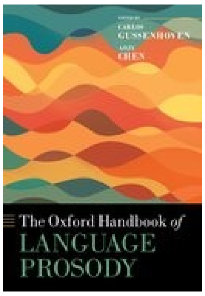
The Autosegmental-Metrical theory of intonational phonology

Arvaniti, A.; Fletcher, J.

2020, Part of book or chapter of book (Gussenhoven, C.; Chen, A.J. (ed.), The Oxford Handbook of Language Prosody, pp. 78-96)

Doi link to publisher: https://doi.org/10.1093/oxfordhb/9780198832232.013.4



The Oxford Handbook of Language Prosody

Carlos Gussenhhoven (ed.), Aoju Chen (ed.)

https://doi.org/10.1093/oxfordhb/9780198832232.001.0001

Published: 2020 **Online ISBN:** 9780191870880 **Print ISBN:** 9780198832232

Search in this book

CHAPTER

6 The Autosegmental-Metrical Theory of Intonational Phonology

Amalia Arvaniti, Janet Fletcher

https://doi.org/10.1093/oxfordhb/9780198832232.013.4 Pages 78–95

Published: 10 February 2021

Abstract

The chapter outlines the basic principles of the autosegmental-metrical (AM) theory of intonational phonology. AM posits that at the phonological level intonation consists of a string of L(ow) and H(igh) tones (i.e. a string of tonal autosegments) that associate with metrical heads and phrasal boundaries. Phonetically, tones are realized as tonal targets, specific f0 points defined by their scaling and alignment; scaling refers to the pitch height of the tonal target, and alignment to the synchronization of the target with the segmental material that reflects its phonological association (typically stressed syllables and boundary-adjacent syllables). The chapter explains these essential tenets of AM in some detail and discusses how they differ from those of other models of intonation and what consequences these differences have for modelling and predicting the realization of pitch contours. The chapter presents the basics of phonological representation and phonetic modelling in AM, and briefly touches on intonational meaning and AM applications.

Keywords: autosegmental-metrical theory, intonational phonology, phonological representation, phonetic modelling, intonational meaning

Subject: Phonetics and Phonology, Linguistics

Series: Oxford Handbooks

Collection: Oxford Handbooks Online

6.1 Introduction

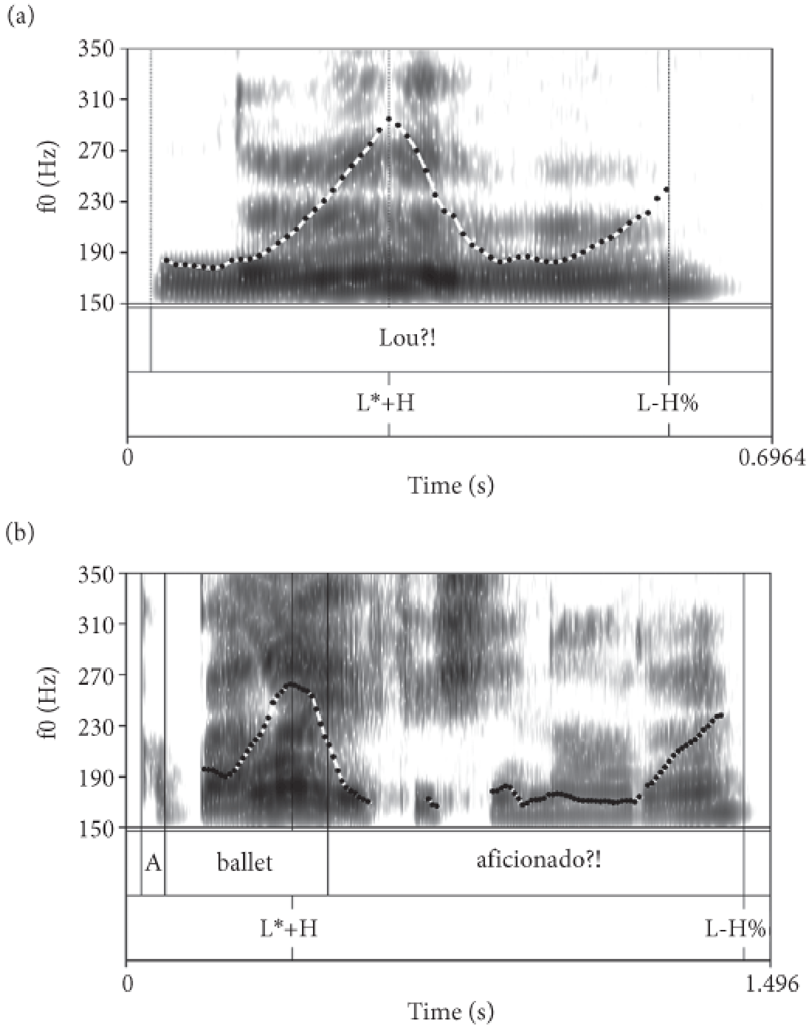
THE autosegmental-metrical theory of intonational phonology (henceforth AM) is a widely adopted theory concerned with the phonological representation of intonation and its phonetic implementation. The term 'intonation' refers to the linguistically structured modulation of fundamental frequency (f0), which directly relates to the rate of vibration of the vocal folds and gives rise to the percept of pitch. Intonation is used in all languages and specified at the 'post-lexical' (phrasal) level by means of a complex interplay between metrical structure, prosodic phrasing, syntax, and pragmatics; these factors determine where f0 movements will occur and of what type they will be. Intonation serves two main functions: encoding pragmatic meaning and marking phrasal boundaries. In addition to intonation, f0 is used for lexical purposes, when it encodes tonal contrasts in languages traditionally described as having a 'lexical pitch accent', such as Swedish and Japanese, as well as languages with a more general distribution of 'lexical tone', such as Mandarin, Thai, and Igbo. Both types are modelled in AM together with tones that signal intonation (see e.g. Pierrehumbert and Beckman 1988 on Japanese). In addition to these linguistic uses, f0 is used to signal 'paralinguistic' information such as boredom, anger, emphasis, or excitement (on paralinguistic uses of f0, see Gussenhoven 2004: ch. 5; Ladd 2008b: ch. 1; see also chapter 30).

Several models for specifying f0 contours are available today, such as Parallel Encoding and Target Approximation (PENTA) (Xu and Prom-On 2014, inter alia), the International Transcription System for Intonation (INTSINT) (Hirst and Di Cristo 1998), and the Fujisaki model (Fujisaki 1983, 2004). However, many aim at modelling f0 curves rather than defining the relation between f0 curves and the phonological structures that give rise to them. In contrast, AM makes a principled distinction between intonation as a subsystem of a language's phonology and f0, its main phonetic exponent. The arguments for this p. 79 distinction ↳ are similar to those that apply to segmental aspects of speech organization. Consider the following analogy. In producing a word, it is axiomatic in linguistic theory that the word is not mapped directly onto the movements of the vocal organs. There is instead an intervening level of phonological structure: a word is represented in terms of abstract units of sounds known as 'phonemes' or articulatory 'gestures', which cause the vocal organs to move in an appropriate way. According to AM, the same applies in intonation. If a speaker wants to produce a meaning associated with a polar question (e.g. 'Do you live in Melbourne?'), this meaning is not directly transduced as rising pitch. Instead, there is an intervening level of 'abstract tones' (which can, like phonemes, be represented symbolically); these tones specify a set of pitch targets that the speaker should produce if this particular melody is to be communicated. This relationship between abstract tones and phonetic realization also applies in languages that have lexically specified tone. In both types of language, only the abstract tones form part of the speaker's cognitive-phonological plan in producing a melody, with the precise details of how pitch changes are to be realized being filled in by phonetic procedures. AM thus integrates the study of phonological representation and phonetic realization (for details, see §6.2 and §6.3 respectively). The essential tenets of the model are largely based on Pierrehumbert's (1980) dissertation (see also Bruce 1977), with additional refinements built on experimental research and formal analysis involving a large number of languages (see Ladd 2008b for a theoretical account; see Gussenhoven 2004 and Jun 2005a, 2014a for language surveys).

