9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Acoustic distinctions between speech and singing: Is singing acoustically more stable than speech?

Beatriz Raposo de Medeiros [1], *João Paulo Cabral* [2]

[1]University of São Paulo, Brazil
[2]Trinity College Dublin, Ireland
biarm@usp.br, cabralj@scss.tcd.ie

Abstract

In this paper we study how spoken and sung versions of the same text differ in terms of the variability in duration and pitch. These two modalities are usually studied separately and few works can be found in the literature that report results about the comparison of their acoustic properties. In this work, recordings of both speech and singing of Brazilian Portuguese popular songs were conducted. Then, the variability was measured by statistical analysis of the fundamental frequency and speech rate, specifically the mean and variance. In a first study this was done at the syllable and sentence levels and latter at the phone level for further analysis. In general, results show that speech and singing variability cannot be differentiated in terms of the variance. We expected different results because singing is more constrained than speech both in terms of pitch (small variation within the note) and duration (metrical constraint). It seems that the results of higher pitch stability for singing reported in the literature cannot be generalised, particularly for the popular genre in which there is a prosodic proximity between singing and speech. These interesting findings also motivate to analyse other aspects of dynamic pitch and duration to better understand the prosodic differences between the two modalities.
Index Terms: speech and singing, prosody, acoustic analysis

1. Introduction

In general, humans can perceptually distinguish speech from singing and it seems we do this in a very intuitive way. However, it is also true that there are spontaneous spoken utterances that can sound as if they were sung. At first glance it seems difficult to unravel the phenomena that are common in both song and speech, but a reasonable number of scholars have already compared speech and music regarding memory, melody perception, intelligibility, rhythm and intonation [1, 2, 3, 4, 5, 6]. Although these studies reveal differences between musical and speech elements, they shed light on the interplay between pitch and rhythm. The illusory transformation from speech to song in [3] demonstrates the interesting similarities between them. This paper describes two perceptual experiments in which the repetition of the same utterances, as well as the insertion of modified utterances, made listeners judge the spoken phrases as a song.

The fundamental frequency (F0) is an important acoustic parameter that differentiates speech and singing [3, 7]. A singer sustains F0 at an approximately constant value over relatively long durations, such as during the musical note, which tend to follow each other in a controlled way. In contrast, a speaker generally produces more rapid and frequent F0 transitions. Some authors refer to F0 stability in the comparison between speech and song, e.g. [8, 9, 10, 11]. The stability definition we use in this work is aligned with the Tonal Hypothesis [9], that is, two tonal properties lead to a perceptual shift from speech to song: more stable tone targets and musical scalar intervals. Those works propose that song has a more isochronous rhythm and greater F0 stability within each syllable, which is in concordance with the assumption that a musical note approximately matches the syllable unit. Their experiments are mainly or solely based in speech stimuli, for example by using the speech to song illusion demonstrated in [3].

In our work, the aim is to distinguish speech from singing and we compare directly stimuli obtained by recordings of these two modalities. Evidences such as those brought by [3, 8, 9] indicate that low level acoustic characteristics need to be taken into account to explain the intimate relations between music and speech, as well as their differences. The aim of the present work is then to answer to the specific question: *How to distinguish speech from singing taking into account acoustic parameters related to duration and pitch?* The hypothesis is that for both prosodic aspects, singing will show greater stability than speech.

We chose statistical measures commonly used to analyse the stability of F0 and duration, which are their mean and variance. Our analysis of stability also depends on the time window chosen, because the acoustic stability is directly related to the variation of acoustic parameters along a certain time interval. For instance, an entire song is likely to be perceived as singing and hardly as speech. But we expect that the phenomena of ambiguous behaviour at issue can be better captured in shorter time intervals. Our stimuli preparation is similar to that in [8]: First, sentences were selected as linguistic units and secondly syllables were chosen as another unity in which F0 and rhythm were analysed. Following this study, we also conducted the measurements at the phone level for all the vowels. The reason for this was that we observed that dynamic acoustic effects may occur within the syllable (between phonetical or even musical elements) for singing, leading to results that would conceal our F0 stability hypothesis. The remainder of this paper is structured as follows. Section 2 describes the experiment elaboration, mainly how spoken and sung texts (which we name from now on as songs) were chosen, and its development. Section 3 presents results obtained for both duration and pitch aspects. Finally, Sections 4 and 5 are dedicated to discussion and conclusion, respectively.

