Prosody, melody and rhythm in vocal music

The problem of textsetting in a linguistic perspective

Teresa Proto
Universiteit Leiden/Meertens Instituut

The last decades have witnessed a shift from anecdotal remarks concerning the “marriage” of music and lyrics in songs towards a more scientific approach to the matter. Textsetting has thus become the object of more formal analyses accounting for the regularities observed in individual singing traditions with regard to the mapping of linguistic material on musical structures. This paper illustrates the nature of the problem and reflects the status of the research on textsetting in living traditions. It is addressed to a wide audience of linguists interested in the relationship between language and music and points to the challenges that await the further development of this field of studies under the umbrella of linguistics.

Keywords: textsetting, singing idioms, prominence matching, misalignments, durational discrepancies, tone-to-tune alignment, moraic setting, quantitative metrics, tone language, stress-accent

1. Why study textsetting

Setting a text to a tune is not a random process, but one that is regulated according to the phonetic, phonological and syntactic properties of the language in which the lyrics are composed (Dell & Halle 2009). As a matter of fact, when lyrics are set to music, speech units are assigned to musical pitches in such a way as to conform to specific requirements of the singing idiom. Therefore, a systematic study of the way features of speech such as stress, pitch and vowel length interact with analogous features in music, e.g. beats, intervals and durations, should provide useful insights into the very nature of prosody, intonation and rhythm in language and its relation to its cognate in the musical domain.¹

¹ The focus will be on the phonetic and phonological aspects involved in text-to-tune mapping; issues pertaining to the alignment of syntactic constituents fall outside the scope of this study.
2. Commonalities between music and language

It has often been observed (Lerdahl and Jackendoff 1983, Repp 1991, Brown 2001, Gilbers and Schreuder 2002, Schreuder 2006) that language and music are structurally similar. Both display some kind of hierarchical organization, alongside a temporal and a melodic structure; moreover, they have a common time-frame of acquisition. In this section, I will review the main areas of overlap between the two systems and broadly examine their differences.

2.1 The rhythmical level

In music, rhythmic organization is the product of two independent hierarchical structures, meter and grouping. Meter, understood as the regular alternation of downbeats and upbeats, only exists in music. In language, there is no such a thing as an isochronous beat: research has failed to establish the existence of isochronous intervals in spoken language. Experimental literature about isochrony in speech (e.g. Bolinger 1981, Dauer 1983, Terken & Hermes 2000) indicates that time intervals between rhythmically prominent elements are highly variable. Isochrony seems to be subjective rather than objective: listeners expect intervals to be isochronous and tend to adapt their perception to their expectations. In order to perceive regular patterns, absolute precision is not required and considerable latitude is allowed without destroying the sense of isochrony. In speech, the ability to predict the location (in time) of prominent elements may be beneficial under many respects: particularly in perception, it may help to detect the presence of prosodic boundaries and it may assist the listener in segmenting connected speech (Patel 2008: 147–148). Furthermore, the expectations for regular patterns result in perceiving sequential events as grouped together into higher-level patterns. Such grouping is common to both language and music and it constitutes an essential step in the interpretation of complex sound sequences. In music, it can be expressed by cues as varied as the relative proximity of note onsets, symmetry, thematic parallelism and changes in pitch. In its alignment (or misalignment) with linguistic groups, such as feet, prosodic words and phonological phrases, musical grouping plays an important role in shaping rhythm in songs.

2.2 The melodic level

A melody can be defined as a sequence of relative pitches. In this sense, it is common to both music and language; however, while musical melodies are based on a stable set of pitch intervals (discrete, fixed and scale-based), F0 contours in language are continuous and largely dependent on the context. Moreover, the stable
system of intervals observed in music allows melodies to make use of a tonal centre (the tonic), which is “felt as the focus of pitch stability in the piece” (Jackendoff & Lerdahl 2006: 45). There is no sense in which this might apply to language: in ordinary speech no tone is more stable or central than others in a structural sense. Speakers seem to rely not on exact interval size, but on changes in the melodic contour as cues in discriminating meaning at least at the postlexical level. The idea that speakers of tone languages would acquire words with a precise and absolute pitch associated to them (Deutsch et al. 2004) is controversial and still much debated.

