999; Lee 2007; Ye & Connine 1999), and there has been an upsurge of work on, for example, pitch cues in conveying lexical and other meanings in typologically diverse languages (Genzel & Kügler in press; Kula & Braun 2015; Ramachers 2018; Sjerps et al. 2018; Wang, Xu & Ding 2018; Yamamoto & Haryu 2018). 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 among 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 vs. 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 focusses 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') (Ding et al. 2017; Giraud & Poeppel 2012). Nevertheless, much work still needs to be done to specify the brain mechanisms which 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 if 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 recogniser to track production statistics. There are undoubtedly constraints on which statistics are tracked (e.g. with respect to the size of the structures which 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 processing) but also 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.

References

Ainsworth, W. A. (1972). Duration as a cue in the recognition of synthetic vowels. *Journal of the Acoustical Society of America* 51: 648-651.

Baese-Berk, M., L. Dilley, M. Henry, L. Vinke, and E. Banzina (2019). Not just a function of function words: Distal speech rate affects perception of prosodically weak syllables. *Attention, Perception, & Psychophysics* 81: 571-589.

Baese-Berk, M., C. Heffner, L. Dilley, M. Pitt, T. Morrill, and J. D. McAuley (2014). Long-term temporal tracking of speech rate affects spoken-word recognition. *Psychological Science* 25: 1546-1553.

Baese-Berk, M. M., A. R. Bradlow, and B. A. Wright (2013). Accent-independent adaptation to foreign-accented speech. *Journal of the Acoustical Society of America* 133: 174-180.

Banzina, E., L. Dilley, and L. Hewitt (2016). The role of secondary stressed and unstressed unreduced syllables in word recognition: acoustic and perceptual studies with Russian learners of English. *Journal of Psycholinguistic Research* 45: 813-831.

Bard, E. G., R. C. Shillcock, and G. T. Altmann (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception & Psychophysics* 44: 395-408.

Beckman, M. (1986). *Stress and Non-stress Accent*. Dordrecht: Foris.

Beckman, M.E., and J. Edwards (1994). Articulatory evidence for differentiating stress categories. In P. Keating (Ed.), *Papers in Laboratory Phonology III: Phonological Structure and Phonetic Form* (pp. 7-33). Cambridge University Press, Cambridge.

Beňuš, Š., and J. Šimko (2014). Emergence of prosodic boundary: continuous effects of temporal affordance on inter-gestural timing. *Journal of Phonetics* 44: 110-129.

Bishop, J., and P. A. Keating (2012). Perception of pitch location within a speaker's range: Fundamental frequency, voice quality, and speaker sex. *Journal of the Acoustical Society of America* 132: 1100-1112.

Blazej, L. J., and A. M. Cohen-Goldberg (2015). Can we hear morphological complexity before words are complex? *Journal of Experimental Psychology: Human Perception and Performance* 41: 50-68.

Bond, Z. S., and S. Garnes (1980). Misperceptions of fluent speech. In R. A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, NJ: Erlbaum.

Bond, Z. S., and L. H. Small (1983). Voicing, vowel, and stress mispronunciations in continuous speech. *Perception & Psychophysics* 34: 470-474.

Borrie, S. A., M. J. McAuliffe, J. M. Liss, C. Kirk, G. A. O'Beirne, and T. Anderson (2012). Familiarisation conditions and the mechanisms that underlie improved recognition of dysarthric speech. *Language and Cognitive Processes* 27: 1039-1055.

Bradlow, A. R., and T. Bent (2008). Perceptual adaptation to non-native speech. *Cognition* 106: 707.

Breen, M., L. Dilley, J. D. McAuley, and L. Sanders (2014). Auditory evoked potentials reveal early perceptual effects of distal prosody on speech segmentation. *Language, Cognition and Neuroscience* 29: 1132-1146.

