Singing the Syllables: Translating Spelling into Music in Tibetan Spelling Chant
Phil Rose
School of Language Studies, Faculty of Arts, ANU

Ich kann das Wort so hoch unmöglich schätzen,
Ich muß es anders übersetzen, …

Goethe Faust: erster Teil

1.0 Introduction
Every literate culture has ways of saying how the written words of their main language are made up of units of writing. In English, this is called spelling. This paper looks at spelling in Tibetan, where it is called sbyor klog (pronounced a little like “jaw lock”). However, it adopts the special perspective, in keeping with the theme for this volume of “sound in translation”, of how the spelling is actually chanted by a Lama. This is called Tibetan Spelling Chant.

How is Tibetan Spelling Chant (henceforth TSC) an example of the theme ‘sound in translation’? In many fascinating ways. Although ‘sound’ appears self-explanatory - the Lama is, after all, creating sound when he chants - the reader will find that in TSC, sound is not a straightforward notion. Firstly, there are two different kinds of sound involved, namely speech sounds and musical sounds (and it is sometimes not easy to tell them apart). Secondly, there are many different ways in which these sounds can be looked at – acoustically, for example, or historically. These differences are important for understanding TSC.

In what sense is ‘translation’ involved in TSC? Translation involves the purposeful movement of an essential feature of something by changing one or more of its original parts. In translating from one language into another, for example, (the first sense of ‘translation’ that springs to mind) one moves meaning from one place (the source of the message, be it book, speaker at the UN, or whatever) into another (the head of the listener or reader) by preserving as much of the meaning as possible while changing the linguistic form which carries it. In Euclidean geometry, translation is to move a figure or body while preserving some of its features (it must not be rotated or dilated, for example). In TSC, unlike language, it is not meaning, but structure that is being moved, from one domain – language – into another – music. The structure involved is that of the spelling. Just like other examples of translation, not all aspects of this structure are preserved in TSC (since languages differ not in what they can say but in what they must say, in translating from one language into another, you cannot preserve all the original meaning).

In TSC, then, spelling structure is translated into music. To what purpose? In many cultures, considerable significance attaches to being able to spell: in Anglo culture it is still considered a mark of education, and is one of the myriad ways we, as social beings, use to discriminate between ourselves. Tibetan Spelling Chant is culturally significant for several reasons. Firstly, spelling is considered a necessary preliminary to fluent reading of the
Buddhist Scriptures, and novice monks are taught to spell by chanting the spelling of individual syllables: it is felt that chanting helps to make a difficult task – learning to spell – less onerous. It is also likely that chanting helps to memorise important aspects of the spelling. The primary significance of TSC, however, is that creating verbal energy in this way is considered a virtuous act, thus benefiting both chanter and listener(s).

This paper will thus illustrate some of the many aspects of sound involved in translating the spelling structure of Classical Tibetan into music. At the same time it will demonstrate the necessity of combining classical techniques of linguistic analysis with musical and acoustic analysis to yield a proper description of the genre. Some examples will be given of the complex mapping between the linguistic structure which encodes the orthographic structure, and the musical structure of the chant. The description is based on Rose (2001), a detailed and fairly exhaustive analysis of a recording of a corpus of 142 orthographic syllables chanted by an acknowledged practitioner: Lama Choedak Yutok, a Tibetan Buddhist Lama of the Sa skya tradition.

Spelling presupposes some kind of written representation of language. In the top part of Figure 1 is shown the first few words in a passage of Classical Tibetan written in Tibetan script. Like most South- and South-east Asian scripts, the Tibetan script developed from an Indic precursor and was adopted to write Tibetan as spoken in the 6th - 9th centuries (Beyer 1992: 41; van der Kuijp 1996: 372). It is probably best known in the West as the script in which the Tibetan versions of the Buddhist Canon have been transmitted (Miller 1956:1). The script runs from left to right, and the actual text starts after the two vertical lines. Syllables are separated by dots, so this excerpt contains 18 syllables.

The fragment of script in figure 1 is actually the start of the passage the spelling of which was chanted by Lama Choedak. It is possible, by ctrl+clicking where indicated in the figure, to get an idea of what Tibetan Spelling Chant sounds like. I have cut out the chanted spelling of some of the syllables (those shown in grey) to keep the recording to a manageable length and to include the spelling of several different types of syllables. So that the reader can follow the chant, the bottom part of figure 1 contains a representation

Figure 1. Part of a Classical Tibetan passage. Click here to listen to selected parts (shown in black) of the chanted form. The chanted syllables of the passage are shown below the script in quasi phonetic representation. Vertical lines indicate boundaries between spelling sentences; commas indicate boundaries between spelling phrases. Forms in italics indicate spoken syllables. ~ = nasalised vowel.
of the speech sounds of the chanted passage. The boundaries between the spelling sentences which spell the individual syllables are marked with vertical lines: the reader should be able to hear from the Lama’s pausing and breathing where these boundaries occur – this is where spelling of a particular syllable ends. The boundaries internal to the spelling sentences which mark the phrases that spell different components of the syllable are also shown.

