USER-CENTERED EVALUATION OF LYRICS-TO-AUDIO ALIGNMENT

Ninon Lizé Masclef[1] **Andrea Vaglio**[1,2] **Manuel Moussallam**[1]

[1] Deezer Research
[2] LTCI, Télécom Paris, Institut Polytechnique de Paris
research@deezer.com

ABSTRACT

The growing interest for Human-centered *Music Information Retrieval* (MIR) motivates the development of perceptually-grounded evaluation metrics. Despite remarkable progress of lyrics-to-audio alignment systems in recent years, one thing which remains unresolved is whether the metrics employed to assess their performance are perceptually grounded. Even if a tolerance window for errors was fixed at 0.3s for the *Music Information Retrieval Evaluation eXchange* (MIREX) challenge, no experiment was conducted to confer psychological validity to this threshold. Following an interdisciplinary approach, fueled by psychology and musicology insights, we consider the lyrics-to-audio alignment evaluation from a user-centered perspective. In this paper, we call into question the perceptual robustness of the most commonly used metric to evaluate this task. We investigate the perception of audio and lyrics synchrony through two realistic experimental settings inspired from karaoke, and discuss implications for evaluation metrics. The most striking features of these results are the asymmetrical perceptual thresholds of synchrony perception between lyrics and audio, as well as the influence of rhythmic factors on them.

1. INTRODUCTION

Nowadays, the machine learning community is raising the question of how to design explainable [1] and human-grounded algorithms [2]. Especially in the field of MIR, user studies and evaluation metrics plays a pivotal role in this shift towards Human-centered MIR. Subjective listening tests [3–5] and ethnomusicological studies [6] previously demonstrated the feasibility of including tasks in the real setting context of user experience. Regarding metrics, we are witnessing the transition from exclusively system-centered evaluation to user-aware evaluation. In the reference toolkit mir_eval, the lack of Human-centered metrics was justified by the complexity and cost required to develop robust subjective evaluation methods [7]. However there are a few limitations of system-based evaluation, such as their inability to capture the inherently subjective experience of MIR and the absence of necessary correlation between system-centered evaluation and users' perceptions [8]. One could, and indeed should, ask what is the meaning of the effectiveness of an algorithm without the presence of an embodied experience of human perception? In epistemic terms, how is the distance to the ground truth translated into an error measurement without the mediation of an individual? Since the advent in 2005 of the system-centered evaluation approach by the MIREX, there were several attempts at creating perceptually grounded metrics, notably among the field of music transcription [9–11], source separation [12] and audio similarity [8].

One application at the frontier of music perception and human machine interaction is karaoke. Currently, the majority of alignments used by karaoke systems are fully manually achieved, or partially corrected by human annotators. Obtaining manual annotations of lyrics-to-audio alignment is costly and time-consuming. To obtain such annotations automatically, one could turn to automatic lyrics to audio alignment system. Such system takes as input lyrics text and outputs timed position of their appearance in the audio signal, at the word, line, or paragraph level. Several recent automatic lyrics-to-audio alignment systems have achieved high performance taking inspiration from automatic speech recognition [13–15] and using large public singing voice annotation dataset like DALI [16]. Among the metrics developed for the MIREX challenge to evaluate lyrics-to-audio alignment, the most commonly used is the *Percentage of correct onsets* (PCO) ρ^k_τ, illustrated in [17], using a tolerance window for the perception of lyrics-to-audio alignment errors defined by a threshold τ [17].

$$\rho^k_\tau = \frac{1}{N_k} \sum_{word\ i} 1_{|\hat{t_i} - t_i| < \tau} \times 100 \qquad (1)$$

where N_k is the number of words in the track k, t_i the ground truth start of the word timestamp of the lyrics unit and $\hat{t_i}$ the predicted timestamp. It suggests that listeners tolerate errors falling within this window, and still perceive as synchronous lyrics and audio whose onsets are separated by this offset. A tolerance window for errors was fixed at 0.3s for the MIREX, albeit no psychology experiment was conducted to confer validity to this threshold.