The term 'autosegmental-metrical', which gave the theory its name, was coined by Ladd (1996) and reflects the connection between two subsystems of phonology: an autosegmental tier representing intonation's melodic part as well as any lexical tones (if part of the system), and metrical structure representing prominence and phrasing. The connection reflects the fact that AM sees intonation as part of a language's 'prosody', an umbrella term that encompasses interacting phenomena that include intonation, rhythm, prominence, and prosodic phrasing. The term 'prosody' is preferred over the older term 'suprasegmentals' (e.g. Lehiste 1977a; Ladd 2008b), so as to avoid the layering metaphor inherent in the latter (cf. Beckman and Venditti 2011): prosody is not a supplementary layer over vowels and consonants but an integral part of the phonological representation of speech.

Crucial to AM's success has been the central role it gives to the underlying representation of tunes as a series of tones rather than contours. Specifically, AM analyses continuous (and often far from smooth) pitch contours as a series of abstract primitives. This is a challenging endeavour for two reasons. First, intonational primitives cannot be readily identified based on meaning (as tones can in tone languages, such as Cantonese, where distinct pitch patterns are associated with changes in lexical meaning). In contrast, the meaning of intonational primitives is largely pragmatic (Hirschberg 2004), so in languages like English choice of melody is not constrained by choice of words. Second, f0 curves do not exhibit obvious changes that readily lead to positing distinct units; thus, breaking down the f0 curve into constituents is not as straightforward as identifying distinct patterns corresponding to segments in a spectrogram. This is all the more challenging, as a melody can spread across several words or be realized on a monosyllabic utterance. To illustrate this point, consider the pitch contours in Figure 6.1. The utterance in panel a is monosyllabic, while the one in panel b is eight syllables long. The f0 contours of the two utterances are similar but not identical: neither can be said to be a stretched or squeezed version of the other. Nevertheless, both contours are recognized by native speakers of English as realizations of the same melody, in terms of both form and pragmatic ↳ function, the aim of which is to signal incredulity (Ward and Hirschberg 1985; Hirschberg and p. 80 Ward 1992). The differences between the contours are not random. Rather, they exhibit what Arvaniti and Ladd (2009) have termed 'lawful variability', i.e. variation that is systematically related to variables such as the length of the utterance (as shown in Figure 6.1), the position of stressed syllables, and a host of other factors (see Arvaniti 2016 for a detailed discussion of additional sources of systematic variation in intonation). Besides understanding what contours like those in Figure 6.1 have in common and how they vary, a central endeavour of AM is to provide a phonological analysis that reflects this understanding.

Figure 6.1



Spectrograms and f0 contours illustrating the same English tune as realized on a monosyllabic utterance (a) and a longer utterance (b).

6.2 AM phonology

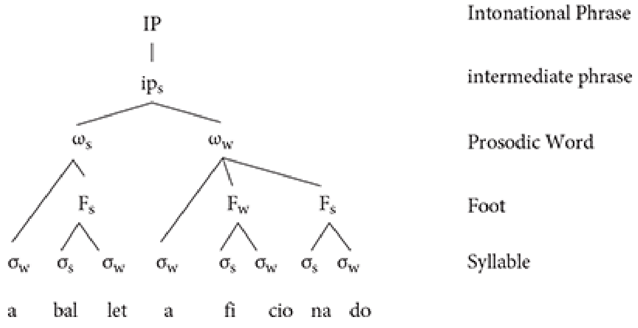
6.2.1 AM essentials

In AM, intonation is phonologically represented as a string of Low (L) and High (H) tones and combinations thereof (Pierrehumbert 1980; Beckman and Pierrehumbert 1986; Ladd 2008b; cf. Leben 1973; Liberman 1975; Goldsmith 1981). Tones are considered ↳ 'autosegments': they are autonomous segments relative to the string of vowels and consonants. Ls and Hs are the abstract symbolic (i.e. phonological) primitives of intonation (much as they are the primitives in the representation of lexical tone). Their identity as Hs and Ls is defined in relative terms: H is used to represent tones deemed to be relatively high at some location in a melody relative to the pitch of the surrounding parts of the melody, while L is used to represent tones that are relatively low by the same criterion (cf. Pierrehumbert 1980: 68–75). Crucially, the aim of the string of tones is not to faithfully represent all the modulations that may be observed in f0 contours but rather to capture significant generalizations about contours perceived to be instances of the same melody (see Arvaniti and Ladd 2009 for a detailed presentation of this principle). Thus, AM phonological presentations are underspecified in the sense that they do not account (and are not meant to account) for all pitch movements; rather they include only those elements needed to capture what is contrastive in a given intonational system. At the phonetic level as well, it is only the tones of the phonological representation that are realized as targets, with the rest of the f0 contour being derived by 'interpolation' (see §6.3.3 for a discussion of interpolation).

6.2.2 Metrical structure and its relationship with the autosegmental tonal string

The relationship between tones and segments (often referred to as 'tune–text association') is mediated by a metrical structure. This is a hierarchical structure that represents (i) the parsing of an utterance into a number of constituents and (ii) the prominence relations between them (e.g. the differences between stressed and unstressed syllables). The term 'metrical structure', as in the term 'autosegmental-metrical theory', is typically used when the representation of stress is at issue; when the emphasis is on phrasal structure, the term 'prosodic structure' is often used instead. Both relative prominence and phrasing can be captured by the same representation (see e.g. Pierrehumbert and Beckman 1988). An example is given in (1), which represents the prosodic structure of the utterance in Figure 6.1b, 'a ballet aficionado?!'. As can be seen, syllables (σ) are grouped into feet (F), which in turn are grouped into prosodic words (ω); in this example, prosodic words are grouped into one intermediate phrase (ip), which is the only constituent of the utterance's only intonational phrase (IP). Relative prominence is presented by marking constituents as strong (s) or weak (w).

