The tune drives the text -
Competing information channels of speech shape phonological systems

Timo B. Roettger¹ & Martine Grice²
¹Northwestern University, ²University of Cologne

Abstract
In recent years there has been increasing recognition for the vital role of intonation in speech communication, both in diachronic and synchronic terms. Although contemporary models of intonation assume that intonation – the tune – and the text that bears it are represented on separate tiers and do not influence each other, this paper distils unrelated findings across diverse languages that point to a dynamic negotiation between these two tiers. The result of this negotiation for the text often involves vowels, which, given their rich harmonic structure, lend themselves particularly well to the transmission of pitch. Existing vowels can be lengthened, new vowels can be inserted and their loss of voicing can be blocked. The negotiation between tune and text ensures that pragmatic information is accurately transmitted and possibly plays a role in the typology of phonological systems.

Keywords: Intonation, epenthesis, devoicing, lengthening
1 Introduction

Speakers simultaneously express multiple meaning dimensions during the act of speech. For example, a simple sentence such as “The tune drives the text” expresses the proposition at hand through a combination of lexical concepts and their morphosyntactic combination. Beyond that, other channels of the speech signal such as intonation can express semantic-pragmatic functions such as illocutionary force and discourse structure.

Traditionally, much attention has been paid to lexical meaning, as expressed through consonants and vowels organised into metrical structures, referred to as the text. More recently, research has increasingly focused on utterance-level meanings as expressed by intonation, referred to as the tune. Utterance-level meaning expressed by intonation involves paralinguistic functions as diverse as emotion, attitude and speaker involvement, and semantic-pragmatic functions, such as the illocutionary force behind a proposition such as declaring the proposition or asking for confirmation (e.g., “The tune drives the text.” vs. “The tune drives the text?”) or the information structure, i.e. contextualizing the proposition within the immediate discourse (e.g., “The TUNE drives the text.” vs. “The tune drives the TEXT.”, as answers to “What drives the text?” and “What does the tune drive?” respectively).

There is no privileged connection between lexical meaning and the text. Tones can also express lexical meaning. Neither is there a privileged relation between utterance-level meaning and the tune - discourse particles (text) can express pragmatic functions as well as intonation. In this paper, however, we concentrate on lexical-level text and utterance-level tune and how they relate to each other.

The intonation contour of an utterance is traditionally thought to be independent of the words that bear it (e.g. Abercrombie, 1967; Pierrehumbert, 1980; Ladd 2008), the two being represented on separate levels, allowing them to be freely combinable: “not only can the same text have many different melodies, the same melody can occur on many different texts” (Pierrehumbert, 1980:8). The separation of tune and text has led to important insights into how prosodic systems are structured, how intonational contours are aligned with the available segmental structure, and how intonational meaning is expressed. However, separating segmental and intonational aspects of speech from each other is at odds with their inherent phonetic integration and has led us to largely ignore the interactions between them.

For intonation to be produced, voiced segmental material is needed to enable the vocal folds to vibrate; for intonation to be perceived, it is important for the segmental material to exhibit periodic energy with a rich harmonic structure (Zhang, 2004; Barnes et al., 2014). Thus, amongst the segments that make up the text, some do not allow for vocal fold modulation at all, and others are not sufficiently periodic to allow for optimal perceptual
retrieval of such modulation. The shortage of tone-carrying segmental material can lead to cases in which a pragmatically relevant tonal contour might be phonetically impoverished, potentially obscuring a pragmatic contrast that would otherwise be signalled by intonation. Given the vital importance of intonation in speech communication and assuming that languages have evolved to support message transmission accuracy (e.g. Hall et al., 2016), we expect sound systems to minimize uncertainty not only at the lexical level but also at the intonational level.

The present paper will argue that languages do exactly that. In section 2, we briefly review assumptions made by contemporary models of intonational phonology, focusing on the autosegmental-metrical model (section 2). Although these models represent the tune and the text on separate tiers, such that the tiers cannot influence each other, there are situations in which they do in fact appear to do so. This is when there are conflicts between tune and text, for instance when there is not enough segmental material to bear a complex sequence of tones. We will discuss cross-linguistic evidence suggesting that functional conflicts between signalling lexical and intonational meaning are common, and when they happen, linguistic systems have several mechanisms to accommodate such functional conflicts. Some of these mechanisms lead to adjustments of the tune (the text drives the tune, section 3), others lead to the adjustment of the text (the tune drives the text, section 4). The latter case has important implications for phonological typology and should help us further our understanding of why sound systems are the way they are (section 5).