Despite these differences, some close parallels can be drawn between language and music. For example, a consistent drop in pitch level is generally observed before musical phrase boundaries, much like the decrement in pitch range and fundamental frequency found in speech utterance-finally. In speech, pitch movements play an important role in marking intonation groups. In doing so, they typically align with the phrasal structure of the sentence in a way that resembles grouping in music (Patel 2008: 182–5). Moreover, it has been hypothesized that learning of rhythmic and tonal patterns in one’s native language influences the creation of rhythmic and melodic patterns in the native instrumental music (Patel & Daniele 2003).

2.3 The temporal level

The temporal level involves the durational patterning of events, which is measured by the time intervals between successive onsets (IOI). If we consider duration alone, a rhythm can be described as a sequence of IOIs between tones/sounds. Perceptually, rhythm appears to be built on temporal structures that rely on longer intervals — as opposed to shorter ones (Fraisse 1956). Longer intervals also play a key role in grouping, as segmentation tends to occur at events with relatively long IOIs. With regard to the temporal dimension of rhythm, music and speech are comparable in many respects, at least insofar as tempo or ‘expressive timing’ is concerned. One of the best known mechanisms both systems use to mark boundaries is phrase final lengthening, i.e. the consistent lengthening of a note or a speech sound at the end of a phrase (Patel 2008: 109–112). Other common mechanisms are the deceleration of tempo towards the end of phrases, and the acceleration at the beginning of each melodic movement (Schreuder 2006: 24–25).

2.4 Timbral contrasts

Timbre, understood as the quality of a sound, is a property of both language and music, although it contributes in different ways to build categories in each system. In language, timbre is the primary dimension for organized sound contrasts;
in music, however, organized systems of timbral contrasts are believed to be rare within the instruments of a culture (Patel 2008: 28). It has been observed that a systematic mapping of linguistic timbre (typically nonsense syllables) onto musical sounds is implemented in the teaching of specific styles of instrumental music (e.g. by tabla drummers, Patel 2008: 62). However, there is no knowledge of any singing tradition in which timbral contrasts of speech are systematically mapped onto the music, namely in songs. As a matter of fact, the role of timbre is usually not considered in discussions of textsetting. One exception is contained in Pescatori et al. (1988: 145–146) where it is suggested that consonants are excluded at the onset of Venetian rowing songs because their articulation would produce an obstruction in the vocal tract that would prevent the complete air intake necessary for simultaneously singing and rowing. It is not unlikely that the investigation of rowing songs in other languages would uncover similar constraints on consonant distribution.

3. The problem of textsetting

Despite the similarities displayed by music and language at a very general level, little is known about their actual interaction in vocal music in general, and in individual singing idioms, in particular. Given the nature of the two systems involved, as it is outlined in the previous section, we would expect interactions to show up at one or more of the structural levels mentioned above. The phenomena that have received most attention are located at the metrical and the melodic levels. Restrictions applying at the metrical level have been the focus of most studies dealing with non-tone languages, whereas the interactions between text and tune at the melodic level have been the main topic of studies focused on tone languages.

3.1 Tone languages

For tone languages, where fundamental frequency carries lexical meaning, research has aimed to establish how strong (or weak) the correspondence is between the linguistic melody of spoken words and the musical melody attached to the words in songs. Initially the debate has revolved around the question of whether there is a tendency to construct tunes such that the melodic contours in music mirror those found in speech. Working on Cantonese opera, Yung (1983) found that singers usually use the spoken melodic contours of words as a guide for improvising new texts on pre-existing melodies. He also noticed that there is often a one-to-one correspondence between particular language tones and particular pitches. The experimental work conducted by Saurman (1996) on 11 Thai informants led to similar results: she found that in Thai songs the more closely the
melodic curve follows the contour designed by the lexical tones, the easier it is for listeners to understand the lyrics when they hear them for the first time. As far as African languages are concerned, it has been suggested that tunes in Igbo and Yoruba not only have a strong tendency to accommodate language tones, but they also enhance phenomena of natural speech such as downstep and downdrift (Ekwueme 1974; Barreto Nogueira 2008; Villepastour 2014).