Additionally, while spectacular progress has been made in the past years, the gap between state-of-the-art systems, as measured in the MIREX competition, has narrowed, with many systems achieving close to perfect PCO scores on the test sets. Therefore, it might now be important to make room for qualitative rather than quantitative metrics. In this work, we are interested in challenging the PCO metric from a user-centric perspective, focusing on how humans perceive asynchrony to derive stricter metrics for the task. To this aim, we expose the design of two perceptual experiments in Section 3 and their respective results in Section 4. We then propose a PCO adaptation in Section 5 and conclude in Section 6.

2. RELATED WORKS

Singing karaoke engages coordination of articulatory movements, music and language processing systems, as well as crossmodal integration of audio and visual stimuli. It is thus a rich context of perception involving complex stimuli. As a consequence, we briefly consider the research on all the domains outlined above to illustrate paradigms and hypotheses relevant to lyrics-to-audio alignment perception. When presented with a pair of audiovisual stimuli, individuals reported an asymmetric perception of asynchrony, with audio lagging preferred over visual lagging [18, 19]. This asymmetry has been correlated with faster transmission of the visual signal over the audio signal [18] or with the auditory dominance in temporal processing [20]. The latter hypothesis asserts that, when emitting a judgment of synchrony, audio would provide individuals a more accurate sensory information in the case of dynamic event such as music, and also a more stable internal representation of periodicity, contrary to the visual modality [20]. The listening experience is a continuous production of rhythmic expectancies [21]. In the case of sensorimotor synchronisation experiment, one effect induced by rhythmic expectancies is the anticipation of the stimuli in a sequence, also called Negative Mean Asynchrony (NMA). First reported by Dunlap [22], it states that the reaction to an audio stimuli tends to precede rather than follow the stimuli. Repp [23] discovered that individuals anticipate audio events up to 100ms ahead of time. Klemmer [24] revealed that the anticipation effect varies with the tempo of the rhythmic stimuli, usually measured in terms of InterOnset Interval (IOI) duration. He found that the reaction time of individuals when attempting to stay in phase with an isochronous stimulus, is a function of the IOI between stimuli. The reaction time was greater for shorter IOI, suggesting that individuals have less sensibility in slow tempo. These observations were further formalized as a function of local and global rhythmic context by McAuley [25]. Besides global rhythmic factors, the listening experience is punctuated by local variations. Metric events are periodic peaks of attention organized into nested hierarchies that coordinate attention to events on various time-scales, allowing for grouping and accentuation of notes [26]. Musical stresses are the cues to infer a general rhythmic pattern [26]. Among the significant factors of stress reported were the duration of syllables [26], loudness [27], alignment with beats [21] and sequence boundaries [21, 28].

Given previous studies, our theoretical hypothesis is based on two points. Firstly, we expect individuals to tolerate more audio lagging than lyrics lagging. Secondly, we expect perception of lyrics-to-audio synchrony to rely both on global and local rhythmic context.

3. METHOD

To investigate the perception of lyrics-to-audio alignment, we designed two psychological experiments inspired from the main application of this task, karaoke. We chose karaoke as it is a popular practice where the participants' rhythmical precision is important, requiring attention to the displayed lyrics as much as to the audio. The first experiment is designed to test the influence of global parameters on human perception of audio/displayed lyrics synchrony and to investigate its symmetrical properties. The second experiment intends to explore local factors influences. To run both experiments we developed a karaoke application prototype, whose displayed textual lyrics were intentionally misaligned with the background audio according to various, controlled conditions, thereby creating an audiovisual offset. The stimuli were presented to individuals who then annotated their perceived quality of alignment in different error scenarios. A snippet of this interface is displayed in Figure 1.

Both experiments were run online, through a web interface that was designed to be correctly displayed on both computer and phone screens, for a total duration of two weeks each, between January and April 2021. The first experiment was conducted only with Deezer employees while the second experiment was public and hence involving a larger and more diverse set of participants. Before engaging in karaoke, participants are asked to fill out a questionnaire allowing us to determine their level of musical expertise and familiarity with the practice of karaoke. We collect, with their consent, a range information of their age, declared gender and native language. We do not have control on their external environment when performing karaoke (external noise) or any other factor which might disturb the readability of the karaoke (low light, uncorrected vision problem). Nevertheless, the instructions of the experiment encourage the participants to use headphones and favor a quiet environment.

In both experiments the dependent variable measured is the perceived synchrony and the amount of offset between lyrics and audio is a within subjects factor. In order to prevent from order effect, the values of audiovisual offset are presented in random order. These two experiments are akin to the Simultaneity Judgment task (SJ) widely used in the literature for studying the synchrony perception of audiovisual stimuli [18, 19].