Tibetan Spelling Chant is thus an expression of orthographic form in linguistic and musical structure. The remainder of this paper will describe and exemplify aspects of both the linguistic and musical structure, and show how they combine in the chant. There are two main parts to the paper. In the first part, which deals with linguistic structure, I explain how a written Tibetan syllable like \( \text{b*Uv} \) is spelled. In the second part I describe aspects of the melodic and rhythmic structure of TSC, and show with the help of descriptive acoustics how the linguistic structure is expressed musically, including some of the embellishments used by the chanter. Before this, however, it is necessary to discuss some interesting problems associated with the representation of speech sounds in TSC.

1.1 Representation

In figure 1 the speech sounds in chant have been represented with quasi-phonetic symbols, since the use of the proper symbols, like \([n]\), \([s]\) or \([\text{s}3]\), would only introduce unnecessary complication: the reader can, after all, actually hear the speech sounds involved. It should be realised, however, that the complexities in representing the speech sounds in TSC are not just associated with the use of unfamiliar phonetic symbols, but in fact go much deeper.

One complexity arises from the important distinction drawn in Linguistics between two levels of speech sound structure called phonetic and phonological. A phonetic representation shows the details of the actual sounds; a phonological representation shows only those aspects of the sounds that are relevant for their organisation in a particular language. For reasons that are too complex to address here – a proper account may be found in Rose (2001: 166) – neither a phonetic nor a phonological representation on its own is particularly useful for the sounds of Modern Tibetan in TSC.

Another complexity in representation comes from the fact that in chant the speech sounds are sung, and that in being sung, they may lose some of their linguistic features. For example, in Modern Tibetan pitch is just as much a part of a word as its vowels and consonants: it is a tone language. But pitch is also manipulated in chant, and the linguistic tone of a word may be modified or lost thereby. It is then a problem whether the linguistic pitch of the chanted word should still be represented. Because of these complexities, I have indicated chanted sounds in a compromise representation. Where it is necessary, I have used oblique slashes to indicate sounds in general, thus: /li/.
2.0 Linguistic Structure of TSC

In much the same way as a written English word is made up of different letters, the shape of a written Tibetan syllable is made up of different components. Both English letters and Tibetan components have conventional names. Spelling involves saying these names using speech sounds. For example, to spell the English word *thought* we would say:

\[ /t\acute{e}i\tilde{t}\,\text{o\,d}z\acute{t}\,\text{i}\,\text{spelz \ 0\acute{\j}\,t}/ \]
\[ “t” “h” “o” “u” “g” “h” “t” spells “thought”. \]

Here I have used the proper symbols to emphasise that the name of each of the components is itself an English word, made up of English speech sounds, for example \(/eit\tilde{t}/\) is an English word meaning the name of the letter “h”. The word *thought* has been used to demonstrate that there can often be a big difference between the spelling and the actual sound. There are only three sounds in *thought* – /θ o t/ – but seven letters. As will be seen, this is the case also with Tibetan, and for the same reason. How words are said in a particular language – what the speech sounds are and how they go together – is referred to as phonological structure. Phonological structure is therefore a necessary component in the description of spelling.

But there is more to spelling than just phonological structure, since the way the names of the letters, or components, are combined – their syntax – is also part of the convention of spelling. The syntax of English spelling is relatively straightforward. In an English spelling sentence, for example “c” “a” “t” spells “cat”, the letters are first simply named in sequence, and this sequence then usually constitutes the subject noun phrase: [ si e t i ]NPsubj.. This is followed by a verb *spells*, and then the name of the word being spelt follows in its entirety as the object noun phrase [ kæt ]NPobj.. Both *spells* and *cat* constitute a syntactic constituent called a verb-phrase. This syntactic structure, which is hierarchical, is conventionally represented in Linguistics with labelled brackets, thus:

\[ { [ \text{si e t i }]_{\text{NPsubj.}} \text{ [ spelz [ kæt ]_{NPobj.}} \text{]VP} }_{\text{sentence}} \]

\[ “c” “a” “t” spells “cat”. \]

It is interesting to note from the –s on the verb *spells* that the subject noun phrase, although it consists of separate noun-like units, is considered singular, thus representing semantic rather than syntactic structure (implied is: the letters “c” “a” “t” taken together as a unit spells “cat”. Compare this with the use of a plural verb in the sentence “c” “a” “t” are all letters of English.)

Note that this is of course not the only possible spelling syntax. Different languages have different conventions. In Ancient Chinese, for example, (simplifying a little) a possible spelling for a character pronounced *cat* would have been “cat cob hat cut”. What this stands for is: the Chinese character *cat* is spelt with the initial consonant of the character pronounced *cob* and the Rhyme of the character *hat* (Malmquist 1994: 10). (“Cut” stands for the Chinese name for this particular method of spelling: 反切 fânjî or reverse cutting.) We shall see a much more complicated spelling syntax for Tibetan.
Thus to describe the Linguistic structure of spelling we must refer to both phonological structure and syntactic structure. The same applies to Tibetan spelling, except that its syntax is far more complicated, and interesting than English.