2 Models of intonation

Earlier work on intonation distinguished between tune and text by representing the tune as diacritics on and around the orthographic or phonetic transcription of the consonants and vowels, representing the text (Trager and Smith, 1951; Halliday, 1967; Crystal, 1969; Tench, 1996). Although these early models still considered the tune - in the form of phonological tones - to be a property of segments, it was acknowledged that tones could stretch over multiple segments, leading to the term suprasegmental. However, work on a number of African languages in which tones have lexical and grammatical functions revealed that tones could not only span more than one segment but that the very same segments could bear different tones in different grammatical contexts (Leben, 1973). Moreover, tones expressing certain grammatical functions might not have a host segment at all, leading to a situation where the tone is floating. Floating tones can also be the result of the deletion of a segment, either diachronically or synchronically, but where the tone remains. This apparent autonomy of the tune called for a new model which could account for the relation between tune and text, where in the early studies the tune consisted of lexical/grammatical tones.
The new model was *Autosegmental Phonology* (Goldsmith, 1976), whereby tones are not properties of segments but instead they are autonomous units on a par with segments, but represented on their own tier. Not every segment is a tone-bearing unit (TBU); initially this function was given to vowels, and later to the syllable (Kahn, 1976) of which the vowel was the nucleus. The link between the segmental tier and the tonal tier is represented as association lines linking tones to TBUs. These association lines can be deleted, just as new lines can be added, allowing tones to spread to TBUs that have no tone, operations that are subject to universal well-formedness conditions. This architecture allows the tune to be represented independently of the text, yet provides a mechanism to capture the phenomena observed in a reasonably constrained way.

The autosegmental model was applied to intonation (Goldsmith, 1974; Liberman, 1975; Leben, 1976), in particular to English, in which tones are more sparsely distributed than in the languages previously analysed with this model. Goldsmith (1974) showed that the same tune could be produced on strings of different lengths. Thus, the tonal sequence H*L representing a falling intonation typical of neutral statements could be produced on words of varying length and structure, such as *Sam, Canada* and *balloon*, as in (1). However, it is not sufficient to indicate that the fall occurs on a particular word, its location within the word is also important. This is achieved with the star (asterisk), marking a metrically strong syllable for association with the tonal sequence. Moreover, the timing of the fall in relation to the metrically strong (starred) syllable could be captured by a star on one of the two tones. In (1) the star is on the H tone, indicating that the fall begins on, rather than before, the starred syllable.

(1) 

```
|   *   |   *   |   *   |
| H L   | H L   | H L   |
|       |       |       |
|   *   |   *   |   *   |

Sam    Canada    Balloon
```

Thus, in autosegmental models of English, the tune and the text were on independent tiers but a mechanism was put in place not only to associate tones with words but also to privilege certain tones and certain TBUs (using the star notation) to ensure an appropriate association, even if the number of tones and TBUs was not equal on each level.

A further development involved the structuring of the text into higher level constituents, such that TBUs were designated terminal elements of metrical trees. *Autosegmental-metrical Phonology* developed from these two approaches, providing a framework for representing association to TBUs, represented as the heads of certain
constituents, and to edges of those and higher constituents. Within this framework, the privileged status of certain tones in a tonal complex such as H*L could also be represented as a branching structure with strong-weak relations (Pierrehumbert & Beckman, 1988; Grice, 1995). Autosegmental-metrical Phonology has been applied to the analysis of intonation in a broad and typologically varied range of languages, not only well researched ones, but also understudied languages and varieties. It is the most widespread theory of intonation at present.

Crucially, neither Autosegmental Phonology nor the subsequently more elaborate Autosegmental-Metrical Phonology have made any provision for the influence of one tier over the other. This means that there are two autonomous tiers, the tune and the text, that can be linked by association, but, crucially, there is no mechanism in place to allow for these tiers to influence each other: Neither is the choice of phonological tune (in the form of tones) dependent on the corresponding text (in the form of segments, syllables, and words), nor is the choice of text dependent on the tune assigned to it. However, in recent years, this strict separation of tune and text has been empirically challenged.

For intonational tones to be realised as a perceived pitch contour, it is generally agreed that they need to be realized on segmental material that is voiced. For the contour to be perceived effectively, it needs to occur on segmental material of high periodic energy (Zhang, 2004; Barnes et al., 2014). However, some of the segments used to express lexical meaning are neither well suited to producing pitch modulation nor to enabling a perceptual retrieval of such modulation. The acoustic correlate of perceived pitch is fundamental frequency (f₀), which can only be reliably calculated if there is sufficient strength in the periodic energy of the speech signal.

The sparse distribution of intonational tones and the tendency for multiple tones to occur at edges of constituents (e.g. if the head is close to an edge), can lead to so-called tonal crowding, in which a number of different tones are all simultaneously associated with one syllable. This multiple association leads to a sequencing of the tones (e.g. LHLH) representing a dynamic intonation pattern (in this case rise-fall-rise). Depending on the makeup of the syllable, it may not be possible for the entire movement to be realised - all other things being equal. In such cases, important acoustic information for the recognition of the speaker intention could be lost. Consider the phrases "I know the tune." and "I know the text.", both with an intonation pattern indicating uncertainty (Ward & Hirschberg, 1985), hypothetical pitch contours of which are illustrated in Fig. 1.
Figure 1: Schematic representation of the f0 contour of the first two sentences “I know the tune.” and “I know the text.”, indicating the meaning of uncertainty, expressed by a rise-fall-rise (H* L-H%) on both phrases. In (b), significant parts of the final rise-fall-rise would be required to co-occur with voiceless segments.

Whereas the rhyme of the final word in (a), ‘tune’, is fully voiced, the rhyme of the final word in (b), ‘text’, is composed of a short vowel and a voiceless coda cluster. This segmental make-up heavily constrains the realization and perceptual retrieval of intonational movements, especially when they are complex, such as in the rise-fall-rise. This scenario marks a functional conflict, potentially endangering a pragmatic contrast signalled by intonation.