(1)

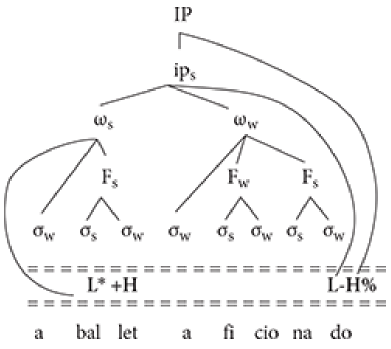


p. 82 The prosodic structure in (1) is based on the model of Pierrehumbert and Beckman (1988), which has been implicitly adopted and informally used in many AM analyses. This model is similar to other well-known models (cf. Selkirk 1984; Nespor and Vogel 1986) but differs from them in some critical aspects. First, the number and nature of levels in the prosodic hierarchy are not fixed but language specific. For instance, Pierrehumbert and Beckman (1988) posit three main levels of phrasing for Tokyo Japanese: the accentual phrase (AP), the intermediate phrase (ip) and the Intonational Phrase (IP). However, they posit only two levels of phrasing for English: the ip and the IP, as illustrated in (1), since they found no evidence for an AP level of phrasing (Beckman and Pierrehumbert 1986). Further, the model assumes that it is possible to have headless constituents (i.e. constituents that do not include a strong element, or 'head'). In the analysis of Pierrehumbert and Beckman (1988), this applies to Japanese AP's that do not include a word with a lexical pitch accent; in such AP's, there are no syllables, feet, or prosodic words that are strong. The same understanding applies by and large to several other languages that allow headless constituents, such as Korean (Jun 2005b), Chickasaw (Gordon 2005a), Mongolian (Karlsson 2014), Tamil (Keane 2014), and West Greenlandic (Arnhold 2014a); informally, we can say that these languages do not have stress. In addition, the model of Pierrehumbert and Beckman (1988) relies on *n*-ary branching trees (trees with more than two branches per node); grouping largely abides by the Strict Layer Hypothesis, according to which all constituents of a given level in the hierarchy are exhaustively parsed into constituents of the next level up (Selkirk 1984). However, Pierrehumbert and Beckman also accept limited extrametricality, such as syllables that are linked directly to a prosodic word node (Pierrehumbert and Beckman 1988: 147 ff.); this is illustrated in (1), where the indefinite article *a* and the unstressed syllable at the beginning of *aficionado* are linked directly to the relevant ω node. (For an alternative model of prosodic structure that allows limited recursiveness, see Ladd 2008b: ch. 8, incl. references.)

Independently of the particular version of prosodic structure adopted in an AM analysis, it is widely agreed that tones associate with phrasal boundaries or constituent heads (informally, stresses) or both (see §6.2.3 for details on secondary association of tones, and Gussenhoven 2018 for a detailed discussion of tone association). Tones that associate with stressed syllables are called 'pitch accents' and one of their roles is prominence enhancement; they are notated with a star (e.g. H*). The final accent in a phrase is called the 'nuclear pitch accent' or 'nucleus', and is usually deemed the most prominent. Pitch accents may consist of more than one tone and are often bitonal; examples include L*+H- and L-+H* (after Pierrehumbert's 1980 original notation but also annotated as L*H or L*+H, and LH* or L+H* respectively). Pitch patterns have been analysed as reflexes of bitonal accents in a number of languages, including English (Ladd and Schepman 2003), German (Grice et al. 2005a), Catalan (Prieto 2014), Arabic (Chahal and Hellmuth 2014b), and Jamaican Creole (Gooden 2014). Grice (1995a) has also posited tritonal accents for English. (See also §6.2.3 on secondary association.)

In Pierrehumbert (1980), the starred tone of a bitonal pitch accent is metrically stronger than the unstarred tone, and the only one that is phonologically associated (for details see §6.3.1); the unstarred weak tone that leads or trails the starred tone is 'floating' (i.e. it is a tone without an association). Research on a number of languages since Pierrehumbert (1980) indicates that additional types of relations are possible between the tones of bitonal accents. Arvaniti et al. (1998, 2000) have provided experimental evidence from Greek that
p. 83 ↳ tones in bitonal accents can be independent of each other, in that neither tone exhibits the behaviour of an unstarred tone described by Pierrehumbert (1980). Frota (2002), on the other hand, reports data from Portuguese showing that the type of 'loose' bitonal accent found in Greek can coexist with pitch accents that show a closer connection between tones, akin to the accents described by Pierrehumbert (1980) for English.

Tones that associate with phrasal boundaries are collectively known as 'edge tones' and their main role is to demarcate the edges of the phrases they associate with. These may also be multitonal; for example, for Korean, Jun (2005b) posits boundary tones with up to five tones (e.g. LHLHL%), while Prieto (2014) posits a tritonal LHL% boundary tone for Catalan. Following Beckman and Pierrehumbert (1986), many analyses posit two types of edge tone, ‘phrase accents’ and ‘boundary tones’, notated with - and % respectively (e.g. H-, H%). Phrase accents demarcate ip boundaries and boundary tones demarcate IP boundaries. For example, *Nick and Mel were late because they missed the train* is likely to be uttered as two ip’s forming one IP: [[Nick and Mel were late]$_{ip}$ [because they missed the train]$_{ip}$]IP; the boundary between the two ip’s is likely to be demarcated with a H- phrase accent. An illustration of the types of association between tones and prosodic structure used in AM is provided in (2) using the same utterance as in (1).(2)

> (2)
>
> 
>

IP
> ip$_s$
> ω$_s$
> ω$_w$
> F$_s$
> F$_w$
> F$_s$
> σ$_w$ σ$_s$ σ$_w$ σ$_w$ σ$_s$ σ$_w$ σ$_s$ σ$_w$
> L* +H
> L-H%
> a bal let a fi cio na do

All of the languages investigated so far have edge tones that associate with right boundaries. Left-edge boundary tones have also been posited for several languages, including English (Pierrehumbert 1980; Gussenhoven 2004), Basque (Elordieta and Hualde 2014), Dalabon (Fletcher 2014), and Mongolian (Karlsson 2014). However, the specific proposal of Beckman and Pierrehumbert (1986) linking phrase accents to the ip and boundary tones to the IP has not been generally accepted, although it has been adopted by many analyses, including those for Greek (Arvaniti and Baltazani 2005), German (Grice et al. 2005a), Jamaican Creole (Gooden 2014), and Lebanese and Egyptian Arabic (Chahal and Hellmuth 2014b). Some AM analyses dispense altogether with phrase accents either for reasons of parsimony—positing only two types of primitives, pitch accents and boundary tones—or because they adopt a different conception of how the f0 contour is to be broken into constituent tones (see, among others, Gussenhoven 2004, 2005 on Dutch; Frota 2014 on Portuguese; Gussenhoven 2016 on English). Thus, phrase accents are not necessarily included in all AM analyses.

p. 84 Following Pierrehumbert and Hirschberg (1990), pitch accents and edge tones are treated as intonational morphemes with pragmatic meaning that contribute compositionally to the pragmatic interpretation of an utterance (Pierrehumbert and Hischberg 1990; Steedman 2014; see chapter 30 for a discussion of intonational meaning). Although this understanding of intonation as expressing pragmatic meaning is generally accepted, it may not apply to the same extent to all systems. For example, in languages like Korean and Japanese, in which intonation is used primarily to signal phrasing, tones express pragmatic meaning to a much lesser extent than in languages like English (Pierrehumbert and Beckman 1988 on Japanese; Jun 2005b on Korean).

