Speech and song synchronization: A comparative study

Beatriz Raposo de Medeiros[1], *Fred Cummins*[2]

[1]Universidade de São Paulo
[2]UCD School of Computer Science and Informatics, University College Dublin

biarm@usp.br fred.cummins@ucd.ie

Abstract

Does synchronization among speakers or singers require the presence of a beat? Is an implied underlying pulse or meter relevant? We set out to explore synchronization among speakers and singers as they speak or sing a variety of texts. We compare metrically strong nursery rhymes with non-metered prose. We compare singing in genres with two very different types of rhythm (samba and rock), and we compare sung and spoken versions of texts. In each case, we ask whether the rhythmic qualities of the texts facilitate synchronization. The metrical structure of the nursery rhyme does not facilitate synchronization compared to prose, while the simple beat of rock music does help. Further comparisons are provided in the text.

Index Terms: synchronous speech, song, syncopation, stress timing

1. Introduction

The ability of speakers to synchronize when reciting a common text is clearly seen in the ubiquitous practices of protest and prayer worldwide. Many such texts are over-practiced, as in the mantra-like repetition found in the Catholic rosary, or the texts are very short and are repeated rhythmically, as in most protest chants. In the former case, over-practice may help to support synchronization as the temporal patterns, while somewhat irregular, become predictable by virtue of familiarity. In the second, the speech borders on the musical through the use of emphatic beats, and perhaps also simply through repetition, as in the speech-to-song illusion [1].

Studies of synchronization among speakers have hitherto used the kind of pragmatically vacuous text so familiar from such classic corpora as the TIMIT or CSLU speaker recognition corpora [2, 3, 4]. While these studies have revealed much about the formal characteristics of synchronized speaking, much remains to be explored in assessing the influence of one or other text type, and the influence of music-like regularities in the process of synchronization. The role of periodicity and meter in particular warrant attention, as most theories of the temporal control of action would suggest that periodicity is beneficial, or even necessary, for synchronization [5], and yet prior studies have verified that synchronization of speaking is possible in the absence of any demonstrable periodic structure [6, 4].

We adopt a stance with respect to speech that sees it as temporally structured, coordinated movement. In this, our explorations lie within a long tradition stretching back to Stetson who famously characterised speech as “movement made audible” [7]. We also adopt a dynamical perspective, viewing synchronization among speakers as a form of entrainment [8, 4]. If we use a relatively strict definition of synchronization as “doing the same thing at the same time”, there are relatively few truly synchronized behaviours. These include military marching, some sports such as swimming, rowing, trampolining and diving, some forms of dancing, and music making in unison. All of these activities have at least one of the following two features, and some have both: There may be a clearly perceptible beat or isochrony to the behaviour, and/or the temporal evolution of the activity is strongly scaffolded by inertial, elastic, or gravitational constraints. Synchronous speaking, we note, not only frequently lacks overt periodicity, but it is also achieved in the absence of any such strong physical scaffolding [9].

While the use of a dynamical systems vocabulary provides us with a rich set of tools for approaching such coordinated behaviour within and between speakers, it leaves us with something of a conundrum, as synchronization among speakers can take place in the absence of any clear periodic structure [4], although, as noted, group chants frequently tend towards music-like repetition. We are therefore motivated to explore synchronization phenomena in which we vary the underlying temporal anchoring of the utterances, including utterances that are clearly periodic in underlying form, and those that are clearly non-periodic. To this end, we here explore both sung and spoken texts, with varying amounts of simple periodicity. A basic question we pose is whether periodicity actually facilitates synchronization, as a common-sense intuition suggests it ought.

The work presented here examines synchronization among speakers as we vary the speech material being spoken. In a first comparison, we look at synchronization of metrically regular versus metrically irregular speech, using a prose text and a nursery rhyme as central examples. In the second comparison, we use songs, shorn of accompanying music. We compare the synchronization in samba with synchronization found in rock. We also obtain spoken versions of the sung texts to allow a direct comparison of speech and song.