3 The text drives the tune

In such a conflict of tune-text association, the tune can be adjusted (Erikson & Alstermark, 1972; Bannert & Bredvad-Jensen, 1975; Grønnum, 1989; Grice, 1995; Grabe, 1998; Grabe et al., 2000; Lickley, Schepman, & Ladd, 2005; Rathke, 2009; Grice et al., 2015). The nature of these adjustments depends on syntagmatic, paradigmatic, and language-specific factors (see Hanssen, 2017; or Roettger, 2017, for recent overviews).

Adjustments of the tune were first reported for Swedish, a language that has two lexically distinctive accent categories. Both accent types, Accent 1 and Accent 2, are characterised by a falling pitch contour, represented as a sequence of high and low tones, differing only in the alignment of the high tone: the high tone in Accent 1 words reaches its target earlier than the high tone in Accent 2 words (Bruce, 1977). Erikson and Alstermark (1972) discussed the adjustment of these accent patterns in light of the segmental structure of the words bearing them. On the one hand, they show that f0 movement is reduced with decreasing vowel length. In this case, the tonal target is undershot, with the fall not reaching what is assumed to be its low target compared to words with sufficient sonorous segments. In the most extreme case, e.g. if the syllable contains a short vowel and a voiceless coda consonant, there may not remain any trace of a fall in f0. This mechanism has been called “truncation”. Alternatively, the full pitch movement is sometimes realised more rapidly on shorter vowels. These local “rate adjustments” were subsequently coined “compression” by Bannert and Bredvad-Jensen (1975).
These two mechanisms, truncation and compression, have since been attested for intonational tune-text conflicts across many languages. Grønnum (1989) reports truncation for Danish rise-falls and Northern German falls. Grabe (1998) reports on truncation and compression for standard varieties of German and English: In her study, Grabe systematically manipulated the segmental material on which speakers had to produce different intonational contours, ranging from a disyllabic word with a long vowel (/ʃiːfɐ/) through a monosyllabic word with a long vowel (/ʃiːf/) to a monosyllabic word with a short vowel (/ʃɪft, ʃɪf/). Because the consonants in her stimuli were voiceless, the phonetic opportunity for realising tonal movements was restricted to the short vowel. Her results indicate that Southern Standard British English compresses both falls and rises, whereas Northern Standard German truncates falls and compresses rises. This is in line with truncation reports for high fall-rise contours in German and Dutch (Lickley, Schepman, & Ladd 2005; Ladd 2008) in which the falling component of the contour may be missing, resulting in a simple high rise. A series of studies by Rathcke (2009) demonstrate truncation for rising-falling contours in German and partial compression of falling contours. Although often reported on as two separate strategies, truncation and compression have been found to involve mixed strategies, with tunes being to some extent both truncated and compressed (Prieto & Ortega-Llebaria, 2009; Roettger, 2017).

In some cases, truncated patterns have been described as having a different representation from non-truncated patterns. For example, Grice (1995) discusses truncation in Palermo Italian in cases where there is tonal crowding: Complex rising-falling contours entirely lack their falling component if the final syllable carries a pitch accent (see Fig. 2). Grice analysis this as a final Low tone without an association to a TBU. The rising portion of the rise fall is thus intact but the fall to the L tone is affected.

![Figure 2](image)

Grice, Ridouane, and Roettger (2015) discuss tune-text conflicts in Tashlhiyt Berber, a language that allows exceptionally long sequences of voiceless consonants. One of the observed conflict resolutions is a complete truncation of a rise-fall, resulting in a loss of the entire tonal complex (see Fig. 3c).
Alternative strategies to deal with conflicts in tune-text-association involve temporal shifts of tonal targets. If the segmental material is insufficient to realise a sequence of tonal targets, they may occur instead on more suitable text (Grice et al., 2015; Roettger, 2017). Thus, instead of losing the entire tonal complex, Tashlhiyt speakers have the option of placing the tones on a preceding word that contains sufficiently sonorous segmental material for their realisation (see Fig. 3b). The tendency of tones to be anticipated in cases of tonal crowding at phrase edges has been observed in many other languages (Steele, 1986; Casper & van Heuven, 1993; Prieto, van Santen, & Hirschberg, 1995; D'Imperio, 2001; Prieto, D'Imperio, & Fivela, 2005; Schepman, Lickley, & Ladd, 2006; Mücke et al., 2009). For example, in Neapolitan Italian, questions and statements can be expressed with similar tonal events, i.e. a rise-fall in pitch phrase finally (if the final word is in narrow focus). These tonal events usually differ in the alignment of the high tone, with the high tone being reached later in questions than in statements (D'Imperio & House, 1997). However, questions can also be produced with an additional final rise, leading to a more complex tonal sequence (a rise-fall-rise, as in Fig. 4b). Looking at the alignment of the high target of the pitch accent across instances with and without this utterance-final rise, the high target shifts to the left when a phrase-final rise is present (Cangemi & Grice, 2016, compare Figs. 4a and b).
Beyond cases of tonal crowding at phrase edges, tones can also be shifted in so-called stress clash contexts. For example, in words such as “Heathrow”, the part of London in which the largest airport is located, the primary stress is located on the final syllable, which usually functions as the TBU for a pitch accent. When this word is directly followed by a word with primary stress on the initial syllable (e.g. “Airport”), however, the first syllable is promoted to TBU status, so as to bear an accent (cf. Fig. 5, e.g. Bolinger, 1965; Shattuck-Hufnagel et al., 1994).