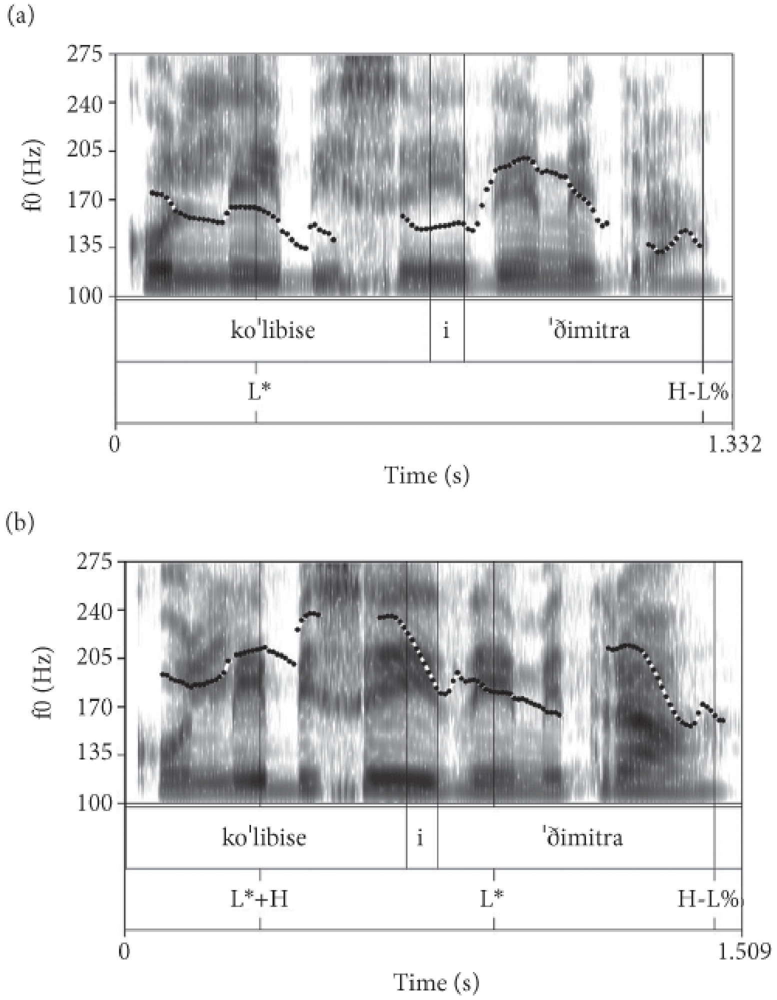
6.2.3 Secondary association of tones

In addition to a tone’s association with a phrasal boundary or constituent head, AM provides a mechanism for ‘secondary association’. For instance, according to Grice (1995a: 215 ff.), leading tones of English bitonal accents, such as L in L+H*, associate with the syllable preceding the accented one (if one is available), while trailing tones, such as H in L*+H, occur a fixed interval in ‘normalized time’ after the starred tone. The former is a type of secondary association (for discussions of additional association patterns, see Barnes et al. 2010a on English; van de Ven and Gussenhoven 2011 on Dutch; Peters et al. 2015 on several Germanic varieties).

Although secondary association has been used for a variety of purposes, it has come to be strongly associated with phrase accents. Pierrehumbert and Beckman (1988) proposed the mechanism of secondary association to account for the fact that phrase accents often spread (see also Liberman 1979). Specifically, Pierrehumbert and Beckman (1988) proposed that edge tones may acquire additional links (i.e. secondary associations) either to a specific tone-bearing unit (TBU), such as a stressed syllable, or to another boundary. For example, they posited that English phrase accents are linked not only to the right edge of their ip (as advocated in Beckman and Pierrehumbert 1986) but also to the left edge of the word carrying the nuclear pitch accent. An example of such secondary association can be found in Figure 6.1b, in which the L- phrase accent is realized as a low f0 stretch. This stretch is due to the fact that the L- phrase accent associates both with the right ip boundary (and thus is realized as close as possible to the right edge of the phrase) and with the end of the accented word (and thus stretches as far as possible to the left). The general mechanism whereby edge tones have secondary associations has also been used by Gussenhoven (2000a) in his analysis of the intonation of Roermond Dutch, which assumes that boundary tones can be *phonologically* aligned both with the right edge of the phrase and with an additional leftmost position.

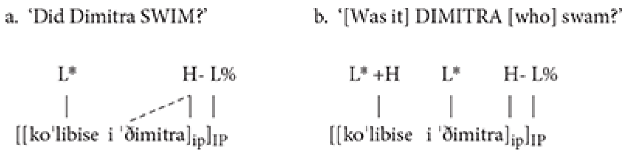
The analyses of Pierrehumbert and Beckman (1988) and Gussenhoven (2000a) were the basis for a wider use of secondary association for phrase accents developed in Grice et al. (2000), who argue that the need for positing phrase accents in a given intonation system is orthogonal to the need for the ip level of phrasing. Grice et al. (2000) examined putative phrase accents in a variety of languages (Cypriot Greek, Dutch, English, German, Hungarian, Romanian, and Standard Greek). They showed that the phrase accents they examined are realized either on a peripheral syllable, as expected of edge tones, or an earlier one, often one p. 85 that is metrically strong; which of the two realizations prevails depends on whether the ↳ metrically strong syllable is already associated with another tone or not. This type of variation is illustrated in Figure 6.2 with the Greek polar question tune L* H-L% (Grice et al. 2000; Arvaniti et al. 2006a). As can be seen in Figure 6.2, both contours have a pitch peak close to the end of the utterance. This peak co-occurs with the stressed antepenult of the final word in the question in Figure 6.2a ([ˈði] of [ˈðimitra]), but with the last vowel in the question in Figure 6.2b (the vowel [a] of [ˈðimitra]). (Note also that the stressed antepenult of [ˈðimitra] has low f0 in Figure 6.2b, as does the stressed syllable of [koˈlibise] in Figure 6.2a; both reflect an association with the L* pitch accent of this tune.) Grice et al. (2000) attribute this difference in the alignment of the pitch peak to secondary association: the peak is the reflex of a H- phrase accent associated with a phrasal boundary, but also has a secondary association to the last metrically strong syvllable of the utterance. This association is phonetically realized when this metrically strong syllable is not associated with a pitch accent; this happens when the focus is on an earlier word, which then attracts the L* pitch accent. The phonological structures involved are shown in (3a) and (3b); (3a) shows the primary and secondary association of the H- phrase accent; (3b) shows that the secondary association of H- is not possible because [ˈði] is already associated with the L* pitch accent.

Figure 6.2



Spectrograms and f0 contours of the utterance [koˈlibise i ˈðimitra] with focus on [koˈlibise] ‘swam’ (a) and on [ˈðimitra] (b), translated as ‘Did Dimitra SWIM?’ and ‘[Was it] DIMITRA who swam?’ respectively.