2. The Experiment

2.1. Selection of Songs

The texts chosen to be spoken and sung were those of four songs that constitute sung versions of excerpts from the literary work *Macunaima*. Due to its importance in the Brazilian modern literature, the book was followed by its spoken versions in a film and a play, around the 60s, whose titles were the same of the

10.21437/SpeechProsody.2018-110

literary piece. Many years later extracts became songs composed by the popular musician Iara Renno and recorded in the compact disc *Macunaima Opera Tupi* (MOT) released in 2008. Songs were mainly performed by a female singer, the composer herself, which was the reason for the choice of female subjects in our study. The set of texts originated from Andrade's literary work fitted very well to our purpose of comparison between song and speech, because they had spoken and sung versions of the same linguistic content. The song titles are: Conversa (*Conversation*), Jardineiro (*Gardner*), Macunaima (*Macunaima*) and Valei-me (*Betake me*).

For each title, a musical score was created by experienced musicians, for that information to be available for singers during the recordings.

2.2. Recordings

Three actresses and three female singers were recorded in an acoustical isolated booth equipped with an AKG microphone and a Boss recorder. The singers sung the chosen texts while the actresses read the same texts. In a preliminary trial to choose the actresses, we made sure that they did not know the sung versions of the chosen texts. This was a requirement condition in the experiment to avoid possible interference between any familiarity with the songs and the speaking task. Before the recording session, singers were asked to respond a short questionnaire in order to ensure that they could learn the songs both by listening to them and reading the score. Written instructions were given to the actresses so that they could produce acted speech reading the four texts. In this experiment, acted speech is the condition to be compared with the free tempo condition of singing. All participants had a time for training so they could become familiar with the task. Each song was performed at least twice by each participant and each session lasted around two hours.

2.3. Data preparation

First, the best speech and song recording of each text title were chosen. In total, there were 12 songs and 12 spoken utterances produced by two sets of subjects respectively (3 singers and 3 actresses). Next, the text was segmented into sentences following the criterion that they should be complete clauses (subject + verb + complement). But when clauses were formed by many phrases we attempted to divide them in a away that their semantic meaning was preserved. These criteria allowed to obtain in total 25 relatively short sentences that were not longer than 11 seconds, from the 4 texts. These sentences resulted in a total of approximately 285 syllables. Note that a small variation in the number of syllables could occur between recordings produced by different subjects.

Afterwards, the sentences were segmented into syllables. First, the syllable annotations were created in the form of Praat textgrids, for two of the subjects (one for singing and the other for speech). Then, the resulting segmentation was used as references to automatically generate the segmentations for the sentences of the other subjects using a syllable alignment algorithm based on the Dynamic Time Warping algorithm (DTW) implemented in Matlab. Finally, the resulting segmentations were manually revised. In addition, manual phone annotations were also obtained to carry the second experimental measurements on vowels. The rationale for choosing the syllable and phone units is that they are linguistic unities that carry a musical note.

