INTELLIGIBILITY OF SUNG LYRICS: A PILOT STUDY

Karim M. Ibrahim[1] **David Grunberg**[1] **Kat Agres**[2]
Chitralekha Gupta[1] **Ye Wang**[1]
[1] Department of Computer Science, National University of Singapore, Singapore
[2] Institute of High Performance Computing, A*STAR, Singapore
karim.ibrahim@comp.nus.edu.sg, wangye@comp.nus.edu.sg

ABSTRACT

We propose a system to automatically assess the intelligibility of sung lyrics. We are particularly interested in being able to identify songs which are intelligible to second language learners, as such individuals often sing along the song to help them learn their second language, but this is only helpful if the song is intelligible enough for them to understand. As no automatic system for identifying 'intelligible' songs currently exists, songs for second language learners are generally selected by hand, a time-consuming and onerous process. We conducted an experiment in which test subjects, all of whom are learning English as a second language, were presented with 100 excerpts of songs drawn from five different genres. The test subjects listened to and transcribed the excerpts and the intelligibility of each excerpt was assessed based on average transcription accuracy across subjects. Excerpts that were more accurately transcribed on average were considered to be more intelligible than those less accurately transcribed on average. We then tested standard acoustic features to determine which were most strongly correlated with intelligibility. Our final system classifies the intelligibility of the excerpts and achieves 66% accuracy for 3 classes of intelligibility.

1. INTRODUCTION

While various studies have been conducted on singing voice analysis, one aspect which has not been well-studied is the *intelligibility* of a given set of lyrics. Intelligibility describes how easily a listener can comprehend the words that a performer sings; the lyrics of very intelligible songs can easily be understood, while the lyrics of less intelligible songs sound garbled or even incomprehensible to the average listener. People's impressions of many songs are strongly influenced by how intelligible the lyrics are, with one study even finding that certain songs were perceived as 'happy' when people could not understand its lyrics, but was perceived as 'sad' when the downbeat lyrics were made comprehensible [20]. It would thus be useful to enable systems to automatically determine intelligibility, as it is a key factor in people's perception of a wide variety of songs.

We are particularly interested in measuring the intelligibility of songs with respect to second language learners. Many aspects of learning a second language to the point of fluency have been shown to be difficult, including separating the phonemes of an unfamiliar language [30], memorizing a large number of vocabulary words and grammar rules [22], and maintaining motivation for the length of time required to learn the language. Consequently, many second language learners need help, and music has been shown to be a useful tool for this purpose. Singing and language development have been shown to be closely related at the neurological level [24, 32], and experimental results have demonstrated that singing along with music in the second language is an effective way of improving memorization and pronunciation [12, 19]. However, specific songs are only likely to help these students if they can understand the content of the lyrics [11]. As second language learners may have difficulty understanding certain songs in their second language due to their lack of fluency, they could be helped by a system capable of automatically determining which songs they are likely to find intelligible or unintelligible.

We therefore seek to design a system which is capable of assessing a given song and assigning it an intelligibility score, with the standard of intelligibility biased towards people who are learning the language of the lyrics but have not yet mastered it. To gather data for this system we compiled excerpts from 50 songs and had volunteering participants listen to the song in order to discover how intelligible they found the lyrics. Rather than simply having the participants rate the intelligibility of the song, we had the participants transcribe the lyrics that they heard and then calculated an intelligibility score for each excerpt based on the statistics of how accurately the students transcribed it. Excerpts that were transcribed more accurately on average were judged to be more intelligible than those transcribed less accurately on average. A variety of acoustic features were then used to build a classifier which could determine the intelligibility of a given piece of music. The classifier was then run on the same excerpts used in the listening experiment, and the results of each were compared.

The remaining outline of this paper is as follows: Sec-

tion 2 lists relevant literature in the field. Section 3 describes the transcription experiment performed to gather data. Section 4 discusses the features and the classifier. Finally, Sections 5 and 6 shows the evaluation of our proposed model and our conclusions, respectively.

