Studying the Intervenience of Lyrics Prosody in Songs Melodies

Jose Fornari

NICS, University of Campinas (UNICAMP), Brazil
tutifornari@gmail.com

ABSTRACT

Songs are made of two intrinsically connected parts: poetry (in the form of songs lyrics) and music. The proper fitting between these parts seems to be made by acoustic features that encompass the relationship between them, representing two fields of sonic communication: musical and verbal communication. While lyrics convey semantic meaning, music enhances its emotional intention, filling informational gaps and enhancing its signification that otherwise would make the poetic meaning of lyrics incomplete of even misleading. This work presents an introductory research about the influence of lyrics on their accompanying melodies. The experiment here presented analyzes three famous popular songs. Computational predictions, given as time series of eight acoustic descriptors, were retrieved from pairs of audio files; one solely with the speech of the lyrics, and another solely with its corresponding melody. In order to avoid data tainting from human emotional interpretation, the audio files with the speech were generated by a text-to-speech voice synthesizer. For the same reason, melodies are generated by MIDI files. These pairs were analyzed by computational models of higher-level acoustic descriptors that output time series representing the development of a particular acoustic aspect on time. The correlation of each acoustic feature for each pair of audio file are here presented, in the form of the correlation coefficient. R The experimental results are here presented, explained and discussed, in order to introduce a study on the acoustic features that better describe the intervenience of lyrics prosody in song melodies.

I. INTRODUCTION

Songs are here defined as pieces of singing music comprising two well-connected fundamental parts: poetry and melody. Poetry conveys the semantic meaning of this artistic communication. Melody supports and enhances the emotional meaning intended by the song-writer. This is done through the usage of strict musical (thus non-verbal) musical features. The meaning conveyed by melody is often difficult or even impossible to be achieved solely by words. We may compare the role of music,in songs with the emoticons in textual messages. Emoticons[1] are the contemporary pictorial representations of facial expressions or body movements, built with ASCII punctuation marks and letters, usually written along with normal text to express a person's mood or emotional intention. Nowadays they are commonly found in instant text messages, emails or online chats. These nonverbal messages are often used to alert the reader of the tenor or temper of textual statements, and can change or improve interpretation of plain text, that could otherwise have dubious meaning, such as in ironic remarks. As emoticons fulfill emotional meaning of plain text messages, music melody fulfill the emotional context of speech.

This work presents a preliminary study on common acoustical aspects of speech and melody of songs that may present a degree of correlation. The experiment here described compares two types of audio files. One, from the speech recorded out of the lyrics of three selected songs. The other one is from their corresponding melodies, without verbal cues, which is the audio containing only the instrumental melodic line, or accompanying harmony. These pairs of audio files are here analyzed by a group of eight higher-level acoustic descriptors (brightness, harmonic complexity, key clarity, mode, event density, pulse, articulation and repetition). Their retrieved data is here named “prediction” that has the variation in uniform periods of time of one specific acoustical feature. These predictions are compared for each audio file pair, to track possible interdependences of each particular feature, between speech and melody. In order to avoid personal interpretation of verbal and musical contexts, speech and melody audio data were generated by computer models. A voice synthesizer generated audio speech from lyrics text, and a MIDI sequencer generated audio files from its melody.

For this experiment, melody audio files were taken from the transcription of the melody lines of lead sheets, publicly available at Wikifonia (www.wikifonia.org). For melody, it was used MuseScore (www.musescore.org), a simple and open-source musical notation editor software enabled to write melodies and save them as MIDI files or to be converted as audio WAV files. Speech audio files came from the spoken text of the lyrics, converted by *VoiceOver*; a MAC OS X built-in application that provides several synthesized voices to read aloud texts written in English. The voice used here to convert text to speech was “Alex”, considered to be an advanced TTS (text-to-speech) application that even includes incidental sounds of reading, such as fine breath and automatic pause control to enhance speech naturalness and understandability. TTS applications, such as the one used here, have been also used by blind people to read aloud texts or even as a replacement for the ones who lost their ability of speaking, such as the case of the famous film critic Roger Elbert[2]. In this experiment, this TTS was used to convert to speech a small portion of about 5 seconds of the lyrics of three well-known popular songs: 1) “You've Got a Friend” by Carole King, 2) “Yesterday” by John Lennon and Paul McCartney, and 3) “You Make Me Feel so Young” by Josef Myrow and Mack Gordon.

The acoustic features here studied were retrieved by computer models written as MATLAB script files. They are named: Pulse Clarity (PC), Key Clarity (KC), Harmonic Complexity (HC), Articulation (AT), Repetition (RP), Mode (MD), Event Density (ED) and Brightness (BT). Each feature is calculated by an algorithm that returns as output a time series describing its variation in the time domain. Although

[1] http://en.wikipedia.org/wiki/Emoticon

[2] http://blogs.voices.com/voxdaily/2011/04/roger_ebert.html

some of these features seem as not related with musical prosody measurement, they may present an indirect or hidden linkage with aspects of speech and melody. For instance, HC is here used as a measurement of inharmonicity of these audio pairs. KC is used to measure partial distribution entropy, which relates to the presence and variation of early harmonics (i.e. F0, F1, F2, etc.).

