Comparing PENTA to Autosegmental-Metrical Phonology

Janet B. Pierrehumbert
University of Oxford

Feb 16, 2017

1 Introduction

In this volume, Xu, Prom-on and Liu provide an overview of the recent work on the PENTA (parallel encoding and target approximation) model of prosody. This is a third-generation model of prosody and intonation. In characterizing the model this way, I am taking classic works based on auditory transcriptions, such as Bolinger (1958); Trager & Smith (1951); Crystal (1969), as first-generation models, and the autosegmental-metrical models (AM models) launched in the 1970's and 1980's as second-generation models. Second-generation models benefited enormously from the rise of computer workstations with specialized software for speech processing, which enabled researchers to examine thousands of f0 contours and to create experimental stimuli in which melodic characteristics of speech were varied in a controlled manner. However, the developers of AM models did not yet have the inference and optimization methods that have played such a central role in the development of PENTA. Generative AM algorithms, such as those laid out in Anderson et al. (1984); Beckman & Pierrehumbert (1986), were very seat-of-the pants efforts, compared to the multi-parameter trajectories that are achieved with PENTA. So, we can ask whether PENTA has superseded the AM approach? To what extent has it built on insights of the earlier approach? What aspects of the AM research program remain even now topics for future research?

A central issue in addressing these questions is the comparison between prosody and word phonology. AM theory, a development within generative linguistics, had as one of its goals a unified formalism for describing segmental, rhythmic, and melodic aspects of speech at both the lexical and the phrasal levels. AM theory adopted from phonemic theory the claim that

each language has a relatively small inventory of phonological units, that these are combined with each other in accordance with language-specific constraints (which define the phonological grammar of the language), and the mapping between form and meaning has a great deal of arbitrariness. *Go* is a verb of motion in English, but a board game in Japanese. *Puck* is a disc used in ice-hockey, but with the change of a single feature, the word is obscene.

This arbitrariness, in combination with the relative simplicity of phonology, means that the phonological level defines a bottleneck between the phonetic realizations of words and their meanings. This bottleneck helps the language learner to acquire a large vocabulary by allowing articulatory and perceptual patterns exhibited in one word to be reused in other words. It promotes fast, accurate speech perception because words that are highly confusable from a phonetic point of view are unlikely to be similar in meaning, and thus unlikely to be lexical competitors in any given context. These benefits accrue precisely because the flow of information from the extremely rich, high-dimensional, world of meanings and communicative functions to the phonetics is severely restricted. (See Beckman & Pierrehumbert (2000) for further discussion). Insofar as intonational patterns resemble word forms, then, they should be describable in terms of combinations of phonological units that could convey different meanings in some other combinations. Pierrehumbert & Hirschberg (1990) is an effort to provide a semantics for intonational patterns that has such a structure. Xu et al. (in press) represents a fundamental departure. Starting from a negative assessment of the extent of convergence in the field about the units of analysis, it proposes direct mappings from prosodic or intonational meanings ("communicative functions") to phonetic parameters. This mapping does not have the same modules as those assumed in AM theory.

AM differed from much earlier work in phonology, such as classical phonemic theory or Chomsky & Halle (1968), in positing units that are phonetically realized at many different time scales. This development was a response to observations about the relationship of sound structure to meaning. The job of phonology is to describe systems of contrast – differences in form that speakers can use to communicate differences in meaning. It became evident that some contrasts are available in fewer positions than others. For example, it is very common for languages to support a full set of consonantal contrasts in syllable onset position, but only a restricted set in syllable coda position. English and Russian both have a larger set of vocalic contrasts in stressed syllables than in unstressed ones. In analyzing African tone languages, Leben (1971); Goldsmith (1976) found that, while only vowels could

carry distinctive tones, the number of distinctive tones in a word did not correspond to the number of vowels. Instead, the number of melodies for an entire in word in languages such as Mende or Hausa is quite restricted, and in long words a tone is spread out over two or more syllables. This observation naturally leads to the suggestion that the melody is underlyingly a characteristic of the word (and not of the phoneme or syllable), with general principles of tonal mapping intervening between the underlying form and the phonetic outcome. Intonation patterns involve an even larger time scale, as many intonational distinctions can be made only once per phrase. While syllables typically have durations in the range of 100 to 200 msec, depending on their stress and position in the phrase (Arai & Greenberg, 1997; Dankovicova, 1999), intonation phrases are typically longer. The shortest intonation phrases are monosyllables that are amongst the longest syllables (being stressed and phrase-final), while phrases with two to four content words plus any associated clitics are more common, often taking a second or more to produce. A central goal of AM theory is to articulate what contrasts are available at what temporal scales.

In pursuit of this goal, AM theorists proposed units and principles for combining them. Some of the most widely accepted units are: the segment, the mora, the syllable, the metrical foot, and the intonational phrase. Essentially all AM grammars had at least one unit in between the metrical foot and the intonational phrase, such as the prosodic word or the accentual phrase, but the name and nature of this unit varied with the research group and the target language. For the theory of prosody and intonation, it is obvious that a time scale larger than the segment is needed, since distinctive tonal contrasts are sparser than segmental contrasts in every language that has been studied. AM phonology also asked whether the inventory of units is universal, whether tonal contrasts are assigned at the same temporal scale in all languages, and how the units used to describe melody and rhythm relate to those used in the lexical phonology.

The target paper refers to three different temporal scales (the utterance, the word, and the syllable) although the paper does not take a clear stand on what sizes of units are available. The syllable is held to be a privileged universal unit for the planning and execution of speech output, while some communicative functions are selected at a much larger time scale, such as the utterance.

To summarize, then, the target paper has fewer levels of representation for prosody and intonation than AM theory does. It relates communicative functions directly to phonetic outcomes. AM theory asserts one set of relations between meanings and the phonological representation, and a second

and qualitatively different set of relations between the phonological representation and the phonetic outcome. Further, Xu et al. (in press) has a less elaborated theory of the units of sound structure, assigning a role to the syllable that is more privileged than its role in AM theory.