2. LITERATURE REVIEW

That sung lyrics could be more difficult to comprehend than spoken words has long been established in the scientific community. One study showed that even professional voice teachers and phoneticians had difficulty telling vowels apart when sung at high pitch [7]. Seminal work by Collister and Huron found listeners to make hearing errors as much as seven times more frequently when listening to sung lyrics than spoken ones [3]. Such studies also noted lyric features which could help differentiate intelligible from unintelligible songs; for instance, one study noted that songs comprised mostly of common words sounded more intelligible than songs with rarer words [9]. However, lyric features alone are not sufficient to assess intelligibility; the same lyrics can be rendered more or less intelligible depending on, for instance, the speed at which they are sung. These other factors must be taken into account to truly assess lyric intelligibility.

Studies have been conducted on assessing the overall *quality* of singing voice. One acoustic feature which multiple studies have found to be useful for this purpose is the power ratio of frequency bands containing energy from the singing voice to other frequency bands; algorithms using this feature have been shown to reliably distinguish between trained and untrained singers [2,23,34]. Calculation of pitch intervals and vibrato have also been shown to be useful for this purpose [21]. However, while the quality of singing voice may be a factor in assessing intelligibility, it is not the only such factor. Aspects of the song that have nothing to do with the skill of the singer or the quality of their performance, such as the presence of loud background instruments, can contribute, and additional features that take these factors into account are needed for a system which determines lyric intelligibility.

Another related task is that of singing transcription, in which a computer must listen to and transcribe sung lyrics [18]. It may seem that one could assess intelligibility by comparing a computer's transcription of the lyrics to a ground truth set of lyrics and determining if the transcription is accurate. But this too does not really determine intelligibility, at least as humans perceive it. A computer can use various filters and other signal processing or machine learning tools to process the audio and make it easier to understand, but a human listening to the music will not necessarily have access to such tools. Thus, even if a computer can understand or accurately transcribe the lyrics of a piece of music, this does not indicate whether those lyrics would be intelligible to a human as well.

3. BEHAVIORAL EXPERIMENT

To build a system that can automatically process a song and evaluate the intelligibility of its lyrics, it is essential to gather ground truth data that reflects this intelligibility on average across different listeners. Hence, we conducted a study where participants were tasked with listening to short excerpts of music and transcribing the lyrics, a common task for evaluating intelligibility of lyrics [4]. The accuracy of their transcription can be used to assess the intelligibility of each excerpt.

3.1 Method

3.1.1 Participants

Seventeen participants (seven females and ten males) volunteered to take part in the experiment. Participants were between 21 to 41 years (mean = 27.4 years). All participants indicated no history of hearing impairment and that they spoke some English as a second language. Participants were rewarded with a $10 voucher for their time. Participants were recruited through university channels via posters and fliers. The majority of the participants were university students.

3.1.2 Materials

For the purpose of this study, we focused solely on English-language songs. Because one of the main applications for such a system is to recommend music for students who are learning foreign languages, we focused on genres that are popular for students. To identify these genres, we asked 48 university students to choose the 3 genres that they listen to the most, out of the 12 genres introduced in [4], as these 12 genres cover a wide variety of singing styles. The twelve genres are: Avante-garde, Blues, Classical, Country, Folk, Jazz, Pop/Rock, Rhythm and Blues, Rap, Reggae, Religious, and Theater. Because the transcription task is long and tiring for participants, we limited the number of genres tested to only five, from which we would draw approximately 45 minutes worth of music for transcription. We selected the five most popular genres indicated by the 48 participants: Classical, Folk, Jazz, Pop/Rock, and Rhythm and Blues.

