

FUNDAMENTAL FREQUENCY STABILITY IN SPEECH AND SINGING: A PERCEPTION STUDY

ESTABILIDADE DA FREQUÊNCIA FUNDAMENTAL NA FALA E NO CANTO: UM ESTUDO DE PERCEPÇÃO

Beatriz Raposo de Medeiros¹
Alexsandro Meireles²
André Baceti³

Abstract: Fundamental frequency (f0) stability is generally assumed to be an attribute of the singing voice, though it can also be found in speech. In a previous work, it was found that sung sentences performed with relative freedom within MPB (*Musica Popular Brasileira*) style show less f0 variability in comparison with their spoken counterparts and, for this reason, we account for f0 stability to be an important acoustic feature that distinguishes singing from speech. The current study is complementary to the previous study whose results statistically confirmed the hypothesis that singing has a more stable f0 than speech. Our aim with the perpetual experiment is to verify whether the stability feature is perceived by naïve listeners. 36 listeners were asked to classify sentences spoken by actresses and sung by singers from four available answers, choosing only one of them. The dataset used for the perceptual stimuli was formed by four textual units comprising 25 spoken and 25 sung sentences uttered respectively by an actress and a singer and was a representative part of the previous study full dataset. Listeners' classification obtained for each sentence was compared to the f0 variation of this same sentence. Statistical analysis was done by means of a discrete choice model (DCM) aiming to verify whether f0 variance could be an attribute for listeners to choose between one of the four responses. The result indicates a statistically significant ($Pr(>|z|) 2.22e-16$) relationship between f0 stability and the singing classification. The discussion takes listeners' impressions about the test and the conclusion indicates that the current finding is empirical evidence of cognitive aspects of stability in music.

Keywords: singing, speech, f0 stability, perception, Brazilian song

Resumo: A estabilidade da frequência fundamental (f0) é geralmente considerada como um atributo da voz cantada, embora possa também ser encontrada na fala. Em um trabalho anterior, sentenças cantadas produzidas com liberdade relativa no estilo MPB mostraram menos variabilidade de f0 em comparação com suas contrapartes faladas e, por esse motivo, consideramos ser a estabilidade de f0 uma característica acústica importante na distinção entre fala e canto. O presente estudo é complementar a este estudo pregresso, pois os resultados estatísticos confirmaram a hipótese de que o canto possui um f0 mais estável do que a fala. Nosso objetivo com o experimento perceptual foi verificar se a estabilidade é percebida por ouvintes inexperientes. 36 ouvintes classificaram sentenças faladas por atrizes e cantadas por cantores com base em quatro respostas possíveis, escolhendo apenas uma delas. Os dados usados para os estímulos perceptuais foram formados por quatro unidades textuais compostas de 25 sentenças faladas e 25 sentenças cantadas enunciadas, respectivamente por uma atriz e uma cantora e representavam grande parte dos dados

¹ Professor at the University of São Paulo, USP, São Paulo, SP, Brazil. biarm@usp.br
Orcid: <https://orcid.org/0000-0001-8298-0070>

² Professor at the Federal University of Espírito Santo, UFES, Brazil. meirelesalex@gmail.com
Orcid: <https://orcid.org/0000-0003-1901-9329>

³ Development Director of Murabei Data Science, São Paulo, SP, Brazil. a.baceti@murabei.com

do estudo anterior. A classificação dos ouvintes para cada sentença foi comparada com a variação de f0 para a mesma sentença. A análise estatística foi feita por um modelo de escolha discreta (DCM), a fim de verificar se a variação de f0 poderia ser um atributo para os ouvintes escolherem entre uma das respostas. Os resultados indicam uma relação estatisticamente significativa ($\Pr(>|z| 2.22e-16)$) entre estabilidade de f0 e classificação de canto. A discussão considera as impressões dos ouvintes sobre o teste e a conclusão indica que o achado atual é evidência empírica para aspectos cognitivos de estabilidade de f0 na música.

Palavras-chave: canto, fala, estabilidade de f0, percepção, MPB

INTRODUCTION

Recently, research on the relations between speech and singing was conducted focusing on experimental measurements of an aspect of voice that has not been given sufficient attention before, the fundamental frequency (f0) stability (RAPOSO-DE-MEDEIROS; CABRAL, 2018; RAPOSO-DE-MEDEIROS et al., 2021). The stability in producing voice can be understood as sustained vocal fold oscillation or sustained phonation and is commonly recognized as a singing characteristic, though it can occur in speech (TITZE, 1988; YANAGIHARA; KOIKE, 1967; LINDBLOM; SUNDBERG, 2007).

The initial study (RAPOSO-DE-MEDEIROS; CABRAL, 2018) presented a brief review of the literature on f0 stability and performed an analysis of the assumptions presented by other authors of stability in singing as an important feature to differentiate it from speech (DEUTSCH *et al.*, 2011; FALK; RATHCKE, 2010; TIERNEY *et al.*, 2012). The results confirmed that stability of f0 can be assigned to singing and distinguish it from speech. In a more recent work (RAPOSO-DE-MEDEIROS et al, 2021), the same was confirmed, using the same dataset of the initial study and performing a deeper analysis. The dataset contained different texts recorded in both a spoken and a sung version and were used for the purpose of a comparison between speech and singing. These texts or textual units belonged to four different songs to which a more detailed description is provided in section 1.1 of the present paper. The statistical analysis showed that for each textual unit (sung text versus spoken text) the spoken version showed more f0 variability. An F-test presented significant results ($p < 0.05$) for all comparisons, as it was run for each text in its sung and spoken version. Also, the question “Which modality (speech and singing) does have greater f0 variability throughout the vowel?” was answered: “Speech is the modality that shows greater f0 variability compared to singing”. This study focused on using f0 acoustic measures extracted at each vowel in order to obtain a f0 average per vowel, using one millisecond window, and can be designated from now on as the production study. The f0 estimation method was the one available in the software VoiceSauce (SHUE, 2010)⁴

As for the distinction between speech and singing observed in perceptual studies, the aspects focused, in general, are signal energy and register (SUNDBERG, 2013), vowel and consonant sounds (KOLINSKY, 2009), musical tones, and words recalling (VAN BESOUW *et al.*, 2009), and even processing frequencies and spectral patterns in the brain (ZATORRE *et al.*, 2002). However, regardless of a particular research field, we did not find a study directly dedicated to understanding the stability of f0 in speech and singing in the context of perceptual experiments.