Studies in support of a parallelism between speech tones and musical pitches greatly outnumber those supporting the opposite view. In contrast to Schneider (1961), Agawu (1988) found no direct correlation between the melodic contours of speech and song in Ewe. Similarly, Jähnichen’s study of classical Vietnamese singing (2014) failed to establish a correlation.

A third way of looking at this issue has consisted in admitting that different singing styles or vocal genres use different degrees of correspondence between lexical tones and musical pitches. List (1961) found in Thai nursery rhymes a high degree of correspondence between the spoken and sung contours, whereas in folksongs this relationship is weaker (59% to 60%). Similar observations were made for Kammu, a Mon-Khmer language spoken in northern Laos: it appears that the two speech tones surface in different ways in songs depending on the vocal genres, “spanning from close to 100% correlation to around 50%” (Karlsson et al. 2014: 174).

Some studies have established a relationship between spoken tones, musical pitches and rhythmical structure of the song. With regard to grouping, Morey (2010) reports that when lexical tone and musical pitch align in Tai Phake songs, grammatical units need not match the grouping structure of the music. However, when they do not align, there is strict matching of grammatical units with the half-line unit. With regard to meter, Wee (2007) claimed that in Mandarin Chinese folksongs the melodic contours of the spoken and sung texts tend to be parallel only in strong metrical positions, i.e. on syllables that show up at the beginning of the measure as in (1a). Such syllables are called ‘heads’.

(1) a.

A schematic representation of these two measures is given in (1b) where the melodic tier provides the musical pitches and the tonal tier contains information on the linguistic tones carried by the underlying syllables.
Crucial to Wee’s account is a series of correspondence rules that take care of the alignment between the melodic and the tonal tier. According to them, the tonal contour of the head syllables in (1a) is preserved because the leftmost note associated with the syllable following each head is respectively not lower and not higher than D (in measure 2 this condition is vacuously fulfilled).

Reporting on Shona, a Bantu language of Zimbabwe, Schellenberg (2009) claimed that the locus of the parallelism between spoken tones and musical notes is not predictable, although the degree of correspondence was statistically significant across songs.

The alignment of tones and pitches at the temporal level is still relatively understudied. Saurman (p.c.) allows that in Thai songs vowel duration, as well as pitch, may play a role in the mapping of text to tune, at least in some genres. Schellenberg (2013) engaged to study if the durational phonetic correlates of tone were maintained in sung Cantonese. Interestingly, he found that durational distinctions of spoken tone are neutralized in songs.\(^1\)

More subtle interactions between linguistic tone and musical pitch have been further observed. As explained in Sollis (2010), in \textit{pikono} songs from Papua New Guinea a melodic phrase cannot descend further than an interval of a fifth. Within these limits, spoken contours and melodic contours match. In Cantonese, singers usually include an extra rising contour when singing words that have a rising tone. However, these extra rising contours do not appear when singing in the lower range of the voice (Schellenberg 2013: 91).

All these findings seem to suggest that the relationship between sung melody and spoken contour is not a linear one and must be studied while taking into account the multi-dimensionality of sound organization.

\(^1\) Of the six Cantonese tones, tone 2 is the longest one while tones 1 and 4 are the shortest.
3.2 Non-tone languages

In the domain of non-tone languages, a distinction should be drawn between stress-accent and non-stress-accent languages (Beckman 1986). Studies carried out on stress-accent languages have focused on the so-called “prominence matching” principle, an expression which usually refers to the way stressed syllables in words are assigned to strong beats in the music.

This principle is illustrated in (2) by means of an English example. In (2), one line from the traditional song “What shall we do with the drunken sailor” is illustrated (taken from Halle & Lerdahl 1993). At the bottom of the diagram is the conventional musical notation of that particular line, in the upper part is the corresponding metrical grid with its alternation of downbeats and upbeats, represented by the different heights of the x-columns (columns of three or four x’s are downbeats whereas columns of one or two x’s are upbeats). In the middle, are the lyrics set to the tune. The alignment of the text is done in such a way that the stressed syllables of polysyllabic words (drúnken, sáilor) are matched to the downbeats of the music (highest x-columns).