After selecting the genres, we collected a dataset of 10 songs per genre. Because we were interested in evaluating participants' ability to transcribe an unfamiliar song, as opposed to transcribing a known song from memory, we focused on selecting songs that are not well-known in each genre. We approached this by selecting songs that have less than 200 ratings on the website Rate Your Music (rateyourmusic.com). Rate Your Music is a database of popular music where users can rate and review different songs, albums and artists. Popular songs have thousands of ratings while less known songs have few ratings. We used this criteria to collect songs spanning the 5 genres to produce our dataset. The songs were randomly selected, with no control over the vocal range or the singer's accent, as long as they satisfied the condition of being in English and having few ratings.

Because transcribing an entire song, let alone 50 songs, would be an overwhelming process for the participants, we selected short excerpts from each song to be transcribed. Two excerpts per song were selected randomly such that each excerpt would include a complete utterance (e.g., no excerpts were terminated mid-phrase). Excerpts varied between 3 to 16 seconds in length (average = 6.5 seconds), and contained 9.5 words on average. The ground-truth lyrics for these songs were collected from online sources and reviewed by the experimenters to ensure they matched the version of the song used in the experiment. It is important to note that selecting short excerpts might affect intelligibility, because the context of the song (which may help in understanding the lyrics) is lost. However, using these short excerpts is essential in making the experiment feasible for the participants, and would still broadly reflect the intelligibility of the song. The complete dataset is composed of 100 excerpts from 50 songs, 2 excerpts per song, covering 5 genres, and 10 songs per genre. Readers who are interested in experimenting on the dataset can contact the authors.

3.1.3 Procedure

We conducted the experiment in three group listening sessions. During each session, the participants were seated in a computer lab, and recorded their transcriptions of the played excerpts on the computer in front of them. The excerpts were played in randomized order, and each excerpt was played twice consecutively. Between the two playbacks of each excerpt there was a pause of 5 seconds, and between different excerpts a pause of 10 seconds, to allow the participants sufficient time to write their transcription. The total duration of the listening session is 46:59 minutes. Two practice trials were presented before the experimental trials began, to familiarize participants with the experimental procedure.

3.2 Results and Discussion

To evaluate the accuracy of the participants' transcription, we counted the number of words correctly transcribed by the participant that match the ground truth lyrics. For each transcription by each student, the ratio between correctly transcribed words to the total number of words in the excerpt was calculated. We then calculated the average ratio for each excerpt across all 17 participants to yield an overall score for each excerpt between 0 and 1. This score was used to represent the ground-truth transcription accuracy, or *Intelligibility score*, for each excerpt. The distribution of Intelligibility scores in the dataset is shown in Figure 1. From the figure, we can observe that the intelligibility scores are biased towards higher values, i.e. there are relatively few excerpts with a low intelligibility score. This may be caused by the restricted set of popular genres indicated by students, as certain excluded genres would be expected to have low intelligibility, such as Heavy Metal.

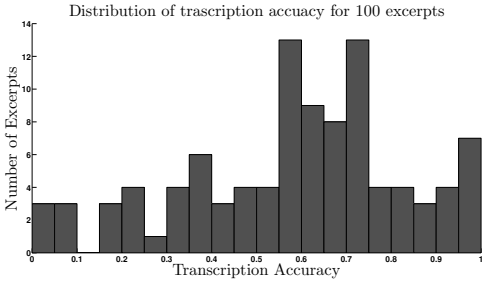

Figure 1. The distribution of the transcription accuracies (Intelligibility score).