2.1 A simple example

The best way of giving an idea of how Tibetan spelling works is to start from the simplest example, and then introduce more complexity. Figure 2.1 is an example of the simplest type of syllable. It is simple because it is composed of a single orthographic unit (although written with three strokes). For obvious reasons it is convenient to represent written Tibetan in romanised form. In the Wylie romanisation used in this paper (Wylie 1959) is romanised ba. It is important to realise that the romanisation does not represent the pronunciation of the component, i.e. the way its name is said. The Tibetan script was adopted to represent the sounds of Tibetan in the 6th to 9th centuries. All languages change over time, and the sounds of Tibetan have changed enormously since then. Since it is the modern sounds that are used to name the components of the syllable, and how the syllable itself is pronounced, it is necessary to distinguish between the romanisation, which is a representation of the orthographic components of the syllable, and their phonological representation, which represents the actual sounds used in naming the components as well as the final result. This situation is similar to English spelling: because of some very complicated changes in the pronunciation of vowels, the letter “e” is said with an /i/ vowel, as in the word she, not a vowel corresponding to the way it was said when the spelling was devised.

is one of a set of thirty symbols, called radicals, most of which are romanised with a sequence of a consonant plus an a vowel, thus: ba. To spell the radical ba on its own in TSC, it is simply spoken. (There are two instances of single, spoken, radicals in figure 1, indicated by italics). This means we have to know how the name of the component ba is pronounced in modern Tibetan, and this means talking about linguistic (as opposed to musical) tone.

As already indicated, Modern Standard Tibetan is a tone language, which means that pitch is just as much a part of the words or syllables as their consonants and vowels. In the kind of Tibetan spoken by Lama Choedak the tonology is very simple: there are just two tones: high and low (symbolised with acute and grave accents respectively). Ba is pronounced with a low tone, thus: /pà/. Ctrl+click here to listen to an example. Ctrl+click here to listen to an example with a high tone: /pá/. This high toned example is actually the name for the radical pa. It can be heard that both high and low tones are realised with a falling pitch, but the low tone has in addition an initial level or slightly rising pitch component which gives it an overall slightly convex pitch.

Figure 2.2 shows, using the two syllables /pá/ and /pà/, the acoustics that correspond to the pitch percepts of these two tones. Along the top is shown the wave-form of the two syllables - the rapid fluctuations in air-pressure that is sound. It can be seen that the high
tone syllable lasts for about 30 hundredths of a second, and the low tone syllable is a little bit longer, at about 40 centiseconds. Below the wave form are the fundamental frequency curves for the two syllables. Fundamental frequency (F0) is the basic rate of repetition of the complex changes in air pressure. It is the primary acoustical correlate of pitch, and corresponds to the rate of vibration of the speaker’s vocal cords. It can be seen that the F0 shapes correspond to the pitch shapes. The F0 of the high tone syllable falls from an onset at about 145 Hz to about 105 Hz. The F0 of the low tone syllable rises from an onset at about 120 Hz to peak about 130 Hz, and then falls to the same value as the offset of the high tone syllable.

Figure 2.2. Wave-form (above) and Fundamental frequency (below) for high and low tone syllables /pá/ (left) and /pà/ (right).

Note that the tonal difference between the two syllables is represented in the Wylie romanisation by a difference in the initial consonant: pa vs ba. At the time the spelling was devised, it is assumed that the main difference between the two syllables was in the initial consonant. One of the major changes that Tibetan has undergone is tonogenesis, whereby contrastive linguistic tone has developed from syllable-initial consonants, and then the original difference between the consonants has been lost. This is a change found in the history of many languages. For example the precursor of Vietnamese, a modern tone language, was not tonal. The tone of Modern Tibetan is preserved only in a few well defined cases in TSC, but, as will be demonstrated, it is sometimes reflected iconically in the musical structure, in the choice of embellishments to notes.
2.2 Vowel symbols

The next level of complexity is when a radical is combined with a separate symbol to indicate a vowel other than a. Figure 2.3 shows the syllable \( \text{bo} /pò/ \). It can be seen that it is composed of the radical \( \text{ba} \) with a \( \text{naro} \) symbol on top. This symbol represents the \( o \) vowel in \( \text{bo} \) and is called \( /nàro/ \) (there are three other vowel symbols, with names, representing the vowels \( i \), \( u \) and \( e \)). The syllable \( \text{bo} \) is spelt \( /pà nàro pò/ \), which can be translated as “the radical \( \text{ba} \) with \( naro \) vowel spells \( \text{bo} \)”. This is an example of a vowel phrase, a syntactic constituent which consists of three parts: the name of the radical \( \text{ba} /pà/ \); the name of the \( naro \) vowel: \( /nàro/ \), and the result of combining them: \( /pò/ \). We represent this formally as:

\[
[ \text{pà nàro pò} ]_{\text{vowel phrase}}
\]

This example starts to give a picture of how the spelling works. An input is specified: \( /pà/ \), then an operator: \( /nàro/ \), and then the resulting output: \( /pò/ \), the whole being combined in one phrase. As the vowel phrase \( /pà nàro pò/ \) also constitutes a spelling sentence on its own, a more complete syntactic representation would be:

\[
\{ [ \text{pà nàro pò} ]_{\text{vowel phrase}} \}_{\text{spelling sentence}}
\]

2.3 Postfix

One of a small selection of radical symbols, called a postfix, can be added after the radical. Figure 2.4 demonstrates this. It shows the syllable \( \text{bod} \) (which is actually the word for \( \text{Tibet} \)). The \( \text{ba} \) radical, with \( \text{naro} \) on top, can be easily recognised. The final symbol – the postfix – is the radical \( \text{da} \).