In surveying the large and poorly mapped lands between speech and song, this study is necessarily exploratory in spirit. The essential questions to be addressed are these:

- Will synchronization be facilitated by the presence of strong metrical structure in spoken texts (Nursery rhyme versus prose)
- Will rhythmic complexity affect the degree of synchronization observed (rock versus samba)
- Will synchronization be facilitated by the presence of an underlying, implied, musical beat (singing versus speaking)

2. Different Types of Speech

Four types of source text in Brazilian Portuguese provide the material of this study: one prose text, a nursery rhyme, a samba song and a rock song. The text in prose was extracted from a

10.21437/SpeechProsody.2014-137

novel (*Um Sopro de Vida (A Breath of Life)*, by Clarice Lispector) and was chosen for possessing short and long sentences, as well as short and long words and topicalization, aspects commonly founded in oral speech. Prose's principal feature is to offer an irregular sequence of accents without any metrical structure. We use a nursery rhyme as a form of poetry in which metrical structure is more regular and in which beat expectations are maximally strong. Nursery rhyme recitation in group is a very common situation in infant-caretaker play in many cultures.

As regards the songs, we chose two different rhythmical types: a rock song (*Aluga-se*) and a samba song (*Preciso me encontrar*)[1]. For each song we obtained both a sung version and a spoken version, in order to see whether the musical meter would facilitate synchronization. We perceive, intuitively, that the rhythm of samba differs from that of rock, marking them as quite distinct. We refer to the rhythm of samba as syncopated, which positions the genre within a family of Latin American rhythms (e.g. salsa, habanera). In the case of rock, syncopation is not obligatory, and even, in some cases, this is not desirable (e.g. heavy metal).

Singing is understood in the present study, as well as in a previous study [10], as a specific variant of speech we can call sung speech. Despite being relatively rare among linguistic studies, comparisons between singing and speaking can raise interesting questions about how apparently very different systems, such as language and music, fuse so well in song.

3. Singing as a Type of Speech

Singing is a ubiquitous phenomenon, though its function can greatly vary among human groups. Popular song is widespread in Western culture and, broadly speaking, is consumed as entertainment. In the particular case of Brazilian song, there are many types of songs that could be categorized by theme or rhythm, however the present study will focus on only two types: samba and rock. Popular song is the most prevalent genre in Brazil and is profoundly integrated into everyday social life. It is no overstatement to say that the song is Brazil's music. A typical feature of several types of Brazilian song is singing together: From the very popular rural work songs to the urban *rodas de samba*[2]. The first samba composers, those from the early 20th century, did not have any formal musical education, and so did not read or write scores. Samba composition was oral, writing down only the lyrics and repeating the melody many times in order to memorize it. Some Brazilian song scholars have suggested that the melody of samba bears the hallmarks of the intonation patterns of Brazilian Portuguese, though this remains largely untested. It is a point we revisit in the discussion.

Singing is a kind of hybrid phenomena in which speech and music meet. The songs chosen for use in this study are from two different rhythmic genres: samba and rock. Despite their differences, both genres allow an easy fusion of text and music. The principal difference lies in how the rhythmic phrases in these songs are related to the underlying stream of pulses, as discussed in the next section. Comparisons between speech and singing may shed a light on both segmental and prosodic characteristics of speech, showing how musical constraints adapt to language constraints and vice-versa [11, 12, 10, 13]. In this sense, it appears entirely plausible to consider the song a kind of speech.

Fig. 1 illustrates some of the radical temporal differences found between the sung and spoken manifestations of the same text. The extreme prolongation of the final word only makes sense within a musical context.

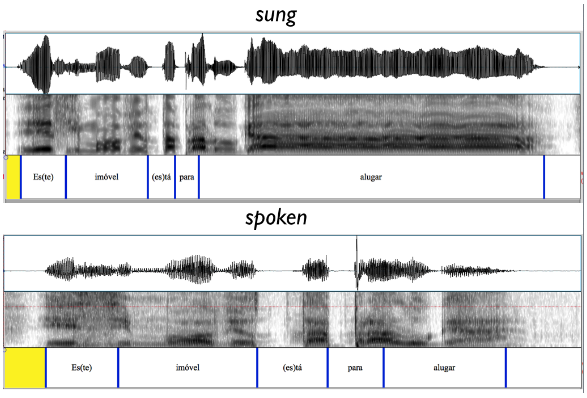

Figure 1: Samples of sung (top) and spoken (bottom) utterances from the rock song. Timescales differ between panels.

3.1. Different Song Rhythms

In our exploratory study here, we consider the relation between speech timing and metrical structure in both a poetic and musical context. Furthermore, we enrich the exploration of the relation between speech and musical rhythm by considering two types of musical rhythm: a simple 4/4 rock rhythm and a more complex samba rhythm.