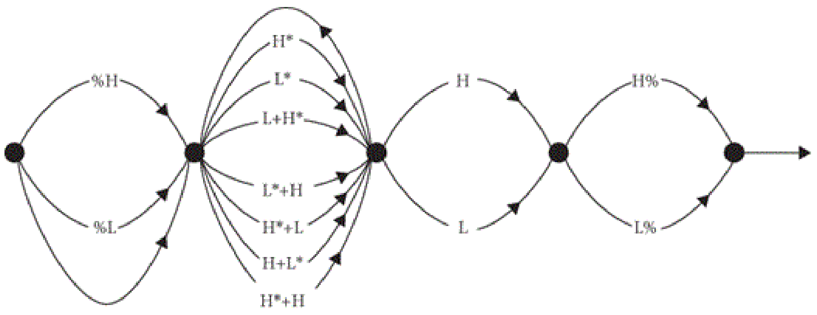
(3)


a. ‘Did Dimitra SWIM?’
L*
H- L%
[[koˈlibise i ˈðimitra]ip]IP
b. ‘[Was it] DIMITRA [who] swam?’
L* +H
L*
H- L%
[[koˈlibise i ˈðimitra]ip]IP

6.2.4 The phonological composition of melodies

In Pierrehumbert (1980), the grammar for generating English tunes is as shown in Figure 6.3. Primitives of the system can combine freely and, in the case of pitch accents, iteratively. With the exception of left-edge boundary tones, which are optional, all other elements are required. In other words, a well-formed tune must include at least one pitch accent followed by a phrase accent and a boundary tone (e.g. H* L-L%). The fact that elements combine freely is connected to Pierrehumbert’s position that there is no hierarchical structure for tunes (they are a linear string of autosegments, as illustrated in (2)). It follows that there are no qualitative differences between pitch accents, as in other models of intonation, and no elements are privileged in any way. This conceptualization of the tonal string also allows for the integration of lexically specified and post-lexical tones (i.e. intonation) into one tonal string.

Figure 6.3



The English intonation grammar of Pierrehumbert (1980); after Dainora (2006).

Not everyone who works within AM shares this view. Gussenhoven (2004: ch. 15, 2005, 2016) provides analyses of English and Dutch intonation that rely on the notion of nuclear contours as units comprising what in other AM accounts is a sequence of a nuclear pitch accent followed by edge tones. Gussenhoven's nuclear contours are akin to the nuclei of the British School. Gussenhoven (2016) additionally argues that a grammar along the lines of Figure 6.3 makes the wrong predictions, since not all possible combinations are grammatical in English, while the grammar results in both over- and under-analysis (capturing dubious distinctions while failing to capture genuine differences between tunes, respectively). Dainora (2001, 2006) also showed that some combinations of accents and edge tones are much more likely than others (though p. 87 the frequencies she presents may be skewed as they ↳ are based on a news-reading corpus). Other corpus studies of English also find that there are preferred combinations of tones in spoken interaction (e.g. Fletcher and Stirling 2014). Overall, the evidence indicates that some combinations are preferred and standardized, possibly because they reflect frequently used pragmatic meanings. This is particularly salient in some languages in which tune choice is limited to a small number of distinctive patterns (e.g. see chapter 26 for a description of intonation patterns in Indigenous Australian languages) by contrast with languages such as Dutch or English, where a range of tonal combinations are available to speakers (e.g. Gussenhoven 2004, 2005).

6.3 Phonetic implementation in AM

As noted in §6.1., AM provides a model for mapping abstract phonological representations to phonetic realization. Much of what we assume about this connection derives from Pierrehumbert (1980) and Bruce (1977). Phonetically, tones are said to be realized as 'tonal targets' (i.e. as specific points in the contour), while the rest of an f0 curve is derived by interpolation between these targets. That is, f0 contours are both phonologically and phonetically underspecified, in that only a few points of each contour are determined by tones and their targets (see Arvaniti and Ladd 2009 for empirical evidence of this point). Tonal targets are usually 'turning points', such as peaks, troughs, and elbows in the contour; they are defined by their 'alignment' and 'scaling' (see §6.3.1 and §6.3.2). Scaling refers to the value of the targets in terms of f0. Alignment is defined as the position of the tonal target relative to the specific TBU with which it is meant to co-occur (e.g. Arvaniti et al. 1998, 2006a, 2006b; Arvaniti and Ladd 2009). The identity of TBUs varies by language, depending on syllable structure, but we can equate TBUs with syllable nuclei (and in some instances with morae and coda consonants; see Pierrehumbert and Beckman 1988 on Japanese; Ladd et al. 2000 on Dutch; Gussenhoven 2012a on Limburgish). The TBUs with which tones phonetically co-occur are related to the metrical positions with which the tones associate in phonology: thus, pitch accents typically co-occur with stressed syllables (though not all stressed syllables are accented); edge tones are realized on peripheral TBUs, such as phrase-final vowels.

6.3.1 Tonal alignment

In AM, tonal alignment is a phonetic notion that refers specifically to the temporal alignment of tones with segmental and/or syllabic landmarks. Alignment can refer to the specific timing of a tone, but it may also reflect a phonological difference, in which case the timing of tones relative to the segmental string gives rise to a change in lexical or pragmatic meaning. For example, Bruce (1977) convincingly showed that the critical difference between the two lexical pitch accents of Swedish, Accent 1 and Accent 2, was due to the relative temporal alignment of a HL tonal sequence. For Accent 1, the H tone is aligned earlier with respect to the accented vowel than for Accent 2, a difference that Bruce (2005) encoded as a phonological difference between H+L* (Accent 1) and H*+L (Accent 2) for the Swedish East Prosodic dialect (see Bruce 2005 for a full overview of dialect-specific phonological ↳ variation in Swedish). Pierrehumbert (1980) similarly proposed L+H* and L*+H in English to account for the difference between early versus late alignment, with the H of the L*+H being realized after the stressed TBU (see also the discussion of trailing and leading tones in §6.2.3). While alignment differences are to be encoded in tonal representations when they are contrastive, in cases where variation in tonal alignment is not contrastive, a single representation, or 'label', is used. For instance, in Glasgow English, the alignment of the rising pitch accent, L*H in the analysis of Mayo et al. (1997), varies from early to late in the accented rhyme, without any apparent difference in meaning. In such cases, the tonal representation may allow for more options without essentially affecting the analysis. For instance, since the rising pitch accent of Glasgow English is variable, a simpler representation as H* instead of L*H may suffice, as nothing hinges on including a L tone in the accent's representation or on starring one or the other tone (cf. Keane 2014 on Tamil; Fletcher et al. 2016 on Mawng; Arvaniti 2016 on Romani).

One point that has become very clear thanks to a wide range of research on tonal alignment is that the traditional autosegmental idea that phonological association necessarily entails phonetic co-occurrence between a tone and a TBU does not always hold (e.g. see Arvaniti et al. 1998 on Greek; D'Imperio 2001 on Neapolitan Italian). This applies particularly to pitch peaks. Indeed, one of the most consistent findings in the literature is that of 'peak delay', the finding that accentual pitch peaks regularly occur after the TBU they are phonologically associated with. Peak delay was first documented by Silverman and Pierrehumbert (1990), who examined the phonetic realization of prenuclear H* accents in American English. It has since been reported for (among many others) South American Spanish (Prieto et al. 1995), Kinyarwanda (Myers 2003), Bininj Gun-wok (Bishop and Fletcher 2005), Catalan (Prieto 2005), Irish (Dalton and Ní Chasaide 2007a), and Chickasaw (Gordon 2008). The extent of peak delay can vary across languages and pitch accents, but it remains stable within category (Prieto 2014). This stability in known as 'segmental anchoring'.