Our aim with the perceptual experiment is, therefore, to verify whether this stability feature is perceived by naive listeners, assuming that there is some gradience between speech and singing, as discussed earlier in the production paper. We had already

⁴ The method is the STRAIGHT vocoder as explained in KAWAHARA *et al.* (1999).

highlighted that the speech and singing signals can be understood in an axis in which the edges can be well categorized as different, while the middles carry traits of both sound signals. Thus, we can think of talking as a more "pure" speech style, followed by reading aloud and chants as intermediate level, and baroque choir as a more "pure" singing style. We can represent this continuum in a line (see Figure 1). Obviously, these categorizations will change depending on the musical culture and society.

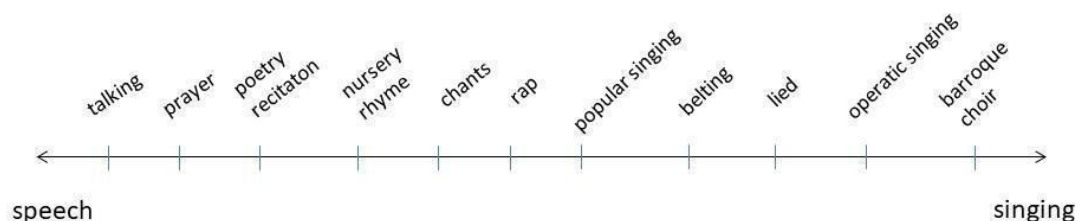


Figure 1: The speech-singing continuum reproduced from RAPOSO-DE-MEDEIROS et al, 2021.

The article structure has other four sections from 1 to 4 that expose and explain the actual experiment. Section 1 presents the perceptual test and its results, including a musical analysis of units that compose our dataset. Section 2 describes the statistical analysis, giving its result and searching to clarify the model, Discrete Choice Model or DCM. Some of the participants agreed to talk about the perceptual test task. Thus, this allowed us to gather qualitative impressions about the interplay between singing and speech. In section 3, we will discuss these impressions and how they relate to the quantitative results. Finally, we present the conclusion in section 4.

1. PERCEPTUAL TEST

1.1 Stimuli description

The stimuli used in the perception test are part of the dataset⁵ that is formed by recordings made for the production study (RAPOSO-DE-MEDEIROS et al, 2021). Each stimulus corresponds to a sentence belonging to four different songs, namely *Conversa*, *Jardineiro*, *Macunaíma* and *Valei-me*. We adopted the term textual unit to designate both the singing as well as the spoken version of the song. The word "song" alone can also be a substitute for the term "textual unit". Each song possesses different musical genres influences, being this a common feature in MPB (abbreviation for *Música Popular Brasileira* [Brazilian Popular Music]). A brief description of each song can be useful for a more complete approach to the phenomena at issue.

The song *Conversa* is in G minor, ranging from G3 to G4 with 116 beats per minute tempo (196 Hz to 392 Hz; 116 bpm). Showing a modal character and exploring a dominant and diminished scale in repeated notes, this song shows a confluence of various musical genres and, at the same time, can be heard as a typical northeastern Brazilian melody as for its modal character. *Jardineiro* in A minor (E3 164 Hz to A4 440 Hz; 96 bpm) contains a harmonic structure closer to the tonal one in relation to its cadences and no modulations occur. Thus, fifth and fourth chords, with the A chord suffice to accompany the melody. *Macunaíma* is in B minor of a modal nature (F3 174 Hz to F4 349 Hz/ 120 bpm). The melody is built on two main modes: the pentatonic-blues scale at the beginning and the B

⁵ Dataset is available at <https://lafalin.fflch.usp.br/bancodedados-I>

doric mode that dominates most of the song. What brings the feeling of bluesy sound is the presence of the blue-note, the natural F that forms the fourth-augmented interval in relation to the B note. However, this song possesses long phrases with repeated notes and small intervals between them with a format that can hardly be associated with a certain genre, but most of the time shows speech-like singing characteristics. *Valei-me* is composed in C minor and was built on the cadence Cm Cdim and Fm6 (F3-174 Hz to G4-392 Hz; 112 bpm). The melody is developed in the respective modes generated by the chords. Natural C minor or Aeolian mode appears in C (C D Eb F G Ab Bb C) from the beginning to the chorus, which is elaborate in the blues pentatonic scale (C Eb F (F#) G Bb), and with the melody emphasizing the tritone or augmented fourth.

In terms of speech style in the spoken counterparts of these songs, we can say that actresses, at the time of data collection for the production experiment, were successful in performing the reading tasks with free interpretation, based on the fact that texts are well known in their spoken versions for theater and cinema⁶.

Figure 2 below illustrates pitch contours of the same sentence, spoken and sung. We deliberately avoided bringing the musical scores of songs, because, despite being useful to show the melodic ascendent-descendent positions of notes in the score, it was not evident for non-musicians.