Although the syllable is spelt \( \text{bod} \), it is actually pronounced \( /pö/ \), without a /d/ and with a vowel different from /o/: /ö/ represents a front rounded vowel similar to the vowel in French \( \text{fleur} \) flower, or German \( \text{Götter} \) gods (the “\( \text{	extbackslash} \)”, for reasons of font incompatibility, represents the low tone ). The reason for the discrepancy between the spelling and the sound is historical, like the example of tonal development mentioned above. The syllable was originally pronounced something like \( /\text{bod}/ \). The consonant /d/ is made in the front part of the mouth, just behind the teeth, by the front part of the tongue. The vowel /o/ is made with the back part of the tongue in the back part of the mouth, and with the lips rounded. Historically, the originally back rounded vowel /o/ has been pulled forwards in anticipation of the front consonant /d/, but has retained its lip rounding and thus become a front rounded /ö/. This is another historical change called umlaut. It is probably best known from its occurrence in the history of Germanic, where its effects can still be seen today in Modern German vowel alternations in words like \( \text{Fuß} \sim \text{Füße} \) and their English cognates \( \text{foot} \sim \text{feet} \). In the history of Tibetan, the final consonant \( \text{d} \) then disappeared. A low tone has also arisen from the original syllable-initial consonant \( \text{b} \).
The syllable *bod* is spelt by first giving the vowel phrase to derive *bo*, then adding a postfix phrase to get *bod*, thus:

\[
\{ \{ \text{pà nàro pò} \} \text{vowel phrase} \{ \text{pò \ tò pò} \} \text{postfix phrase} \} \text{spelling sentence}
\]  

(5)

In (5) both vowel phrase and postfix phrase are shown combined as constituents of a spelling sentence. The vowel phrase and its structure in (5) are clear, and it can be seen that the postfix phrase, like the vowel phrase, contains three parts. The second and third parts of the postfix phrase appear analogous to the operator and the output components of the vowel phrase, but the first part – /pò/ - needs explaining. Given the input-operator-output structure, we would expect the first term in the postfix to be the same as the result of the preceding vowel phrase – here /pò/. However, it is one of the regularities of TSC that in the postfix phrase the input often anticipates the output. There is a possible reason for this, but it has to do with the performance of chanting itself and will be taken up again in the section below on musical structure.

2.4 Subjoined component

A very common component in the Tibetan syllable is found immediately under the radical. This so-called subjoined, or subfixed, component often corresponds to a *y* or *r* following the consonant spelt by the radical. So in the syllables *kya* or *tra* the *y* and *r* would be spelt by subjoined components. Figure 2.5 shows the syllable *byung*, with a subjoined *y*. The radical in *byung* is recognisable as *pà ba*, and immediately underneath it is \_*- the subjoined *y*. Below the subjoined *y* is \_*- the symbol for the vowel *u*, called /zhèbkyu/. The postfix is \_*- nga, corresponding to the final *ng*. The subjoined portion is spelt, in a subscript phrase, with the same input-operator-output structure as the vowel and postfix phrase, thus:

\[
[pà yàda jà] \text{subscript phrase}
\]  

(6)

The /da/ means joined, so /yàda/ means with ya joined, and the phrase at (6) could be translated as “the radical *ba* with ya joined to it spells /jà/”. Why is the output /jà/ and not /byà/? This is because of another historical change, called labial palatalisation, whereby syllable-initial bilabial consonants like /p/ /b/ and /m/ changed to corresponding palatal sounds like /ch/, /j/ and /ny/ (Ohala 1978: 370-373), so the syllable written *bya*, and originally said /bya/, is now said /jà/. (In some parts of Tibet, where the change did not take place, the pronunciation is still /bya/). Labial palatalisation is also the reason why the first syllable in the name for Tibetan spelling – *sbyor klog* – is said /jo/ like “jaw”, with a palatal, but is written with a *b*. The labial palatalisation was caused by the following /y/ sound, which is also palatal, but it was an acoustically rather than an articulatorily motivated change. It happened because the acoustics of /b/ before /y/ in a syllable like /bya/ are similar enough to the acoustics of /j/ in /ja/ for listeners to mishear /bya/ as /ja/ and then to repeat their mistake in saying what they heard, namely /ja/.
The whole syllable byung is spelt with three phrases: a subscript phrase, a vowel phrase and a postfix phrase, in that order. The process is cumulative, the output of each phrase acting as the input for the next (with the exception of the postfix phrase, which, as already explained, shows phonological anticipation), thus:

\[
\{ pà yàda jà \text{ subscript phrase} \ [ jà zhèbkyu jù \text{ vowel phrase} \ [jùng ngà jùng] \text{ postfix phrase} \} \text{ spelling sentence} \ (7)
\]