2 Semantics versus phonology

The flat representional apparatus of PENTA entails that prosodic meanings have qualitatively different behaviour from word meanings. For word meanings, there is widespread agreement that an intermediate level of representation, in between semantics and phonetics, is needed. This agreement is based on the dissociations observed between phonetic similarity and semantic similarity. By defining an intermediate level of representation, two distinct systems of relations are made available in the model. One associates meanings with phonological forms, while the other associates phonological forms with parametric outcomes. The phonological level reduces the dimensionality of the system to be learned. It acts as a bottleneck, blocking arbitrary mappings of semantic features to phonetic parameters. For example, features of meaning such as spiciness, permanency, or biodegradability have no direct claims on f0 or voice quality. To express these meanings, a speaker must select a suitable word, and then access the phonological representation of the word, whose limited properties in turn control the phonetic realization. In PENTA, it is claimed that communicative functions are mapped directly to quantitative parameters. The framework has two levels of representation instead of the three. The PENTA encoding system, described as being similar to morphemes, has the job of capturing the relationship of these two levels. By using only a single set of relations, PENTA is claiming that the prosodic system lacks the properties that have historically motivated an intermediate level of representation for the segmental makeup of words. To assess this claim, we can ask to what extent intonational meanings do, or don't, resemble the meanings of words.

Probably the easiest example of natural language semantics is imageable nouns: a word form for an imageable noun is associated with the set of objects or events that provide examples of the relevant category. Despite the undeniable existence of some idiophonic words, in general the similarities amongst the forms and referents of nouns are poorly correlated, a property of linguistic systems referred to as "duality of patterning". *Spat* has no particular semantic relationship to nouns made by minimal changes in its form (e.g. *spam, scat*). In sound, *pine* might be confusable with *tine*, but

in meaning, a pine is more similar to a cedar or a spruce than to a tine. Other types of words have more elusive meanings that are very discourse dependent. Words like *indeed* and *even* are interpreted with regard to the space of possible events or states of affairs that the speaker assumes to be in the listener's mind. However, words with such slippery semantics still provide plenty of examples of dissociations between form and meaning. For example, the three intensifers *really, super, awfully* are each more similar in sound to other semantically unrelated words (eg *lily, tuber, Aussie*) than they are to each other. That is, they share a communicative function – asserting that something has some characteristic to a high degree – while differing greatly in form.

In languages with lexical tones, these participate in similar dissociations. They can be identified by finding minimal pairs that differ only in their tone, but refer to unrelated concepts. The well-known *ma* set of Chinese even provides a minimal quadruplet. This behavior is what leads Beckman & Pierrehumbert (1986) to treat Japanese, traditionally characterized as a "pitch accent language" as a tone language with a sparse assignment of tones. Two otherwise identical words with unrelated meanings may differ only in the presence or absence of a HL contour in the lexical representation. In languages such as English where the location of lexical stress can vary, similar observations can be made about stress. *Coral, vermillion* and *tangerine* are very similar colors, but they have the main stress in different positions: initial, medial and final position (with *tangerine* actually having two stresses). The trochaic pattern of *coral* is the single most common stress pattern of English, meaning that words with this stress pattern have no shared communicative functions. In short, lexical tone and lexical stress act like phonological elements in that they are poorly correlated with semantic dimensions of meaning, while being highly predictive of the phonetic realization of a word. Lexical tone plays an important part in determining the f0 contour and voice quality, while lexical stress has complex ramifications for duration, force of articulation, and timing. In PENTA, lexical tone and lexical stress are described as "lexical communicative functions" because they distinguish words from each other. By this standard, all of segmental phonology should also be included in "the lexical communicative functions". For example, the phonological feature [+voice] distinguishes *pin* from *bin* and *coat* from *goat*. However, Xu et al. (in press) explicitly denies that the communicative function level corresponds to the phonological level, instead drawing a parallel between the PENTA encoding schemes and morphemes.

The focal communicative function may at first appear to be a more promising candidate for direct mapping from function to phonetic outcome.

I will discuss focus in detail because it is both a key example of a communicative function in PENTA, and is a central topic in semantic and pragmatic theory. A starting point for the semantic treatment of focus is the fact that it can change the truth conditions for sentences, just as substituting one word for another can change the truth conditions for a sentence. An example from Rooth (1992) illustrates this fact.

(1) I only said that Carl likes HERRING.

(2) I only said that CARL likes herring.

Sentence (1) is false if I also said that Carl likes some relevant food other than herring, such as kippers. Sentence (2) is false if I also said that some other relevant person, such as Sam, also likes herring. These differences are related to the differing presuppositions of the sentences. Sentence (1) presupposes I said that Carl likes something. Sentence (2) presupposes I said that somebody likes herring. The instantiation of "something" and "somebody", respectively, are produced with narrow focus. The many other words whose contribution to sentential semantics depends on the focus placement include *even, not, which, always.*

PENTA would not have serious trouble with these particular examples. In the model, focus is treated as a unified communicative function, by which novel information is made phonetically prominent and other information is backgrounded. In these examples the novel information and the phonetic prominence coincide. But is this the case in general?

It turns out that the locations of novel information and of phonetic prominence do not always coincide. In AM theory, focus provides one of the key arguments for a modular theory distinguishing the semantic/pragmatic, phonological, and phonetic levels of representation. The focus feature +F is a semantic feature, associated with new information that instantiates variables in the presuppositions of a sentence. The accentedness within the phrase is a phonological feature, associated with phonetic prominence. The association +F with the locations of accents is rather complex.