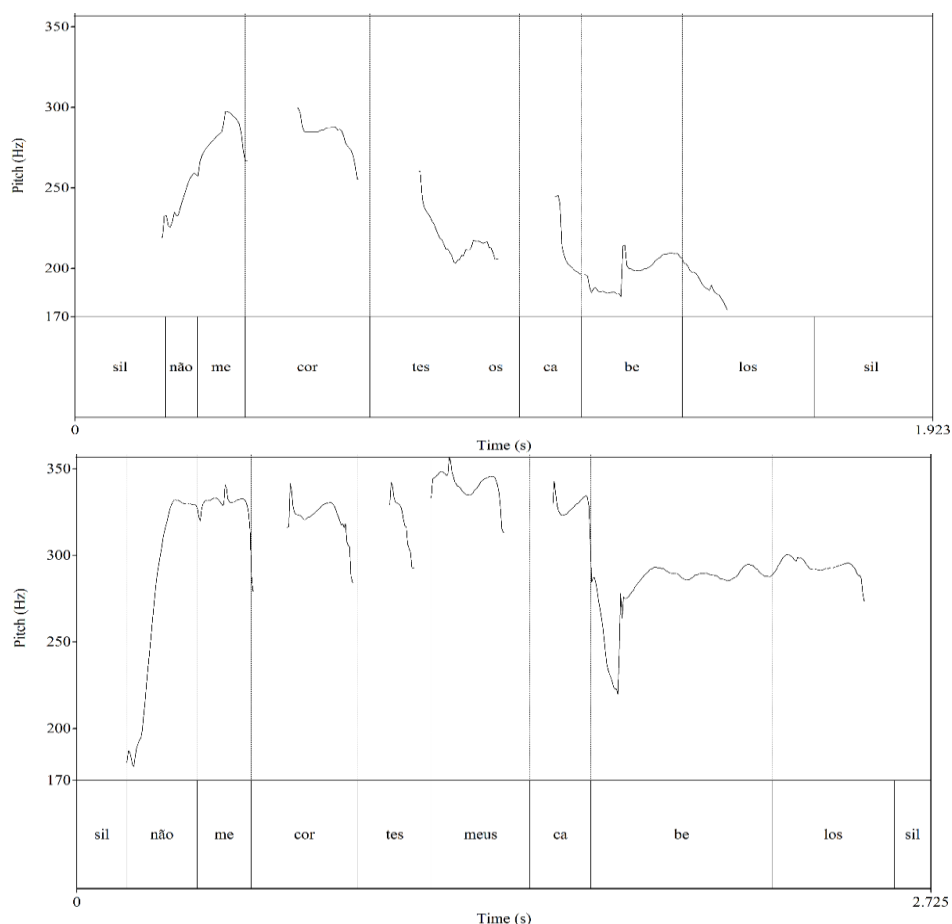


Figure 2: Pitch contours of one sentence of the title *Jardineiro* that was read by one of the actress (top) and sung by one of the singers (bottom).

⁶ The literary piece from which songs by Iara Rennó and versions of acted speech for the movies and theater derived is *Macunaíma* by Mário de Andrade. The album *Macunaíma Opera Tupi* by Rennó can be found on music platforms.

These examples depict different patterns of intonational contours. Except from the stylistic glissandos in singing produced at the beginning of the sentence and before the syllable /be/, vowels tend to sustain the f0 as they are the music notes themselves. In speech, on the other hand, there is much more f0 variation, as can be seen by the f0 contours in figure 2.

1.2 Data collection

The experimental design comprehends 36 listeners, 25 sentences, two modalities (speech and singing), and one response for each sentence that was listened. This results in a total of 1800 occurrences, considering 900 speech stimuli and 900 singing stimuli.

The set of sentences, therefore, which was reproduced to serve as a stimulus to the listener, consisted of 25 spoken sentences and 25 sung sentences, purposefully confusing all the songs and presented only once to the listener. The sentences were no longer than 10 seconds and the short spoken ones could have a 1.5 second length, as their sung counterpart could be a quarter long (2s). In fact, even sung sentences rarely exceeded 6 s. The number of words per sentence was variable, and the number of syllables ranged from 7 to 18, approximately. On average, spoken syllables were 187 ms long and sung syllables were 362 ms long. The rate of syllables per second was 5.3 in speech and 2.8 in singing (RAPOSO-DE-MEDEIROS; CABRAL, 2018). The song lyrics, or text units, are available at the Appendix.

We considered that presenting each stimulus just once would enable the subjects to remain as naïve as possible, since it would reduce chances of learning the task.

Listeners had four options as answers (see Figure 1): Only Speech, Speech Similar to Singing (SpSSi), Singing Similar to Speech (SiSSp) and Only Singing. These options were being presented on the computer monitor while participants were listening to the stimuli.

For each one of the sentences listened as a stimulus, participants had to choose only one of the four options (see Figure 3). This is a MFC (Multiple Forced Choice) test available at Praat. We run this test in Praat (version 6.0.25) as it is part of its scripts. The computer screen with the response option as depicted in Figure 1 was presented throughout the test duration, while stimuli were played to the listener.