4. COMPUTATIONAL SYSTEM

The purpose of this study is to select audio features that can be used to build a system capable of 1) predicting the intelligibility of song lyrics, and 2) evaluating the accuracy of these predictions with respect to the ground truth gathered from human participants. In the following approach, we analyze the input signal and extract expressive features that reflect the different aspects of an intelligible singing voice. Several properties may contribute to making the singing voice less intelligible than normal speech. One such aspect is the presence of background music, as accompanying music can cover or obscure the voice. Therefore, highly intelligible songs would be expected to have a dominant singing voice compared with the accompanying music [4]. Unlike speech, the singing voice has a wider and more dynamic pitch range, often featuring higher pitches in soprano vocal range. This has been shown to affect the intelligibility of the songs, especially with respect to the perception of sung vowels [1, 3]. An additional consideration is that in certain genres, such as Rap, singing is faster and has a higher rate of words per minute than speech, which can reduce intelligibility. Furthermore, as indicated in [10], the presence of common, frequently occurring words helps increase intelligibility, while uncommon words decrease the likelihood of understanding the lyrics. In our model, we aimed to include features that express these different aspects to determine the intelligibility of song lyrics across different genres. These features are then used to train the model to accurately predict the intelligibility of lyrics in the dataset, based on the ground truth collected in our behavioral experiment.

4.1 Preprocessing

To extract the proposed features from an input song, two initial steps are required: separating the singing voice from the accompaniment, and detecting the segments with vocals. To address these steps, we selected the following approaches based on current state-of-the-art methods:

4.1.1 Vocals Separation

Separating vocals from accompaniment music is a well-known problem that has received considerable attention in the research community. Our approach makes use of the popular Adaptive REPET algorithm [16]. This algorithm is

based on detecting the repeating patten in the song, which is meant to represent the background music. Separating the detected pattern leaves the non-repeating part of the song, meant to capture the vocals. Adaptive REPET also has the advantage of discovering local repeating patterns in the song over the original REPET algorithm [26]. Choosing Adaptive REPET was based on two main advantages: The algorithm is computationally attractive, and it shows competitive results compared to other separation algorithms, as shown in the evaluation of [14].

4.1.2 Detecting Vocal Segments

Detecting vocal and non-vocal segments in the song is an important step in extracting additional information about the intelligibility of the lyrics. Various approaches have been proposed to perform accurate vocal segmentation, however, it remains a challenging problem. For our approach, we implemented a method based on extracting the features proposed in [15], then training a Random Forest classifier using the Jamendo corpus [1] [27]. The classifier was then used to binary classify each frame of the input file as either vocals or non-vocals.

4.2 Audio features

In this section, we investigate the set of features we used in training the model for estimating lyrics intelligibility. We use a mix of features reflecting specific aspects of intelligibility plus common standard acoustic features. The selected features are:

1. **Vocals to Accompaniment Music Ratio (VAR)**: Defined as the energy of the separated vocals divided by the energy of the accompaniment music. This ratio is computed only in segments where vocals are present. This feature reflects how strong the vocals are compared to the accompaniment. High VAR suggests that vocals are relatively loud and less likely to be obscured by the music. Hence, higher VAR counts for higher intelligibility. This feature is particularly useful in identifying songs that are unintelligible due to loud background music which obscures the vocals.

2. **Harmonics-to-residual Ratio (HRR)**: Defined as the the energy in a detected fundamental frequency (f0) according to the YIN algorithm [5] plus the energy in its 20 first harmonics (a number chosen based on empirical trials), all divided by the energy of the residual. This ratio is also applied only to segments where vocals are present. Since harmonics of the detected f0 in vocal segments are expected to be produced by the singing voice, this ratio, like VAR, helps to determine whether the vocals in a given piece of music are stronger or weaker than the background music which might obscure it.

3. **High Frequency Energy (HFE)**: Defined as the sum of the spectral magnitude above 4kHz,

$$HFE_n = \sum_{k=f_{4k}}^{N_b/2} a_{n,k} \tag{1}$$

where $a_{n,k}$ is the magnitude of block n and FFT index k of the short time Fourier transform of the input signal, f_{4k} is the index corresponding to 4 kHz and N_b is the FFT size [8]. We calculate the mean across all frames of the separated and segmented vocals signal, as we are interested in the high energy component in vocals and not the accompanying instruments. We get a scalar value per input file reflecting high frequency energy. Singing in higher frequencies has been proven to be less intelligible than music in low frequencies [3], so detection of high frequency energy can be a useful clue that such vocals might be present and could reduce the intelligibility of the music, such as frequently happens with opera music.