These features are retrieved both from the synthetic speech of lyrics, without melody, as from the melodic line, without lyrics. This is done to eliminate the intersection between melody and lyrics, given by singing (melody with lyrics) so the correlation between the acoustic features of speech prosody with the ones from the pure melodic line can be better investigated. In this sense, at a first glance, some of these seems in fact odd features to be retrieved, such as BT and HC, that can help the understanding of how the pure sonority (without verbal meaning) of speech may influences (or be influenced by) the structural composition of its corresponding melody.

Lately, music is increasingly assuming the role of mood regulator [Gabrielsson 2001]. Due to the ubiquity of gadgets, able to store and play huge collections of music, people have used music to regulate their emotions. At the same time, an increasingly part of human verbal communication is becoming distributed, remote and even offline, done through emails, blog postings, and so forth. Speech inflections, that deliver emotional cues and help the verbal understanding, are not pproperly handled by text and thus lack its presence in such media. In principle, textual remote communication is void of such inflections, with makes more difficult to grasp the actual meaning of a phrase, such as if it is: sarcastic, caustic, cynic, mocking, satiric, sardonic or sincere. For this matter, emoticons, as previously mentioned, help to sign the intended meaning. Maybe, these two factors helped to stablish the preponderance of songs in our contemporary society. Singing a song becomes a social way of helping regulate mood at the same time that delivers precise contextual meaning, brought by verbal information conveyed in lyrics. This work studies a selected group of acoustical aspects of singing, through the means of their influence from lyrics to melodies.

Next session presents an overview on the origins, similarities and distinctions between singing and talking, which is the basis of the main topic on the influence of lexical meaning, found in lyrics, on musical meaning, found in melody. Section III presents an overview on the computer models used to retrieve the acoustical aspectsused here. Section IV describes the experiment and its retrieved data. Section V presents the results, followed by section VI that brings the discussion of the experimental results. Lastly, section VII presents the conclusion and possible further research on this subject.

II. SINGING AND TALKING

What is the difference between talking and singing? Surely, anyone can immediately recognize if someone is talking or singing, although many wouldn't be able to provide an explanation or identify the acoustic or auditory aspects that set them apart. The ability of speaking and singing are features of mankind since its origins. [Mithen 2005] explains that in hominids, such as the neanderthals, it was present the vocal tract and respiratory control to enable speech. However they lacked neural pathways in their brains required for language, as they were still to be developed by natural selection. Neanderthals, and other hominids, didn't have the cognitive fluidity or metaphorical thought, which refers to the ability of concurrently perceiving, identifying and managing information simultaneously retrieved by the sensory systems and processed by several different cognitive domains. The absence of archeological evidences of symbolic artifacts in such tribes from 120,000 to 35,000 years ago, seems to correspond to the absence of symbolic thought and therefore symbolic utterance, such as spoke language. The need amongst hominids to take care of dependent infants, seems to have influenced the development of music-like communication skills, more complex and sophisticated than any other specie had had before; one that also included iconic gestures, dance, onomatopoeia, vocal imitation and sound synesthesia. Nowadays, the joy of music making still includes similar bodily movement, such as: tapping toes, shaking heads, clapping hands, stepping the feet, and dance, These are predominantly and universally activities related to music listening that are shared with other humans within their communities, and make music to be a form of social activity, commonly spread throughout all human cultures, in all historic periods. In fact, the ubiquity of communal activities related to music-making is well documented in a broad range of anthropological and ethnomusicological works, such as [Meurant 1995] and [Feld 2004].

According to [Brown 2000] the two most socially prominent features found in musical communication, specially when contrasting singing with talking, are its ability of promoting: Temporal synchronization and Pitch-blending. In fact, we humans are accounted as the only species able to perform temporal synchronization to an external pulse. This ability seems to be part of our nature, instead of being nurtured (i.e. developed). Prelinguistic infants exhibit that capacity in utterances, known as IDS (infant-directed speech), as shown in several situations, such as during mother-infant interaction, which is often pointed out as the source of musical development, as well as a promoter for emotional tuning, cultivating confidence and empathy between infant and caretaker. This seems to bring the conclusion that singing has a developmental, if not emotional, priority over talking.

This article analyses the utterances of lyrics, disregarding their semantic meaning, but only considering the variation in the time domain of their acoustic aspects. With this approach the aim here is to study acoustic features from speech of lyrics that present time envelope similar to the same features from their corresponding melody, thus pointing to which aspects of lyric prosody influences the corresponding song melody.