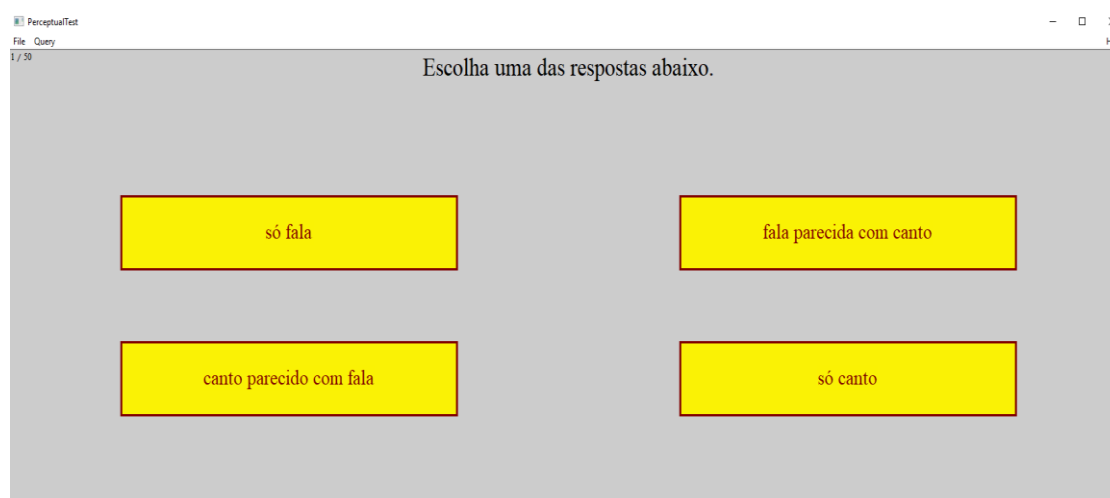


Figure 3: Top: Choose one of the answers below (*Escolha uma das respostas abaixo.*). Four options in the yellow rectangles: Only Speech (*só fala*); Speech Similar to Singing (*fala parecida com canto*); Singing Similar to Speech (*canto parecido com fala*); only singing (*só canto*).

Participants were 36 students of the University of São Paulo (*Universidade de São Paulo*, Brazil), with an average age of 25 years, who signed an informed consent document agreeing to participate as volunteers. Gender/sex was not taken into account for the test hypothesis, thus just for information there were 21 females and 15 males. None of them reported having formal training in music. At the end of each section, the participant was invited to talk freely about his or her impressions about the test. None of the participants knew Macunaíma Ópera Tupi, the CD that contained the songs used as stimulus in the perception test. A compilation of the impressions is made and a qualitative analysis is proposed to link them to obtained results in the discussion section.

Analyses from the results that follow account for answers related to two sets of data, speech data and singing data, as stimuli belonging to different textual units were scrambled. Although this work's focus is on differentiating speech from singing, regardless of the individuality of each textual unit, we considered it to be useful in achieving an auxiliary analysis taking into account each song's results.

1.3 Results

We start by presenting simple percentage results, as we consider them useful to reason about the answers distribution.

Table 1 (General): Percentages of responses to speech (n = 900) and singing (n = 900) stimuli. All subjects and text units.

	Only speech	SpSSi	SiSSp	Only singing
Speech	75	23.2	1.8	0
Singing	0.1	11.2	28.7	60

The general result presented in Table 1, above, relates to all songs mixed. It is worth observing that when the stimulus is speech, it was never perceived as singing. However, the opposite is not true: when the subject listens to a singing sentence (no matter which song) he/she would perceive this as just singing in only 60% of cases. Thus, we propose an explaining analysis regarding each song. Each of the following tables refers to a different song and their results will be discussed in the next two sections, one focusing on the speech version, and the other on the singing version.

Table 2: Percentages of responses to speech and singing stimuli for the unit *Conversa*.

	Only speech	SpSSi	SiSSp	Only singing
Speech	91.7	8.3	0	0
Singing	0.40	18.3	31.35	50

Table 3: Percentages of responses to speech and singing stimuli for the unit *Jardineiro*.

	Only speech	SpSSi	SiSSp	Only singing
Speech	69.4	27.8	2.8	0
Singing	0	7.9	26.2	65.9

Table 4: Percentages of responses to speech and singing stimuli for the unit *Macunaíma*.

	Only speech	SpSSi	SiSSp	Only singing
Speech	76.4	21.3	2.3	0
Singing	0	9.3	34.3	56.5

Table 5: Percentages of responses to speech and singing stimuli for the unit *Valei-me*.

	Only speech	SpSSi	SiSSp	Only singing
Speech	57.8	40	2.2	0
Singing	0	8.3	21.7	70

1.3.1 About the perception of speech stimuli

Table 2 indicates that *Conversa* was the textual unit that was most often perceived as speech, when the stimulus was exactly speech (91.7%). Several factors can contribute to this, especially the lexicon and syntax in an informal register, in addition to the topic being commonplace, exactly as in a conversation. It should be noted that the listener was not aware of the song's titles. *Conversa* was the only song that presented 0% for the SiSSp response and also the only one that presented a very low value for the SpSSi option (8.3%).

In contrast, *Valei-me* (Table 5) was heard a few times as speech when the stimulus was a spoken sentence: roughly speaking, it is only heard half of the time as a Only Speech. The reason for this is perhaps because most of its sentences are chanted as a prayer and, in relation to this, we can think of its rhythmic aspect in which the regularity would be greater. Prayers and nursery rhymes are considered to offer more regular temporal structures (RAPOSO-DE-MEDEIROS; CUMMINS, 2014), and, for this aspect, can differ from speech in its more canonical form, which we name here only speech and would be located at the very first type of the speech-singing continuum (Figure 1).

The songs *Jardineiro* (Table 3) and *Macunaíma* (Table 4) present similar results in the percentages of responses Only Speech and SpSSi, while these results are similar to those in the general table (Table 1). The sum of the percentages of Only Speech and SpSSi in each table, namely *Jardineiro*, *Macunaíma*, and the general Table 1, is respectively 97.2%, 97.7% and 98.2%. This indicates that speech stimuli are almost entirely heard as speech as the general trend is addressed, though with the benefit of the doubt around 24% regarding speech sounding similar to singing (SpSSi).