Brown, M., L. Dilley, and M. Tanenhaus (2014). Probabilistic prosody: Effects of relative speech rate on perception of (a) word(s) several syllables earlier. In N. Campbell, D. Gibbon, & D. Hirst (Eds.), *Proceedings of the 7th International Conference on Speech Prosody* (pp. 1154-1158). Dublin, Ireland.

Brown, M., A. P. Salverda, L. Dilley, and M. Tanenhaus (2011). Expectations from preceding prosody influence segmentation in online sentence processing. *Psychonomic Bulletin and Review* 18: 1189-1196.

Brown, M., A. P. Salverda, L. Dilley, and M. Tanenhaus (2015). Metrical expectations from preceding prosody influence perception of lexical stress. *Journal of Experimental Psychology: Human Perception and Performance* 41: 306-323.

Bruce, G. (1977). *Swedish word accents in sentence perspective*. Lund: Gleerups.

Buxó-Lugo, A., and D. Watson (2016). Evidence for the influence of syntax on prosodic parsing. *Journal of Memory and Language* 90: 1-13.

Carlson, K., C. J. Clifton, and L. Frazier (2001). Prosodic boundaries in adjunct attachment. *Journal of Memory and Language* 45: 58-81.

Cho, T. (2016). Prosodic boundary strengthening in the phonetics–prosody interface. *Language and Linguistics Compass* 10: 120-141.

Cho, T., and P. Keating (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics* 29: 155-190.

Cho, T., J. M. McQueen, and E. A. Cox (2007). Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics* 35: 210-243.

Christiansen, M. H., and N. Chater (2016). The Now-or-Never bottleneck: A fundamental constrain on language. *Behavioral and Brain Sciences* 39: e62.

Christophe, A., S. Peperkamp, C. Pallier, E. Block, and J. Mehler (2004). Phonological phrase boundaries constrain lexical access: I. Adult data. *Journal of Memory and Language* 51: 523-547.

Cole, R. A., and J. Jakimik (1978). Understanding speech: How words are heard. In G. Underwood (Ed.), *Strategies of information processing* (pp. 67-116). London: Academic Press.

Connine, C. M., D. Blasko, and M. Hall (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraints. *Journal of Memory and Language* 30: 234-250.

Cooper, N., A. Cutler, and R. Wales (2002). Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners. *Language and Speech* 45, No. 3: 207-228.

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics* 20: 55-60.

Cutler, A. (1986). *Forbear* is a homophone: Lexical prosody does not constrain lexical access. *Language and Speech* 29: 201-220.

Cutler, A., and S. Butterfield (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language* 31: 218-236.

Cutler, A., and C. E. Clifton (1984). The use of prosodic information in word recognition. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and Performance X* (pp. 183-196). Hillsdale, NJ: Erlbaum.

Cutler, A., D. Dahan, and W. van Donselaar (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech* 40: 141-201.

Cutler, A., and D. Foss (1977). On the role of sentence stress in sentence processing. *Language and Speech* 20: 1-10.

Cutler, A., and D. G. Norris (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance* 14: 113-121.

Cutler, A., and T. Otake (1999). Pitch accent in spoken-word recognition in Japanese. *Journal of the Acoustical Society of America* 105: 1877-1888.

Cutler, A., and W. van Donselaar (2001). *Voornaam* is not (really) a homophone: Lexical prosody and lexical access in Dutch. *Language and Speech* 44: 171-195.

Davis, M. H., I. S. Johnsrude, A. Hervais-Adelman, K. Taylor, and C. McGettigan (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General* 134: 222-241.

Davis, M. H., W. D. Marslen-Wilson, and M. G. Gaskell (2002). Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance* 28: 218-244.

Dilley, L. 2005. *The phonetics and phonology of tonal systems.* (Ph.D. dissertation), MIT, Cambridge, MA.

Dilley, L. (2008). On the dual relativity of tone. In *Proceedings from the Annual Meeting of the Chicago Linguistics Society* (Vol. 41, pp. 129-144). Chicago, IL.

Dilley, L. (2010). Pitch range variation in English tonal contrasts is continuous, not categorical. *Phonetica* 67: 63-81.