The idea of segmental anchoring is based on the alignment patterns observed by Arvaniti et al. (1998) for Greek prenuclear accents and further explored in subsequent work by Ladd and colleagues on other languages (e.g. Ladd et al. 2000 on Dutch; Ladd and Schepman 2003 and Ladd et al. 2009b on English; Atterer and Ladd 2004 on German). Segmental anchoring is the hypothesis that tonal targets anchor onto particular segments in phonetic realization. The idea of segmental anchoring spurred a great deal of research in a variety of languages that have largely supported it (e.g. D'Imperio 2001 on Neapolitan Italian; Prieto 2009 on Catalan; Arvaniti and Garding 2007 on American English; Gordon 2008 on Chickasaw; Myers 2003 on Kinyarwanda; Elordieta and Calleja 2005 on Basque Spanish; Dalton and Ní Chasaide 2007a on Irish).

Finally, research on tonal alignment also supports the key assumption underpinning AM models in which tonal targets are levels rather than contours (i.e. rises or falls). This idea was put to the test in Arvaniti et al. (1998), who found that the L and H targets of Greek prenuclear accents each have their own alignment properties. A consequence of this mode of alignment is that the rise defined by the L and H targets has no invariable properties (such as duration or slope), a finding used by Arvaniti et al. (1998) to argue in favour of levels as intonational primitives. Empirical evidence from tone perception in English (Dilley and Brown 2007) showing that listeners perceptually equate pitch movements with level tones supports this view (see also House 2003).

p. 89 6.3.2 Tonal scaling

Since Ladd (1996) a distinction has been made between ‘pitch span’, which refers to the extent of the range of frequencies used by a speaker, and ‘pitch level’, which refers to whether these frequencies are overall high or low; together, level and span constitute a speaker’s ‘pitch range’. Thus, two speakers may have the same pitch span of 200 Hz but one may use a low level (e.g. 125–325 Hz) and the other a higher level (e.g. 175–375 Hz). A speaker’s pitch range may change for paralinguistic reasons, while, cross-linguistically, gender differences have also been observed (e.g. Daly and Warren 2002; Graham 2014).

Three main linguistic factors affect tonal scaling: declination, tonal context, and tonal identity. Declination is a systematic lowering of targets throughout the course of an utterance (’t Hart et al. 1990), though declination can be suspended (e.g. in questions) and is reset across phrasal boundaries (Ladd 1988; see also Truckenbrodt 2002). Listeners anticipate declination effects and adjust their processing of tonal targets accordingly (e.g. Yuen 2007). Within AM, the understanding of declination follows Pierrehumbert (1980): the scaling of tones is modelled with reference to a declining baseline that is invariant for each speaker (at a given time). The baseline is defined by its slope and a minimum value assumed to represent the bottom of the speaker’s range, which tends to be very stable for each speaker (Maeda 1976; Menn and Boyce 1982; Pierrehumbert and Beckman 1988). L and H tones (apart from terminal L%s) are scaled above the baseline and with reference to it.

Tonal context relates to the fact that the scaling of targets is related to the targets of preceding tones. For sequences of accentual H tones in particular, Liberman and Pierrehumbert (1984) have argued that every tone’s scaling is a fraction of the scaling of the preceding H. Tonal scaling is influenced by tonal context: for example, according to Pierrehumbert (1980: 136), the difference between the vocative chant H*+L- H- L% and a straightforward declarative, H* L- L%, is that the L% in the former melody remains above the baseline (and is realized as sustained level pitch), while the L% in the latter is realized as a fall to the baseline. In Pierrehumbert’s analysis, this difference is due to tonal context: in H*+L H-L%, L% is ‘upstepped’ (i.e. scaled higher) after a H- phrase accent; this context does not apply in H* L-L%, so L% reaches the baseline. One exception to the view that each H tone’s scaling is calculated as a fraction of the preceding H is what Liberman and Pierrehumbert (1984) have called ‘final lowering’, the fact that the final peak in a series is scaled lower than what a linear relation between successive peaks would predict. It has been reported in several languages with very different prosodic systems, including Japanese (Pierrehumbert and Beckman 1988), Dutch (Gussenhoven and Rietveld 1988), Yoruba (Connell and Ladd 1990; Laniran and Clements 2003), Kipare (Herman 1996), Spanish (Prieto et al. 1996), and Greek (Arvaniti and Godjevac 2003); see Truckenbrodt (2004, 2016) for an alternative analysis of final lowering.

Tonal identity refers to different effects of a number of factors on the scaling of H and L tones. In English, for instance, L tones are said to be upstepped following H tones, while the reverse does not apply (Pierrehumbert 1980). Further, changes in pitch range affect the scaling of H and L tones in different ways: L tones tend to get lower when pitch span expands, while H tones get higher (e.g. Pierrehumbert and Beckman 1988; Gussenhoven and Rietveld 2000).

An aspect of tonal scaling that has attracted considerable attention is ‘downstep’, the lower-than-expected scaling of H tones. In Pierrehumbert (1980) and Beckman and Pierrehumbert (1986), the essential premise is that downstep is the outcome of contextual rules. Thus, Pierrehumbert (1980) posits that downstep
p. 90 applies to the second H tone in a ↳ HLH sequence, as in the case of the second H in Pierrehumbert’s (1980)

representation of the vocative chant H*+L- H- L% above. In Beckman and Pierrehumbert (1986), tonal identity is also a key factor: all bitonal pitch accents trigger downstep of a following H tone. This position has been relaxed somewhat as it has been found that bitonal accents do not always trigger downstep in spoken discourse in American English (Pierrehumbert 2000). Others have argued that downstepped accents differ in meaning from accents that are not downstepped and thus that downstep should be treated as an independent phonological feature to mark the contrast between downstepped and non-downstepped accents, such as !H* and H* respectively (Ladd 1983, 2008b). The issue of whether downstep is a matter of context-dependent phonetic scaling or represents a meaningful choice remains unresolved for English and has been a matter of debate more generally.

In some AM systems, additional notations are used to indicate differences in scaling. For example, 0% and % have been used to indicate an intermediate level of pitch that is neither high nor low within a given melody and is often said to reflect a return to a default mid-pitch in the absence of a tonal specification (e.g. Grabe 1998a; Gussenhoven 2005; see Ladd 2008b: ch. 3 and Arvaniti 2016 for discussion). Still other systems incorporate additional variations in pitch, such as 'upstep', or higher-than-expected scaling. For instance, Grice et al. (2005a) use ^H% in the analysis of German intonation, and Fletcher (2014: 272) proposes ^ as 'an upstepped or elevated pitch diacritic' in her analysis of Dalabon. The use of symbols such as 0% reflects the awkwardness that mid-level tones pose for analysis, particularly if evidence suggests that such mid-level tones contrast with H and L tones proper, as has been argued for Greek (Arvaniti and Baltazani 2005), Maastricht Limburgish (Gussenhoven 2012b), Polish (Arvaniti et al. 2017), and German (Peters 2018). The use of diacritics more generally reflects the challenge of determining what is phonological and what is phonetic in a given intonational system, and thus what should be part of the phonological representation; see Jun and Fletcher (2014) and Arvaniti (2016) for a discussion of field methods that address these issues.