It should be noted that *Macunaíma* could offer an even higher percentage than 76.4% for the answer Only Speech, since it contains long sentences, but it seems that this is not a factor that will influence the choice of these answers, or it should be taken parsimoniously as a prevailing speech feature, as one listener would report that long sentences would not necessarily lead to chose the option "only speech".

1.3.2 About the perception of singing stimuli

The song most identified as Only Singing was also less perceived as Only Speech: *Valei-me*. The opposite is true for the song *Conversa*. The two songs, as already seen, show a different behavior from the general trend.

Conversa has the highest value for SpSSi responses when the stimulus is singing, as well as a high percentage of SiSSp responses. The sum of these two indicators (SpSSi and SiSSp) is almost 50% (49.6%). Thus, it is observed that individuals tend to perceive this song as influenced by speech characteristics. This result, which differs greatly from the general trend, is not, however, unexpected.

The songs presented are classified as popular songs due to specific musical features such as the tessitura, both tuning and tempo fluctuation, and the freedom to vary the rhythmic design of some musical phrases. Regarding the features at the intersection of speech and singing, we can consider their articulatory configuration similar to that of speech. This last aspect makes popular singing more intelligible than, for example, classical singing, whose typical articulatory maneuvers tend to centralize the vowels in the acoustic space, reducing their distinctivity.

About *Jardineiro*, it is important to notice that 92,1% of cases (SiSSp + Only Singing) listeners perceived the sung version as singing, which makes this unit similar to *Valei-me*, due to the fact that this song possess more evident musical elements such as different melodic intervals of sixth, fifth and fourth and, even when in a descent or ascendent scale, these are formed by no repeated notes.

Macunaíma is the second least perceived textual unit as Only Singing. A great number of responses (34.3%) for the sung sentences in *Macunaíma* are from the SiSSp option. This textual unit, in its sung version, features monotonic melodies and faster tempo, which can lead to a poor distinction between strong and weak beats, therefore, bringing it closer to speech characteristics. The result of 34.3% for SiSSp seems to be influenced by these characteristics. However, it is interesting to note that the percentage for SpSSi is not as high (9.3%) for this song as it is for *Conversa* (18.3%). This motivates us to claim that there must be another feature belonging to *Macunaíma* that makes it be perceived as singing most of the time when the stimulus is singing. Such a characteristic may be the f0 low variation phenomenon that we consider to be a distinctive song feature.

Looking at both the results of the compared modalities (speech and singing) and the results of the songs separately, it is possible to say that there is a very evident trend: listeners perceive the speech stimulus more often as speech, while they distribute their responses more between Only Singing and SiSSp, when the stimulus is singing.

2. STATISTICAL ANALYSIS

2.1 Interaction between the answers to the perception test and f0 variation

In exploratory terms, a first observation of tabulated values was made relating the number of times an answer was chosen by the listeners and the variation values of the fundamental frequency of each sentence. The f0 extraction, using VoiceSauce as explained in the introduction, provided us with f0 means and variation at each vowel and hence also for each sentence. Observations showed that there could be a relationship between the choice made and the f0 variation. This variation was extracted per sentence and related to the four responses: Only Speech, Speech Similar to Singing (SpSSi), Singing Similar to Speech (SiSSp) and Only Singing.

For example, the sentence 1 in *Valei-me*, has a 0.0323 for its f0 variation in its spoken version. Listeners classified this sentence 35 times as speech, 1 as Speech similar to singing (SpSSi) and 0 for the other two singing categories. From the perspective of singing stimulus, *Jardineiro*'s first sentence with a low f0 variation (0.0176) and got 31 classifications as Only Singing against zero responses for the speech categories. The

difference between singing and speech f_0 variation was attested as significant in the previous study using a residual model based in a log-linear regression model. The residual model is explained in details in the production study (Raposo-de-Medeiros et al., 2021).

Gathering data from our findings per sentence originated in the production study and the number of responses of the present study allowed us to assume that the f_0 variation could be an attribute that was perceived by listeners to classify a type of speech or type of singing.

2.2 Statistics: the Discrete Choice Model

Once we had, on the one hand, the f_0 variances of each sentence and, on the other hand, the answer given by the listener to a sentence, statistical tests were run based on a discrete model. Such a model deals with discrete random variables as the answers given in the perception test, which were qualitative. The model used was the Discrete Choice Model (DCM)⁷, which aims to answer the following question: what is the probability of an individual to choose an alternative due to the fact that a certain characteristic can be attributed to this alternative. In the case of the relationship we wanted to establish, the question to be asked was the following:

- How likely are the listeners to choose one of the four response alternatives (Only Speech, SpSSi, SiSSp and Only Singing) when hearing sentences with greater or lesser variation of f_0 ?

The perception experiment results performed were submitted to the DCM, which comprises a multinomial analysis, in which the different choices compete with each other and are necessarily exclusive (it is not possible to choose more than one of the options). This type of model can use the characteristics of both the sentence and the individual to verify which of them increase (variables with positive estimates) or decrease (variables with negative estimates) the probability of each of the choices. The predictor variable used was the sentence, or more specifically, the mean of the quadratic residuals of the f_0 stability model per vowel, for each of the sentences.

Our hypothesis, therefore, is that, although the listeners did not demonstrate being aware of the f_0 stability for singing, this characteristic was perceived by them.

It is interesting to point out that they are aware of some sound aspects and even name such aspects, albeit imprecisely, as we can see in section 3. A plausible explanation for the fact that listeners do not express an impression about the maintenance of a more stable f_0 that would correspond to the sung vowels is that in order to describe and/or be aware of this phenomenon it is necessary to acquire specific knowledge about music and/or phonetics.