Dilley, L., and M. Breen (to appear). An *enhanced* autosegmental-metrical theory (AM^{+}) facilitates phonetically transparent prosodic annotation: A reply to Jun. In J. Barnes & S. Shattuck-Hufnagel (Eds.), *Prosodic Theory and Practice*. Cambridge, MA: MIT Press.

Dilley, L., and M. Brown (2007). Effects of pitch range variation on F0 extrema in an imitation task. *Journal of Phonetics* 35: 523-551.

Dilley, L., and C. Heffner (2013). The role of f0 alignment in distinguishing intonation categories: Evidence from American English. *Journal of Speech Sciences* 3: 3-67.

Dilley, L., S. Mattys, and L. Vinke (2010). Potent prosody: comparing the effects of distal prosody, proximal prosody, and semantic context on word segmentation. *Journal of Memory and Language* 63: 274-294.

Dilley, L., and J. D. McAuley (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language* 59: 294-311.

Dilley, L., and M. Pitt (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science* 21: 1664-1670.

Dilley, L., S. Shattuck-Hufnagel, and M. Ostendorf (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24: 423-444.

Ding, N., A. D. Patel, L. Chen, H. Butler, C. Luo, and D. Poeppel (2017). Temporal modulations in speech and music. *Neuroscience and Biobehavioral Reviews* 81: 181-187.

Dorman, M. F., P. C. Loizou, and D. Rainey (1997). Speech understanding as a function of the number of channels of stimulation for processors using sine-wave and noise-band outputs. *Journal of the Acoustical Society of America* 102: 2403-2411.

Dupoux, E., and K. Green (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology: Human Perception and Performance* 23: 914-927.

Farmer, T. A., M. H. Christiansen, and P. Monaghan (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences* 103: 12203-12208.

Fougeron, C., and P. A. Keating (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America* 101: 3728-3740.

Francis, A. L., V. Ciocca, N. K. Y. Wong, W. H. Y. Leung, and P. C. Y. Chu (2006). Extrinsic context affects perceptual normalization of lexical tone. *Journal of the Acoustical Society of America* 119: 1712-1726.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech* 1: 126-152.

Garellek, M. (2014). Voice quality strengthening and glottalization. *Journal of Phonetics* 45: 106-113.

Garellek, M., and P. A. Keating (2011). The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *Journal of the International Phonetic Association* 41: 185-205.

Garellek, M., P. A. Keating, C. Esposito, and J. Kreiman (2013). Voice quality and tone identification in White Hmong. *The Journal of the Acoustical Society of America* 133: 1078-1089.

Genzel, S., and F. Kügler (in press). Production and perception of question prosody in Akan. *Journal of the International Phonetic Association*.

Gerfen, C., and K. Baker (2005). The production and perception of laryngealized vowels in Coatzospan Mixtec. *Journal of Phonetics* 33: 311-334.

Gibson, E., L. Bergen, and S. T. Piantadosi (2013). The rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences* 110: 8051-8056.

Giraud, A.-L., and D. Poeppel (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience* 15: 511-517.

Gordon, M., and P. Ladefoged (2001). Phonation types: A cross-linguistic overview. *Journal of Phonetics* 29: 383-406.

Gow, D., and P. C. Gordon (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance* 21: 344-359.

Grossberg, S., and C. Myers (2000). The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review* 107: 735-767.

Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge: Cambridge University Press.

Halle, M., and K. N. Stevens (1962). Speech recognition: A model and a program for research. *IEEE Transactions on Information Theory* 8: 155-159.

Hayes, B. (1995). *Metrical Stress Theory: Principles and Case Studies*. Chicago: University of Chicago Press.

Heffner, C., L. Dilley, J. D. McAuley, and M. Pitt (2013). When cues combine: how distal and proximal acoustic cues are integrated in word segmentation. *Language and Cognitive Processes* 28: 1275-1302.