4. **High Frequency Component (HFC)**: Defined as the sum of the amplitudes and weighted by the frequency squared,

$$HFC_n = \sum_{k=1}^{N_b/2} k^2 a_{n,k} \tag{2}$$

where $a_{n,k}$ is the magnitude of block n and FFT index k of the short time Fourier transform of the input signal and N_b is the FFT size [17]. This is another measure of high frequency content.

5. **Syllable Rate**: Singing at a fast pace while pronouncing several syllables over a short period of time can negatively affect the intelligibility [6]. In the past, Rao et al. used temporal dynamics of timbral features to separate singing voice from background music [28]. These features showed more variance over time for singing voice, while being relatively invariant to background instruments. We expect that these features will also be sensitive to the syllable rate in singing. We use the temporal standard deviation of two of their timbral features: sub-band energy (SE) in the range of ([300-900 Hz]), and sub-band spectral centroid (SSC) in the range of ([1.2-4.5 kHz]), defined as

$$SSC = \frac{\sum_{k=k_{low}}^{k_{high}} f(k)|X(k)|}{\sum_{k=k_{low}}^{k_{high}} |X(k)|} \tag{3}$$

$$SE = \sum_{k=k_{low}}^{k_{high}} |X(k)|^2 \tag{4}$$

where $f(k)$ and $|X(k)|$ are frequency and magnitude spectral value of the k^{th} frequency bin, and k_{low} and k_{high} are the nearest frequency bins to the lower and upper frequency limits on the sub-band respectively.

[1] http://www.mathieuramona.com/wp/data/jamendo/

According to [28], SE enhances the fluctuations between voiced and unvoiced utterances, while SSC enhances the variations in the 2^{nd}, 3^{rd} and 4^{th} formants across phone transitions in the singing voice. Hence, it is reasonable to expect high temporal variance of these features for songs with high syllable rate, and vice versa. Thus, this feature is able to differentiate songs with high and low syllable rates. We would expect that very high and very low syllable rates should lead to low intelligibility score, while rates in a similar range to that of speech should result in high intelligibility score.

6. **Word-Frequency Score**: Songs which use common words have been shown to be more intelligible than those which use unusual or obscure words [10]. Hence, we calculate a word-frequency score for the lyrics of the songs as an additional feature. This feature is a non-acoustic feature that is useful in cases where the lyrics of the song are available. We calculate the word-frequency score using the `wordfreq` open-source toolbox [31] which provides an estimates of the frequencies of words in many languages.

7. **Tempo and Event Density**: These two rhythmic features reflect how fast the beat and rhythm of the song are. Event density is defined as the average frequency of events, i.e., the number of note onsets per second. Songs with very fast beats and high event density are likely to be less intelligible than slower songs, since the listener has less time to process each event before the next one begins. We used the `MIRToolbox` [13] to extract these rhythmic features.

8. **Mel-frequency cepstral coefficients (MFCCs)**: MFCCs approximates the human auditory system's response more closely than the linearly-spaced frequency bands [25]. MFCCs have been proven to be effective features in problems related to singing voice analysis [29], and so were considered as a potential feature here as well. For our system, we selected the 17 first coefficients (excluding the 0th) as well as the deltas of those features, which proved empirically to be the best number of coefficients. The MFCCs are extracted from the original signal without separation, as it reflects how the whole song is perceived.

By extracting this set of features for an input file, we end up with a vector of 43 features to be used in estimating the intelligibility of the lyrics in this song.