A samba meter typically uses two beats per measure, and is thus a binary rhythm. When samba is transcribed, the conventional time signature is 2/4. Around these two beats, an off-beat system is built that is traditionally and commonly known as the samba syncopated rhythm. Specifically, there is a recurrent accent displacement that prevents a note from aligning with the second beat of the bar. This shift is caused by an sixteenth-note (semi quaver) of short duration that occurs just before the second beat, extending into and beyond the second beat. Another possible (and very frequent) accent shift is created by a bar-final sixteenth-note elongated and extending into the beginning of the following bar. Both accent shifts may be seen in Bar 3 of the score in Fig. 2.

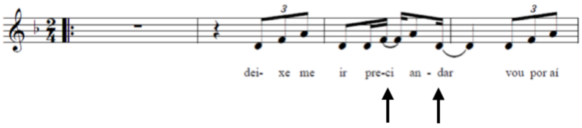

Figure 2: First four bars of Preciso me encontrar (*I need to find myself*). Brazilian *samba* by Candeia. Bar 3 depicts the typical rhythmic gestalt of samba and two common kinds of accent shifts.

As with other syncopated rhythm structures found in music, samba off-beats, as shown here, constitute the scaffold of a complex rhythm, as it can be characterized as shifting the initially proposed accent in a very specific way. There is some controversy as to whether the accent shifts found in Brazilian music stem from an African origin, or whether they are better

[1] The rock song *Aluga-se*, by Raul Seixas and the samba song *Preciso me encontrar*, by Candeia are very popular in Brazil.

[2] *Roda de Samba*: name given to a *samba* session, where people sit in circle or around tables, playing instruments and singing together.

understood as a Westernization of the complex polyrhythms of the African sources [14].

It would be overly simplistic to describe rock rhythms as non-syncopated. Syncopation is not unusual in rock. It is a well-known and salient characteristic of related genres such as ska and reggae, and is found to some degree in many mainstream examples of rock music [15]. However, a great deal of rock music displays a fixed 4-beats-to-the-bar meter, in which musical notes are aligned with the beats, strong accents occur in metrically strong positions, and the position of the beats is clearly signalled by the drum track (See Fig. 3). Where samba has an obligatory syncopation, it remains a stylistic option in rock.

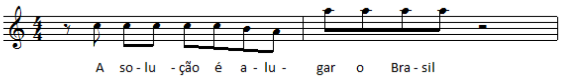

Figure 3: First two bars of Aluga-se (*For rent*). Brazilian rock by Raul Seixas.

4. Methods

Ten dyads were recorded, of which five were all female, three were male-male, and two were mixed sex. Subjects were aged between 25 and 45. All subjects reported no known problems with speaking or hearing. All were self-professed competent singers, and all were native speakers of Brazilian Portuguese.

Six texts provided six different experimental conditions. Texts are described in Section 2 above, and include a nursery rhyme, a prose excerpt, a sung samba song, a sung rock song, and spoken versions of the samba and rock lyrics. For each text, subjects were recorded in pairs, and were asked to remain in synchrony with one another. Choice of key and tempo was left to the subjects themselves. All singing/speaking was unaccompanied by music or an overt beat of any sort. Prior to recording, subjects signed an informed consent and read a text in prose (not recorded) in order to practice. Recordings were made using head mounted microphones (Shure SM10A) connected to a Marantz (PMD 661). Subjects stood facing each other inside the booth (Whisperroom 4872S) 1.3 meters from each other and began speaking/singing after the words “One, two, three. Ok” said by the researcher.

All recordings were segmented into sentence-length units: 7 for prose, 6 for the nursery rhyme, 13 for samba (sung and spoken) and 18 for rock (ditto). All 750 dyadic recordings were used in the subsequent analysis. The longest sentence belonged to prose, containing 36 syllables, and the shortest sentence was a rock verse, having 4 syllables. None of the sentences exceeded 11 seconds in duration.

A quantitative estimate of asynchrony was computed for each sentence using the method introduced in [4]. Two time aligned utterances are compared. Each is first represented as a sequence of Mel-Frequency Cepstral Coefficients. The sequences of MFCC vectors are then subjected to time warping, and the amount of warping necessary to map one utterance onto the other provides a quantitative estimate of the asynchrony between them. The measure is normalised by the number of time windows in the sequences, so that asynchrony values for shorter and longer sentences are comparable. The algorithm has been found to be most reliable when the calculation of the amount of warping, as indexed by the area under the warping curve, is restricted to voiced portions of the speech [4].