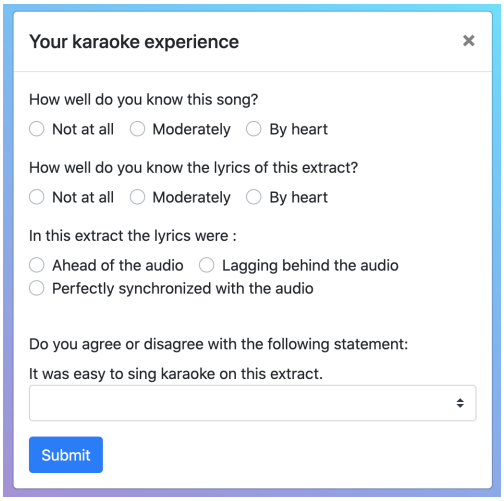

Figure 1. Questionnaire used to evaluate lyrics-to-audio alignment.

3.1 Dataset

Since the measured effects should be valid irrespective of the song, we allow participants to choose their song for karaoke within a set of 80 songs from various genre (pop, rock, rap and metal) and language (English, French, German). We selected popular songs in the DALI dataset [16] with alignment done at word level. The first criterion for the choice of songs was their popularity, so that we can expect a large proportion of participants to be knowledgeable of their lyrics and melody. Other important point guiding our choice was the correct lyrics-to-audio alignment and the absence of syntactical problems. We manually controlled the alignment quality of this subset by visualizing their lyrics in the karaoke prototype and eliminated poorly aligned songs from our selection. To avoid a learning effect of the song, each song can be selected once for a trial and can only be listened to twice during a trial. Moreover, the order of the songs in the selection menu for karaoke is randomized for each trial.

3.2 Influence of global factors

3.2.1 Experiment design

In this experiment, each participant is asked to choose 14 songs from the dataset from which karaoke excerpts are presented. Each audio extract lasts 35 seconds and consists of a sequence of words within lyrical lines, highlighting each word subsequently according to their aligned onset times. A lyrics-to-audio alignment error is generated for each user-song pair randomly from a set of positive and negative offsets between the audio and the lyrics displayed on screen. The offset is fixed for the whole sequence, which means all words in the stimulus are shifted by the same amount. At the end of each trial, participants are asked to report whether they perceive an asynchrony between lyrics and audio with a ternary response ("lyrics ahead", "lyrics lagging", "synchronous"). This experiment has a repeated measure design, with lyrics-to-audio synchrony perception as a dependent variable, and the lyrics-to-audio error offset as the independent variable having 14 modalities. It aims to measure an overall threshold of lyrics-to-audio synchrony perception and to study the influence of global rhythmic factors on this threshold, such as the tempo and word rate. If our theoretical hypothesis is confirmed, we expect to observe a greater proportion of "synchronous" responses for lyrics ahead than lyrics lagging, as well as a modulation of the perceptual threshold with the global rhythmic context (tempo, word rate).

3.2.2 Choice of offsets

In order to precisely define a threshold, we use a wide range of 14 offsets from -1s to 1s with negative offsets corresponding to lyrics ahead and reversely positive offsets mean lyrics lagging behind audio. We intentionally keep this number as small as possible, since this value is equal to the number of annotated songs required for each participant. Meanwhile, we wish to highlight effects around the commonly used threshold of 0.3s and -0.3s. Thus we use smaller steps around these values. We also included larger offsets (1s, 0.75s) as control values, to test that individuals systematically report those as asynchronous. In the same spirit, we expect the offset value 0 to trigger "synchronous" answers. The full experimental protocol was carefully tested beforehand with user testing sessions on six people. Based on these test results, we evaluated that completing the annotation required approximately 12 minutes per participant. Overall, the experiment involved 53 participants who completed the task.