III. ACOUSTIC FEATURES

When we identify a sound merely by listening to it (for instance, distinguishing singing from talking) there are features in this sound that our auditory perception automatically retrieve, remember and compare There are also other acoustic features are not important for its recognition. In fact, they can be eliminated from audio files and the average listener would be barely unable to sense the change. That is the case of lossy audio compression algorithms, such as MP3, that eliminates huge amounts of information from audio files, compressing them in rates of less than 10% of its original size, such as its default; a compression of 12:1. Yet, the recurrent acoustic perceptual change is nearly imperceptible. This experiment attempts to do the opposite, which is, to find and retrieve the most prominent acoustic features that are fundamental (and unique) for the auditory identification of music. These features can be contextual (i.e. dependent upon listeners' memory) or free of context (independent of it). The acoustic descriptors used here came from the computer models developed in [Eerola 2009], and adapted to the specific needs of this experiment. The literature of Music Information Retrieval (MIR) also refers to acoustic features free-of-context as Low-Level features, and contextual features as the Higher-Level ones. Low-level descriptors retrieve psychoacoustic features (e.g. Loudness, Pitch, Duration and Spectral features, such as its centroid, etc.). Higher-level descriptors retrieve independent features based upon previously retrieved data, such as: Tonality, Pulse, or Harmonic features. In [Eerola 2009] was presented the first version of eight computational models designed to predict contextual acoustic features found in music. These descriptors were created aiming the prediction of specific Higher-Level musical features associated with emotional arousal evoked by music stimuli. This was part of the project "Tuning your Brain for Music"[3]. These descriptors are: Pulse Clarity (PC), Key Clarity (KC), Harmonic Complexity (HC), Articulation (AT), Repetition (RP), Mode (MD), Event Density (ED) and Brightness (BT). Their predictions are given as time series representing the variation of each feature in uniform periods of time. They were used in several experiments, such as in [Higuchi 2011] that tested audio files of pianistic performances, in the attempt of characterizing and distinguishing between technical and expressive pianistic performances. Following there a brief introduction to each acoustic descriptor.

PC is a descriptor that measures the sensation of pulse in music. Pulse may also be seen as related to agogics; a fluctuation of musical periodicity that is perceived as the fluctuation of sub-tonal frequencis, usually below 20Hz. In this sense, what is perceived is not as tone, in the frequency domain, but as pulse, in the time domain. This is due to the fact that around 20 Hz, the auditory system migrate from rhythmic to tonal perception. PC returns a time series of predictions, each one describing a value of pulse, for one specific time, for a pulse of any musical nature (melodic, harmonic or rhythmic). The measuring of each prediction goes from zero, meaning that there is no sensation of musical pulse, till one, meaning that there is a clear sensation of musical pulse, disregarding its frequency (i.e. there is no distinction between "slower" or "faster" pulses) [Lartillot 2008]. For this experiment, PC is used to measure how similar is the pulse of lyrics speech and melody line.

KC is a descriptor that measures the sensation of tonality, or musical tonal center. This is related to the sensation of "how tonal" the analyzed music seems to be. Exemplifying, if the sequence of notes: C, D, E, F, G, A, B, C is played in ascending order, there will be, for most people, a clear sense of tonality. In opposition, for another sequence of notes, such as: C, F#, F, B, A#, A would not be easily related to any tonality. The prediction of KC ranges from zero, meaning atonal, until on, meaning "clear tonal. Intermediate regions tend to refer to sudden tonal changes, or dubious tonalities. For the experiment, KC is used to measure spectrum distribution entropy, or the degree of organization of partial components in the audio spectrum. This is related to the existence and degree of variation of early harmonics, starting from its fundamental F0, and higher ones (F1, F2, F3, etc.), such as the research done by [Stegemoller 2008], in her study on the similarities between speech and song.

HC is a descriptor that measures the sensation of how complex is the harmony of a music. Information Theory relates the degree of complexity in a signal with the amount of entropy, which can be interpreted as the amount of disorder, or how stochastic is a signal [Shannon 1948]. However, here the interest is in measuring the "auditory perception" of entropy, instead of acoustic entropy of a musical sound. For instance, in acoustical terms, white-noise is a very complex signal, however its auditory perception is the one of an unchanging noisy sound. This descriptor aims to measure the perception of complexity or the perceptual entropy of musical harmony, without considering melodic and/or rhythmic complexity. Its prediction goes from zero, meaning no harmonic complexity is perceived, till one, meaning the existence of clear perception of harmonic complexity [Fornari 2008]. Here HC is used to measure the entropy of inner partials of the speech and melody, once that there is no harmony in the analyzed audio files. In this sense, the prediction goes from zero, "no inharmonicity" to one, "clearly inharmonic".

AR prediction refers to the musical articulation of sequential events (notes), as how they are tied together. If there is no perceptual pause between notes of a melody, the articulation is *Legato*, or connected. If there are perceptual pauses between notes, the articulation is *Staccato*. This descriptor retrieves the articulation from musical audio files, by detecting frequent pauses or sudden drops of intensity, attributing to it an overall rank that continuously ranges from zero (legato) till one (staccato). Its prediction is used in this experiment to measure how similar are the pauses between events of speech and melody.