One complication arises from the existence of examples in which given material is accented. In a study of the Boston Radio News corpus, Shattuck-Hufnagel et al. (1994) show that the locations of prenuclear accents are determined mainly by phonological constraints and not by information status, and Gussenhoven (1999) also argues that prenuclear accents are not semantically informative in English. The issue is not confined to prenuclear accents, however. Schwarzschild (1999) discusses cases such as (3). Here capitalization is used to indicate the nuclear accents, and one of these, the second instance of *MOON*, falls on given material – indeed its entire phrase is given.

(3) The rising of the TIDES % depends on the MOON being full, % and the MOON being full % depends on the position of the SUN %.

A production study by German et al. (2006), following up observations by Ladd (1980), shows that speakers often place the nuclear accent on a given noun in preference to a new sentence-final preposition. These same cases also show that new information can fail to be accented. Finally, the theory also needs to cover cases in which focussed information is old, but still displays the characteristic behaviour of focus in defining the scope of a semantic operator. This situation is illustrated in (4).

(4) Speaker A: Everybody already knew that Mary only eats VEGETABLES.

Speaker B: If even PAUL knew that Mary only eats vegetables %, why didn't he suggest a different restaurant?

Here, the nuclear accent in Speaker B's first clause falls on *Paul*, and *vegetables* is deaccented even though it is the focus for *only*. As shown by Beaver et al. (2007), weak phonetic reflexes of the +F feature support the idea that focus is present, but they are very slight compared to those for a first-occurrence focus with nuclear stress.

Further complications arise from the fact that nuclear accent on a word is interpreted in some cases, but not others, as a focus marking for an entire phrase containing that word. This phenomenon, the contrast between the "narrow focus" and "broad focus" interpretations of an utterance, displays a strong dependence on syntactic factors, as in the study by Birch & Clifton (2002). Birch & Clifton (2002) explores how nuclear accents map differentially to broad focus, depending on whether the accented word is a syntactic head, an argument, a modifier, or an adjunct. Vallduvi & Engdahl (1996) documents the ways in which the interaction of focus with syntactic constraints can be be language-specific.

Such dissociations between the semantic feature +F and the location of the nuclear accent in the phrase provide a classic argument for an intermediate level of representation. In view of the role of syntactic factors, I should say "at least one" intermediate level of representation – but here I will concentrate on the phonology. Words that are emphasized even through they are given, such as the second occurrence of MOON in (3), are produced with heightened prominence. In English, words that follow them in the same intonational phrase (here, "being full") are subordinated. Though

the assignment and expression of focus remains an active research area, all competitive approaches to the issue follow Schwarzschild (1999) in using a constraint satisfaction architecture in which the accent placement in any individual phrase results from the interaction of semantic focus with additional factors stated at other levels of linguistic representation.

Because PENTA does not impose any restrictions on the number of communicative functions, nor on the formal power or complexity of mappings between communicative functions and phonetic outcomes, it is in principle capable of describing this rich variety of phenomena. For example, in response to German et al. (2006), the PENTA authors could split apart the communicative functions of "focus on a noun", and "focus on a preposition". In the case of a final preposition, the encoding function could push the phonetic prominence back onto the preceding noun. This approach would miss significant generalizations. It would make it appear coincidental that the outcome is pretty much the same as having focus on the noun itself. It would miss the fact that that the phonological and syntactic entities that are alluded to in a more standard treatment are independently needed to capture many other linguistic patterns. A more conservative response would be for the PENTA authors to allow that the "communicative function" of focus as it PENTA is actually a phonological feature – the location of the nuclear accent in the phrase. This response would pave the way for PENTA to be integrated with modern linguistic theories of focus and accent placement that cover a wider range of phenomena.

A well-established communicative function of intonation that Xu et al. (in press) does not discuss is the scalar implicature. Constructions that involve implicit scales provide one of the most studied areas of semantics and pragmatics. Lexical choices and intonational choices both provide key examples. So I will begin with some examples of lexical choices. The meanings of commonplace adjectives like *tall, good, hot* are rooted in the discourse structure, because each makes a claim about the position of the modified noun with regard to a contextually evoked scale (Kennedy, 2007). In the examples in (5), the standards for *tall, good* and *hot* are relativized to the scale for the relevant comparison set. Sentence (5B) is not contradictory, because the relevant dimensions along which goodness is assessed differ for jack-o-lanterns and pies.

(5A) Max is tall for an American, but short for a basketball player.

(B) Howdon Biggy pumpkins are good for jack-o-lanterns, but poor for pies.

(C) 100 Celsius is hot for a sauna, but not for an oven.

Now if *warm* means having a relatively high temperature, what does a

sentence like (6) mean?

(6) It's warm.

The speaker relies on mutual knowledge of a contextually relevant scale (temperatures of saunas? temperatures of ovens?) and then s/he also implies that the temperature is lower than *hot* on that scale: If it really was hot, it would have been more cooperative and informative to say so. Thus if the sauna is heating up, (6) would be appropriate if the temperature is 70 but not yet 100, but if the oven is heating up, the sentence might be used if the temperature is 120 but not yet 180.

In English intonational phonology, the interpretation of the rise-fall-rise contour (transcribed as L*+H L H% in the ToBI system (Silverman et al., 1992)) also depends on the contextually relevant scale of comparison. A naturalistic study (Ward & Hirschberg, 1985), a theoretical analysis (Pierrehumbert & Hirschberg, 1990), and a perception experiment (Hirschberg & Ward, 1992) all converge on the conclusion that speakers use this contour to implicate uncertainty about the scale or about a scalar value. The speaker may be genuinely uncertain, or may implicate uncertainty to express irony or disbelief. The following example, which I used as part of a classroom exercise developed with Mary Beckman, illustrates this communicative function.

(7) Speaker A: Great America lets you ride the Vortex only if you're over 5' 3.

Speaker B: MELANIE L*+H rides the vortex, and she's 5' 2.