3.3 Influence of local factors

In this second experiment, we make some changes in the karaoke interface. This time, we require each participant to choose one song from the dataset from which 10 audio excerpts are presented with different audiovisual offsets. Each sample is composed of three lyrical lines from the given song. The experience can be repeated multiple times with additional songs if desired. Each song takes around 3 to 5 minutes to annotate. Whilst in the first experiment the alignment errors were located on all the words of the sentence, in the second experiment, the position of the error may be located on the first, the last word of the sentence, or close to a beat. These choices are driven by some of the significant factors of stress described in Section 2 namely alignment with beats [21] and sequence boundaries [21,28]. We decided to discount the influence of long syllables [26] and loudness [27] for this study. In fact, long syllables and loud words are found to be overlapping respectively with the last word of the sentence and words closed to beats. The perceived synchrony is reported as a binary response ("yes", "no") with confidence on a 5-point Likert scale. This experiment intends to quantify the interaction of the error location in the sentence and the offset on the perceived alignment. It has a factorial design with the lyrics-to-audio offset and the position of the error as within subject factors. If our theoretical hypothesis is confirmed, we expect to observe a modulation of the percep-

tual threshold with the location of the error in the sequence.

Proximity of a word to a beat is defined as at a distance less than a *sixteenth note* from the beat, computed as $\text{♪} = 15/$ Beats Per Minute (BPM). The tempo estimation relies on Anssi Klapuri's algorithm, which showed 80% accuracy with constant tempo during the International Society for Music Information Retrieval (ISMIR) 2004 tempo induction challenge [29]. Starting from the baseline threshold of synchrony perception established in the previous experiment, the second experiment focuses only on lyrics lagging with 3 offsets (0.25, 0.5, 0.75) and a control sample with no offset. We chose only positive offsets because of practical constraints. Indeed, applying a negative offset at the word level can (and does frequently) result in overlapping with previous words, at least for beat-aligned and end words. Filtering out cases of overlapping words resulted in an important selection bias toward very slow songs. To avoid that, we could apply linearly decreasing offsets to precedent words until no overlap remains, as a naive way to "catch-up" with the true annotation. Such behaviour is consistent with what is observed in errors made by lyrics-to-audio alignment systems, multiple errors on consecutive words being recurrent. The concern was that we would not control which first offsetted word the participant would be confronted with. We decided to not consider negative offsets in this experiment but the problem of overlapping words for positive offset remains. However, after applying linearly decreasing offsets to consecutive words until no overlap occurs, the first offsetted word to which each participant is confronted remains the word of interest. Ultimately, we collected 2458 annotations from 193 participants.

4. RESULTS

As we intend to compute an overall threshold of synchrony perception, we perform the analysis at the level of the aggregated results, considering all annotations from all users. We removed all the trials from participants who did not answer correctly to our control levels i.e. "non synchronous" at 1s and "synchronous" at 0s. This represented precisely 11% of answers for the first experiment. We conducted a similar cleaning phase for users of the second experiment using the control offset of 0 and removed 8% of answers.

4.1 Asymmetry of Lyrics-to-Audio Alignment Perception

Using the data collected in the first experiment, we compute an aggregated proportion of respondents who indicate that lyrics and audio are "synchronous", and display it as a function of the lyrics offset in Figure 2. We see that synchrony perception is typically asymmetric, positive offsets being more easily detected than negative ones. This was expected as it resonates with previous findings [18, 19]. Beyond aggregated data, we also looked at individual responses and found that the thresholds were indeed asymmetric for 72% of individuals.

To give perspective, we plot the window function that correspond to the PCO metric scoring as used in MIREX, with an absolute threshold value of 0.3s. We also fit a function akin to a scaled skew normal distribution function to the data points. Among several attempts with asymmetrical continuous functions, this was the best fit we obtained, although it does not respect the maximality at 0. Parameters of the fitted function are a skewness factor of 1.12, a location of -0.22 and a scale of 0.29, a multiplicative factor is also applied in order to have a value of $1.$ at the maximum. Using this function we can derive new perceptive thresholds for synchrony using a simple rule of 50% of respondents being able to detect the offset. For lyrics ahead and lyrics lagging we respectively identify the offsets -0.33s and 0.22s. Given the amount of noise in the data, we can reduce these to -0.3s and 0.2s and examine if the differences of perception are significant for these values. Indeed, pairwise tests revealed a significant difference of proportions of response "synchronous" on the levels -0.3 and 0.3s ($\chi^2(1) = 4.26$, p = .038), while proportions on the levels -0.3s and 0.2 are not statistically different ($\chi^2(1) = 0.04$, p = .08).