Heldner, M., and E. Strangert (2001). Temporal effects of focus in Swedish. *Journal of Phonetics* 29: 329-361.

Hillenbrand, J., and R. A. Houde (1996). Role of F0 and amplitude in the perception of intervocalic glottal stops. *Journal of Speech and Hearing Research* 39: 1182-1190.

Holt, L. L. (2006). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *Journal of the Acoustical Society of America* 120: 2801-2817.

Ito, K., and S. R. Speer (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language* 58: 541-573.

Johnson, K., E. A. Strand, and M. D'Imperio (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics* 27: 359-384.

Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review* 83: 323-355.

Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review* 99: 349-364.

Kelly, M. H., and J. K. Bock (1988). Stress in time. *Journal of Experimental Psychology: Human Perception and Performance* 14: 389-403.

Kidd, G. R. (1989). Articulatory rate-context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance* 15: 736-748.

Kim, S., and T. Cho (2013). Prosodic boundary information modulates phonetic categorization. *Journal of the Acoustical Society of America* 134: EL19–EL25.

Kleinschmidt, D. F., and F. T. Jaeger (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review* 122: 148-203.

Krivokapić, J., and D. Byrd (2012). Prosodic boundary strength: An articulatory and perceptual study. *Journal of Phonetics* 40: 430-442.

Kula, N. C., and B. Braun (2015). Mental representation of tonal spreading in Bemba: Evidence form elicited production and perception. *Southern African Linguistics and Applied Language Studies* 33: 307-323.

Kuperberg, G. R., and F. T. Jaeger (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience* 31: 32-59.

Kurumada, C., M. Brown, and M. K. Tanenhaus (2018). Effects of distributional information on categorization of prosodic contours. *Psychonomic Bulletin and Review* 25: 1153-1160.

Ladd, D. R. (2008). *Intonational Phonology* (2nd ed.). Cambridge: Cambridge University Press.

Lai, W. (2018). *Voice gender effect on tone categorization and pitch perception.* Paper presented at the Sixth International Symposium on Tonal Aspects of Language, Berlin.

Lee, C.-Y. (2007). Does horse activate mother? Processing lexical tone in form priming. *Language and Speech* 50: 101-123.

Lee, C.-Y. (2009). Identifying isolated, multispeaker Mandarin tones from brief acoustic input: A perceptual and acoustic study. *Journal of the Acoustical Society of America* 125: 1125-1137.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy (1967). Perception of the speech code. *Psychological Review* 74: 431-461.

Liberman, M., and A. Prince (1977). On stress and linguistic rhythm. *Linguistic Inquiry* 8: 249-336.

Lisker, L., and A. S. Abramson (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *WORD: Journal of the International Linguistic Association* 20: 384-422.

Martin, J. G. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological Review* 79: 487-509.

Mattys, S. L. (2000). The perception of primary and secondary stress in English. *Perception and Psychophysics* 62: 253-265.

Mattys, S. L. (2004). Stress versus coarticulation: Toward an integrated approach to explicit speech segmentation. *Journal of Experimental Psychology: Human Perception and Performance* 30: 397-408.

Mattys, S. L., L. White, and J. F. Melhorn (2005). Integration of multiple speech segmentation cues: a hierarchical framework. *Journal of Experimental Psychology: General* 134: 477-500.

McClelland, J. L., and J. L. Elman (1986). The TRACE model of speech perception. *Cognitive Psychology* 18: 1-86.

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science* 317: 631-631.

McQueen, J. M., A. Cutler, and D. Norris (2006). Phonological abstraction in the mental lexicon. *Cogn Sci* 30: 1113-1126.

Miller, J. L. (1981). Phonetic perception: evidence for context-dependent and context-independent processing. *Journal of the Acoustical Society of America* 69: 822-831.

Miller, J. L., and A. M. Liberman (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics* 25: 457-465.

Mitterer, H., Y. Chen, and X. Zhou (2011). Phonological abstraction in processing lexical-tone variation: evidence from a learning paradigm. *Cogn Sci* 35: 184-197.