Here, Speaker A evokes a scale of acceptability as a Vortex passenger, based on height. Speaker B expresses their reservations about the scale by challenging the entailment that Melanie falls too low. S/he uses the the L*+H accent to underline this reservation. In our classroom exercise, we have found that beginning linguistics students reliably associate such rejoinders with a phrase expressing disbelief ("Really?") as opposed to a word expressing agreement ("Indeed!"). "Indeed!" is perceived as consistent with a H* accent. This example provides a good example of a pragmatic meaning that can be conveyed with a word choice, an intonational choice, or both. Other well-known examples of such meanings include turn-taking (does the speaker mean to hold the floor, or yield it?) and topic shifts. These can also be signalled with words (*Now, ...*, *On the other hand ...*), with manipulations of the pitch range and accent choice, or with combinations of these (Hirschberg & Litman, 1993).

Such observations mean that when the speaker wishes to convey any of

these meanings, s/he makes a selection amongst the means available. Some combinations of meanings and forms of expression are mutually exclusive. For example, the speaker cannot end a sentence with two different words simultaneously. (8) provides an illustrative of mutual exclusivity in the intonational domain.

(8) Are you coming or NOT H* L L%.

This is a yes/no question, and according to Xu et al. (in press), its communicative function should be conveyed with a rising f0 pattern in English. However, the sentence is also an exhaustive disjunction. In (8), which has the most typical intonation for this construction (Pruitt & Roelofsen, 2013), the falling terminal pattern indicates that the list is complete. Its use supersedes the option of using a rising pattern to indicate that the truth value of the proposition under discussion is not known, and input from the listener is expected (c.f. Pierrehumbert & Hirschberg (1990)). Because of the phonological bottleneck provided by the small number of tonal elements available at the end of the phrase, the speaker cannot simultaneously signal a complete list and an open proposition.

Because the encoding functions in PENTA are so unrestricted, and indeed are designed to handle cases in which gestures associated with different functions seem to be overlaid, PENTA says nothing about how some choices may be mutually exclusive. AM theory, in contrast, specifies a particular set of bottlenecks, including for each language the inventory of contrastive pitch accents or lexical tones, the constraints on metrical structure, and restrictions on the pileup of contrastive tonal events at phrase boundaries. These have the effect of limiting how many distinct dimensions of meaning can be conveyed by the prosody and intonation of any given expression.

I have sketched some ways in which lexical tone and intonation resemble segmental features in the lexicon, and the relation of focus to nuclear accent points to a distinction between the semantic feature +F and the phonological feature of nuclear accent. I've also suggested that there is significant overlap between prosodic and intonational meanings, on the one hand, and the meanings that words have, on the other hand. It is not clear, however, that prosodic and intonational meaning are really as arbitrary as what we see for segmental phonology. This question is difficult to address, particularly in the light of recent work on lexical iconity and the emergence of conventions in sign languages (Meir et al., 2013; Dingemanse et al., 2015; Dingelmanse et al., 2016). However, Xu et al. (in press) and AM theorists are in agreement that prosodic and intonational systems are language specific, and therefore must be learned. For example, a key reference in Xu et al. (in press), Rialland (2009), documents the association between low

tone and questioning in a group of African languages, the opposite of what is found in many other language families. Although it is frequently assumed that stressed syllables universally are marked with high tones, this is not the case for Indic languages such as Bengali (Hayes & Lahiri, 1991). Nor does the high tone have a constant meaning even within English. Depending on where a high tone (or f0 peak) occurs, it may be associated with a contrastive emphasis (the L+H* accent), a yes-no question (the terminal H%), a scalar implicature (the L*+H discussed above), or a change of topic (see Hirschberg & Litman (1993)).

3 Phonology and phonetic realization

Acknowledging that tones are specified more sparsely than segmental features, PENTA takes the syllable to be the domain of tonal specification. A substantial body of research in phonology and phonetics indeed treats the syllable as a privileged unit for language acquisition, motor planning, and speech perception (MacNeilage, 1998). AM theory has not, in general, viewed the syllable as privileged to this extent. In AM theory, the syllable is one of a hierarchy of different prosodic units controlling contrasts at different time scales, and these interact to determine the f0 contour for any given time span. This hierarchy is motivated by assessment of contrastiveness, on the one hand, and by different time scales for phonetic realization, on the other.

The component of AM theory that deals with contrastiveness is covered by the rubric "prosodic licensing". The general idea is that phonological representations include a structural skeleton, and different pieces of the content are associated with different parts of the skeleton. In much the same way, a house is a structure with an entrance and an optional stairwell; a doorknocker must go with an entrance, and a banister goes with a stairwell. Each structural position thus provides an opportunity to make some set of choices – a doorknocker or not? a high vowel or a low vowel? an offglide or not? Admittedly sometimes the choice is a Hobson's choice, as in the case of Model T Fords (structures which were always painted black) and the Tokyo Japanese accentual phrase (which always begins with a L tone). But in any case, if the rate at which some feature is selected is relatively rapid, then the feature is licensed by some smaller unit. If it is specified less often, then it is licensed by a bigger unit. Units that do not in themselves license a feature may still carry information about it that can be exploited in speech perception. Xu et al. (in press) defends the choice of the syllable as the tone-

bearing unit in PENTA in part on the grounds that unstressed syllables in post-focal position have different contours than stressed syllables in English, indicating the presence of f0 targets where AM theory does not posit any pitch accents. This is an interesting phenomenon. Previously, Grice et al. (2000) reported that the phrase accent in the ToBI framework can dock on a post-nuclear stressed syllable instead of at the phrase boundary, leading to differences between stressed and unstressed syllables in post-nuclear position. Beaver et al. (2007) observe a marginal effect of secondary focus in postnuclear position on the F0 range and minimum f0 of the focussed syllable. Such observations indicate that phonetic realization systems like Anderson et al. (1984) are overly simplistic. They do not, however, demonstrate that the post-focal stressed syllables support an independent choice of contour type. Insofar as the postnuclear stressed syllables display traits that are licensed at the phrasal level, an AM theorist might compare the situation to seeing a reflection of the banister in the hall mirror. The banister is a property of the stairway. If there is a hall mirror, it can indirectly reveal what the banister looks like, but it is not possible for the decorator to make distinct selections for the banister and the reflection of the banister in the mirror.