5. Results

Asynchrony values are not distributed normally, but are skewed right, due, in part, to some relatively large outliers. All asynchrony values were therefore first log transformed. Fig. 4 shows the distribution of asynchrony scores for the six conditions. The units of the asynchrony calculation are derived from the area under the warping function.

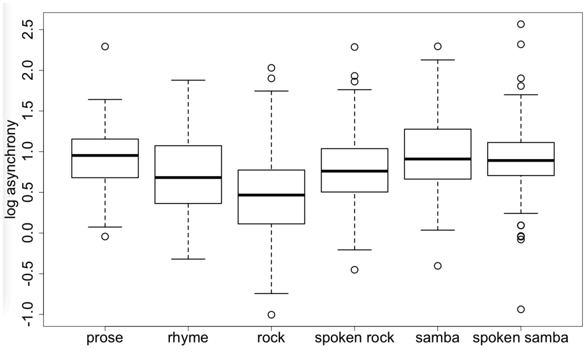

Figure 4: Boxplot of the distribution of log-transformed asynchrony scores for six different texts.

We are interested in specific planned comparisons that are of theoretical interest. The choice of comparisons is driven by our curiosity about the role of an underlying beat which may be pronounced and unambiguous (rock, nursery rhyme), present but subtle (samba) or absent (prose). Furthermore, we have available to us a comparison between sung and spoken lyrics within each genre. For each comparison, we conduct a simple t-test, and all t-tests are subject to a conservative Bonferroni correction to protect family-wise error rates.

The first comparison of interest is between prose and the nursery rhyme. Unexpectedly, the small difference in asynchrony observed between these two conditions was not significant (t(116)=2.4, n.s.), so there was no clear benefit of the regular meter in the nursery rhyme compared to the prose.

The next comparison examines synchronization in the sung versions of the rock song and the samba. Here, there is a marked difference. t(294)=8.1, $p < .001$. As expected, synchronization is greater in the rock condition than the samba, though we withhold interpretation of these results until the discussion.

The rock song was both sung, and spoken. The difference between these two is also highly significant, and as expected, synchrony is greater in the sung version. t(345)=5.8, $p < .001$. A similar comparison for the samba did not yield any difference between the sung and spoken conditions, however. t(255)=1.0, n.s.

A final comparison was done to look at the two conditions that contain the clearest metrical structure: the sung rock song and the spoken nursery rhyme. Here, a small difference was significant after correction of the p-value: t(118)=3.4, $p < .05$. Synchrony was greater for the rock song than the nursery rhyme.

6. Discussion

A naive assumption that we set out to test is that periodicity, overt or implied, would facilitate synchronization among speakers/singers. Although this hypothesis seems to be entirely in accord with common sense, there is room for doubt. The remarkable ability of speakers to synchronize in the absence of any overt or implied periodicity has now been well documented [6, 4]. Most strongly synchronized activities that humans engage in make use of external constraints to facilitate synchronization, and these take the form of a regular pulse and/or the structuring of the behaviour through the presence of strong inertial or gravitational constraints [9]. Synchronous speech exhibits neither property, while singing without musical accompaniment, as here, makes use of an implied beat.

To start with the spoken domain proper, we were somewhat surprised to find that the regular meter of the nursery rhyme did not seem to facilitate synchronization compared with prose reading. This result is consistent with past findings that speech synchronization does not require periodicity, but is at odds with the simple hypothesis that periodicity will necessarily facilitate staying in time with one another.

The picture changes when we compare rock to samba. Now the strong and clear periodicity underlying the rock produces a clear advantage compared with the more fluid and complex rhythms of samba. The underlying musical pulse also ensures that the sung rock is considerably more synchronous than the spoken version. The same comparison for the samba yields no advantage at all for the musical, sung version.