![Figure 5: Schematic representation of the f0 contour on “Heathrow” and “Heathrow airport”, in which Heathrow co-occurs with a pitch accent on the primary stressed syllable in (a) and with a pitch accent on the secondary stressed syllable in (b).](image)

Such shifts in accent placement have been observed across unrelated languages (see Tilsen, 2012, for a recent overview). All of the mechanisms that we have discussed so far, i.e. truncation, compression, and tonal shift, can be considered negotiations between competing meaning-signalling channels of speech. In the case of truncation and compression, the information channel that is responsible for intonational meaning, the f0 contour, is impoverished to a certain extent. Truncating or compressing the f0 contour leads to loss of information and results in weaker perceptual retrievability of the pragmatic contrast. Tonal shift, on the other hand can be seen as upholding the intonational channel, with the caveat that a shifted tonal event may be misinterpreted as signalling meaning that was not intended (e.g., in the case of Tashlhiyt, shifting a focus-marking pitch accent to a constituent that is not focused; or giving the listener the impression that the primary stress is on a different syllable in the word, possibly impeding lexical access, and at worst activating a different lexical item from the one intended).

The idea that the text affects the tune is not new. There is a substantial literature on consonant-tone interactions in lexical tone systems (e.g. Bradshaw, 1999; Yip, 2002; Tang, 2008, for surveys). For example, there are strong statistical relationships between tone height and laryngeal properties of prevocalic consonants, which has been the driving force behind diachronic models of tonogenesis (e.g. Haudricourt, 1954; Hombert et al., 1979). The converse pattern, i.e. the impact of tone on its segmental environment has been largely ignored. In fact, it has been claimed that lexical tone does not affect consonants at all.
(Hyman, 1973, 1976). Other authors have argued that tone can have a systematic influence on consonants (Maddieson, 1974, 1976; Gandour & Maddieson, 1976). In the remainder of this paper, we will discuss evidence indicating that the realisation of post-lexical tonal movements is systematically correlated with the adjustment of the segments involved. We will argue that investigating tune-text interactions will further our understanding of selected diachronic aspects of phonotactic systems and point to ways in which phonological systems negotiate functional conflicts between different meaning-bearing channels.

4 The tune drives the text

Another way to decrease the functional impact of tune-text conflicts is the adjustment of the text to create a segmental environment that is better suitable to realize $f_0$ movements. Although, such an adjustment of the text has received less attention in the literature, there are a number of languages exhibiting such patterns in phrase final position, typically when there is a complex intonation contour, involving at least one rising element, and when this contour is either on a monosyllabic word or on a word with final stress, leading to tonal crowding. In the following we examine three cases in which segmental patterns have been reported to adjust, all of which involve vowels, elements of high intensity and rich harmonic structure, making them, from a perceptual standpoint, the ideal carriers of pitch information: (i) Existing vowels can be lengthened, (ii) new vowels can be inserted or (iii) the devoicing or deletion of existing vowels can be blocked. It is argued that all three modifications make the segmental string more suitable for the production of pitch movements. They also make the pitch contour easier to perceive, facilitating a more robust retrieval of intonational meaning.

4.1 Vowel lengthening

One way to create a segmental environment that is better suited for realizing tonal movements is lengthening of existing segments that are already good carriers of the tune. Such lengthening has been observed for Bari Italian yes-no questions, which are typically realised with an accentual rise followed by a fall-rise. Phrase-final accented syllables bearing the rise-fall-rise are considerably lengthened in comparison to the same syllable in neutral statements, characterised by a simple fall (Grice Savino, & Refice, 1997; Refice, Savino, & Grice, 1997). Frota (2002) reports on similar lengthening of phrase final vowels in fall-rises in European Portuguese. Lengthening has also been observed for different varieties of German (Gartenberg & Panzlaff-Reuter, 1991; Gilles, 2005). For example, Gilles (2005) found that monosyllables are longer with a fall-rise intonation than with falling contours. Ladd (2008: 183) discusses rise-fall tunes in Hungarian questions, which can only be realised in monosyllabic words when they are accompanied by prolongation of the segmental material.
Similar observations have been made for Spanish and Catalan by Prieto and Ortega-Llebaria (2009). They found that, in words with final stress, syllables are longer in rise-falls than in Biondi (2011) reports on unusual phrase-final vowel lengthening in accusatory speech, characterised by high pitch in Kashibo-Kakataibo. Vanrell and Cabré (2011) and Cabré and Vanrell (2013) report that the vocative intonation in Sardinian which is characterised by a rise-fall pattern is accompanied by lengthening in monosyllabic words. Cruz (2013) reports on utterance-final vowel lengthening in questions for Southern European Portuguese. Heston (2014) reports on lengthening of utterance-final vowels when they are accompanied by a final rise-fall in Fataluku, a Papuan language. Heston notes that in these cases, vowels are approximately twice as long as comparable vowels accompanied by a simple fall.