The concept of licensing is applicable both in the lexical phonology and in the phrasal phonology. In many languages, the set of consonantal contrasts available in the coda position is smaller than in onset position. For example, in the middle of the word in Japanese, the coda consonant (if any) must either be a nasal sharing the same place of articulation as the following onset, or else the first part of a geminate. This means that the coda position licenses only the feature [+/- nasal] (Ito et al., 1995). In English, the full range of vowels is found only in stressed syllables in a foot. Unstressed syllables do not license all the vowel features. But the metrical foot is also the minimal domain for a pitch accent. This means that if a phrase is just one word with two metrical feet, such as *tangerine*, it can have both a nuclear and a prenuclear accent. But a sentence such as *It's a plane*, when produced with all the the function words cliticized (and therefore unstressed), has only one accent. In the African tone languages I already discussed, it is not possible to make a fresh choice of tone with every syllable; a limited number of melodies is assigned to the whole word. In Japanese, the word is also the domain for the pitch accent; a word either has a HL melody, or it doesn't. Such licensing constraints mean that Mende, Hausa, and Japanese contrast with languages such as Thai, in which every syllable may have its own tone. At both lexical and the phrasal level, the edges of domains may have a special status for licensing. While English has rather restricted consonant

clusters within monomorphemic words, extra coronal obstruents are allowed at the end of the word. This gives us word-final consonantal pile-ups, as in *fifths* and *folds*. Similarly, the end of an intonation phrase provides an opportunity for an extra tonal event. In the middle of the phrase, it is impossible to have so much f0 movement on a single syllable. As a result, the presence of a complex contour can provide a cue to listeners that a phrase boundary exists, as shown in Streeter (1978).

The licensing system and the phonetic requirements are not always well-aligned. Misalignments can occur because the licensed elements are too dense, in relation to the speed of implementation that is possible from an articulatory point of view. They can also occur because the elements are too sparse, failing to specify some phonetic properties that are essential for the speech to be produced. Both of these discrepancies are addressed in AM theory and in PENTA. The approaches have some similarities and differences, and both approaches leave some problems unresolved.

If a feature is licensed at a very high density, the realization of one value may overlap that a preceding or following value. A much-studied case of this situation is nasal coarticulation, in which lowering of the velum for a nasal consonant begins during a preceding non-nasal segment. The exact timing and extent of this effect differs from language to language, providing an important case of language-specific detailed phonetic learning (Beddor & Krakow, 1999). In the domain of tone and intonation, Bruce (1977) already drew attention to the articulatory undershoot found in Swedish when a phrasal tone is crowded with a lexical accent. For English, Anderson et al. (1984) uses linear smoothing to approximate some consequences of tonal crowding. PENTA builds on the insights of these AM papers in its treatment of tonal crowding, but goes much further with its approach to characterizing the detailed f0 dynamics.

However, an important unsolved problem in AM theory for handling tonal crowding also remains in PENTA. As shown in Silverman & Pierrehumbert (1990), tonal crowding may be ameliorated by temporal displacement of tones instead of by undershoot (an in-depth discussion of the issue may be found in the same volume: Bruce (1990)). The extent and interaction of these two options is not fully elucidated in AM theory nor in PENTA, and may await a more explicit motor theory of tonal production. A particular difficulty for PENTA in handling tonal displacement as a solution to crowding is the claim (MS p. 5) that the inertia of the articulators only comes into play at one level – there is no representation of inertia in the brain. This claim appears to preclude the use of feedforward motor control strategies that are known to be acquired with intense practice, and that appear nec-

essary to model language-particular segmental coarticulation patterns (see Rosenbaum (2010) for a review of closed loop, open-loop and feedforward motor control models.

Xu et al. (in press) also admits that PENTA does not yet handle cases in which more than one distinctive tonal target is assigned to a single syllable. This occurs in AM theory when a pitch accent and one or more boundary tones are all assigned to a monosyllabic utterance, such as *Anne L*+H L%*. By virtue of its licensing theory, AM theory defines equivalence classes amongst the f0 contours of utterances with very different numbers of syllables. *Anne L*+H L H%* , *Melanie L*+H L H%* and *Melanie L*+H did it L H%* all count as having instances of the same contour type, with the scalar implicature discussed above. PENTA does not make this equivalence because it does not yet cover the shortest utterances with complex contours.

If the licensed features are misaligned with the phonetic requirements because they are too sparse, AM theory provides two different responses. One is categorical copying or spreading at the phonological level. The other is phonetic interpolation, where the time in between fully specified locations is taken up by making transitions with more or less alacrity.

Turkish vowel harmony provides an example where the first approach is justified; the harmonizing features that are lacking on the vowels in a the underlying representation of a suffix are spread from the stem. In the nonconcatenative Arabic morphological system, both vowels and consonants can be copied to fill in all the structural positions supplied by the morphological pattern. In languages with reduplication, it is common for an affix to be merely a structural template, which is filled by copying material from the stem. The original AM treatments of African tone languages similarly asserted that tones were spread or copied in order to supply f0 targets for otherwise toneless vowels, as vocalic articulations require a specification of the source. The second approach is taken in Keating (1988)'s treatment of the tongue position for the phoneme /h/. The tongue position differs greatly between /ihi/ and /aha/, because /h/ lacks intrinsic oral specifications such as [high] or [front]. In Anderson et al. (1984); Silverman et al. (1992); Pierrehumbert & Beckman (1988), the option of underspecifying syllables for tones is heavily exploited. For both English and Japanese, according to these models, the alacrity of transitions between contrastive tonal specifications depends systematically on how far separated these specifications are; if people have plenty of time to make the transition, they adjust the articulators more slowly. The status of some of these analyses seems clearcut, but not all. The unrestricted character of phonetic realization rules in AM theory sometimes makes it difficult to distinguish between phonological and

phonetic analyses of the same pattern. Notably, in Hobson's choice cases I alluded to above, nothing really prevents the suggestion that the entire f0 contour is supplied by the phonetic realization component, without any phonological representation of the melodic features at all.