RP is a descriptor that measures how clear is the repeating of musical (melodic, harmonic or rhythmic) patterns. This is done by calculating the similarity of time-framed windows. Its prediction goes from zero (without noticeable repetition within the musical excerpt) to one (clear presence of repeating musical patterns). For this experiment, RP is used to measure the similarity in the development of the repetition of non-verbal patterns within speech and melody.

MD is a descriptor that refers to the major, or Ionian musical scale; one of the eight modes of the diatonic musical scale. The most identifiable ones are: major (Ionian) and minor scales (such as the Aeolian). They are distinguished by the presence of a tonal center associated to intervals of major / minor thirds in the harmonic and melodic structure. The first usage of this descriptor was to retrieve an overall measure continuously ranging from zero (minor mode) to one (major mode). Later, it was used to predict modal hierarchy, referring to a perceptual scale of "obscurity / clarity" related to the ordering of the diatonic modes, from minor to major modes [Ramos 2011]. Here, MD is used to measure the similarity between the shapes of spectral composition over time, in

[3] www.braintuning.fi

terms of their “obscurity / clarity”, between speech and melody.

ED is a descriptor that measures the perception of simultaneous and distinguishable musical (melodic, harmonic or rhythmic) events. Its scale goes from zero (only one identifiable musical event) to one (maximum amount of simultaneous events that an average listener can distinguish). For this experiment, ED is used as an indicator of data synchrony, once that both melody and speech, in theory, are a queue of single events, one at each time instant. This is later used in the experiments as alignment measure.

BT is a descriptor that measures the sensation of how “bright” a musical moment is perceived. This is related to the spectral centroid positioned in higher frequencies, but it is also dependent upon spectral shape and variation, such as: attack, articulation and the balance between deterministic and stochastic components. BT measurement goes from zero (opaque or “muffled) to one (clearly bright). For this experiment BT is used to predict the similarities between the occurrence of bright sounds of speech and melody.

IV. THE EXPERIMENTAL DATA

As mentioned in the Introduction, the experimental part of this work analyzed three famous popular songs with eight descriptors described in section III. To each song, a pair of audio files were created; one for the speech of its lyrics (done by a text-to-speech voice synthesizer) and another for its melody (done by a MIDI sequencer). The predictions of each descriptor for each pair (speech and melody) were compared in their similarity. This was done by the calculation of the coefficient of correlation between the predictions of each pair.

The first song analyzed was “Yesterday”. There are two verses for the same chorus. The first round starts with “*Yesterday, all my troubles seemed so far away*,”. The second one starts with: “*Suddenly, I'm not half the man I used to be.*” As seen in Figure 1, both lyrics have the same melody. In this part of the experiment, it was calculated the time series of the acoustic features retrieved by the descriptors, for one melody excerpt and two lyrics excerpts (all shown in Figure 1). As seen, only the first phrase of this song was analyzed.

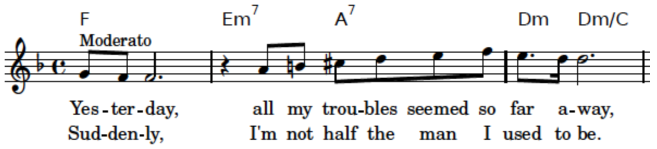
F
Moderato
Em7
A7
Dm
Dm/C
Yes-ter-day, all my trou-bles seemed so far a-way,
Sud-den-ly, I'm not half the man I used to be.

Figure 1. First 3 bars of the lead sheet of “Yesterday”.

The second experiment analyzed the song “You've Got a Friend”. This melody is larger than the melody excerpt used in the first part of the experiment, and there is no repetitions of the same melody with two different lyrics, as before. The lyric speech that was analyzed here is: “*When you're down, and troubled, and you need some loving care,and nothing, nothing is going right. Close your eyes, and think of me, and soon I'll be there, to brighten up, even your darkest night.*”. This is shown Figure 2, together with its melody.

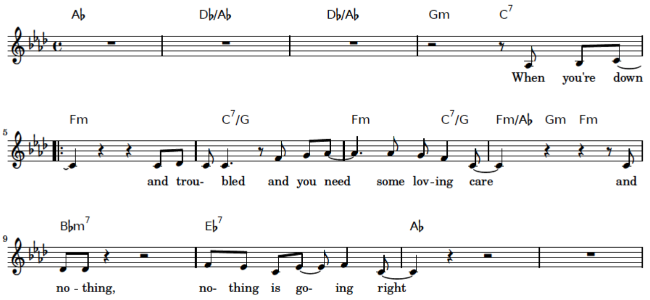
A♭
D♭/A♭
D♭/A♭
Gm
C7
When you're down
5
Fm
C7/G
Fm
C7/G
Fm/A♭
Gm
Fm
and trou- bled and you need some lov-ing care and
9
B♭m7
E♭7
A♭
no - thing, no- thing is go- ing right

Figure 2. Extended excerpt of the lead sheet of “You've Got a Friend”.