In all these cases, there is lengthening when the phrase-final syllable is associated with high pitch and all of the above cases are concerned with lengthening at or near prosodic edges. Similarly, lengthening has also been observed for scenarios in which pitch accents are close together, resulting in a sequence of H and L tones in close proximity. In such contexts, the segmental material can be lengthened, resulting in more space between consecutive tonal targets (Liberman, 1975; Tilsen, 2012).

These durational adjustments are often gradient in nature and considered phonetic adjustments. However, it is also possible to interpret these adjustments as the insertion of metrical beats (Selkirk, 1984). For example, in Sanskrit, vowels are sometimes orthographically marked as ‘overlong’ (with the numeral 3), indicating that they consist of three moras. These overlong vowels have been described as occurring mainly in questions (Strunk, 1983). This example not only suggests a relationship between questions and vowel length but also a categorical lengthening pattern that has necessitated orthographic marking.

All of these cases, observed for a wide range of languages, can be interpreted as adjustments of the text to enable the realisation of communicatively relevant pitch movements.

4.2 Vowel insertion

Another way to create a segmental environment that is more suitable for realizing tonal movements is inserting a vowel. In the following, we use the term *vowel insertion* to loosely refer to any addition of non-lexical vowels including both ‘intrusive vocoids’ and ‘epenthetic vowels’, two proposed typological categories that mainly differ with respect to their status within the phonological system of the respective language (e.g. Harms, 1976; Levin, 1987; Hall, 2006, 2011; Silverman, 2011). Our description of the observations below will not make strong assumptions about what status the reported vowels have in their respective phonological system.
Although, there are claims that vowel insertion is not affected by higher prosodic or intonational factors (Hall, 2011), there are a number of languages in which a vowel is more likely to be inserted if there are tune-text conflicts. For example, research on Romance languages has mentioned vowel insertion patterns conditioned by prosodic and intonational aspects.

In one of the more detailed acoustic studies, Grice, Savino, and Roettger (2018) report on experimental evidence from a production study of Bari Italian speakers: In a number of varieties of Italian, the pronunciation of loan words ending in consonants (e.g. internet, blog) is often characterised by a word final schwa (e.g. Bertinetto, 1985; Krämer, 2009; Repetti, 2012; Broniś, 2016; Cavarra, 2016). In Grice et al.’s study, speakers produced monosyllabic and disyllabic names ending in word-final consonants (e.g. Matt, Tim, Carol), a pattern that is phonotactically rare in their native vocabulary. Targets were produced in different prosodic contexts associated with tonal contours that differed in their nature and complexity. Grice et al. find that the frequency of occurrence and duration of schwa is modulated by metrical and intonational factors. Monosyllabic words elicited overall more schwas and if schwa was present, overall longer schwas than disyllabic words. The probability of schwa occurrence and its duration was robustly affected by the intonational tones to be realized on the word. Rising tones and rise-falls elicit more and longer schwas than falling tones. For example, the word Carol in (2) is more likely to be produced with a word-final schwa when the word occurred in a question than in a statement. This tune-dependent pattern was stronger for monosyllabic words than for disyllabic words.

(2) ha kjamato karol(ə) ‘Carol called’ (Grice et al., 2018)

These quantitative assessments are in line with many cross-linguistic observations throughout the literature. For instance, Martínez-Gil (1997) reports on Galician vowel insertion in phrase-final position, a pattern that is very similar to the Bari Italian case. He reports an optional [i] that can occur in phrase-final position. However, he notes that this optional vowel can only occur in monosyllabic words or polysyllabic words with final stress. While he formulates this pattern as categorical, it is strikingly similar to the Italian case in which there are more schwas in monosyllabic words than in disyllabic words. Martínez-Gil attributes these distributional asymmetries to structural requirements to create a well-formed bimoraic trochee at the edge of each intonational phrase. However, there are certain correlations with intonation: Martínez-Gil states that “sentential prominence in Galician falls on the last word bearing primary stress” and that “epenthesis is heavily favoured […] whenever a word bears phrasal prominence” (p. 288) which can be interpreted as a pitch accent.
Similarly, in Sardinian, there is word-final vowel insertion which has been described as repairing illicit structures defined by either having a final consonant or final word stress. According to a recent survey in Torres-Tamarit, Linke, and Lee (2017), the insertion of an epenthetic vowel does not always take place. Epenthesis is robust in phrase-final position, which is the locus of certain complex tonal movements associated with questions (Vanrell et al., 2015), but it is variable in phrase-medial position. For Carrarese and Pontremolese, two Northern varieties of Italian, Cavirani (2016) mentions variable realisation of word-final vocoids whose presence and duration are determined by ‘emphatic conditioning’, which, again, can be interpreted as a prominent (and therefore particularly dynamic) type of pitch accent.