2.3 DCM results

Tables 6 and 7 below present the statistical results. Each test answer is a discrete variable associated with f_0 variability, depicted in the first column in Table 6. $\text{Pr}(>|z|)$ indicates that there is a statistically significant relationship between f_0 stability and the choice made.

⁷ <https://cran.r-project.org/web/packages/mlogit/index.html>

We can verify, in table 7, that the mean square residue (MSR) of f0 are gradually negative for the three cases SpSSi > SiSSp > Only Singing. Thus, it appears that the lower the residue, the higher the probability of the listener classifying the sentence as singing and not as speech. The variable Only Speech is not shown in any of the 2 tables because it is used as a reference to estimate the other class intercept.

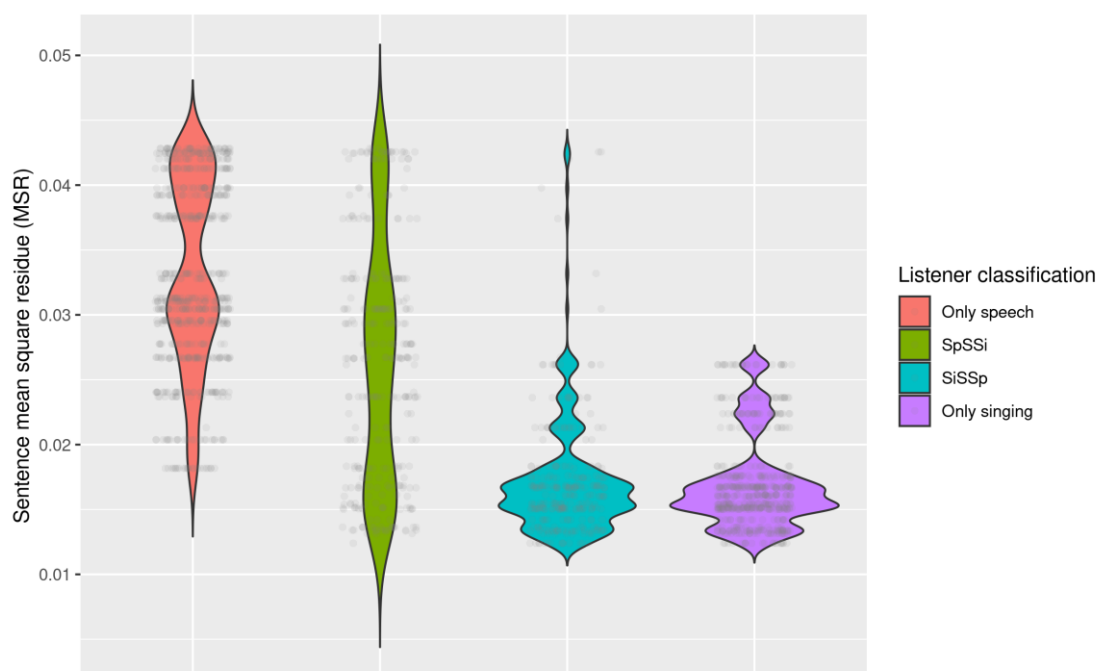
Table 6: Parameter estimation of discrete choice model's intercepts. Intercept is the point at which the y-axis is cut by the regression line, it can be interpreted as the bias of the listener to choose an specific answer. Only Speech is used as reference, so it was not estimated.

Variable	Estimate	SD	z-value	Pr > z
SpSSi (intercept)	2.33	0.28	8.26	2.22e-16
SiSSp (intercept)	6.92	0.39	17.40	2.22e-16
Only Singing (intercept)	8.02	0.36	21.82	2.22e-16

Table 7: Parameter estimation of mean square residual (MSR) regressor. Negative values indicate a probability reduction of choosing the answer as MSR increases.

Variable	Estimate	SD	z-value	Pr > z
SpSSi (MSR)	-104.01	9.50	-10.94	2.22e-16
SiSSp (MSR)	-330.61	18.66	-17.71	2.22e-16
Only Singing (MSR)	-355.27	16.72	-21.24	2.22e-16

The graph below shows the distribution of MSR according to the classification made by the listener. It is possible to verify that the Only Singing and the SiSSp answers present lower MSR distribution, so less variability of f0, while speech presents higher values. This results corroborate MSR parameter estimates though discrete choice model.



Graph 1: Violin plot of MSR according to the classification made by the listeners. It is possible to verify correlation between low MSR and SiSSp/Singing classifications.

Intra vowel f0 stability that is represented by the low mean square residual (MSR) seems to perform an important role in the choice made. The DCM parameter estimation showed statistical significance and the results corroborated the hypothesis that greater stability would lead listeners to recognize the recording as singing.

3. DISCUSSION: LISTENERS' IMPRESSIONS ABOUT THEIR PERCEPTION

It would be interesting to think about approaching phonologists and phoneticians for a simple test in which they would answer yes or no to the question "Which is more stable in terms of f0, speech or singing?". Depending on how they understood the question, the specialists could say that it is speech, since singing presupposes intonation of distant intervals between them, such as a fifth interval for example (C to G) providing variability of f0 between musical notes. If they understood that it is about the stability of f0 intra notes and therefore intra vowels, they would think as we hypothesized here: that singing is more stable.

Therefore, the f0 stability of singing should be seen as a crucial feature of sung or instrumental music without losing sight of the fact that another question can be formulated in a musical study having to do with variations between pitches, or musical notes. However, we emphasize that this is not our quest. Still, how do naive people, or probably naive people in music and phonetics, express their perception of what is speech and what is singing?