2.4 Further expansions

Above I have illustrated four orthographic components in a Tibetan syllable: radical, vowel, postfix and subjoined unit. There can be up to seven components, however. Figure 2.6 shows a Tibetan syllable with all seven: bskyugs (it means vomited). Its radical is \( \text{ŋ} \) ka, in addition to which a subjoined ya, and a zhebkyu vowel can be recognised, making up the \( े \) kyu in bskyugs. To this a postfix \( \text{ŋ} \) ga is added, making up \( \text{ŋ} \) \( \text{k} \) yug. The new components are: a post-postfix \( \text{k} \) sa after the \( \text{ŋ} \) ga; a prefix \( \text{k} \), recognisable as ba, before the \( \text{k} \) yug; and a superfix \( \text{k} \) sa, on top of the \( े \) kyu complex. The syllable is pronounced /kyú\( \text{ŋ} \)/ (\( \text{ŋ} \) is a glottal-stop, as in Cockney bu’er for butter, and corresponds to the postfix \( \text{ŋ} \) ga). Like silent letters in English, the three new components – the prefix ba, superfix sa and post-postfix sa – are not represented in the pronunciation. The syllable bskyugs ब्स्कुञ्ज is spelt as follows (the /o/ is an obligatory part of the prefix phrase):

\[
\{[pà o] \text{ prefix phrase} \ [sá káda ká] \text{ superfix phrase} \ [ká yàda kyá] \text{ subjoined phrase} \ [kyá zhèbkyu kyú] \text{ vowel phrase} \ [kyú? gà sá kyú?] \text{ postfix phrase} \} \text{ spelling sentence} \ (8)
\]

3.0 Musical Structure of TSC

One important fact about the musical structure of TSC is that much of it can be described in terms of the linguistic structure. This means that if you have to chant, say, the spelling of a vowel phrase, you know the range of possible notes on which to chant the vowel phrase components, and the metrical structure in which the beats carrying them must be organised. For example, the vowel phrase of the rather complicated first syllable in the chanted passage in figure 1 is /nga naro ngo/. This means adding a naro vowel to a nga radical spells /ngo/. The metrical structure assigned to such a vowel phrase in TSC would be a single measure of three quarter-note beats (3/4), with the input /nga/, operator /naro/, and output /ngo/ components each assigned a quarter note value. Furthermore, the operator component /naro/ would be chanted on a double-eighth note. This is shown in the top part of figure 3.1.

3.1 Rhythm of TSC

The acoustic basis of perceived rhythm in speech is elusive. However, the rhythm of basic beats in TSC does correlate fairly straightforwardly to acoustic duration. To demonstrate this, the bottom part of figure 3.1 shows a so-called wide-band spectrogram of the vowel
phrase /nga naro ngo/ of the first syllable. Duration, in centiseconds, is shown along the bottom and frequency, in Hz, up the right-hand side.

![Spectrogram](image)

Figure 3.1. Above: Representation of rhythm of vowel phrase /nga naro ngo/ of first chanted syllable. Below: Wide band spectrogram of /nga naro ngo/ showing duration of phrase constituents.

A spectrogram like this shows the distribution of acoustic energy in vocal output, and how it changes over time. It shows especially the energy associated with supralaryngeal gestures like vowels and consonants. The thick horizontal bands of energy, for example, are the supralaryngeal vocal tract resonances, called formants, and vowel quality is determined by the frequency of the first two or three of these formants (the band of energy just above 1 kHz in the syllable /ngo/ is the second formant of its /o/ vowel).

A wide-band spectrogram has good enough time resolution to allow any acoustic boundaries between the units to be identified, so that the units’ duration can be measured. (Quite often there are no clear-cut boundaries between speech sounds, but in this case there are.) The duration of the vowel-phrase constituents in figure 3.1 can be measured as: 53.6 csec. for /nga/, 52.5 csec. for /naro/, and 50.5 csec. for /ngo/. This shows they all have very similar duration. Measurements on 39 quarter-note beats throughout the chanted corpus (Rose 2001: 178) showed their mean duration to be 48.8 csec. (sd = 3.9 csec.), so the beats in figure 3.1 are just a little longer than average. The mean beat duration of 48.8 csec. can be conveniently rounded-off to a value of 50 centiseconds, which corresponds to a moderato tempo of 120 quarter notes per minute.
3.2 Melody of TSC
In Rose (2001), five pitches were distinguished for the chant melody: B♭, C, D, E♭ and F (B♭ is in the second octave below middle C; the others in the next higher octave), with the Lama appearing thus to be chanting in B♭.

Three of these pitches – D, C and B♭ – can be heard on the /nga naro ngo/ vowel phrase, with its D, C B♭, C melody (in this linear representation, beats are separated by periods). Ctrl+click here to listen to the phrase. Figure 3.2 shows the F0 of the vowel phrase. It has been superimposed on a wide band spectrogram to show how the changes in F0 signalling the pitch are produced relative to the vowels and consonants of the phrase. The most important features in the F0 time course – those corresponding to my D - C - B♭ - C pitch percepts – are the quasi level portions of F0 marked as T (target) 1 through 4. /nga/ D can be seen to have a quasi level F0 target at about 150 Hz. After this the F0 falls to the second target at 133 Hz on /na/ C, then falls again to the third target at about 113 Hz on /ro/ B♭. Finally, the F0 rises again to the fourth target at about 132 Hz on /ngo/ C, at a level, it can be seen, only slightly lower that of the second target on /na/ C.

Figure 3.2. Above: musical representation of chanted vowel phrase /nga naro ngo/. Below: fundamental frequency of vowel phrase superimposed on its wide-band spectrogram.
The F0 of speech or singling is never really static, because the vibration rate of the vocal cords producing it is affected by a myriad of factors in addition to the pitch target the speaker is aiming at. The most important of these are that the vocal cords have to change their tension in moving from one target to another; and that their rate of vibration is also affected by the vowels and consonants they are said with.