The third experiment uses the the lyrics and melody of “You Make Me Feel So Young”. This is the song with a most complex rhythmic information, as seen in the Figure 3 that presents the melody and corresponding lyrics speech

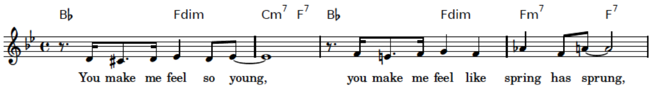
B♭
Fdim
Cm7
F7
B♭
Fdim
Fm7
F7
You make me feel so young,
you make me feel like spring has sprung,

Figure 3. First phrases of “You Make Me Feel So Young”.

V. RESULTS

After calculating all time series for the pairs of audio files (speech and melody) of each one of the three songs excerpts (described in section IV), is was calculated the Correlation Coefficient (R) between each pair of predictions, for each one of the eight descriptors. This is given by the matrix of correlation coefficients R calculated by the Matlab function “corrcoef”. R = corrcoef(X) is related to the covariance matrix C = cov(X) by:

$$R(i,j) = \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}} \qquad (1)$$

The first song analyzed has two verses for the same melody. Table 1 shows the results of R for the predictions each descriptor between the pairs of melody and 1st verse (2nd column) and melody and 2nd verse (3rd column).

Table 1. Correlation Coefficients (R) for the predictions of each acoustic descriptor. Song: “Yesterday”.

Descriptors	R between predictions of Melody and 1st Verse	Idem, for the 2nd Verse
AR	0.08	0.05
BT	0.31	0.26
HC	0.03	0.08
ED	0.16	0.14
KC	0.53	0.21
MD	0.19	0.19
PC	0.33	0.23
RP	0.30	0.56

As shown by Table 1, the prediction with highest correlation was RP, specially in the second verse, where this R=0.56. Figure 4 shows the comparison for these two time series of RP prediction. Left bars (in blue) are the prediction for the melody. Right bar, in red, are the predictions for the speech of the second verse of this lyric.

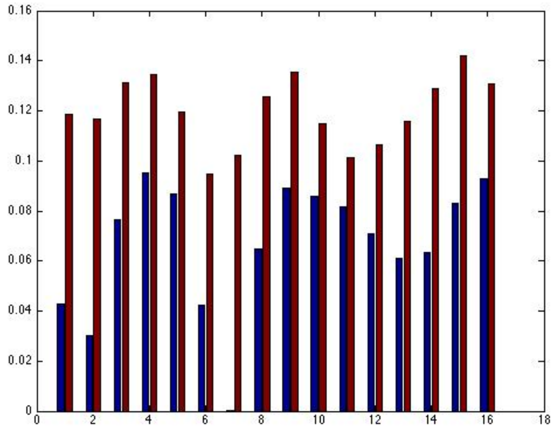

Figure 4. RP, for "Yesterday". Left bars (in blue) are the prediction of RP for the melody. Right bar, in red, are the ones for the speech of the 2nd verse of this lyric.

In the second experiment, the descriptor AR presented the highest correlation. The results are show in Table 2.

Table 2. Correlation Coefficients for the predictions of each acoustic descriptor. Song: "You've Got a Friend".

Descriptors	Predictions of Melody and Lyric First Verse
AR	0.48
BT	0.27
HC	0.04
ED	0.05
KC	0.08
MD	0.01
PC	0.27
RP	0.33

The results of the third experiment, for the song "You Make Me Feel So Young" are shown in the Table 3. The first column presents the results for the prediction measured for the entire melody. They are initially smaller probably because the audio files are larger than other former experiments, and consequently, the correlation coefficient is smaller because of the data shifting. In the second column only the first phrase was measured ("you make me feel so young, you make me feel like spring has sprung"). The audio file of the speech was edited in order to synchronize with the audio of melody. The coefficients of correlation for the descriptors of this pair are show in the secon column of this table.

Table 3. Correlation Coefficient for the predictions of each acoustic descriptor. Song: "You Make Me Feel So Young".

Descriptors	Predictions of Melody and Speech	Predisctions of first phrase w/ time correction
AR	0.22	0.03
BT	0.02	0.23
HC	0.29	0.70
ED	0.13	0.88
KC	0.29	0.53
MD	0.04	0.02
PC	0.04	0.23
RP	0.30	0.58

Figure 5 shows the comparison between predictions of ED for the audio files of the first phrase of the song "You Make Me Feel So Young", with manual alignment of onsets between the events of speech and melody. It is noticeable the prediction of ED for both audio files (melody and speech) are similar. As said before ED prediction does not delivers an information that can be used to analyze the intervenience of lyrics in melody as both are single source material, and ED is a descriptor intended to measure the perception of event multiplicity. However, ED prediction can be used to measure the degree of synchronism between files. As shown in Figure 5 and Table 3, when the prosody of single events of speech and melody are in sync, the correlation coefficient of the time series of ED prediction is very high (for this experiment, R=0.88)