Of the 26 listeners who reported their impressions about the perception test, 19 expressed having difficulty choosing between SpSSi and SiSSp. This was expected, since the stimuli were collected ideally as belonging to different types of speech: the speech uttered by the actresses and the singing performed by the singers. The actresses were asked to act freely, based on the given text. The singers, in turn, sang in the tempo that suited them and produced the melody as they memorized it best, as requested, fulfilling the task of producing a song.

Table 8: Listeners' most repeated impressions

Impressions	Number of listeners
Difficulty with intermediate options	19
Length/Duration of linguistic units	11
Intonation and rhythm description	8

The specific instruction of the perception test was quite simple and sought to direct the listener as little as possible: listen to the stimuli and point out a response. Taking into account that there is no categorical distinction between singing and speech, the test aimed to verify in which sentence – and therefore, in which song or textual unit – the perception tended towards a more intermediate indicator.

We named as *Length/Duration of linguistic units* the impressions that reported or tried to describe with some precision if sentences were short or long, or if utterances were elongated. This is how we understand descriptions such as: “words were very long” and “when the sentences were longer, the sentences...”. Both these descriptions and those related to prosody (intonation and rhythm description) were problematic, as it was impossible to know for sure what the listeners were trying to mean, since they are not experts on the subject. Some used terms properly and others used them in such a way that the meaning was quite problematic (although very interesting), such as the use of the term “accelerated musicality”.

It is interesting to note that when speaking about temporal aspects of sentences such as the lengthening of a linguistic unit, the listener justified that the latter led him to perceive the passage as singing. In relation to prosody, the idea of a more cadenced rhythm and the search to establish differences in the intonation of speech and singing, the latter being more melodious, also made the option fall on singing.

Other less mentioned impressions were: not having thought about the subject before, associating sung speech with a language accent; and the influence of the linguistic register on the choice of response.

We did not detect any terms in the listeners' comments that indicated the description of fundamental frequency stability. We did not even expect this, because it is a very specialized knowledge about the functioning of speech and singing. However, we highlight some appreciations such as:

1. “it seemed to have a more melodious rhythm”
2. “tendency to opt for singing if the utterance ended with a treble”
3. “even when it is a longer sentence, we think it is singing...”
4. “when all the words were too long, I considered it Only Singing .”

Such appreciations indicate the intention to express a f0 stability phenomenon, although the first of them seems to be inadequate, as it assigns melody quality to rhythm. However, it can also indicate that the listener wanted to express himself about a perception of a more musical melody, precisely because the f0 was more stable over a certain period of time. The same can be said about the second appreciation, since ending a sentence in high pitch means keeping the f0 high, even if it is in a small portion of the whole utterance, for example, in the final syllable. In the third and fourth comments, the sentence length and segment duration issues are at stake, both of which involve the idea of time. As might be expected, long sentences (that is, formed with many words) could be more typical of speech, but we have seen that sentence lengths do not represent a

feature from which one decides to name it as speech. Thus, we speculate that if even when faced with a more typical feature of speech, the listener chose to name the utterance as singing, it is plausible to say that the stability of f0 in vowels was the perceived feature that influenced the listener's choice. Lengthening words (as we have seen, the duration of syllables and/or phones is longer in singing than in speech), in addition to being a typical feature of vowel duration, is not separate from the phenomenon of intonation and musical note sustaining, therefore, from the stability of f0.

4. CONCLUSION

Making use of sentences extracted from four Brazilian text units (spoken and sung) we carried out a perceptual test whose aim was to verify whether there was a relation between f0 stability in singing and the listener's choice for responses that would relate it to singing. The test responses related to singing were Only Singing and Singing Similar to Speech (SiSSp).

It was statistically demonstrated that listeners are more likely to choose Only Singing and SiSSp options when the stimulus contains less variation in f0. This result corroborates the previous result obtained in a production study in which f0 stability is a feature that can be assigned to singing to differentiate it from speech.

The model proposed to test our hypothesis in the present study, the DCM, was designed from observing the relationship between a sentence f0 variation and the number of times it was chosen as one of the possible answers, as described in section 1.1. The link between the production study (RAPOSO-DE-MEDEIROS et al., 2021) and the perceptual (current) study was crucial to ensure that f0 stability was a likely characteristic to induce the listener to his decision upon the possible answers. Thus, we conclude that the question "How likely are the listeners to choose one of the four response alternatives (Only Speech, SpSSi, SiSSp and Only Singing) when hearing sentences with greater or lesser variation of f0?" was successfully answered.

The explanation about the stimuli musical aspects, the qualitative analysis from the percentage results and the discussion took into consideration details that could be addressed as not generic enough to prove a crucial difference between speech and singing. We advocate that such aspects as revealed in different parts of this article are related to two vocal modalities that offer diversity and universality at the same time. Regardless of the fact of how we categorize music and speech, we would like to propose that we do have an emerging characteristic for each of them across cultures. Deciding what is sung voice or uttered speech depends on specific acoustic parameters and one of them is certainly f0 stability.

The results obtained in this work represents a "closing the loop" on our investigation about f0 stability and can be considered as an important empirical evidence for cognitive studies and views about the interaction between speech and singing. About this interaction, a theoretical proposal (RAPOSO-DE-MEDEIROS, 2020) is the idea of a "musical larynx" that would explain human ability to produce very stable phonation, thus achieving sound stability integrated to musical elements such as scale and temporal structures.

We acknowledge that a more encompassing study could lead us to more emphatic results, as we would consider a bigger database with more singing genres. The issue then remains upon a good choice. Which genres were we to choose? Surely hundreds of genres would be left aside. The best solution would be to use a greater number of songs of the same genre (MPB), even if the issue of the spoken version produced from an existing song remained a problem to be solved. This not the case of the proposed design whose

intention was precisely to avoid this, as it comprised speech from Brazilian dramaturgy that was only later turned into songs.