As already mentioned, the primary acoustical correlate of pitch is fundamental frequency (F0), which is the number of times per second, or Hertz (Hz) the complex sound wave repeats. In Rose (2001: 183 – 185) it was investigated to what extent the pitches of the chant melody matched their expected F0 values. In figure 3.2, for example, it can be seen that the match is very good. The expected frequencies for Bb, C and D are 117 Hz, 131 Hz and 147 Hz respectively (Baken 1987: 487); as pointed out above, the observed values were: 113 Hz (“Bb”), 133 Hz and 132 Hz (“C”) and 150 Hz (“D”).

It was found in Rose (2001) that of the five melody pitches, overall “Bb”, “C” and “F” corresponded well to the expected frequencies. However, there was a problem with the “D” and “Eb” pitches, with mean F0 values of 150 Hz and 152 Hz (the expected values would be 147 Hz and 157 Hz). Although the mean F0 values corresponding to the Lama’s “Eb” and “D” pitches do not differ too much from their expected frequencies, and are in fact statistically significantly different, the very small difference between them, and their bracketing by the expected F0 values, suggests that two separate pitch targets are not involved. This means, in turn, that it is not clear whether my inference above as to key (“Bb”) is justified. As always, too, in cases like this where only one performance by one subject is being analysed, it is not possible to determine whether the performance is a good exemplar of the genre, or whether there is a tiny lack of precision in the Lama’s pitch control.

In fact, the indeterminacy may be irreducible, given what is known about the relationship in professional singing between F0 and pitch targets, and the way deviations from expected values are professionally judged. Sundberg et al. (1996), for example, investigated the performance of a well known aria in commercially available recordings of 10 singers, and the appraisal by professional adjudicators of the extent to which certain notes in the aria were achieved. It was found (ibid. 294, 305) that the professional singers differed considerably in whether their pitch targets corresponded to expected F0 values. This means that discrepancies between pitch and F0 in this performance of chant are not necessarily indicative of inaccuracy. It was also found (ibid. 295) that there was not a great deal of agreement in the professional assessors’ judgements of the accuracy of the singers’ notes; and further that their judgements did not relate in a clear way to the singers’ accuracy as measured by their F0. This means that it may not be possible to rely, as one must do in questions of performance correctness in phonetic field-work, on the judgement of other practitioners as to the quality of the chant performance.

In view of this, a reasonable interpretation of the pitch targets in TSC is that the chant melody appears to be constructed from three basic pitches: HIGH MID and LOW, with an
additional SUPERHIGH pitch of restricted occurrence. LOW corresponds to B♭, MID to C, HIGH to D or E♭ in apparent free variation, and SUPERHIGH to F.

4.0 Embellishments
The melody of TSC is not solely composed of plain quarter and eighth notes, of course. The Lama makes use of several embellishments, four of which - rising and falling appoggiatura; and up and down triplets - are illustrated below.

4.1 Rising and falling appoggiatura.
Quarter notes are sometimes embellished with rising or falling appoggiatura. About ten percent of the beats have a sharp rise in pitch onto a level target. Ctrl + click here to hear an example. It is on the first syllable of a spelling sentence for Obsolete de, consisting of the vowel phrase /ta trengbo tè/: “the radical da with drenbo spells /tè/” (/trèngbo/ is the name of the vowel symbol e, transcribed e. For phonetic reasons it often sounds like /ringbo/ in chant).
It can be heard that the first syllable is chanted with a rising appoggiatura onto a HIGH pitch target (Eb).

Figure 4.1 shows the musical representation of the spelling sentence, and the corresponding acoustics. It can be seen that the F0 onsets low, at about 120 Hz, but rapidly rises to the HIGH target at about 152 Hz by mid syllable (cf Eb F0 at 156 Hz). It can also be seen that the duration of the embellished beat is a little less than 40 centiseconds, showing that the embellishment is accomplished within the duration of a single beat.

Syntactically, rising appoggiaturas are almost exclusively associated with constituent-initial position, and most of these occurrences are in fact at the beginning of a spelling sentence. Phonologically, most rising appoggiaturas also occur on syllables with low tones (figure 4.1 is an example). However, it is also found to a limited extent on high tone syllables with aspirated stops /ph th kh/ and voiceless fricative /s/ (Rose 2001: 186). The very first syllable (sa) of the chant in figure 1.2 is an example. Aspirated stops and voiceless fricatives, and Tibetan low tone, are speech sounds which are made with a relatively spread glottal aperture, and belong to the Halle and Stevens’ (1971) natural phonological class of so-called spread-glottis sounds. This suggests that rising appoggiatura has to do primarily with glottal configuration, and that the otherwise nicely iconic tonal conditioning (low tones can take rising appoggiatura) is, at least to a certain extent, epiphenomenal.

The spelling sentence in figure 4.1 also shows another typical feature of chant, namely the spoken final syllable (this is transcribed with cross noteheads, and a slur to represent the gradually falling pitch of speech; in the linear musical representation, it is indicated by italics, thus: C Bb). Spoken syllables have well-defined occurrence in chant (Rose 2001: 179). They are obligatory in utterance-final position, both in constituents like vowel phrases, as in the example in figure 4.1, and also when the spelling sentence consists of a single radical, of which there are two examples, indicated by italics, in figure 1.2. In figure 4.1 the F0 on the final syllable falls, in a way typical for speech, from 132 Hz to 109 Hz, and also has a duration shorter than the normal quarter beat.