![Diagram](image)

It has been argued that textsetting in English folksongs is governed by the strict application of the stress-to-beat matching principle (Halle & Lerdahl 1993, Hayes & MacEachern 1996, Dell & Halle 2009, Hayes 2009a/b). When this principle is violated, a so-called ‘stress-beat mismatch’ arises. This type of mismatch is considered unacceptable by native speakers of English, and it is in fact rare in traditional folksongs.2

Other languages, such as Spanish and French, whose prosodic systems differ substantially from English, may disregard this rule and allow mismatches between prominence in speech and music. In French folksongs this rule is believed to be inviolable only at line end positions (Dell 1989, Dell & Halle 2009); misalignments

---

2. For prominence matching in Early Modern English songs, see Proto (2013a).
occur line-internally, but even in this context there is a significant tendency to avoid them (Temperley & Temperley 2013). In Spanish, the stress-beat matching principle doesn’t seem applicable in any position, since stress-beat misalignments are freely distributed across the line (Rodríguez-Vázquez 2010); moreover, they tend to result in lexical stress being perceived as shifted (Morgan & Janda 1989). The study of Italian textsetting has revealed that violations to this rule are allowed between syllables that are adjacent at the lower levels of the musical hierarchy as it is expressed by the metrical grid; however, at higher levels other factors such as directionality and end-line proximity come into play and influence the acceptability of misalignments (Proto & Dell 2013).

A study of German Volkslieder (Proto 2013b) has shown that violations of the stress-to-beat matching are allowed in this singing idiom as well, and some of them are not even perceptible to native speakers. In fact, manipulations of the musical contour in correspondence of the mismatch locations have revealed that native speakers’ perception of the stress-beat discrepancies relies not only on rhythmical factors, but also on melodic features. Namely, misalignments that are not perceived as such in the original setting become perceptible when switching the pitch values assigned to the two syllables involved.

With regard to the temporal patterning in songs, a ‘rule schema for textsetting’ was formulated by Hayes & Kaun (1996: 260), stating that the natural phonetic durations of syllables should be reflected in the number of metrical beats they receive. This rule schema seems to be at work in English folksongs, where syllables tend to receive a proportional number of metrical beats, which is a reflection of their natural phonetic duration. Proto & Dell (2013) found that in Italian folksongs there exists a special rhythmical configuration in which violating this rule would result in an ill-formed setting. This configuration involves a syncopated musical rhythm in which a shorter note on a downbeat is followed by a longer one, as shown in measure 2 in (3). Mapping the tonic and post-tonic syllables of the proparoxytone fávole to this pattern results in a durational mismatch that is unacceptable in Italian.

3. In any given metrical grid, the strength of any beat is determined by the number of levels at which the beat appear.

4. The absence, in Italian songs, of rhythmical patterns involving a short accented note followed by a long one (‘scotch snap’) had already been observed by Temperley & Temperley (2011).
It is not the case that syncopation is always unacceptable in Italian, nor is it the case that durational mismatches are incompatible with syncopation: a setting like the one shown in (4), where the tonic and post-tonic syllables of the paroxytone bambino are mapped to the same pattern as in (3), sounds perfectly natural in singing, although it violates Hayes & Kaun’s rule.5

(4)

At present, we do not have any clear explanation for this difference between proparoxytones and paroxytones. Duration appears to be the most stable correlate of stress in both cases (D’Imperio & Rosenthal 1999), and yet discrepancies at the durational level are allowed in (4) and excluded in (3).6

Turning now to non-stress-accent languages, research has focused especially on Japanese textsetting and the question whether it is moraic (i.e. one mora per note) or syllabic (i.e. one syllable per note). Since the prosodic system of Japanese is based on the mora, rather than the syllable, you would expect textsetting style to be moraic. Manabe (2009) found that settings in both traditional and Western-like children’s songs are consistently moraic. This tendency is confirmed by Hayes (2008) and Dell (2011), although the former observes that a limited role is also played by syllables. A more complex picture emerges from an on-going study by Starr & Shih (2012). While testing the perception of native Japanese speakers and Japanese learners with regard to text-setting, they found that Japanese choose moraic over syllabic settings when possible; furthermore, they noted that syllabic setting is more common in foreign stratum words than Sino-Japanese words.