4.3 Model training

We used the dataset and ground-truth collected in our behavioral experiment to train a Support Vector Machine model to estimate the intelligibility of the lyrics. To categorize the intelligibility to different levels that would match a language student's fluency level, we divided our dataset to three classes:
`High Intelligibility`: excerpts with transcription accuracy of greater than 0.66.
`Moderate Intelligibility`: excerpts with transcription accuracy between 0.33 and 0.66 inclusive.
`Low Intelligibility`: excerpts with transcription accuracy of less than 0.33.
Out of the 100 samples in our dataset, 43 are in the High Intelligibility class, 42 are in the Moderate Intelligibility class, and the remaining 15 are in the Low Intelligibility class. For this pilot study, we tried a number of common classifiers, including Support Vector Machine (SVM), random forest and k-nearest neighbors. Our trials for finding a suitable model led to using SVM with a linear kernel, as it is an efficient, fast and simple model which is suitable for this problem. Finally, as a preprocessing step, we normalize all the input feature vectors before passing them to the model to be trained.

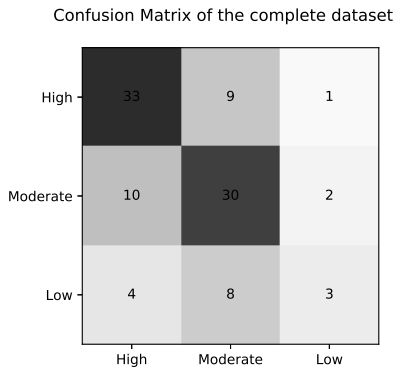

Figure 2. Confusion Matrix of the SVM output.

5. MODEL EVALUATION

Because this problem has not been addressed before in the literature, and it is not possible to perform evaluation using other methods, we based our evaluation on classification accuracy from the dataset. Given the relatively small number of samples in the dataset, we used leave-one-out cross-validation for evaluation. To evaluate the performance of our model, we compute overall accuracy, as well as the Area Under the ROC Curve (AUC). We scored AUC of 0.71 and accuracy of 66% with the aforementioned set of features and model. The confusion matrix of validating our model using leave-one-out cross-validation on our collected dataset is shown in Figure 2. The figure shows that the classifier has relatively more accuracy in predicting high and moderate than low intelligibility, which is often confused with the moderate class. Given that our findings are based on a relatively small segment of excerpts with low intelligibility, the classifier was found to be trained to work better on the high and moderate excerpts.

Following model evaluation on the complete dataset, we were interested in investigating how the model performs on different genres, specifically how it performs when tested

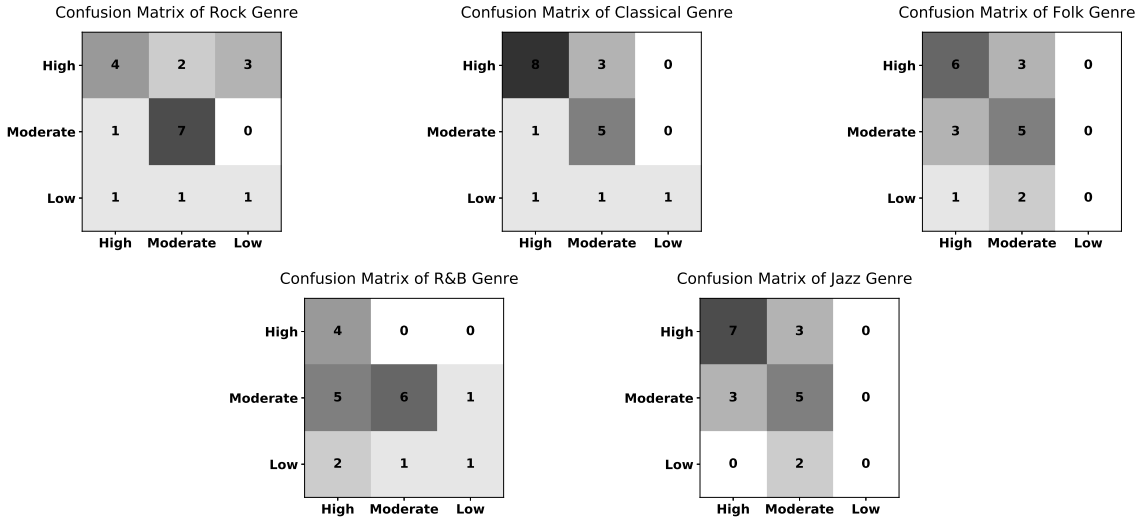

Figure 3. Confusion matrix of the different genres

Genre	Classification Accuracy
Pop/Rock	60%
R&B	55%
Classical	70%
Folk	55%
Jazz	60%