Funding

This work was supported by FAPESP (São Paulo Research Foundation), grant number 2015/0683 and grant number 2022/07143-0, the first author as the main researcher. The second author was supported by FAPES (Espírito Santo Research Foundation), grant number 581/2021.

Declaration of Conflicting Interests

The authors declare that there is no conflict of interest.

REFERENCES

- DEUTSCH, Diana.; HENTHORN, Trevor.; LAPIDIS, Rachael. Illusory transformation from speech to song. **Journal of the Acoustical Society of America**, 129 (4), p. 2245-2252, 2011. doi:10.1121/1.3562174
- FALK, Simone.; RATHCKE, Tamara. On the Speech-To-Song Illusion: Evidence from German. **Proceedings of the 5th Conference on Speech Prosody**, p. 164-169, 2010.
- KAWAHARA, Hideki; Masuda-Katsuse, Ikuyo; de Cheveigne, Alain. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction. **Speech Communication**, 27, 187-207, 1999.
- KOLINSKY, Régine; LIDJI, Pascale, PERETZ; Isabelle, BESSON, Mireille; MORAIS, José. Processing interactions between phonology and melody: vowels sing but consonants speak. **Cognition**, 112, p. 1-20, 2009.
- LINDBLOM, Bjorn.; SUNDBERG, Johan. **The human voice in speech and singing**. In: T. D. Rossing (Ed.), Springer Handbook of Acoustics, pp. 669-706. Stanford: Springer, 2007.
- VAN BESOUW, Rachel. M; HOWARD, David M.; TERNSTRÖM, Sten. Towards an understanding of speech and song perception. **Logopedics Phoniatrics Vocology**, 30 (3-4), p. 129-13, 2005. doi.org/10.1080/14015430500262160
- RAPOSO DE MEDEIROS, Beatriz; CUMMINS, Fred. Speech and song synchronization: a comparative study. **Proceedings of Speech Prosody**, p. 748-751, 2014.
- RAPOSO DE MEDEIROS, Beatriz; CABRAL, João Paulo. Acoustic distinctions between speech and singing: Is singing acoustically more stable than speech? **Proceedings of Speech Prosody**, p. 542-546, 2018.
- RAPOSO DE MEDEIROS, Beatriz. **Singing and speech as comparable phenomena: a dynamical approach**. In: Frank A. Russo, Beatriz Ilari, Annabel J. Cohen. (Ed.), The Routledge Companion to Interdisciplinary Studies in Singing. Volume I: Development. (1ed). p. 97-107. New York: Routledge Taylor & Francis Group, v. 1, 2020.
- RAPOSO DE MEDEIROS, Beatriz, CABRAL, João Paulo, MEIRELES, Alexsandro. R., BACETI, André. A comparative study of fundamental frequency stability between speech and singing. **Speech Communication**, 128, p. 15-23, 2021.
- SHUE, Yen-Liang. **The voice source in speech production: data, analysis and models**. UCLA. PhD Thesis, 2010.
- SUNDBERG, Johan. **Perception of singing**. In: D. Deutsch (Ed.), The psychology of music (3rd ed.), p. 69-105. Cambridge, MA: Academic Press, 2013. doi: 10.1016/b978-0-12-381460-9.00003-1
- TIERNEY, Adam; DICK, Fred; DEUTSCH, Diana; SERENO, Marty. Speech versus Song: Multiple Pitch-Sensitive Areas Revealed by a Naturally Occurring Musical Illusion. **Cerebral Cortex**, 2012. doi: 10.1093/cercor/bhs003
- TITZE, Ingo. The physics of small-amplitude oscillation of the vocal folds. **Journal Acoust. Soc. Am**, 83, p. 1536-1552, 1988.
- YANAGIHARA, Naoaki; KOIKE, Yuki. The regulation of sustained phonation. **Folia Phoniatr**; 19, p.1-18, 1967. doi: 10.1159/000263117
- ZATORRE, Robert, J; BELIN, Pascal; PENHUNE, Virginia. B. Structure and function of auditory cortex: music and speech. **Trends in Cognitive Sciences**, 6 (1), p.37-46, 2002. doi: 10.1016/S1364-6613(00)01816-7

APPENDIX

Conversa

1. Faz três dias que não como
2. Semana que não escarro
3. Adão foi feito de barro
4. Sobrinho me dá um cigarro
5. Me desculpe meu parente
6. Se cigarro não lhe dou
7. A palha, o fosfre e o goiano caiu n'água e se molhou

Jardineiro

- 1 Jardineiro de meu pai
- 2 Não me cortes meus cabelos
- 3 que o malvado me enterrou pelo figo da figueira
- 4 que passarinho comeu
- 5 Cho cho passarinho cho cho
- 6 Todos os passarinhos choraram de pena gemida nos ninhos
7. e o herói gelou de susto

Macunaíma

1. No fundo do mato virgem nasceu Macunaíma
2. Era preto retinto e filho do medo da noite
3. Houve um momento que em que o silêncio foi tão grande
4. escutando o murmurejo do Uraricoera
5. que a índia tapanhumas pariu uma criança feia

Valei-me

1. Subiu esperto pela capistrana prá não cansar
2. porém a vaca era de raça brava: Guzerá
3. Valei-me nossa senhora
4. Santo Antônio de Nazaré
5. A vaca mansa dá leite
6. a braba dá se quisé

Recebido: 1/8/2023

Aceito: 3/9/2024

Publicado: 7/11/2024