Mitterer, H., T. Cho, and S. Kim (2016). How does prosody influence speech categorization? *Journal of Phonetics* 54: 68-79.

Mitterer, H., and J. M. McQueen (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PloS One* 4: e7785.

Mitterer, H., E. Reinisch, and J. M. McQueen (2018). Allophones, not phonemes in spoken-word recognition. *Journal of Memory and Language* 98: 77-92.

Morrill, T. (2012). Acoustic correlates of stress in English adjective-noun compounds. *Language and Speech* 55: 167-201.

Morrill, T., M. Baese-Berk, C. Heffner, and L. Dilley (2015). Interactions between distal speech rate, linguistic knowledge, and speech environment. *Psychonomic Bulletin and Review* 22: 1451-1457.

Morrill, T., L. Dilley, and J. D. McAuley (2014). Prosodic patterning in distal speech context: Effects of list intonation and f0 downtrend on perception of proximal prosodic structure. *Journal of Phonetics* 46: 68-85.

Morrill, T., L. Dilley, J. D. McAuley, and M. Pitt (2014). Distal rhythm influences whether or not listeners hear a word in continuous speech: support for a perceptual grouping hypothesis. *Cognition* 131: 69-74.

Nakatani, L. H., and J. A. Schaffer (1978). Hearing "words" without words: Prosodic cues for word perception. *Journal of the Acoustical Society of America* 63: 234-245.

Newman, R. S., and J. R. Sawusch (1996). Perceptual normalization for speaking rate: effects of temporal distance. *Perception & Psychophysics* 58: 540-560.

Norris, D., and J. M. McQueen (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review* 115: 357-395.

Norris, D., J. M. McQueen, and A. Cutler (2003). Perceptual learning in speech. *Cognitive Psychology* 47: 204-238.

Norris, D., J. M. McQueen, and A. Cutler (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience* 31: 4-18.

Norris, D., J. M. McQueen, A. Cutler, and S. Butterfield (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology* 34: 191-243.

Pickering, M. J., and S. Garrod (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences* 36: 329-392.

Pitt, M., C. Szostak, and L. Dilley (2016). Rate dependent speech processing can be speech specific: Evidence from the perceptual disappearance of words under changes in context speech rate. *Attention, Perception, and Psychophysics* 78: 334-345.

Pitt, M. A., and A. G. Samuel (1990). The use of rhythm in attending to speech. *Journal of Experimental Psychology: Human Perception and Performance* 16: 564-573.

Poellmann, K., H. R. Bosker, J. M. McQueen, and H. Mitterer (2014). Perceptual adaptation to segmental and syllabic reductions in continuous spoken Dutch. *Journal of Phonetics* 46: 101-127.

Poeppel, D., W. J. Idsardi, and V. van Wassenhove (2008). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society B* 363: 1071-1086.

Quené, H., and R. F. Port (2005). Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica* 62: 1-13.

Ramachers, S. T. M. R. (2018). *Setting the tone: Acquisition and processing of lexical tone in East-Limburgian dialects of Dutch*. Utrecht: LOT.

Reinisch, E., A. Jesse, and J. M. McQueen (2010). Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately. *Quarterly Journal of Experimental Psychology* 63: 772-783.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin* 92: 81-110.

Rohde, H., and C. Kurumada (2018). Alternatives and inferences in the communication of meaning. In C. Fedemeier & D. Watson (Eds.), *Psychology of Learning and Motivation* (Vol. 68, pp. 215-252).

Salverda, A. P., D. Dahan, and J. M. McQueen (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition* 90: 51-89.

Samuel, A. G., and T. Kraljic (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics* 71: 1207-1218.

Sawusch, J. R., and R. S. Newman (2000). Perceptual normalization for speaking rate II: Effects of signal discontinuities. *Perception and Psychophysics* 62: 285-300.

Selkirk, E. O. (1984). *Phonology and Syntax: The Relation Between Sound and Structure*. Cambridge, MA: MIT Press.