PENTA avoids the vexed distinction between phonological copying or spreading, and phonetic realization, by denying the existence of a separate phonological level. In cases where a communicative function is sparsely specified (eg, focus is assigned to only one word per phrase), but has wide-ranging effects (both the focussed word, and neighboring words, are affected), the encoding of the communicative function calculates a contour for each syllable in the affected sequence. The syllable has a special status for planning the output, but does not have the same status in the set of communicative functions, which are specified at the syllable, word, and phrasal levels, depending on the function.

The encoding manipulates several gradient parameters. It shares with Fujisaki & Hirose (1984) and Garding (1987) the claim that local gestures (such as lexical tones) interact with larger scale planning, which has its own dynamics. However, unlike these works, PENTA does not superimpose local contours on a phrasal contour. Instead, phrasal choices (eg focus and modality) induce modulations of the pitch range in a manner that is reminiscent of Liberman & Pierrehumbert (1984) and Pierrehumbert & Beckman (1988). Like Kochanski & Shih (2003), the strength or weakness of different gestures – representing the strength of the requirement to achieve the specified target – is also gradiently manipulated. Very broadly speaking, the PENTA encoding builds on the pitch range control mechanisms proposed in AM, while also building on insights from other work that support more detailed modelling of the f0 dynamics. Thanks to its method for optimizing the parameters for detailed representations of f0 contours, it can achieve much more reliable matches to natural f0 contours than AM did, and accordingly a better quality in speech synthesis. Overall, PENTA's strong reliance on gradient parameters would cause AM theorists to view it as a new and particularly sophisticated example of a phonetic realization model. This view is reinforced by PENTA's agnostic view about the system of underlying contrasts.

4 Generalization

Generative phonology, like generative linguistics, has the goal of predicting the form of examples that were not used in developing the model. Capturing the capabilities of native speakers of a language, the grammar of the

language should be able to predict and analyze the forms of previously unobserved phrases. Predictions about language typology should also follow from varying the parameters in the overall formal framework.

For AM theory, the primary test of with-in language generalization was intonation synthesis for novel phrases and sentences. A number of intonation synthesis algorithms using the framework were demonstrated and used commercially. Contrary to the claim on Xu et al. (in press) p. X, several AM models do generate f0 contours that are detailed enough for comparisons with real speech. For example, Pierrehumbert & Beckman (1988) Chap. 6 provides exact equations for synthesizing f0 contours of Japanese, and such f0 contours can of course be compared with those in recordings. Pitrelli & M (2003) describes a system in which ToBI labels are associated with communicative functions of questioning and contrastive emphasis. Decision trees were then statistically trained to map these features together with text features into f0 and duration patterns. This system achieved good results in a perceptual evaluation using novel sentences.

Validation of PENTA is presented in Xu & Prom-on (2014), cited in Xu et al. (in press). This paper presents details of the corpora used to train PENTA and cross-validate the parameter estimates. The dataset is a highly constrained set of read text designed to elicit contrasts in focus location and in questions vs statements through the use of prompt sentences. The parameters are trained on all but one of the speakers in the dataset and then used to predict the f0 contours for the held-out speaker. This procedure tests generalization across speakers at the parametric level, a dimension of generalization that is little explored in AM theory.

Validating models on held-out test data is the standard in other areas of speech and language engineering, and the application of this method in prosody and intonation is welcome. However, it is also important to recognize the limitations of this training and validation procedure. First, the prompt sentences represent a minimal discourse context, in comparison to the longer contexts that have led to the discoveries of some communicative functions of intonation that were mentioned above. The corpus does not include the full range of intonational patterns that are found in expressive, colloquial speech. Secondly, optimization algorithms that attempt an overall match to a dataset are very vulnerable to frequency effects, with detailed fits for the more frequent cases often achieved at the expense of coverage of less frequent cases. This can lead to problems with robustness and adaptation in new situations where the frequencies of the different cases may be quite different. This vulnerability was addressed by constructing a corpus with balanced representation of all the functions covered by the model – uncutting

the suggestion that the learning method used in PENTA resembles that used by the cognitive system. By means that are not yet fully understood, humans can learn from highly unbalanced training sets and also adapt to different situations. Finally, because all speakers read the same materials, the study does not evaluate the model's ability to generalize to previously unseen sentences.

AM theory generated claims about language typology. The approach predicted that languages could differ in many independent dimensions. Questions for the phonological grammar of any given language included:

- 1. Underlying differences
  - How many tonal elements does the language have? What are the basic tonal units that recombine in complex contours?
  - Are tones part of the representations of words in the lexicon?
  - Which phonological units license tones, whether syllables, feet, or larger units?
  - Which tones are assigned at the the phrasal (post-lexical) level?
  - Which semantic/pragmatic meanings are associated with which tonal sequences?
- 2. Realizational differences
  - How is the realization of each tone is affected by its metrical and tonal context?
  - How are the tones licensed by each phonological unit aligned with the segments licensed by the same unit?
  - What happens when word-level and phrase-level tones pile up on short material, such as monosyllabic words?

Such questions separate factors that were lumped together in the traditional distinction amongst tone languages, pitch accent languages, and intonation languages. The result was extensive research on the typology of prosody and intonation. Studies established cross-linguistic differences on essentially all dimensions specified by the theory, culminating in works such as Jun (2006).