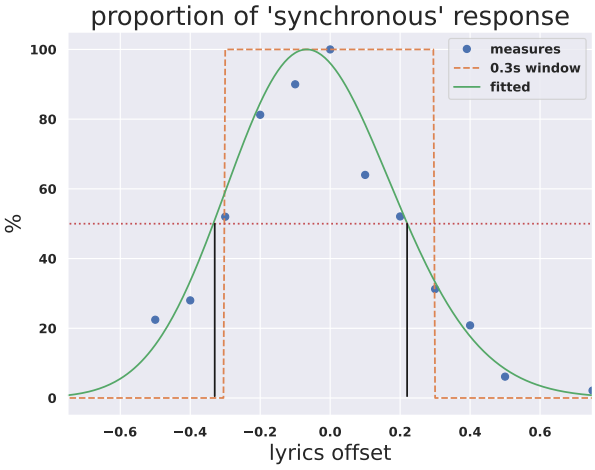

Figure 2. Aggregated results of *synchronous* judgment as a function of the lyrics/audio offset.

4.2 Sensitivity to global rhythmic context

In order to assess whether there is an influence of the global rhythmic context on lyrics-to-audio alignment perception, we compared the distribution of "synchronous" responses at each offset for two rhythmic factors: tempo and Words Per Second (WPS). We split our dataset of songs into two classes of tempo, defined as the upper and lower quartiles of the distribution of tempo, respectively fast ($\geq$138BPM) and slow ($\leq$93BPM). Although it is correlated with tempo, we also consider the average WPS rate of songs as a meaningful global factors. Again, we look at the first and last quartiles as Low ($\leq$1.16WPS) and respectively High WPS ($\geq$1.2WPS) classes. Figures 3 and 4 show the aggregated reported synchrony profiles for the negative offsets for the derived tempo and WPS classes. On both metrics, we observed no threshold discrepancy between the two classes for positive offsets.

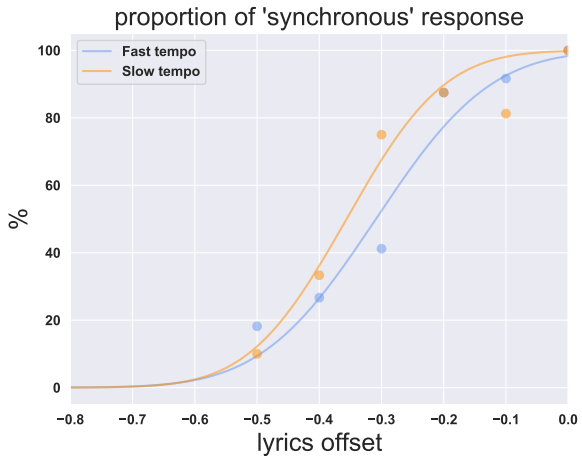

Figure 3. Proportion of response "synchronous" by tempo class.

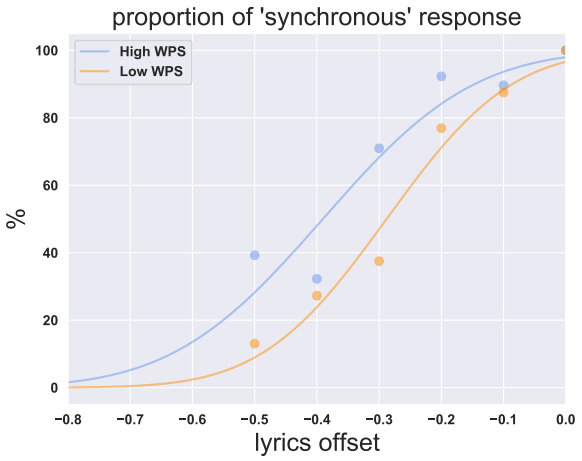

Figure 4. Proportion of response "synchronous" by WPS class.

To highlight the differences between classes, we fit a relatively simple sigmoid function to the data points. Among several candidates, a Gauss error function seemed most appropriate. Fitted functions are also displayed on Figures 3 and 4 and emphasize the different synchrony slopes. As before, we particularly consider the offset value intersecting with an average of 50% of "synchronous" responses as an indicator of participants' sensitivity to temporal asynchronies. Interestingly, the 50% threshold for the perception of synchrony is located at a larger offset (-0.36) for slow tempo than in fast tempo (-0.31). These results show that individuals report more frequently lyrics ahead as synchronous with slow tempo than with fast tempo. The lower sensitivity to lyrics-to-audio alignment errors in slow tempo is consistent with the results of [24]. Significance of these results are tested. The proportions of "synchronous" response at the offset -0.3s show significant difference between songs with high and low tempo ($\chi^2(1) = 5.44$, and $p < .02$).