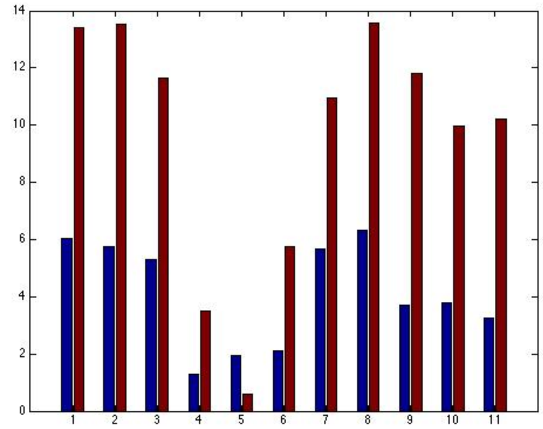

Figure 5. ED for the audio file pair of "You Make Me Feel So Young". Left bars (in blue) are the prediction of ED for the melody of the first phrase. Right bar, in red, are the predictions of ED for the speech of the verse of this respective lyric.

Figure 6 depicts the waveforms pair for this last song of this experiment ("You Make Me Feel So Young"). At the top, it is depicted the waveforms as they were originally, which no synchronisation. At the bottom, the same pair is presented after being manually edited, to align their onsets. This alignment was responsible for the raise of all values of the correlation coefficient, but specially ED, that when from 0.13 to 0.88 (in Table 3).

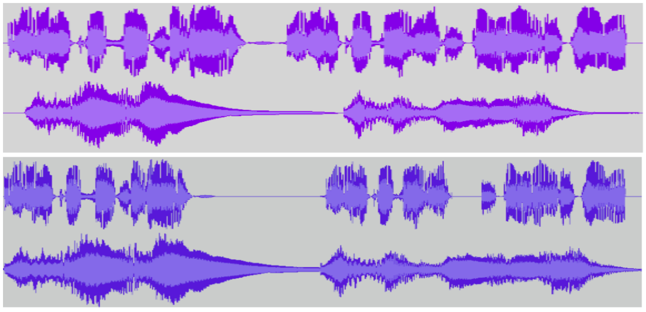

Figure 6. Manual onset alignment between speech and melody audio files of "You Make Me Feel So Young". Top pair presents as they originally are. Bottom pair presents the pair after correction.

VI. DISCUSSION

As seen, this experiment analyzed three popular songs. Due to their intrinsic peculiarities, each analysis had its differential. For the song "Yesterday" (sample 1) the first part of its lyrics has two verses with the same melody, so it was analyzed twice; one for each verse. In the first part of "You've Got a Friend" (sample 2) the lyrics and melody are long and do not repeat, so it was analyzed only once. The first part of "You Make Me Feel So Young" (sample 3) was also analyzed twice. First as in the other previous ones, and secondly, with the manual realignment of each speech and melodic event, so the onsets of both are in sync. This way, there were 3 variations of the same experiment, which expanded the possibilities of analysis for the correlation between acoustic aspects between the spoken (lyrics) and musical (melody) parts of these songs.

In sample 1, the correlation coefficients (R) for the correlation of each descriptor prediction, between melody and 1st and 2nd verses, were about the same in most of cases. For instance, the variation of R, from 1st and 2nd verse for each descriptor is: AR (0.08 to 0.05); BT (0.31 to 0.26); HC (0.03 to 0.08); ED (0.16 to 0.14); KC (0.53 to 0.21); MD (0.19 to 0.19); PC (0.33 to 0.23); RP (0.30 to 0.56). It is seen that the exceptions are KC and RP. For KC predictions R varied from 0.53 (1s verse) to 0.21 (2nd verse). As said before, KC is here related to the variation of early harmonics; from the fundamental frequency F0, to higher harmonics (F1, F2, F3, etc.). This seems to indicate that the 1st verse has harmonics that are more similar to the melody than the ones from the 2nd verse. On the other hand, R for data from RP descriptor varied from 0.3 (1st verse) to 0.56 (the 2nd verse). Here RP is used to measure similarity of non-verbal repetition patterns development, of spoken or melodic audio. It may indicate that 2nd verse has its repetition pattern more similar to the one found in the melody. Looking at Figure 1, it seems possible to infer that this might be due to the fact that 2nd verse has syllables with more attack than 1st verse, therefore can be perceived as separated events.