Table 1: *Average duration results in terms of syllables per second, mean of syllable duration and its variance, obtained for the speech (SP) and singing (SI) conditions.*

	Syl/s		**Average (ms)**		**Variance (%)**	
Song	**SP**	**SI**	**SP**	**SI**	**SP**	**SI**
Conversa	5.6	3.9	178	256	0.51	2.89
Jardineiro	5.3	2.9	188	298	0.55	4.80
Macunaima	5.4	2.1	187	478	0.51	8.50
Valei-me	5.1	2.5	196	397	0.43	2.12
All	5.3	2.8	187	362	0.50	5.61

2.4. Acoustic measures

The syllable duration was obtained from the syllable label files containing the final time of each syllable. Then, we calculated the variance of the syllable duration for each sentence and averaged over all sentences of each song. For example, the mean value of the variance for the song *Conversa* in the speech condition was calculated from the variance obtained for 7 sentences and 3 actresses. We also calculated the mean syllable duration (for both singing and speaking conditions) by averaging over the singers/speakers. The number of syllables per second was also calculated, which is a common measure correlated with rhythm.

The F0 estimation was done with Praat using a 10 ms time window. Then, we calculated the F0 range, mean values of F0, and variance, first at syllable level and in a posteriori experiment at sentence and phone levels. The resulting values were also averaged over all the utterances/songs and the speakers/singers.

We run a t-Test, two-tail test (inequality), in order to verify if the differences in results between speech and singing were statistically significant. We also used a single factor ANOVA to test if the means of the different speakers/singers are equal in the speech/singing conditions respectively.

3. Results

3.1. Duration

The results for syllables per second (syl/s), mean and variance of syllable duration are summarised in Table 1. All the results for syllable duration are statistically significant according with the t-Test ($p < 0.05$). As expected, mean syllable durations were higher for singing and the average values of syl/s were lower for singing accordingly. The mean values of variance of syllable duration were significantly higher ($p < 0.05$) for singing compared with speech. This result was not expected according to our assumptions that duration stability is higher in singing and that the higher the variance the lower the stability.

Figure 1 shows the average syllable durations for one of the sentences of the song *Conversa*. In this example it is possible to observe that the sung sentence has higher variability in syllable duration.

When taking into account the results obtained for all the songs, the ANOVA analysis showed that there was no significant difference among singers for all measurements and among speakers there was a significant difference in mean syllable duration. This difference between the speech and singing condition was expected because singers are subject to the same duration constraints in producing musical notes, unlike speakers in the speech condition.

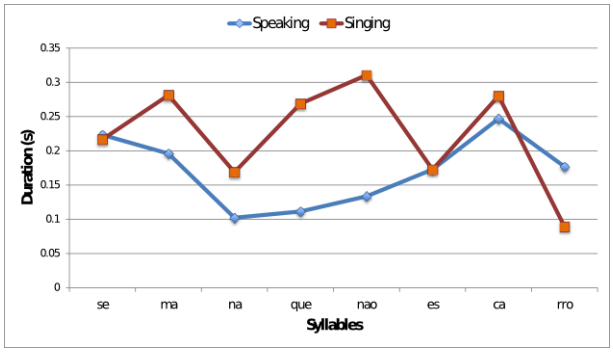


Figure 1: *Average syllable durations of one sentence for the spoken and sung versions.*

Table 2: *Average results obtained for F0 at the syllable level, in the speech (SP) and singing (SI) conditions.*