Analogously, we found out that subjects are more tolerant to lyrics ahead (audio lagging) in high word rate than in low word rate. The 50% threshold for the perception of synchrony is indeed located at a larger offset for high word rate (-0.39) than in low word rate (-0.28) (Figure 4). These results show that subjects are more tolerant to lyrics ahead (audio lagging) in high word rate than in low word rate. We again tested the significance of these results. The proportions of "synchronous" response at the offset -0.3s are significantly different between songs with high and low WPS ($\chi^2(1) = 16.86$, and $p < .00004$).

4.3 Interaction between offset and word position

We designed the second experiment to distinguish perception of asynchrony as a function of the words position in the sentence. As explained in Section 3.3, we are only able to test for positive offsets. As insights from the previous experiment, we can assume that user sensibility is less affected by global factors for positive offsets. As a result, we expected it to be challenging for local factors too. For this reason, we aimed at collecting a much larger set of annotation, with a reduced set of tested offsets.

Figure 5 presents an overview of the results. There is a fairly large amount of noise in the collected data points, and few clear differences between synchrony perceptions for the three classes of word positions. The noise is particularly clear from the displayed level of confidence of participants who were unable to detect the asynchrony even for large values of the offset, but still were quite confident about their choice (average around 3.8). Regarding the location of the alignment error within the sentence, Cochran's Q test did not indicate a notable difference among the proportions of synchrony responses reported for the three error positions, $\chi^2(2) = 5.77$, $p = .056$.

The only visible effect seems to be for words aligned on *beats*, for which the confidence in the "asynchronous" answer at the 0.25 level is markedly higher than for the *end* class. More precisely, a Wilcoxon signed-rank test revealed that lyrics-to-audio alignment comparing error located on the beat with those on the last word did elicit a statistically significant change in the reported confidence of perception of error in individuals at the 0.25 level ($Z = 2.756$, $p < 0.006$). Indeed, mean confidence rating was 4.1 for error on the beat and 3.5 for error on the last word of the sentence. Such phenomenon is not observed for the synchronous case.

5. DISCUSSION

5.1 General discussion

Building on psychological theory and previous studies, we had hypothesized that a perceptual evaluation of lyrics/audio alignment quality would be asymmetrical and depend on both global and local factors. Using a first experiment we did find strong evidence for asymmetry and, to some extent for global factors influence. Despite a much larger experimental setup which involved hundreds of participants, we were not able to exhibit a clear influence of the local factors we tested. This negative result could mean

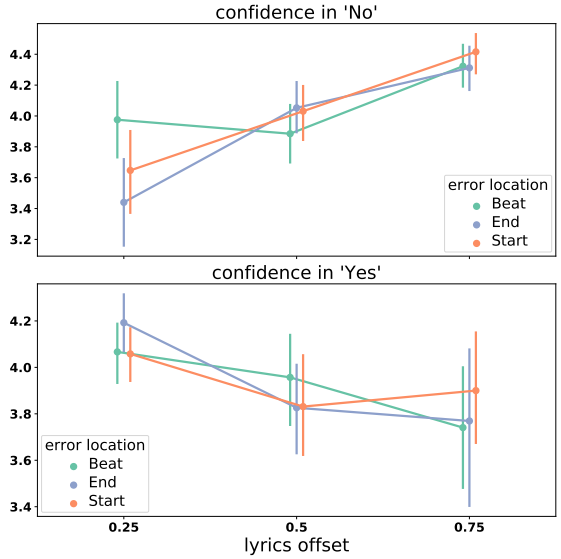

Figure 5. Plot of the reported answer (synchronous is "yes", asynchronous is "no") and confidence score (5 level Likert scale) in the perception of synchrony, by location of alignment error within the sentence.

that the local factors considered, i.e. the word position in the lyrical line are not the relevant ones. It is possible that words grammatical or semantic functions are more subject to human attention in a karaoke context. Indeed, the only significant phenomenon that we observed was on words located on *beats*, for which the asynchrony perception was more acute. Future work should investigate the relationship between rhythmic position and lyrical function of words and test new hypothesis of perceptual differences. Finally, although we did our best to build a realistic yet controlled experimental setup, we acknowledge that as a psychological experiment mostly conducted online, we can not completely rule out the possibility that the measurement noise was too high to allow us to detect signals on local factors.