Table 1. Classification accuracy for different genres

with a genre that was not included in the training dataset. This would imply how the model generalizes when running on different genres that was not present during training, as well as showing how changing genres affect classification accuracy. We performed an evaluation where we trained our model using 4 out of the 5 genres in our dataset, and tested it on the 5th genre. The classification accuracy across different genres is shown in Table 1. The results show variance in classifying different genres. For example, Classical music receives higher accuracy, while genres as Rhythm and Blues and Folk shows less accuracy. By analyzing the confusion matrices of each genre shown in Figure 3, we found that the confusion is mainly between high and moderate classes.

By reviewing the impact of the different features on the classifier performance, we looked into what features have the biggest impact using the attribute ranking feature in Weka [35]. We found that several MFCCs contribute most in differentiating between the three classes, which we interpret to be due to analyzing the signal in different frequency sub-bands incorporates perceptual information of both the singing voice and the background music. This was followed by the features reflecting the syllable rate in the song, because singing rate can radically affect the intelligibility. Vocals-to-Accompaniment Ratio and High Frequency Energy followed in their impact on differentiating between the three classes. The features that had the least impact were the tempo and event density, which does not necessarily reflect the rate of singing.

For further studies on the suitability of the features in classifying songs with very low intelligibility, the genres pool can be extended to include other genres with lower intelligibility, rather than being limited to the popular genres between students. Further studies can also include the feature selection and evaluation process: similar to the work in [33], deep learning methods may be explored to select the features which perform best, rather than hand-picking features, to find the most suitable set of features for this problem. It is possible to extend the categorical approach of intelligibility levels to a regression problem, in which the system evaluates the song's intelligibility with a percentage. Similarly, certain ranges of the intelligibility score can be used to recommend songs to students based on their fluency level.

6. CONCLUSION

In this study, we investigated the problem of evaluating the intelligibility of song lyrics to provide an aid for language learners who listen to music as part of language immersion. We conducted a behavioral experiment to review how the intelligibility of lyrics in different genres of songs are perceived by human participants. We then developed a computational system to automatically estimate the intelligibility of lyrics in a given song. In our system, we proposed features to reflect different factors that affect the intelligibility of lyrics according to previous empirical studies. We used the proposed features along with standard audio features to train a model capable of estimating the intelligibility of lyrics (as low, moderate, or high intelligibility) with an AUC of 0.71. The study provides evidence that the proposed system has promising initial results, and draws attention to the problem of lyrics intelligibility, which has received little attention in terms of computational audio analysis and automatic evaluation.

7. REFERENCES

[1] Martha S Benolken and Charles E Swanson. The effect of pitch-related changes on the perception of sung vowels. *The Journal of the Acoustical Society of America*, 87(4):1781–1785, 1990.

[2] Ugo Cesari, Maurizio Iengo, and Pasqualina Apisa. Qualitative and quantitative measurement of the singing voice. *Folia Phoniatrica et Logopaedica*, 64(6):304–309, 2013.

[3] Lauren Collister and David Huron. Comparison of word intelligibility in spoken and sung phrases. *Empirical Musicology Review*, 3(3):109–125, 2–8.

[4] Nathaniel Condit-Schultz and David Huron. Catching the lyrics. *Music Perception: An Interdisciplinary Journal*, 32(5):470–483, 2015.

[5] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[6] Aihong Du, Chundan Lin, and Jingjing Wang. Effect of speech rate for sentences on speech intelligibility. In *Communication Problem-Solving (ICCP), 2014 IEEE International Conference on*, pages 233–236. IEEE, 2014.