In Standard European Portuguese, vowels are inserted into different prosodic contexts, namely phrase-finally in yes-no questions which are characterised by a complex nuclear contour (fall-rise), and in vocatives characterised by a high plateau (Frota, 2002; Frota et al., 2015). Cruz (2013) compared vowel insertion patterns across different European Portuguese varieties. In Standard European Portuguese, vowel insertion occurs in Intonation Phrase final position when the word co-occurs with interrogative rise-falls as well as with requests, but not in declaratives in which the word co-occurs with simple falls, whereas in the variety spoken in Castro Verde, vowel insertion takes place in Intonation Phrase final position across the board, i.e. regardless of the tonal contour.

Similar observations have been made for genetically unrelated languages. For example, in Tashlhiyt Berber inserted vowels interact with phrase-level intonation in systematic ways. Roettger (2017) discusses a production corpus containing lexically voiceless words occurring in different prosodic positions (phrase medially vs. phrase finally) and in different sentence modalities characterised by different intonational events (rise-fall vs. fall vs. no tonal movement). Besides speaker-specific patterns and certain articulatory constraints, the presence of schwa is correlated with both phrasal position and the nature of the tune that is supposed to be realized on the word (see also Grice et al., 2011, 2015). Schwa was more often observed in phrase-final words than in phrase-medial words and it was more often observed in sentence modalities in which complex tonal movements have to be realised. These tone bearing schwas exhibited high intensity, rich harmonic structure, and surprisingly long durations.

Moroccan Arabic, a language spoken alongside Tashlhiyt Berber, also has positionally conditioned vowel insertion, formulated as a phonological rule by Dell and Elmedlaoui: “In the last syllable of an Intonational Phrase, if the nucleus does not contain a sonorant, make it complex by inserting e before the nuclear consonant.” (2002: 300). Thus, vowel insertion takes place near the end of the intonation phrase. In another variety of Arabic, Tunisian Arabic, schwa is frequently inserted Intonation Phrase finally in polar
questions, where the melody is complex (Hellmuth, to appear), but it is not inserted elsewhere.

Cross-linguistically the literature indicates that vowel insertion is more likely to happen in phrase-final position than in phrase-medial position, and it is more likely to happen in environments in which certain communicatively relevant tonal movements take place. Moreover, it appears that there are certain implicational relationships. If vowel insertion is conditioned by tonal crowding, there is either vowel insertion only in those environments, as is the case for Sardinian, Galician, Moroccan Arabic and Tunisian Arabic, or there is at least quantitatively more frequent vowel insertion in those environments, such as in Bari Italian. Similarly, if vowel insertion is conditioned by prosodic structure, vowel insertion is either observed only phrase finally, as is the case for Galician, Moroccan Arabic, and the Portuguese varieties, or there is at least quantitatively more vowel insertion in phrase-final position, as in Sardinian and Tashlhiyt. Finally, if vowel insertion is conditioned by the tune, it is either restricted to certain contours (mostly contours that contain a rising element) as it is the case for Tunisian and Standard European Portuguese, or there are quantitatively more instances of vowel insertion in those tonal contexts that contain rises in pitch. These observations are in line with the idea that linguistic systems create a segmental environment that is better suited for realizing meaningful $f_0$ movements. In other words, the text has to be suitable for realizing the tune.

4.3 Avoiding vowel devoicing / deletion

Just as certain contexts favour the addition of a vowel to the segmental string, these same sorts of contexts tend to disfavour vowel devoicing or vowel deletion, since vowels (with full voicing) are optimal for the realization of $f_0$ movements. A system that reduces uncertainty at the level of intonational meaning should thus avoid vowel devoicing or deletion in situations where the intonation plays a key role in communication.

Although vowel devoicing is often restricted to high vowels and commonly occurs when the vowel is adjacent to voiceless consonants, prosody can also play a role in determining whether vowels are devoiced or not, although there is much disagreement as to what the relevant prosodic factors might be. Japanese, for example, is described as devoicing high vowels between two voiceless consonants, see example (3). Moreover, several authors have also noted that vowel devoicing is more likely to take place pre-pausally (see example (4)) (e.g. Fujimoto, 2015; Kilbourn-Ceron & Sonderegger, 2017, for recent overviews).
These prosodically conditioned devoicing patterns in Japanese are not categorical. For example, utterance-final devoicing is highly variable (e.g. Maekawa & Kikuchi, 2005; Vance, 2008; Kilbourn-Ceron & Sonderegger, 2017). Crucially, vowel devoicing can be “blocked” by a high pitch accent (e.g. Kuriyagawa & Sawashima, 1989; Hirayama, 2009). In their corpus analysis of Japanese spontaneous speech, Kilbourn-Ceron and Sonderegger (2017) found that, all else being equal, a larger phrase boundary following the vowel was correlated with less devoicing. This variation in devoicing has been interpreted as conditioned by the presence of boundary tones (e.g. Hirayama, 2009; Fujimoto, 2015).

As in Japanese, languages exhibiting lexical tone or lexical pitch accent tend to block final devoicing when a vowel co-occurs with a high tone (Cheyenne: Davis, 1962; Acoma: Miller, 1965; Konso: Orkaydo, 2013). In the Shanghai Wu variety, devoicing has been described as never occurring in the first syllable in a tone sandhi group, which is the unit to which the phonological tone is associated (Beckman & Venditti, 2010; and see Chao, 1968, who describes a similar pattern for a variety spoken in Beijing Chinese).