Song lyrics are conventionally sung, and embedded within the temporal structure of the song. It is possible that asking singers to speak lyrics instead of singing them may generate some confusion as to just how much of the musical structure to reproduce. For example, Fig. 1 shows a sung phrase in which the final word is massively prolonged. Subjects may have been uncertain about whether to reproduce such temporal effects in speaking. One might argue that any such uncertainty due to the instructions ought to influence rock and samba productions alike, and this is clearly not the case. However an alternative account might argue that there is a more intimate link between the rhythms of samba and those of speech, so that the kind of gross temporal exaggeration found in rock (see Fig. 1) is less likely to occur. While it does not settle the matter, the present investigation opens a potential empirical route of approach to such a discussion.

Both the sung rock text, and the spoken nursery rhyme text are metrically structured, yet the presence of an underlying meter does not facilitate synchronization in the same way for the two genres. Further investigation of the differing manifestations of temporal structure in speech and singing will have to be sensitive, not only to meter, but also to the conventions of the genre. Nursery rhymes, it appears, belong squarely in the speech camp where temporal expectations are very different from the musical domain.

The sampling of texts presented in this study is neither comprehensive, nor even representative of the many differentiations one could make in the grey area between speech and music. However, they are sufficiently diverse to illustrate some important characteristics of the relation between overt temporal form and underlying metricality and structure. Perhaps the clearest message to be gleaned is that periodicity by itself is neither essential to, nor required for, highly synchronized coordinated movement among simultaneous speakers. Much of the legacy treatment of rhythm in the study of human behaviour has tended to conflate the distinct concepts of periodicity and rhythmicity, to the extent that the mere presence of periodicity in an observed phenomenon is oftentimes sufficient for it to be labelled "rhythmic". But the rather ill-defined concept of "rhythm" is called upon to do duty in many contexts, from the mysterious oscillations found deep within brains to the aesthetic gyrations of pairs of dancers. From what we have seen here (and elsewhere), it is apparent that the term rhythm picks out quite different things in different domains. The rhythmicity of a nursery rhyme is not the same as that of a rock song, or a samba song, which are, again, mutually distinct. The coordinative relations observed in the temporal signatures of these diverse behaviours must be understood with respect to the skills shared by performers, the conventions of specific genres, and perhaps also the nature of the bond between practitioners.

7. References

[1] D. Deutsch, T. Henthorn, and R. Lapidis, "Illusory transformation from speech to song," *The Journal of the Acoustical Society of America*, vol. 129, p. 2245, 2011.

[2] J. S. Garofolo, *TIMIT: Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, 1993.

[3] R. A. Cole, M. Noel, and V. Noel, "The CSLU speaker recognition corpus." in *Proc. ICSLP*, vol. 98, 1998, pp. 3167–3170.

[4] F. Cummins, "Rhythm as entrainment: The case of synchronous speech," *Journal of Phonetics*, vol. 37(1), pp. 16–28, 2009.

[5] A. Cutler and J. Mehler, "The periodicity bias," *Journal of Phonetics*, vol. 21, no. 1/2, pp. 103–8, 1993.

[6] F. Cummins, "Practice and performance in speech produced synchronously," *Journal of Phonetics*, vol. 31, no. 2, pp. 139–148, 2003.

[7] R. H. Stetson, *Motor Phonetics*, 2nd ed. Amsterdam: North-Holland, 1951.

[8] R. Port and T. van Gelder, Eds., *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: Bradford Books/MIT Press, 1995.

[9] F. Cummins, "Joint speech: The missing link between speech and music?" *Percepta—Revista de Cognição Musical*, 2013, in press.

[10] B. Raposo De Medeiros, "Descrição comparativa de aspectos fonetico-acusticos selecionados da fala e do canto em portugues brasileiro," Ph.D. dissertation, University of Campinas, Campinas, Brazil, 2002.

[11] J. Sundberg *et al.*, *The science of the singing voice*. DeKalb, Il.: Northern Illinois University Press, 1987.

[12] J. Ross and I. Lehiste, *The temporal structure of Estonian runic songs*. Walter de Gruyter, 2001, vol. 1.

[13] R. Kolinsky, P. Lidji, I. Peretz, M. Besson, and J. Morais, "Processing interactions between phonology and melody: Vowels sing but consonants speak," *Cognition*, vol. 112, no. 1, pp. 1–20, 2009.

[14] C. Sandroni, *Feitiço decente: Transformações do samba no Rio de Janeiro, 1917-1933*. Jorge Zahar Editor, 2001.

[15] D. Temperley, "Syncopation in rock: a perceptual perspective," *Popular Music*, vol. 18, no. 01, pp. 19–40, 1999.