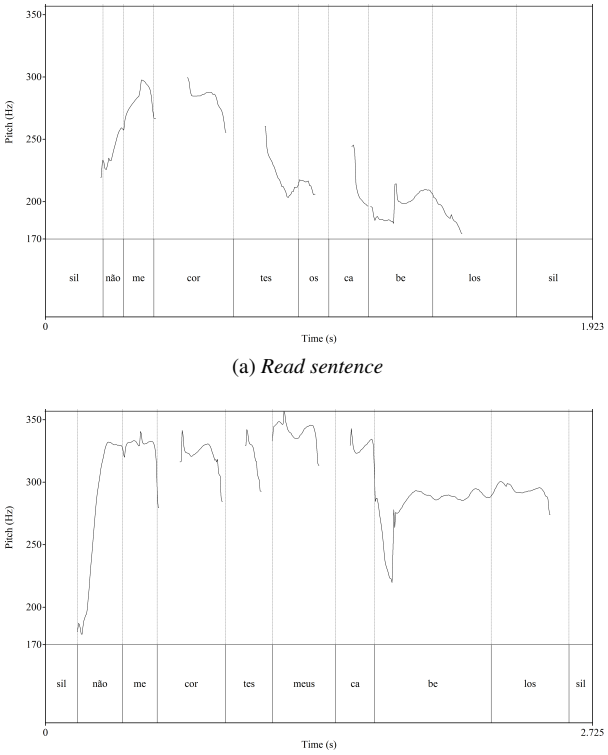
	Mean F0		**Variance**		**F0 Range**	
Song	**SP**	**SI**	**SP**	**SI**	**SP**	**SI**
Conversa	226	264	*292.4*	*197.2*	*38.7*	*33.2*
Jardineiro	209	289	163.8	267.4	26.9	40.1
Macunaima	198	248	107.6	186.0	26.2	35.4
Valei-me	218	284	*230.6*	*198.0*	*36.3*	*36.1*
All	212	270	*191.8*	*212.9*	31.5	36.3

3.2. Pitch

Table 2 shows the average results for F0 calculated at syllable level. The F0 mean was higher for singing, which was expected. For the variance, only the results obtained for *Jardineiro* and *Macunaima* were statistically significant ($p < 0.05$). Although the variance was higher for singing than speech for these songs, we cannot differentiate the F0 stability for the other two titles and all the titles together. In general, the F0 range was significantly higher for singing, with the exception of *Jardineiro* and *Macunaima*. But for these texts, results were not statistically significant, in contrast with the rest of the F0 range results. These results were not expected according with our hypothesis that F0 is more stable (lower F0 variability) in sung syllables than in spoken ones. The overall results at sentence level showed the same trend, although they are not presented here due to the limit in paper length.

In order to further investigate the importance of the high F0 variations in singing we performed another F0 measure at syllable level. It consisted of comparing the values between speech and singing for each syllable. For this, we needed to verify the syllable labels to ensure a one-to-one mapping of the syllables between the speaking and singing recordings. In this process, we performed occasional deletions of syllable labels. The rate of syllables with higher variance in speech compared with singing was $55\% \pm 6\%$ (95% confidence interval). This result is more supportive of the hypothesis that F0 is more stable in singing than speech. That is, a larger number of segments, in this case syllables, have larger variance in speech than singing. Our explanation for this result is that the effect of high F0 variations is attenuated because the number of segments in which there are high F0 variations in singing is small compared with the number of segments that have a more stable pitch.

The ANOVA test showed that there was a significant difference in mean F0 among speakers and singers. Regarding the variability measurements, there was only a significant difference in F0 range among speakers. In singing, both F0 range and variance did not show significant differences among singers, which can be explained by the F0 constraints imposed by musical tones. These results permit to exclude the factor of possible differences in the variance parameter among speakers and singers on the unexpected result that pitch variability was not higher for singing.



(a) *Read sentence*

(b) *Sung sentence*

Figure 2: *F0 contours of one sentence of the title Jardineiro that was read by one of the actors (top) and sung by one of the singers (bottom).*

We explain the high variance of F0 in singing by the effect of abrupt F0 changes that may occur in the transitions between tones in songs. Figure 2 shows an example of the F0 contours for one sentence. In Figure 2b, it is possible to observe abrupt transitions in F0, in particular at the start of the syllable "não" and in the transition between the syllables "ca" and "be". Nevertheless, this F0 contour shows much more stable F0 in other regions, such as in the final part of the sentence. In the F0 contour for speech, shown in Figure 2a, we can also observe segments with significant F0 variations but they appear to be less abrupt or lower in amplitude compared with the highest F0 changes of the singing contour. Thus, singing may have segments which show much more stability in terms of F0 than speech but the effect of F0 perturbations, such as pitch transitions between tones (glissandos) seem to influence the F0 measures of stability that we use.