Languages with intonational pitch accents show similar patterns. For example, Dell and Tangi (1992) report that in Ath-Sidhar Rifian Berber, schwa is usually devoiced between voiceless consonants but remains voiced “under an intonation which requires a final high pitch […]” (ibid: 154). Heath (1987: 184) described a process in Moroccan Colloquial Arabic in which schwa deletion in word-final syllables is blocked by “list Intonation”, characterised by a final rise in pitch. Gordon (1998) reports on vowel devoicing being less likely in accented positions in Bonaara Oromo, and Tunica. Andreeva and Koreman (2003) discuss an interesting case of vowel devoicing in Bulgarian. When a pitch accent occurs on the last word in the phrase (i.e. to signal focus), the amount of vowel devoicing depends on the scaling and alignment of the accent peak, with less devoicing for more prominent accent types. Moreover, they observe no devoicing when the word appears phrase finally and co-occurs with a rising boundary tone. This is in line with observations about vowel devoicing in Greek (Dauer, 1980; Kamaiki, 2015), which is less likely to take place when the target word occurs with rising intonation (as for example in wh-questions), rather than falling intonation.

All these examples strongly suggest a relationship between the application of vowel devoicing / deletion and the tonal conditioning of the environment in which the vowel is to be produced: In positions in which there is a communicatively relevant tonal movement, there is

---

1 Several languages exhibit blocking of vowel devoicing in stressed syllables too (see Gordon 1998 for an overview).
less vowel devoicing. This supports the hypothesis that the text is to some extent driving the tune. Devoiced vowel variants in these positions would lead to the loss of pitch information related to lexical or postlexical meaning, endangering an otherwise important meaning contrast.

5 General discussion

It has been argued that languages have evolved to support optimal transmission of intended meaning, in the sense that sound patterns are shaped by biases supporting message transmission accuracy, set off against resource cost (Hall et al., 2016). In human speech there is always a substantial amount of uncertainty associated with any intended meaning due to, for example, signal ambiguity, i.e. the degree of competition between different meanings as being associated with the same signal. Ambiguity can emerge when a signal contrast is frequently impoverished by physiological, acoustic, or cognitive aspects of the speech transmission process (e.g. Blevins, 2004). Such ambiguities have been argued to lead to sound change, shaping phonological systems through time (e.g. Ohala, 1981, 1993). While the literature on signal ambiguity has mainly focused on aspects of speech signalling lexical meaning, other important meaning signalling aspects of the speech signal have been widely ignored.

Tune transmission accuracy is achieved through a text optimised for pitch contours to be articulated and perceived. This might be at odds with transmission accuracy of the text such as lexical content, that might require preservation of structures, with minimal segmental changes. Adding a vowel to a word might be less desirable in light of lexical retrieval, but significantly increase transmission accuracy of the tune. On the other hand, vowel devoicing or vowel deletion lead to a reduction of the resource cost at the level of the text, but hinder accurate transmission of information about the tune. It is clear that the requirements of tune and text, both in accuracy and resource cost would have to be negotiated.

Despite the importance of intonation for human communication, contemporary models of intonation assume that tones are independent of the segments on which they are realised. This is at odds with the inherent physiological and perceptual connections between the tune and the text. If the segmental material is not suitable for realizing certain tonal movements, important pragmatic information might be not accurately transmitted.

In this paper we have argued that languages show many different mechanisms to minimize uncertainty. Some of these mechanisms lead to adjustments of the tune, such as truncation, compression, and shifting of tonal targets. These mechanisms change the shape or timing of relevant tonal movements, adjusting them to accommodate segmental restrictions. We have also reviewed the literature on textual adjustments, an area which has
received less attention. In conflicting tune-text settings, languages often lengthen existing vowels or insert non-lexical vowels. Other languages avoid vowel devoicing or deletion in such settings. Arguably, all textual adjustments increase the suitability of the text to realize tonal movements by adding or preserving aspects of the text that are optimal for the production and perceptual retrieval of pitch. This small functional bias might play a role in shaping phonological systems.

Assuming that sound systems are shaped by perceptual, articulatory, and acoustic aspects of speech transmission (e.g. Hombert et al., 1979; Ohala, 1981, 1993; Blevins, 2004; see Yu, 2013, for a recent overview), small channel biases can function as diachronic attractors, leading to certain evolutionary pathways that are more common than others. For example, devoicing of obstruents in domain final position, a very common sound pattern, is attested in many unrelated languages across the world. This frequently occurring sound change might reflect several phonetic tendencies that lead to variation favouring voiceless final obstruents. Smith (1997) suggests that insufficient transglottal airflow might hinder vocal fold vibration in domain final positions. Steriade (1999) argues that stop releases in final position are often impoverished or missing, thus obscuring important acoustic cues to voicing distinctions. Kohler (2000) suggests that final devoicing might be due to a progressive opening of the glottis for breathing in prepausal position. Blevins (2004) additionally proposes that final lengthening in phrase-final position might lead to a disproportional lengthening of stops endangering the voicing contrast which is often cued by closure duration (voiced stops have smaller closure durations). Thus, final devoicing demonstrates not only the breadth of phonetic motivations for sound change (aerodynamic, perceptual, articulatory, acoustic), but also the multiplicity of diachronic paths. One and the same outcome (here a voiced sound becoming voiceless in domain-final position) might result from different properties of our speech transmission mechanisms (or a combination thereof).