Shannon, R. V., F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid (1995). Speech recognition with primarily temporal cues. *Science* 270: 303-304.

Shatzman, K. B., and J. M. McQueen (2006). Prosodic knowledge affects the recognition of newly acquired words. *Psychological Science* 17: 372-377.

Shields, J. L., A. McHugh, and J. G. Martin (1974). Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. *Journal of Experimental Psychology* 102: 250-255.

Shuai, L., and J. G. Malins (2017). Encoding lexical tones in jTRACE: a simulation of monosyllabic spoken word recognition in Mandarin Chinese. *Behavior Research Methods* 49: 230-241.

Sjerps, M. J., C. Zhang, and G. Peng (2018). Lexical tone is perceived relative to locally surrounding context, vowel quality to preceding context. *Journal of Experimental Psychology: Human Perception and Performance* 44: 914-924.

Slowiaczek, L. M. (1990). Effects of lexical stress in auditory word recognition. *Language and Speech* 33: 46-68.

Slowiaczek, L. M. (1991). Stress and context in auditory word recognition. *Journal of Psycholinguistic Research* 20: 465-481.

Sluijter, A. M. C., and V. J. van Heuven (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America* 100: 2417-2485.

Söderström, P., M. Horne, P. Mannfolk, D. van Westen, and M. Roll (2017). Tone-grammar association within words: Concurrent ERP and fMRI show rapid neural pre-activation and involvement of left inferior frontal gyrus in pseudowords. *Brain and Language* 174: 119-126.

Soto-Faraco, S., N. Sebastian-Galles, and A. Cutler (2001). Segmental and suprasegmental mismatch in lexical access. *Journal of Memory and Language* 45: 412-432.

Sulpizio, S., and J. M. McQueen (2012). Italians use abstract knowledge about lexical stress during spoken-word recognition. *Journal of Memory and Language* 66: 177-193.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance* 7: 1074-1095.

Tabain, M. (2003). Effects of prosodic boundary on /aC/ sequences: Articulatory results. *Journal of the Acoustical Society of America* 113: 2834-2849.

Tabossi, P., S. Collina, M. Mazzetti, and M. Zoppello (2000). Syllables in the processing of spoken Italian. *Journal of Experimental Psychology: Human Perception and Performance* 26: 758-775.

Tagliapietra, L., and J. M. McQueen (2010). What and where in speech recognition: Geminates and singletons in spoken Italian. *Journal of Memory and Language* 63: 306-323.

Tavano, A., and M. Scharinger (2015). Prediction in speech and language processing. *Cortex* 68: 1-7.

Tremblay, A., M. Broersma, and C. E. Coughlin (2018). The functional weight of a prosodic cue in the native language predicts the learning of speech segmentation in a second language. *Bilingualism: Language and Cognition* 21: 640-652.

Tremblay, A., M. Broersma, C. E. Coughlin, and J. Choi (2016). Effects of the native language on the learning of fundamental frequency in second-language speech segmentation. *Frontiers in Psychology* 7: 985.

Turk, A. E., and White, L. (1999). Structural influences on accentual lengthening in English. *Journal of Phonetics* 27: 171-206.

Turk, A. E., and S. Shattuck-Hufnagel (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics* 28: 397-440.

Wade, T., and L. L. Holt (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception & Psychophysics* 67: 939-950.

Wang, B., Y. Xu, and Q. Ding (2018). Interactive prosodic marking of focus, boundary and newness in Mandarin. *Phonetica* 75: 24-56.

Wong, P. C. M., and R. L. Diehl (2003). Perceptual normalization for inter- and intra-talker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research* 46: 413-421.

Yamamoto, H., and E. Haryu (2018). The role of pitch patterns in Japanese 24-month olds' word recognition. *Journal of Memory and Language* 99: 90-98.

Ye, Y., and C. M. Connine (1999). Processing spoken Chinese: The role of tone information. *Language and Cognitive Processes* 14: 609-630.