PENTA has also been used for cross-linguistic comparisons, with the discussion in the target paper focussed on two dimensions of variation. One is the presence versus absence of post-focus compression, which is found in English and Mandarin but not in Taiwanese. The other is the use of low rather

than high pitch to mark interrogation, as documented in Rialland (2009). These comparisons lead Xu et al. (in press) to suggest that prosody and intonation are exceptionally stable, with patterns possibly being preserved for time periods over 10,000 years. This conclusion is at odds, however, with other studies that have found systematic differences amongst dialects that diverged much more recently. Pierrehumbert & Beckman (1988) note differences in the implementation of low tones between Tokyo and Osaka Japanese; Bruce (2004) finds seven distinct intonational dialects of Swedish, and intonational variation amongst dialects of English has also been extensively investigated (Fletcher et al., 2005; Grabe & Post, 2002). Such differences can be discovered because it is possible to identify them using AM theory, with its toolkit of different factors at different levels of representation. The striking persistence of some prosodic and intonational patterns, together with the recent origins of others, provide another point of resemblance to the rest of phonology, in which long-standing patterns and rapid changes are also found (Coleman, 2016).

5 Conclusions

In comparing PENTA and AM theory, a central issue is how much prosody and intonation resemble words in the lexicon. I have argued that there are considerable resemblances. The meanings of intonational contours are similar to those of many words. Word choices and intonational choices affect the truth values of sentences. They also reflect pragmatic factors, such as the universe of comparisons in a discourse context. In some cases speakers have a choice of communicating their meaning with a word choice, a intonation pattern, or both.

For words, dissociations between form and meaning motivate an intermediate level of representation, namely the phonology. PENTA denies that this level of representation exists for prosody and intonation. This decision creates difficulties for the interpretability of the model. The dataset on which the system was trained conflates the semantic feature of focus with the phonological feature of nuclear accent. These do not necessarily coincide. It is not clear how PENTA would handle examples of focused material that is not accented, or accented material that is not focused. PENTA treats stress and lexical tone as communicative functions, on a par with semantic focus and interrogation. However, these have no general semantic or pragmatic meanings; their function is differentiating words, which in classical linguistic theory is the hallmark of phonological elements. Thus,

PENTA conflates the phonological description (at the lexical level) with the syntactic, semantic, and pragmatic description (at the level of the phrase or utterance). This means that the theory has not tackled many issues that were previously framed within AM theory and generative semantics. Overall, the communicative functions of PENTA resemble components of an underlying phonological representation more than they resemble semantic and pragmatic functions. These observations lead to the conjecture that PENTA is a novel and highly optimized theory of phonetic realization (eg of the mapping between the phonology and the phonetic outcome).

As a phonetic realization model, PENTA adopts some of the advances that AM theory made in the 1980's. It integrates information at different time scales, and it generates f0 contours for all syllables, including those with and without stress, or with and without lexical tone. The AM phonetic realization algorithms reflected both conceptual and technical limitations. Conceptually, they were influenced by a strong presumption that the underlying linguistic system is minimized, using descriptive elements that are as simple as possible and as orthogonal as possible. Since that time, it has been clear that the human cognitive system can learn very detailed patterns and often represents them with a great deal of redundancy. In this connection, the amount of quantitative detail in PENTA is perfectly plausible. In particular, the idea of tonal targets that have both a position in the pitch range and a dynamics is not problematic. The AM realization algorithms were also limited technical by the inability to optimize the values of the parameters that they did have. PENTA carries out an up-to-date and highly effective optimization. It also meets current standards by validating on held-out test materials. This means that it represents a substantial advance in our understanding of phonetic realization.

6 Acknowledgments

References

Anderson, MJ, JB Pierrehumbert & MY Liberman. 1984. Synthesis by rule of English intonation patterns. In *Proc. IEEE Congress on Acoustics, Speech and Signal Processing*, 2.8.1 – 2.8.4.

Arai, T & S Greenberg. 1997. The temporal properties of spoken japanese are similar to those of english. In *Proceedings of Eurospeech, Rhodes, Greece*, vol. 2, 1011?14.

Beaver, D, BZ Clark, E Flemming, T Florian Jaeger & Maria Wolters. 2007. When semantics meets phonetics: Acoustic occurrences of second-occurrence focus. *Language* 83(2). 245–276.

Beckman, ME & JB Pierrehumbert. 1986. Intonational structure in Japanese and English. *Phonology Yearbook* 3. 13 – 70.

Beckman, ME & JB Pierrehumbert. 2000. Positions, probabilities, and levels of categorization. In *Proceedings of the Eighth Australian International Conference on Speech Science and Technology*, 1–18.

Beddor, Patrice Speeter & Rena Arens Krakow. 1999. Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation. *Journal of the Acoustical Society of America* 106. 2868–2887.

Birch, S & C Clifton. 2002. Effects of varying focus and accenting of adjuncts on the comprehension of sentences. *Journal of Memory and Language* 47. 571 – 588.

Bolinger, Dwight. 1958. A Theory of Pitch Accent in English. *Word* 14(2–3). 109–149.

Bruce, G. 1977. *Swedish Word Accents in Sentence Perspective*. Gleerup.

Bruce, G. 1990. Aligment and composition of tonal accents: comments on Silverman and Pierrehumbert's paper. In J Kingston & ME Beckman (eds.), *Papers in Laboratory Phonology I*, 107 – 114. Cambridge University Press.

Bruce, G. 2004. An intonational typology of swedish. *Speech Prosody 2004* .

Chomsky, Noam & Morris Halle. 1968. *The Sound Pattern of English*. Harper & Row.

Coleman, J. 2016. Acoustic-phonetic modelling of historical and prehistoric sound change. In *Paper presented at LabPhon15, cornell university*, .

Crystal, David. 1969. *Prosodic Systems and Intonation in English*. Cambridge University Press.

Dankovicova, J. 1999. Articulation rate variation within the intonation phrase in Czech and English. In *ICPhS99*, 269 – 272.