In sample 2, for ED data, R=0.05, which is the second smallest ones, (only bigger than HC, where R=0.04, which is also very small in sample 1, for 1st and 2nd verses). As seen, ED is here used to measure the alignment of speech and melody events. Thus, the low R for ED seems to express that the events in spoken and melodic data, are clearly out of synchronism, which may have affected all other descriptors R. The highest R in sample 2 is for AR (R=0.48). Here, AR is used to measure the similarities of the presence of pauses among speech and melody events. Figure 2 shows that sample 2 is, in fact, the one with more pauses presented among the melodic line, as seen in bars: 6, 7, 9 and 10. The second descriptors R, is RP, which is R=0.33. RP predicts the amount of non-verbal repetition patterns in audio file. This relative high value of R may imply that there is similarity of repetition patterns among spoken and melodic information. As this repetition can be of any nature (melodic, rhythmic or harmonic) looking at Figure 2, there are groups of three or four quarter notes, followed by long pauses, then again another group of three or four notes is presented. Although of different pitch, this suggested a rhythmic repetition, which is correspondent in speech, thus leading to a higher R of RP.

In sample 3, ED, shown in Figure 5, present a high correlation once that, in theory, melodies and speech should be perceived as only one event happening at each time instant. This is therefore used as an indicator of asynchronous data between speech and melody. As it was corrected, in Figure 6, ED correlation was the highest one.

This sample is a good example of what is noticeable in most popular lyrics; the existence of self-similar non-verbal acoustic features in spoken audio. Lyrics usually have the presence of rimes and other non-verbal acoustic similarities, among words and sentences, that precedes semantic meaning. For instance, its first two phrases are:

You make me feel so young,
you make me feel like spring has sprung,".

As seen, the first 4 words of each sentence are exactly the same (depicted in bold), as well as the ending syllables, creating the poetic rhyme (underlined). These extra-verbal similarities represent acoustic features that are not dependent upon their semantic meaning, but that are closer to its musical similarity, and can eventually be retrieved by acoustic descriptors such as the ones here studied.

As seen, R is sensitive to data alignment. If similar events are shifted in time, R will be small. If an audio data is compounded of several events, such as in sample 3, where the waveform pairs are shown in Figure 6, its was seen that is crucial that these events are aligned to guarantee a accurate measure of R. This was proven with the increase of R for ED, from 0.13 to 0.88, before and after manual event realignment. Besides this descriptor R, others also increased after the realignment. They are: BT (0.02 to 0.23); HC (0.29 to 0.70), suggesting inharmonic similarity between speech and melody; KC (0.29 to 0.53), suggesting a certain similarity of early harmonics; PC (0.04 to 0.23); and RP (0.30 to 0.58), pointing to a possible similarity of non-verbal patterns repetition. On the other hand, other descriptors R had decreased with the realignment. They are: AR (0.22 to 0.03), ; and MD (0.04 to 0.02), but their overall value of R was small, thus inconclusive.

Overall, it seems that previous audio events alignment is fundamental to be performed in studies like that. As seen in sample 3, there is a huge difference in the values of R before and after the realignment. In further experiments, that procedure shall be applied by default, to study the acoustic aspects that work as connectors to the identification between speech and melody. It is also interesting to note that, on the identification of similarity, for psychoacoustic aspects of sound, such as the perception of loudness, pitch and timbre (here considered as the process of identifying a sound object, such as a musical instrument out one single note) for non-verbal context, the begin of the sound (attack time) is the most important identification aspect [Grey 1977]. On the other hand, for the identification of macro-structural aspects, such as verbal or melodic context, similarity seems to be more related to the termination of its macro-structures (suffixes and rimes).

VII. CONCLUSION

This experiment focussed on the analysis of the influence of lyrics on the development of melodies of songs. The hypothesis here considered is that there are non-verbal acoustic aspects that are similar in spoken and melodic audio of the same song. In further works other types of more complex interrelations shall be studied, such as the correlation of contextual meaning of songs on their harmonic or rhythmic aspects. For now, the study here presented only intended to introduce it, starting from the relationship between prosodic aspects of verbal and musical communication found in songs.

The eight acoustic descriptors used here derived from the former studies described in [Eerola 2009]. The version of descriptors used here were adapted to fit the needs of this particular experiment. In further studies, new acoustic descriptors can be used, or developed, to study other non-verbal acoustic aspects that turns to be important. One of them is the similarity of pitch variation of songs speech and melody. This descriptor may retrieve the similarity between melodic intervals and natural intonation variation of speech. Another descriptor that can be further developed is one designed to retrieve rhythmic patterns. This one could be eventually developed from PC, the one here used to retrieved pulsation. A rhythmic pattern could be measured in terms of complexity and clarity. What most of lyrics have in common with poetry is the presence of metric line and rhymes, which creates a rhythmic structure similar to the one found in melody. The measure of its clarity and complexity can identify its similarity with musical metric, with is simpler, as compared to spoken information. This is actually a notorious feature that distinguishes singing from talking. In terms of pitch and rhythmic variation, singing seem to be a simplification of speech. The intensity of pitch variation during a normal conversation is such that it becomes unclear. This can be easily verified. For instance; when one is talking, if this one just freeze in any part of the sentence, while emitting sound, that will immediately turn into a clear pitch, similar to a musical note. The same is true for rhythm. If a short sentenc is looped severl times, we start to perceive its rhythmic pattern, and sometimes, even its pitch envelope. Also, the computer transcription of audio retrieved from normal speech, or even from singing, normally returns extremely complex musical notation.