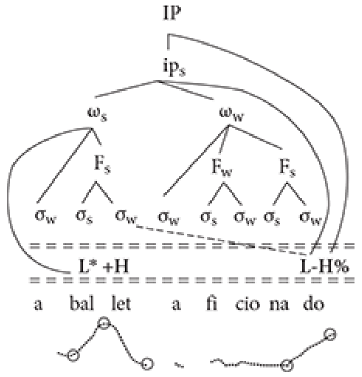
A reason why separating phonological elements from phonetic realization in intonation is such a challenge is the significant amount of variation attested in the realization of intonation. Even speakers of languages that have a number of different contrastive pitch accents may realize the same pitch accent category in different ways. Niebuhr et al. (2011a), for instance, report data from North German and Neapolitan Italian showing that some speakers signal a pitch accent category via f0 shape, whereas others manipulate tonal alignment (see also Grice et al. 2017 on individual variation in the realization and interpretation of pitch accents in German). Tonal alignment may also vary depending on dialect (e.g. Grabe et al. 2000 on varieties of English; Atterer and Ladd 2004 on varieties of German), and even the amount of voiced material available (Baltazani and Kainada 2015 on Ipiros Greek; Grice et al. 2015 on Tashlhiyt Berber). There may also be variation in the degree of rising or falling around the tone target, or general pitch scaling differences depending on where the target occurs in an utterance (IP, ip, or AP), degree of speaker emphasis, dialect, or speaker-specific speaking style. Phrase accent and boundary tone targets also vary in terms of their phonetic realization even across typologically related varieties. The classic fall-rise tune of Australian English H* L-H% is often realized somewhat differently from the same phonological tune in Standard Southern British English. The L-H% represents a final terminal rise in both varieties but scaling of the final H% tone tends to be somewhat higher in Australian English and is often described as a component of 'uptalk' (Warren 2016; see Fletcher and p. 91 ↳ Stirling 2014 and chapter 19 for more detailed discussion). It follows from the preceding discussion that the actual phonetic realization of tonal elements can be gradient with respect to both alignment and scaling. Listeners may not necessarily interpret the different realizations as indicative of different pragmatic effects, suggesting that there is no need to posit additional contrastive categories to represent this variation. It is therefore important that a phonetic model for any language or language variety can account for this kind of realizational variation (for a proposal on how to do so, see Arvaniti 2019 and chapter 9).

6.3.3 Interpolation and tonal crowding

Interpolation is a critical component of AM, as it is directly linked to the important AM tenet that melodies are not fully specified either at the phonological or the phonetic level, in that the f0 on most syllables in an utterance is determined by the surrounding tones. The phonological level involves a small number of tones; at the phonetic level, it is only these tones that are planned as tonal targets, while the rest of the f0 contour is derived by interpolation between them. The advantages of modelling f0 in this manner were first illustrated with Tokyo Japanese data by Pierrehumbert and Beckman (1988: 13 ff.). They showed that the f0 contours of AP's without an accented word could be modelled by positing only one H target, associated with the AP's second mora, and one L target at the beginning of the following AP; the f0 slope from the H to the L target depended on the number of morae between the two. This change in the f0 slope is difficult to model if every mora is specified for f0.

Despite its importance, interpolation has not been investigated as extensively as alignment and scaling. The interpolation between targets is expected to be linear and can be conceived of as the equivalent of an articulator's trajectory between two constrictions (cf. Browman and Goldstein 1992b). An illustration of the mapping between phonological structure and f0 contour is provided in (4), where the open circles in the f0 contour at the bottom represent tonal targets for the four tones of the phonological representation. As mentioned in §6.2.3, the L- phrase accent in this melody shows a secondary association to the end of the word with the nuclear accent, here 'ballet', and thus has two targets, leading to its realization as a stretch of low f0 (for a detailed discussion, see Pierrehumbert and Beckman 1988: ch. 6).

(4)



p. 92 One possible exception to linear interpolation is the 'sagging' interpolation between two H* pitch accents discussed in Pierrehumbert (1980, 1981) for English; sagging interpolation is said to give rise to an f0 dip between the two accentual peaks. It has always been seen as something of an anomaly, leading Ladd and Schepman (2003) to suggest that in English its presence is more plausibly analysed as the reflex of a L tone. Specifically, Ladd and Schepman (2003) proposed that the pitch accent of English represented in Pierrehumbert (1980) as H* should be represented instead as (L+H)*, a notation implying that in English both the L and H tone are associated with the stressed syllable.

Independently of how sagging interpolation is phonologically analysed, non-linear interpolation is evident in the realization of some tonal events. For instance, L*+H and L+H* in English differ in terms of shape, the former being concave and the latter convex, a difference that is neither captured by their autosegmental representations nor anticipated by linear interpolation between the L and H tones (Barnes et al. 2010b). In order to account for this difference, Barnes et al. (2012a, 2012b) proposed a new measure, the Tonal Center of Gravity (see chapter 9).

Further, although it is generally assumed that tones are realized as local peaks and troughs, evidence suggests this is not always the case. L tones may be realized as stretches of low f0, a realization that may account for the difference between convex L+H* (where the L tone is realized as a local trough) and concave L*+H (where the L tone is realized as a low f0 stretch). Similarly, H tones may be realized not as peaks but as plateaux. In some languages, plateaux are used interchangeably with peaks (e.g. Arvaniti 2016 on Romani), while in others the two are distinct, so that the use of peaks or plateaux may affect the interpretation of the tune (e.g. D'Imperio 2000 and D'Imperio et al. 2000 on Neapolitan Italian), the scaling of the tones involved (e.g. Knight and Nolan 2006 and Knight 2008 on British English), or both (Barnes et al. 2012a on American English). Data like these indicate that a phonetic model involving only targets as turning points and linear interpolation between them may be too simple to fully account for all phonetic detail pertaining to f0 curves or for its processing by native listeners. Nevertheless, the perceptual relevance of these additional details is at present far from clear.