We analysed in more detail the F0 variations within syllables for singing by visual inspection of the F0 contours. It permitted to verify that on one side, micro prosodic effects could enhance pitch values in the boundaries between consonants and vowels as phonetic factor. On another side, glissandos, which

Table 3: *Average results obtained for F0 at the phone level (vowels only), in the speech (SP) and singing (SI) conditions.*

	Mean F0		**Variance**		**F0 Range**	
Text Title	**SP**	**SI**	**SP**	**SI**	**SP**	**SI**
Conversa	231	265	135.8	99.4	26.0	20.5
Jardineiro	212	291	*49.3*	*138.3*	16.1.9	24.1
Macunaima	200	249	93.2	93.1	19.7	23.8
Valei-me	221	287	172.5	98.3	28.0	23.6
All	214	272	111	108	22	23

are continuous transitions of musical intervals, could mislead the verification that a tonal target was attained, this being a singing factor. As an attempt to avoid these effects on the pitch stability analysis, we calculated the results for the vowels segments only. Table 3 shows the average F0 results obtained for vowels. The F0 mean was higher for singing, consistent with the results obtained at syllable level. For the variance, only the results obtained for *Jardineiro* are statistically significant ($p < 0.05$). In this case, the higher variance for singing is not in concordance with our stability hypothesis. The mean variances of *Conversa*, *Valei-me* and of all songs together are lower for singing but these results are not statistically significant. Thus, the overall results at phone level were similar to those obtained at syllable level. It seems that the high F0 variations found at certain positions, such as the musical note onset, have a significant effect on the results even when the F0 measurements (F0 range and variance) are averaged over longer or shorter units than the syllable.

4. Discussion

Results shall be discussed in the light of musical aspects of songs that must be categorised as Brazilian popular music. This type of singing is performed in a way that the lyrics can be easily understood by the listener, thus it shows some similarities with speech. For example, in general, the pitch range did not exceed one octave which is similar to the pitch range in speech (196 to 392 Hz). The songs were similar between each other in terms of tempo, since singers based their training on the original tempo that ranged from 96 to 120 beats per minute (bpm). This tempo range, in general, is also comparable to the typical speech rate range. On one hand, these similarities represent a challenge in this work to distinguish singing from speech. On the other hand, the approximation between the two permitted to test the stability measurements when there are subtle differences between the two modalities.

Although the mean F0 range values of the songs were relatively small, the variability in pitch range was higher for singing than speech for some songs. This result was somehow expected at the sentence level, because the variation in tones can be significantly high in a long time window of the song. For the syllable and phone levels, the results were unexpected. One possible explanation are the high F0 variations that only occur in singing caused by the glissandos. Vibratos may also cause this variation and we will investigate this effect in future work. Other possible contribution to this variations are micro prosodic effects, although these can occur in speech as well.

Regarding syllable duration, song's time signature and tempo, which are typical musical constraints, determine regular time intervals between notes, and for this reason, low syllable duration variability for singing was expected. However, the duration variability was higher for singing than speech according to the variance results. We neglected some factors in our hypothesis of higher stability in singing that could explain this result. The first is that consonantal and vocalic intrinsic duration may influence sung syllable duration. Another factor is that syllables as long as one second, which are very common in final phrases in singing, elongate the vowel and this is a clear influence of a musical constraint on the syllable constituency.

Finally, our assumption that stability can be measured by the variability of the pitch and duration measures needs to be further investigated. As we pointed out in the introduction, tonal stability showed to be an important sound property that explains how listeners perceive spoken sentences as songs. Both duration and pitch can be set as cues that differentiate speech from singing, however the common parameters used in this work to measure this variation do not seem to be sufficient to understand this distinction, especially for pitch. As future work, additional phonetic details and singing voice elements will be taken into account, in order to better understand the stability role in a speech-singing comparison. For example, these factors include the consonant and vocalic intrinsic duration in a sung syllable and high F0 variation that can be observed in musical note onset and in the transition between notes.