Here was analyzed only the correlation between some acoustic aspects of speech and melody of songs. The semantical meaning of lyrics seems to be more complex and it still to be taken into account. Further experiments can include the emotional meaning of its poetic structure and its correlation with melody, in order to predict their emotional appraisal, for instance, using a bi-dimensional model of affect.

There are other forms of measuring time series similarity, besides correlation coefficient R, used here. They are also metrics used to measure the distance D between time series. Some of the most famous ones are: Manhattan (D_{Man}), Euclidean (D_E), Mahalanobis (D_{Mah}), Principal Component Analysis and Fourier based (D_{FFT},D_{ξ}, D_{Fk}). [Lhermitte 2011]. In this study, R was chosen for its simplicity. Further experiments in this subject shall perform similarity measurements in other metrics and compare their performance, for the study of influence of lyrics into melodies of songs.

VIII. ACKNOWLEDGMENT

This research is sponsored by FAPESP (www.fapesp.br)

IX. REFERENCES

Krisela Rivera, Nancy J. Cooke, and Jeff A. Bauhs. (1996). The effects of emotional icons on remote communication. In Conference companion on Human factors in computing systems: common ground (CHI '96), Michael J. Tauber (Ed.). ACM, New York, NY, USA, 99-100. DOI=10.1145/257089.257180 http://doi.acm.org/10.1145/257089.257180

Stegemoller, E.L., Skoe, E., Nicol, T., Warrier, C.M., & Kraus, N. (2008). Music training and vocal production of speech and song. Music Perception, 25(5), 419-428.

Mithen, Steven. (2005) The Singing Neanderthals: The Origins of Music, Language, Mind and Body. Weidenfeld & Nicolson Ed. London.

Sousou, Shaden D. (1997) Effects of Melody and Lyrics on Mood and Mamemory. Perceptual and Motor Skills: Volume 85, Issue , pp. 31-40.

Meurant, G., Thompson, R. F. (1995) Mbuti Design: Paintings by Pygmy Women of the Ituri Forest. New York, N.Y.: Thames and Hudson dist. By W.W. Norton, April 1996.--224 p.: ill.--ISBN 0-500-97430-6.

Feld, S. and Brenneis, D. (2004), Doing anthropology in sound. American Ethnologist, 31: 461–474. doi: 10.1525/ae.2004.31.4.461.

Brown, S. (2000). The "musilanguage" model of musical evolution. (N. L. Wallin, B. Merker, & S Brown, Eds.)The origins of music, 271-300. MIT Press.

Eerola, T., Fornari, J. (2009) The Pursuit of Happiness in Music: Retrieving Valence with Contextual Music Descriptors." Book Chapter at Computer Music Modeling and Retrieval. Genesis of Meaning in Sound and Music". Publisher: Springer Berlin / Heidelberg. ISBN: 978-3-642-02517-4.

Fornari, J., Eerola, T., (2008). Estimating the Perception of Complexity in Musical Harmony. The 10th International Conference on Music Perception and Cognition – ICMPC 10. Sapporo, Japan.

Lartillot, O., et al. (2008). An Integrated Framework for Onset Detection, Tempo Estimation and Pulse Clarity Prediction. Ninth International Conference on Music Information Retrieval - ISMIR 2008. Philadelphia, PA USA.

Gabrielsson, A. (2001). Emotions in strong experiences with music. In P. N. Juslin & J. A. Sloboda (Eds.), Music and emotion: Theory and research (pp. 431-449). New York: Oxford University Press.

Lhermitte S., Verbesselt J., Verstraeten W.W. , Coppin P. (2011) A comparison of time series similarity measures for classification and change detection of ecosystem dynamics, Remote Sensing of Environment, Volume 115, Issue 12, 15 December 2011, Pages 3129-3152.

Higuchi, M., Leite, J. P., Graeff, F., Del Ben, C., Fornari, J. (2011) "Reciprocal Modulation of Cognitive and Emotional Aspects in Pianistic Performance. PLoS ONE, Vol. 6, pp.1-10, San Francisco, U.S.A.

Shannon, C.E. (1948), "A Mathematical Theory of Communication", Bell System Technical Journal, 27, pp. 379–423 & 623–656, July & October.

Ramos, D., Fornari, J. (2011) "Um Estudo Sobre a Percepção Musical Afetiva da Hierarquia Modal". Anais do VII Simpósio de Cognição de Artes Musicais - 7SIMCAM, Brasilia, DF, Brasil.

Grey, J. M. (1977) Multidimensional perceptual scaling of musical timbres. J. Acoust. Soc. Am. Volume 61, Issue 5, pp. 1270-1277.

Russell, J. A. (1980). "The circumplex model of Affect. "Journal of Personality and Social Psychology, 39:345-356,