Dingelmanse, M, W Schuerman, E Reinisch, S Tufvesson & H Mitterer. 2016. What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages. *Language* 92. e117 – e133.

Dingemanse, M, D Blasi, G Lupyan, M H Christiansen & P Monaghan. 2015. Arbitrariness, iconity, and systematicity in language. *Trends in Cognitive Sciences* 19(10). 603–615.

Fletcher, J, E Grabe & P Warren. 2005. Intonational variation in four dialects of English: the high rising tune. In *Intonational variation in four dialects of English: the high rising tune*, Oxford University Press.

Fujisaki, H & K Hirose. 1984. Analysis of voice fundamental frequency contours for declarative sentences of japanese. *Journal of the Acoustical Society of Japan* 5. 233 – 242.

Garding, E. 1987. Speech Act and Tonal Theory in Standard Chinese: Constancy and Variation. *Phonetica* 44(1). 13 – 29.

German, JS, JB Pierrehumbert & S Kaufmann. 2006. Evidence for phonological constraints on nuclear accent placement. *Language* 82(1). 151–168.

Goldsmith, John A. 1976. An overview of autosegmental phonology. *Linguistic analysis* 2. 23–68.

Grabe, E & B Post. 2002. Intonational variation in the British Isles. In *Speech Prosody 2002*, .

Grice, M, D R Ladd & A Arvaniti. 2000. On the place of phrase accents in intonational phonology. *Phonology* 17. 143 – 185.

Gussenhoven, C. 1999. On the limits of focus projection in English. In *Focus: Linguistic, cognitive and computational perspectives*, 43 – 55. Cambridge University Press.

Hayes, B & A Lahiri. 1991. Bengali intonational phonology. *Natural Language and Linguistic Theory* 9. 47 – 96.

Hirschberg, Julia & Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics* 19(3). 501–530.

Hirschberg, Julia & Gregory Ward. 1992. The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in english. *Journal of Phonetics* 20. 241–251.

Ito, J, A Mester & J Padgett. 1995. Licensing and Underspecification in Optimality Theory. *Linguistic Inquiry* 26(4). 571–613.

Jun, S-A. 2006. *Prosodic typology: the phonology of intonation and phrasing*. Oxford University Press.

Keating, P A. 1988. Underspecification in phonetics. *Phonology* 5. 275 – 292.

Kennedy, C. 2007. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and philosophy* 30. 1–45.

Kochanski, G & C Shih. 2003. Quantitative measurement of prosodic strength in mandarin. *Speech Communication* 41(4). 625 – 645.

Ladd, DR. 1980. *The structure of intonational meaning: evidence from English*. Indiana University Press.

Leben, W. 1971. Suprasegmental and segmental representation of tone. In *Papers from the Second Conference on African Linguistics, Studies in African Linguistics, Supplement 2*, .

Liberman, MY & JB Pierrehumbert. 1984. Intonational invariance under changes in pitch range and length. In *Language Sound Structure*, 157–233. MIT Press.

MacNeilage, PF. 1998. The frame/content theory of evolution of speech production. *Behavioral and brain sciences* 21(4). 499 – 511.

Meir, I, C Padden, M Aronoff & W Sandler. 2013. Competing iconicities in the structure of langages. *Cognitive Linguistics* 24(2). doi:10.1515/cog-2013-0010.

Pierrehumbert, Janet B & Mary E Beckman. 1988. *Japanese Tone Structure*. Linguistic Inquiry Monographs, MIT Press.

Pierrehumbert, JB & J Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In P Cohen, K Morgan & M Pollack (eds.), *Intentions in Communication*, 271–311. MIT Press.

Pitrelli, J F & Eide E M. 2003. Expressive speech synthesis using American English ToBI: Questions and contrastive emphasis. In *Ieee workshop on automatic speech recognition and understanding*, doi:DOI: 10.1109/ASRU.2003.1318524.

Pruitt, K & F Roelofsen. 2013. The interpretation of prosody in disjunctive questions. *Linguistic Inquiry* 44(4). 632–650.

Rialland, A. 2009. African "lax" question prosody: its realisations and its geographical distribution. *Lingua* 119. 928–949.

Rooth, M. 1992. A Theory of Focus Interpretation. *Natural Language Semantics* 1. 75–116.

Rosenbaum, DA. 2010. *Human Motor Control*. Academic Press, Elsevier.

Schwarzschild, R. 1999. AvoidF and other constraints on the placement of accent. *Natural Language Semantics* 7. 141–177.

Shattuck-Hufnagel, S, M Ostendorf & K Ross. 1994. Stress shift and early pitch accent placement in lexical items in American English. *J. Phonetics* 22. 357–388.

Silverman, Kim, Mary E Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert & Julia Hirschberg. 1992. Tobi: A standard for labeling english prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing, ICSLP*, 12 – 16.

Silverman, Kim & Janet Pierrehumbert. 1990. The timing of prenuclear high accents in english. In J Kingston & ME Beckman (eds.), *Papers in Laboratory Phonology I*, 107 – 114. Cambridge University Press.

Streeter, LA. 1978. Acoustic Determinants of Phrase Boundary Perception. *J. Acoust. Soc. Am* 64(6).

Trager, GL & HL Smith. 1951. *An Outline of English Structure*. Battenberg Press.

Vallduvi, E & E Engdahl. 1996. The linguistic realization of information packaging. *Linguistics* 34. 459 – 519.

Ward, G & J Hirschberg. 1985. Implicating uncertainty: the pragmatics of fall-rise. *Language* 61. 747 – 776.

Xu, Y & S Prom-on. 2014. Towards invariant functional representations of variable frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Communication* 57. 181–208.

Xu, Y, S Prom-on & L Fang. in press. The PENTA model: Concepts, use, and implications. In J Barnes & S Shattuck-Hufnagel (eds.), *Prosodic theory and practice*, MIT Press.