As noted above, the need for interpolation comes from the fact that the phonological representation of intonation is sparse; for example, 'a ballet aficionado' in (2) has eight syllables but the associated melody has a total of four tones. Nevertheless, it is also possible for the reverse to apply—that is, for an utterance to have more tones than TBUs; 'Lou?' uttered with the same L*+H L-H% tune (as in Figure 6.1b) is such an instance, as four tones must be realized on one syllable. In AM, this phenomenon is referred to as 'tonal crowding'. Tonal crowding is phonetically resolved in a number of ways: (i) 'truncation', the elision of parts of the contour (Bruce 1977 on Swedish; Grice 1995a on English; Arvaniti et al. 1998 and Arvaniti and Ladd 2009 on Greek; Grabe 1998a on English and German); (ii) 'undershoot', the realization of all tones without them reaching their targets (Bruce 1977 on Swedish; Arvaniti et al. 1998, 2000, 2006a, 2006b on Greek; Prieto 2005 on Catalan); and (iii) temporal realignment of tones (Silverman and Pierrehumbert 1990 on American English; Arvaniti and Ladd 2009 on Greek). Undershoot and temporal realignment often work synergistically, giving rise to 'compression'. Attempts have been made within AM to pin different resolutions of tonal crowding to specific languages (Grabe 1998a). Empirical evidence, however, indicates that the mechanism used is specific to elements in a tune, rather than to a language as a whole (for discussion see Ladd 2008b; Arvaniti and Ladd 2009; Arvaniti 2016). p. 93 ↳ Arvaniti et al. (2017) proposed using tonal crowding as a diagnostic of a putative tone's phonological status, as it allows us to distinguish optional tune elements (those that are truncated in tonal crowding) from required elements (those that are compressed under the same conditions).

6.4 Applications of AM

The best-known application of AM is the family of ToBI systems. ToBI (Tones and Break Indices) was a tool originally developed for the prosodic annotation of American English corpora (Silverman et al. 1992; see also Beckman et al. 2005; Brugos et al. 2006). Since then several similar systems have been developed for a variety of languages (see e.g. Jun 2005a, 2014a for relevant surveys). In order to distinguish the system for American English from the general concept of ToBI, the term MAE_ToBI has been proposed for the former (where MAE stands for Mainstream American English; Beckman et al. 2005). ToBI was originally conceived as a tool for research and speech technology; for example, the MAE_ToBI annotated corpus can be searched for instances of an intonational event, such as the H* accent in English, so that a sample of its instantiations can be analysed and generalizations as to its realization reached. Such generalizations are useful not only for speech synthesis but also for phonological analysis and the understanding of variation (for additional uses and extensions see Jun 2005c).

ToBI representations consist of a sound file, an associated spectrogram and a pitch track, and several tiers of annotation. The required tiers are the tonal tier (a representation of the tonal structure of the pitch contour) and the break index tier (in which the perceived strength of prosodic boundaries is annotated using numbers). In the MAE_ToBI system, [0] represents cohesion between items (such as flapping between words in American English), [1] represents the small degree of juncture expected between most words, [3] and [4] represent ip and IP boundaries respectively, and [2] is reserved for uncertainty (e.g. for cases where the annotator cannot find tonal cues for the presence of a phrasal boundary but does perceive a strong degree of juncture). A ToBI system may also include an orthographic tier and a miscellaneous tier for additional information, such as disfluencies. Brugos et al. (2018) suggest incorporating an 'alternatives' tier, which allows annotators to capture uncertainty in assigning a particular tonal category. The content of all tiers can be adapted to the prosodic features of the system under analysis, but also to particular research needs and theoretical positions of the developers. For instance, Korean ToBI (K_ToBI) includes both a phonological and a phonetic tier (Jun 2005b), while Greek ToBI (GR_ToBI) marks sandhi, which is widespread in Greek and thus of phonological interest (Arvaniti and Baltazani 2005).

ToBI as a concept has often been misunderstood. Some have taken ToBI to be the equivalent of an IPA alphabet for intonation, a claim that the developers of ToBI have taken pains to refute (e.g. Beckman et al. 2005; Beckman and Venditti 2011). A ToBI annotation system presupposes and is meant to rely on a *phonological* analysis of the contrastive elements of the intonation and prosodic structure of the language or language variety in question. ToBI can, however, be used as an informal tool to kick-start such an analysis on the understanding that annotations will have to be revisited once the phonological analysis is complete (Jun and Fletcher 2014; Arvaniti 2016).

6.5 Advantages over other models

AM offers a number of advantages, both theoretical and practical, relative to other models. A major feature that distinguishes AM is that as a phonological model it relies on the combined investigation of form and meaning, and the principled separation of phonological analysis and phonetic exponence. The former feature distinguishes AM from the Institute for Perception Research (IPO) model ('t Hart et al. 1990), which focuses on intonation patterns but strictly avoids the investigation of intonational meaning. The latter feature contrasts AM with systems developed for the modelling and analysis of f0—such as PENTA (e.g. Xu and Prom-on 2014), INTSINT (Hirst and Di Cristo 1998), or Fujisaki's model (e.g. Fujisaki 1983)—which do not provide an abstract phonological representation from which the contours they model are derived. As argued elsewhere in some detail (Arvaniti and Ladd 2009), the principled separation of phonetics and phonology in AM gives the theory predictive power and allows it to reach useful generalizations about intonation and its relation to the rest of prosody, while accounting for attested phonetic variation.

In terms of phonetic implementation, the target-and-interpolation modelling of f0 allows for elegant and parsimonious analyses of complex f0 patterns, as shown by Pierrehumbert and Beckman (1988) for Japanese. AM can also accommodate non-linear interpolation, unlike the IPO ('t Hart et al. 1990). In addition, although tonal crowding is extremely frequent cross-linguistically, AM is the only model of intonation that can successfully handle it and predict its outcomes (see e.g. Arvaniti and Ladd 2009 and Arvaniti and Ladd 2015 for comparisons of the treatment of tonal crowding in AM and PENTA; see also Xu et al. 2015).

Further, by relying on the formal separation of metrical structure and the tonal string, AM has disentangled stress from intonation. This has been a significant development, in that the effects of stress and intonation on a number of acoustic parameters, particularly f0, have often been confused in the literature (see Gordon 2014 for a review and chapter 5). This confusion applies both to documenting the phonetics of stress and intonation, and developing a better understanding of the role of intonation in focus and the encoding of information structure. Research within AM has shown that it is possible for words to provide new information in discourse without being accented, or to be accented without being discourse prominent

(Prieto et al. 1995 on Spanish; Arvaniti and Baltazani 2005 on Greek; German et al. 2006, Beaver et al. 2007, and Calhoun 2010 on English; Arvaniti and Adamou 2011 on Romani; Chahal and Hellmuth 2014b on Egyptian Arabic).

Finally, since AM reflects a general conceptualization of the relationship between tonal elements on the one hand and vowels and consonants on the other, it is sufficiently flexible to allow for the description of diverse prosodic systems—including systems that combine lexical and post-lexical uses of tone—and the development of related applications. In addition to the development of ToBI-based descriptive systems, as discussed in §6.4., such applications include modelling adult production and perception, developing automatic recognition and synthesis algorithms, and modelling child development, disorders and variation across contexts, speakers, dialects, and languages (see e.g. Sproat 1998 on speech synthesis; Lowit and p. 95 Kuschmann 2012 on intonation in motor speech disorders; ↳ Thorson et al. 2014 on child development; Gravano et al. 2015 on prosodic entrainment and speaker engagement detection; Kainada and Lengeris 2015 on L2 intonation; Prom-on et al. 2016 on modelling intonation; see Cole and Shattuck-Hufnagel 2016 for a general discussion). In conclusion, AM is a flexible and adaptable theory that accounts for both tonal phonology and its relation to tonal phonetics across languages with different prosodic systems, and it can be a strong basis for developing a gamut of applications for linguistic description and speech technology.