The Lama also embellishes notes with falling appoggiatura. Ctrl + click here to hear an example, on the first syllable of the postfix phrase /si ta si/ (si with postfix da spells /si/). This postfix phrase spelt the last part of the much longer syllable srid. Figure 4.2 shows the musical representation and the corresponding acoustics. The first syllable is chanted with a falling appoggiatura onto a MID (C) pitch. It can be seen that the F0 starts very high, at about 160 Hz, and then, over about the first half of the vowel, falls abruptly towards the MID (i.e. C) target. (At least part of this initial fall is due to factors associated with the syllable-initial /s/.) The F0 overshoots the target a little, reaching about 128 Hz. It then recovers to about 132 Hz, almost exactly the frequency of C at 131 Hz. As with rising appoggiatura, the fall is accomplished within the duration for the beat.
The example in figure 4.2 illustrates two more predictable features of TSC postfix phrases relating to the operator (here *da*) and the output (here the final syllable /si/). The operator in a postfix phrase is invariably given a MID (i.e. C) target quarter note. In figure 4.2 it can be seen that the F0 on /ta/ remains fairly stable throughout the vowel. Its mean value is 132 Hz: again very close to C at 131 Hz. The duration of the syllable – 51.6 centiseconds – is also very close to the mean value of 50 csec. mentioned above for chanted quarter notes.

![Moderato](d = c. 120)

Figure 4.2. Example of falling appoggiatura. Above: musical representation of postfix phrase /si ta si/ Eb-C. C. Bb with falling appoggiatura on initial syllable. Below: Fundamental frequency of /si ta si/ superimposed on its wide band spectrogram.

Also predictable is the final syllable. It can be seen from figure 4.2 that the final syllable /si/ is very short: its duration is 23.1 centiseconds, of which the vowel lasts 11.5 csec. This is very close to the mean duration of such postfix-final syllables in the chanted corpus of 21.7 csec. This duration is thus short of half the duration of the quarter notes in the corpus (this is why it has been represented as a sixteenth note with an Abraham-Hornbostel extension symbol (Nettl 1964: 106)). All postfix-final syllables have this truncated duration.
The syllable also has a fully predictable LOW (B♭) pitch. The F0 of this particular example can be seen to be at about 113 Hz, and is typically at a value a little lower than B♭ (116.5 Hz). As with the first syllable F0, the initially falling F0 is a function of the syllable-initial /s/ and can be discounted as a pitch correlate. This extra-short B♭ note is thus fully predictable for the output of a postfix phrase.

It is possible that the very short duration of postfix-final syllables is related to the phonological feature of anticipation, mentioned in section 2.3 above, whereby the input of a postfix phrase often anticipates its output. The anticipation is presumably because the shortness of the final syllable makes it difficult to hear what the resulting form actually is, and therefore it is good to have already signalled it. Of course, this does not explain why the final syllable – the culmination of the whole spelling sentence – is so short in the first case! It may be that the shortness of the final syllable is to allow breath to be taken without compromising the overall rhythm and delivery of the chant.

Figure 4.3. Example of upwards triplet. Fundamental frequency of the vowel phrase /la gigu li/ Eb-F-Eb.C C. D-C superimposed on its wide band spectrogram. The first syllable /la/ carries an upwards triplet.
4.2 Up and down triplets
Upwards and downwards triplets constitute two more embellishments, although they are
rarer than the appoggiatura. Unlike the appoggiatura, there is no phonological restriction
on the type of consonant or tone that triplets occur with, but there may be some iconicity
involved in their clear conditioning by syntactic position: upwards triplets are clearly
associated with input position, and downwards triplets with output (Rose 2001: 189-190).

Ctrl + click here to hear an example of an upwards triplet taken from a spelling sentence in
figure 1. The triplet, from HIGH to SUPERHIGH to HIGH (Eb to F to Eb), is on the first
syllable of the vowel phrase in the spelling sentence for 🅱️ gling: /la kigu li/ (the vowel gigu
with la spells /li/. /kigu/ is the name of the vowel symbol i̯, romanised i). The musical
representation and corresponding acoustics for the phrase are shown in figure 4.3.

Figure 4.3 shows that the F0 on the /la/ is at about 153 Hz with a rapid mid-syllable
obtrusion to about 172 Hz (the rising F0 at the onset of /la/ is from a lower F0 value on the
previous syllable). The expected frequencies of Eb and F are 156 Hz and 175 Hz
respectively, so the F0 in this token is again close – within about 3 Hz – to that expected
from the musical pitch percept, and the centrally located obtrusion, with quasi steady-state
F0 values either side, also agrees nicely with the triplet interpretation.

As far as the rest of the phrase is concerned, the remaining three beats all have a MID (C)
target, and their F0, at ca. 130 Hz, can be seen to be very close to the expected 131 Hz. As
can be heard, the last beat has a falling appoggiatura onto C from D, and the F0
corresponding to this embellishment can easily seen. D corresponds to 147 Hz, so the
agreement is again very good between musical pitch and acoustics. Note that in order
to produce this falling appoggiatura, the Lama times its F0 fall precisely at the onset of the /i/
vowel in /li/, and times the necessary increase in F0 from the preceding C target so that it
occurs during the initial consonant of the /li/. This is a typical strategy in speech whereby
F0 changes that are not intended to signal pitch changes are timed to occur at the same
time as syllable-initial consonants (Rose 1989: 61-79).