There are arguably other factors that could influence this perception, notably at the human level. Indeed, familiarity with the song (e.g. previous knowledge of the lyrics and/or the music), but also participants' facility with the languages, level of musical expertise and even karaoke practice could be important variables to consider. In the conducted experiment, we collected such information from participants. Although we did observe some interesting phenomenon, for the sake of clarity, we chose not to present additional results on these variables here and leave it to a follow-up study.

5.2 Implication for evaluation metrics

Here we would like to present a practical use of our results, as a perceptually motivated evaluation metric for lyrics to audio alignment tasks. Overall, we propose a generalization of the PCO metric in the following form:

$$\psi^k = \frac{1}{N_k} \sum_{word\ i} f(\hat{t}_i - t_i) \times 100 \qquad (2)$$

	PCO	Asym-PCO	Perc-PCO
Gupta [13]	**94.47 (1.52)**	**93.66 (1.59)**	**89.94 (1.71)**
Vaglio [15]	91.85 (1.95)	90.82 (2.04)	86.79 (2.13)
Stoller [14]	87.02 (2.97)	85.23 (3.07)	79.93 (2.90)

Table 1. Averaged metrics over the Jamendo dataset songs. Standard errors are given in parenthesis.

where the function f can be seen as penalty weighting of the annotation offset and other notations are common with Equation 1. We then evaluated 3 state-of-the-art automatic lyrics-to-audio alignment models [14, 15, 30], on the 20 songs of the Jamendo dataset [14]. We have compared using the regular PCO ($f = 1_{[-0.3,0.3]}$), a slightly modified version still using a square window but taking into account the asymmetrical perception ($f = 1_{[-0.3,0.2]}$) and a Perceptual-PCO function that is the one fit from the data collected in our first experiment and depicted in Figure 2. This function can be seen as a smooth relaxation of the square window, taking into account the perceptive asymmetry of the error slopes.

Results are compiled in Table 1. Interestingly, there appears to be little difference between using the standard PCO window and a slightly shifted one. However, scores for the perceptual-PCO are much lower. This is despite the window support being larger (i.e. errors of more than 0.3s are not completely nullified). In our opinion, this new metric is better suited to capture the relative importance of alignment errors and weights them according to human perception. It can also help for comparing between alignment methods that achieve near perfect scores with the standard PCO. It is worth noticing that although we demonstrated it on the PCO, a similar weighting could be applied to other alignment metrics. A step further would be to parameterize the window function f on global song factors such as tempo and WPS. This would arguably require additional experiments with a larger, more diverse set of songs.

6. CONCLUSION

In this work, we challenged the objective evaluation of lyrics-to-audio alignment using hypothesis from psychological theory. We postulated three effects: asymmetry, influence of songs features and influence of words local positions. We were able to demonstrate the first two effects using a large scale online experiment, disguising the synchrony annotation task as a Karaoke experience. This framework proved less efficient for the third effect, despite our efforts to collect up to several thousands annotation points. We nonetheless proposed a readily usable weighting function to allow finer comparison between state-of-the-art alignment methods. Future work will investigate more diverse sets of factors, both on musical attributes and user features.

7. REFERENCES

[1] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[2] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. of the 40th annual meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.

[3] L. Yin-Jyun, C. Ming-Tso, and C. Tai-Shih, "Singing voice correction using canonical time warping," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, pp. 396–400.

[4] C. Pei-Chun, L. Keng-Sheng, and C. Homer, "Emotional accompaniment generation system based on harmonic progression," *IEEE Trans. on Multimedia*, vol. 15, no. 7, pp. 1469–1479, 2013.

[5] A. Huang and R. Wu, "Deep learning for music," *arXiv preprint arXiv:1606.04930*, vol. abs/1606.04930, 2016.

[6] A. Holzapfel and E. Benetos, "Automatic Music Transcription and Ethnomusicology: a User Study," in *Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2019, pp. 678–684.

[7] C. Raffel, B. Mcfee, E. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. Ellis, "mir_eval: A transparent implementation of common mir metrics," in *Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 10 2014.