Another interesting feature of Japanese songs concerns the alignment of pitch accents. In children’s songs, pitch accents apparently tend to be aligned either with strong beats (these songs begin and end on a strong beat) or with higher pitched notes — as in (5) measure 15, where the syllable .RI. carrying a pitch accent falls on a downbeat but is marked by a pitch jump.7 Although the melody may follow the contour of pitch accents, this does not seem to be a consistent pattern.

5. English translations for the lyrics in (3) and (4) are respectively ‘Read the fairytales in the little bed’ and ‘Read, child, in your little bed’.

6. It is interesting to note that the increased duration in stressed open syllables is much greater in paroxytones than in proparoxytones.

7. The diagram represents the Japanese national anthem (Manabe 2009:94). Pitch accents are capitalized.
Clearly, further research is needed to assess the effects of pitch accent on Japanese textsetting.

(5)

Other singing traditions exist in which the mapping of a text to a tune is mediated through the temporal structures of the language. This is the case in languages like Somali, Tashliyt Berber and Hausa, that have a phonological distinction between long and short vowels. Although their prosodic systems are based on the mora, like Japanese, their textsetting style is neither moraic nor syllabic, but rather takes into account both the weight of the individual syllables and the temporal patterns created by their succession within the line. As a result, musical settings do not merely obey the ‘rule schema for textsetting’ mentioned above, but abide by rules that govern the distribution of (groups of) heavy/light syllables across weak/strong positions (Banti & Giannattasio 1996; Dell & Elmedlaoui 2008; Schuh 2011). Quantitative systems of this type differ from others, like Estonian, in which quantity is also connected to stress and partly to a fall in the fundamental frequency (on overlong syllables): with the exception of a study focused on the Estonian Runic songs (Ross & Lehiste 2001), we unfortunately lack a thorough account of how these three features interact in singing in Estonian and related languages.

4. Challenges for linguistics

The problem of textsetting has been approached using different methodologies based on the type of language under investigation. Studies of text-to-tune alignment in tone languages have a rather phonetic character, since they analyse sung melodies as measurable, physical phenomena, in which all the pitches perceived by the listeners must be specified (important exceptions are discussed in Section 3.1). Approaches to non-tone languages are, instead, phonological in nature, in the sense that they aim to formulate abstract rules that pertain to only
a few points in the sung contour that define the relevant constraints. Interestingly, early research on prominence matching included studies based on instrumental analyses of the effects of textsetting on lexical stress in Spanish (Janda & Morgan 1988, and Morgan & Janda 1989), but progressively moved away from phonetics. However, the discovery that in German formal stress-beat mismatches do not necessarily result in a perceptible stress shift brings back the question of what acoustic parameters work as cues for linguistic stress.

The existence, on the one hand, of settings whose acceptability depends exclusively on differences in IOI — like durational discrepancies in Italian —, indicates that the temporal level deserves attention even in languages whose rhythm is not based on quantitative principles. On the other hand, for quantity-based languages that also show tonal contrasts at some level (as is the case in Estonian as well as Hausa and Ngizim, cf. Schuh 2011: 207) it would be important to establish if and how spoken tones contribute to structuring melody and rhythm.

As the above outline suggests, textsetting should be studied while simultaneously taking into account multiple levels of interaction between both the musical and the prosodic structures in a given piece. Most of the analyses carried out have instead been developed along a single dimension of sound organization within the prosodic and/or musical domain. It would be desirable if existing empirical work could be complemented by measurable and/or perceptual data that shed light on those interactions.

Acknowledgements

I would like to thank François Dell and two anonymous reviewers for their valuable comments on an earlier draft of this paper. Their criticisms and suggestions have helped to improve the article significantly. I am also thankful to Jennifer Fynn-Paul for reviewing my English. Any remaining errors are my own responsibility.

References


DOI: 10.1515/9783110874020


Author’s address

Teresa Proto
Universiteit Leiden
Witte Singel-complex
P.N. van Eyckhof 3
2311 BV Leiden
The Netherlands

teresa.proto@gmail.com