5. Conclusions

In this work, we tested the hypothesis that singing is characterised by higher stability than speech in terms of prosodic aspects, specifically pitch and duration. In the experimental part, we conducted recordings of speech and singing so that we could compare them. The duration measurements, particularly the variance of syllable duration, showed that there was more variability for singing compared with speech, in general. Regarding F0 measurements at syllable level, on average the F0 range was significantly higher for singing and there was no significant difference in terms of the F0 variance. These results were unexpected because they do not support the stability hypothesis.

We conducted further analysis to find explanations for the results. First, the choice of the analysis unit had little influence on the results, by comparison of the syllable with sentence and vowel segments. Second, an additional F0 measure showed that there was higher rate of syllables with higher variance in speech compared with singing. This result gives more support to the hypothesis that F0 is more stable in singing than speech. Thus, the choice of the parameters to measure stability is very important and further work is necessary to determine the best parameters that permit to quantify this property.

Although the results did not permit to verify our initial hypothesis, they contribute for a better understanding of the difference in important prosodic parameters between speech and singing. They also motivate further research on the aspect of stability proposed in this paper and the relation between the two types of signal in terms of pith and duration.

6. Acknowledgements

This research is supported by FAPESP (Fundação de Amparo Pesquisa do Estado de São Paulo - Brasil, grant number 2015/06283-0). The second author is supported by the Science Foundation Ireland (Grant 13/RC/2106) as part of ADAPT (www.adaptcentre.ie), at Trinity College Dublin. We also thank the singers and actresses for their availability and Isabel Pie for her work at data preparation.

7. References

[1] R. Crowder, M. Serafine and B. Repp, "Physical interaction and association by contiguity in memory for the words and melodies of songs", *Memory & Cognition*, vol. 18, no. 5, pp. 469–476, 1990.

[2] 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, pp. 1–20, 2009.

[3] D. Deutsch, T. Henthorn and R. Lapidis, "Illusory transformation from speech to song", *Journal of the Acoustic Society of America*, vol. 129, no. 4, pp. 2245-2252, 2011.

[4] R. Johnson, D. Huron and L. Collister, "Music and lyrics interactions and their influence on recognition of sung words: an investigation of word frequency, rhyme, metric stress, vocal timbre, melisma and repetition priming", *Empirical Musicology Review*, vol. 9, no. 1, pp. 2–20, 2014.

[5] B. Raposo de Medeiros and F. Cummins, "Speech and song synchronization: a comparative study", *Proceedings of Speech Prosody*, pp. 748–75, 2014.

[6] A. Simões and A. Meireles, "Speech prosody in musical notation: Spanish, Portuguese and English", *Proceedings of Speech Prosody*, pp. 212–216, 2016.

[7] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing – Acoustical analysis and physiological interpretations", *Proceedings of the Fourth F.A.S.E. Symposium on Acoustics and Speech*, 2:57–70, 1981.

[8] A. Tierney, F. Dick, D. Deutsch and M. Sereno, "Speech versus Song: Multiple Pitch-Sensitive Areas Revealed by a Naturally Occurring Musical Illusion", *CEREBRAL CORTEX*, vol. 23, no. 2, pp. 249–254, 2013.

[9] S. Falk and T. Rathcke, "On the Speech-To-Song Illusion: Evidence from German", *Proceedings of the 5th Conference on Speech Prosody*, Chicago, Illinois, 2010.

[10] D. Gerhard, David, "Computationally measurable temporal differences between speech and song", *PhD Thesis*, School of Computing Science, Simon Fraser University, 2003.

[11] B. Lindblom and J. Sundberg, "The human voice in speech and singing," In T. D. Rossing (Ed.), Springer Handbook of Acoustics, pp. 669–706, 2007.