[8] X. Hu and N. Kando, "User-centered Measures vs. System Effectiveness in Finding Similar Songs." in *Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, Oct. 2012, pp. 331–336.

[9] A. Daniel, V. Emiya, and B. David, "Perceptually-based evaluation of the errors usually made when automatically transcribing music," in *Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2008.

[10] A. Ycart, L. Liu, E. Benetos, and M. Pearce, "Investigating the perceptual validity of evaluation metrics for automatic piano music transcription," *Trans. of the Int. Soc. for Music Information Retrieval (TISMIR)*, vol. 3, pp. 68–81, 2020.

[11] A. Ycart, L. Liu, E. Benetos, and M. T. Pearce, "Musical features for automatic music transcription evaluation," *arXiv preprint arXiv:2004.07171*, 2020.

[12] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*. Berlin, Heidelberg: Springer-Verlag, 2012, p. 430–437.

[13] C. Gupta, E. Yılmaz, and H. Li, "Automatic lyrics transcription in polyphonic music: Does background music help?" in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[14] D. Stoller, S. Durand, and S. Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 181–185.

[15] A. Vaglio, R. Hennequin, M. Moussallam, G. Richard, and F. D'alché-Buc, "Multilingual lyrics-to-audio alignment," in *Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2020.

[16] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters, "Creating dali, a large dataset of synchronized audio, lyrics, and notes," *Trans. of the Int. Soc. for Music Information Retrieval (TISMIR)*, vol. 3, no. 1, 2020.

[17] M. Mauch, H. Fujihara, and M. Goto, "Integrating additional chord information into hmm-based lyrics-to-audio alignment," *IEEE Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 20, no. 1, pp. 200–210, 2012.

[18] R. L. J. van Eijk, A. Kohlrausch, J. F. Juola, and S. van de Par, "Audiovisual synchrony and temporal order judgments: effects of experimental method and stimulus type," *Attention, Perception, & Psychophysics*, vol. 70, no. 6, p. 955–968, 2008.

[19] A. Vatakis, B. Fuat, D. L. Massimiliano, and C. Ángel, *Timing and Time Perception: Procedures, Measures, & Applications*. Brill, 2018.

[20] B. Repp and A. Penel, "Auditory dominance in temporal processing: New evidence from synchronization with simultaneous visual and auditory sequences," *Journal of experimental psychology. Human perception and performance*, vol. 28, pp. 1085–99, 11 2002.

[21] M. R. Jones and M. Boltz, "Dynamic attending and responses to time," *Psychological Review*, pp. 459–491, 1989.

[22] K. Dunlap, "Reaction to rhythmic stimuli with attempt to synchronize," *Psychological Review*, vol. 17, pp. 399–416, 1910.

[23] B. Repp, "Sensorimotor synchronization: A review of the tapping literature," *Psychonomic bulletin & review*, vol. 12, pp. 969–92, 01 2006.

[24] E. T. Klemmer, "Simple reaction time as a function of time uncertainty," *Journal of experimental psychology*, vol. 54, no. 3, pp. 195–200, 1957.

[25] J. D. McAuley and N. S. Miller, "Picking up the pace: Effects of global temporal context on sensitivity to the tempo of auditory sequences," *Perception & Psychophysics*, vol. 69, pp. 709–718, 2007.

[26] F. Lerdahl and R. S. Jackendoff, *A Generative Theory of Tonal Music*. The MIT Press, 06 1996.

[27] A. M. C. Sluijter, V. J. van Heuven, and J. J. A. Pacilly, "Spectral balance as a cue in the perception of linguistic stress," *The Journal of the Acoustical Society of America (JASA)*, vol. 101, no. 1, pp. 503–513, 1997.

[28] T. R. Knösche, C. Neuhaus, J. Haueisen, K. Alter, B. Maess, O. W. Witte, and A. D. Friederici, "Perception of phrase structure in music," *Human Brain Mapping*, vol. 24, no. 4, pp. 259–273, 2005.

[29] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An experimental comparison of audio tempo induction algorithms," *IEEE Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 14, pp. 1832–1844, 2006.

[30] C. Gupta, H. Li, and Y. Wang, "Automatic pronunciation evaluation of singing," in *Int. Speech Communication Association (INTERSPEECH)*, 09 2018, pp. 1507–1511.