An example of a downwards triplet, on the last syllable /se/ in the vowel phrase /sa
trengbo se/ can be heard here. This is from the spelling sentence for 🅱️ sel. The musical
representation and acoustics are in figure 4.4. It can be seen that the F0 in the last syllable
vowel shows a rapid drop in its middle to about 130 Hz, followed by a recovery. The F0 on
either side of the drop is between about 156 Hz and 150 Hz, characteristic of the Lama’s
HIGH (Eb) pitch target.

The rest of the vowel phrase is notable for the rising appoggiatura on the initial syllable
/sa/ (another case of a rising appoggiatura occurring on a syllable with a high tone, but
spread-glottis initial consonant). There is also an occurrence of the SUPERHIGH (F) pitch
target on the second eighth-note of the /trengbo/ vowel operator. As can be seen, the F0
reaches a peak of about 170 Hz, which is near the expected frequency of 175 Hz for F.
5.0 Summary
This paper has illustrated some basic features of Tibetan Spelling Chant – how the syllables of Classical Tibetan are sung. In so doing, and in keeping with the theme of this volume, many aspects of sound have been invoked. Representations of sound, reflecting structure at many different levels, have been exemplified. Sound has been represented acoustically, in spectrograms and fundamental frequency curves; phonologically, in transcriptions of spelling sentences; historically, in the changes that have occurred to speech sounds still preserved in the romanisation; and perceptually, in the melody and rhythm structures of the chant. It is necessary to understand all these levels of structure to be able to describe and analyse Tibetan Spelling Chant.

Figure 4.4. Example of downwards triplet. Above: musical representation of the vowel phrase /sa trengbo se/ Bb-C.C F. Eb-C-Eb with last syllable /se/ carrying a downwards triplet. Below: fundamental frequency of /sa trengbo se/ superimposed on its wide band spectrogram.
Another important thing demonstrated in this paper is how TSC unites the two most important functions of sound for us – sound as speech and sound as music – so that the speech expressing the structure of the spelling is echoed and enhanced by the music. It is not surprising that speech and music relate in this way in TSC. They are both produced by the same vocal tract, and driven by the same brain (although probably not the same parts of the same brain). This is reminiscent of Pinker’s (1997: 535) hypothesis that “… music borrows some of its mental machinery from language.” But the music of TSC is not, as Pinker would have it (ibid. p.529), the communicator of “… nothing but formless emotion”. Quite the contrary: since, as I have pointed out, most of its music is predictable from the spelling structure, the amount of variation free to convey emotion must be limited, whereas the formal component is considerable. One can, for example, infer many aspects of the linguistic structure of a spelling sentence from the rhythm and melody of its accompanying chant.

Here is an appropriate place to comment on my use in this paper of so-called Western notation to represent the music of TSC. Ethnomusicologists generally agree that Western notation is theoretically descriptively inadequate for transcribing some genres of music, but nevertheless continue to use it for pragmatic reasons:

Western notation, …, incorporates some of the characteristics of Western cultivated music and tends to accommodate the transcriber’s subjectivity which is usually rooted in Western cultivated styles. But Western notation can be modified and, because of the facility with which it can be used, it offers the most practical method of presenting new musical data in visual form. It forms the best basis for analysis and description of music. (Nettl 1964: 128).

Its use for TSC, in conjunction with the usual set of additional symbols like grace notes (Nettl 1964: 106) can be justified on very much stronger grounds than just pragmatics, however. There is, firstly, the generally extremely good agreement, exemplified above, between the acoustics and my conventionally represented melodical and rhythmic percepts. The syllables in a measure do indeed have durations that correspond to the notated rhythm. With the exceptions commented on in the text, the notes transcribed with a particular pitch do indeed have the expected fundamental frequency. Secondly, the fact that the spelling structure (as specified by the linguistic analysis) can be put into such good correspondence with the musical structure (as represented in Western notation) can also be advanced as an argument for the adequacy of the notation and the constructs it represents. Given these kinds of agreement, it would be perverse to insist that the notation is not a good model for some (though not all!) aspects of TSC musical structure. Just how good can be appreciated from the quality of an utterance with pitch and duration synthesised from the musical notation. Click here to hear a version of the chant’s first spelling sentence (sa ngada nga, nga naro ngo, ngö~ na ngö~ synthesised from the musical notation using Fujisaki’s (2004) command-response model. This utterance, the middle vowel phrase of which was described in sections 3.1 and 3.2 above, was synthesised assuming a moderato tempo of 120 quarter notes per minute, and a pitch sequence of Bb.
(on sa ngada) C (nga) D(nga) C(na) Bb(ro) C(ngo) D\check{\text{C}}(ngö~) C(na) Bb(ngö~). Click [here](#) to compare with the original. TSC does indeed appear to be largely structured in terms of rhythmic-melodic beats. One should not forget, also, that notation is simply a convention. One defines the convention, as I have done here, to represent as accurately as possible what is assumed to be the essential features of the performance.

There is still much to be learnt from this fascinating interaction of speech and music, not the least how the results of the analysis can best be used to teach students of Tibetan Buddhism how to chant spelling authentically. But the analytic approach exemplified in this paper should not blind us from the holistic significance of Tibetan Spelling Chant as one small part of a culture that is fighting for survival. That is, after all, where its true meaning lies.

6.0 References