Moreover, this example highlights the role of positional effects for phonetically conditioned sound change (e.g. Cole & Hualde, 2013). All phonetic motivations for final devoicing are either conditional on particular prosodic positions (i.e. occur only in certain prosodic positions) or its probability of occurring is increased in these positions.

The present paper has argued that sound systems exhibit certain properties that facilitate the reduction of uncertainty by ensuring that intonational information can be produced and perceptually retrieved. We have discussed three different segmental alternations that can be understood as epiphenomena of this functional pressure. If the tune drives the text, we would expect to see other phenomena interacting with tune-text requirements. The idea that properties of tonal structures may affect their segmental environment is not a new idea. Albeit contested (Hyman, 1976), some authors have argued that the nature of a lexical tone and its accompanying laryngeal configuration can lead to the
diachronic insertion of a segment (e.g. Kock, 1901; Maddieson, 1974, 1976; Gandour & Maddieson, 1976).

The exact mechanisms involved in these tune-text interactions are unclear. In general, there seems to be an asymmetry between falling and rising tonal movements on the one hand, and between complex and simple tonal movements on the other. We have discussed cases in which languages adjust the segmental string when co-occurring with tunes that involve a rise (involving a high tone), implying that rising/high tones are special in some way (see Evans, 2015). Raising and lowering pitch are not achieved in the same way, but instead involve different laryngeal mechanisms (e.g. Moisik et al., 2014). Producing a rise in pitch generally takes longer than a fall of equivalent excursion. Moreover, it has been argued that there are physiological limitations on the maximum speed of pitch change (Ohala & Ewan, 1973, Sundberg, 1979, Xu & Sun, 2002), restricting high/rising tones more than low/falling tones. On the other hand, there are known correlations between pitch level / pitch movement and perceived segment duration. A vowel with a high-level tone is perceived as longer than a vowel with a low-level tone, a vowel with a rising tone is perceived as longer than a vowel with a falling tone (Lehiste, 1976; Wang et al., 1976; Yu, Lee, & Lee, 2014). Thus, raised $f_0$ might be already inherently correlated with longer durations and larger gestural intervals, in both articulation and perception.

Orthogonal to perceptual biases, increased $f_0$ has been argued to activate attention orientation mechanisms (Hsu, Evans, & Lee, 2015). High tones / rising tones may thus drive selective attention, making phrasal positions co-occurring with these pitch dynamics more salient, and thus better targets for reproduction. A small phonetic bias - especially if made salient by pitch prominence - might lead to a drift towards patterns that optimise the transmission of the tune as discussed in this paper.

In terms of Autosegmental-metrical Theory, the separation of the tune and text tiers has been crucial for our understanding of how the same tune can be realised on texts of different lengths. It has not, however, been able to capture interactions between the two tiers. A model of intonation needs to acknowledge that languages need to negotiate cases where the number of tones and tone bearing units are not a perfect match, or where the make-up of the tone bearing unit is phonetically suboptimal for bearing tone. Conflicts may be resolved at the discrete level in terms of adding vowels, blocking their deletion, or adding extra phonological length. They can also be resolved in a more continuous way, making a release stronger and voiced, or increasing the duration of a vowel without changing its phonological length. The line between discrete and continuous strategies is not straightforward; in almost all cases in which discrete strategies have been described, these have to be seen in terms of distributions, i.e. probabilistic patterns rather than rules that apply across the board. An
awareness that properties of the text can determine the choice of tune will provide insights into previously unexplained alternations. Likewise, an awareness of how text may adjust to the tune to enable communicatively relevant intonational cues may further our understanding of the typology of certain segmental changes.

6 Conclusion

In our endeavour to understand the ascent of human language, several scholars have speculated about the phylogenetic origin of speech. Charles Darwin wrote about our early ancestors using singing to mediate sexual and territorial issues. He referred to this as a musical proto language (Darwin, 1871). Darwin’s ideas have been taken on and expanded by several scholars, who proposed similar proto languages (Brown, 2000, Fitch, 2005, Mithen, 2005), all of which assume a central role of rhythm and melody for the evolution of human speech. This paper has argued that grammatical aspects of rhythm and melody - what we refer to as prosodic structure and intonation - may play an important role in shaping phonological systems.

The requirements for the expression of tune and text sometimes conflict with each other, calling for a dynamic negotiation between them. In this paper, we have argued that intonation places functional pressure on its segmental hosts which can facilitate the temporal adjustment and preservation of existing elements - or the insertion of new elements - to reduce uncertainty at the pragmatic level. By distilling unrelated findings across languages and linking them to phonetic aspects of the speech transmission process, we aim to understand why certain phonological processes and alternations observed across languages happen, and, more broadly, invite us to speculate as to why phonological systems are the way they are. Taking the intricate relationship between tune and text seriously will not only advance our knowledge of intonation systems, but it will also add a new dimension to the question as to how the message shapes the phonological form.
References


Kaimaki, Marianna. 2015. “Voiceless Greek Vowels.” In *18th International Congress of Phonetic Sciences (ICPhS)*.