[7] Harry Hollien, Ana Mendes-Schwartz, and Kenneth Nielsen. Perceptual confusions of high-pitched sung vowels. *Journal of Voice*, 14(2):287–298, 2000.

[8] Kristoffer Jensen and Tue Haste Andersen. Real-time beat estimationusing feature extraction. In *International Symposium on Computer Music Modeling and Retrieval*, pages 13–22. Springer, 2003.

[9] Randolph Johnson, David Huron, and Lauren Collister. Music and lyrics interaction and their influence on recognition of sung words: an investigation of word frequency, rhyme, metric stress, vocal timbre, melisma, and repetition priming. *Empirical Musicology Review*, 9(1):2–20, 2014.

[10] Randolph B Johnson, David Huron, and Lauren Collister. Music and lyrics interactions and their influence on recognition of sung words: an investigation of word frequency, rhyme, metric stress, vocal timbre, melisma, and repetition priming. *Empirical Musicology Review*, 9(1):2–20, 2013.

[11] Tung-an Kao and Rebecca Oxford. Learning language through music: A strategy for building inspiration and motivation. *System*, 43:114–120, 2014.

[12] Anne Kultti. Singing as language learning activity in multilingual toddler groups in preschool. *Early Child Development and Care*, 183(12):1955–1969, 2013.

[13] Olivier Lartillot and Petri Toiviainen. A matlab toolbox for musical feature extraction from audio. `https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox`, 2007.

[14] Bernhard Lehner and Gerhard Widmer. Monaural blind source separation in the context of vocal detection. In *16th International Society for Music Information Retrieval Conference (ISMIR), At Malaga, Spain*, 2015.

[15] Bernhard Lehner, Gerhard Widmer, and Reinhard Sonnleitner. On the reduction of false positives in singing voice detection. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7480–7484. IEEE, 2014.

[16] Antoine Liutkus, Zafar Rafii, Roland Badeau, Bryan Pardo, and Gaël Richard. Adaptive filtering for music/voice separation exploiting the repeating musical structure. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 53–56. IEEE, 2012.

[17] Paul Masri and Andrew Bateman. Improved modelling of attack transients in music analysis-resynthesis. In *ICMC*, 1996.

[18] Annamaria Mesaros and Tuomas Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010.

[19] Carmen Mora. Foreign language acquisition and melody singing. *ELT journal*, 54(2):146–152, 2000.

[20] Kazuma Mori and Makoto Iwanaga. Pleasure generated by sadness: Effect of sad lyrics on the emotions induced by happy music. *Psychology of Music*, 42(5), 2014.

[21] Tomoyasu Nakano. An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. In *Proceedings of INTERSPEECH2006*, 2006.

[22] Joan Netten and Claude Germain. A new paradigm for the learning of a second or foreign language: the neurolinguistic approach. *Neuroeducation*, 1(1), 2012.

[23] Koichi Omori, Ashutosh Kacker, Linda Carroll, William Riley, and Stanley Blaugrund. Singing power ratio: quantitative evaluation of singing voice quality. *Journal of Voice*, 10(3):228–235, 1996.

[24] Aniruddh Patel. Language, music, syntax and the brain. *Nature Neuroscience*, 6(7):674–681, 2003.

[25] Lawrence R Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. PTR Prentice Hall, 1993.

[26] Zafar Rafii and Bryan Pardo. Repeating pattern extraction technique (repet): A simple method for music/voice separation. *IEEE transactions on audio, speech, and language processing*, 21(1):73–84, 2013.

[27] Mathieu Ramona, Gaël Richard, and Bertrand David. Vocal detection in music with support vector machines. In *Proc. ICASSP '08*, pages 1885–1888, March 31 - April 4 2008.

[28] Vishweshwara Rao, Chitralekha Gupta, and Preeti Rao. Context-aware features for singing voice detection in polyphonic music. In *International Workshop on Adaptive Multimedia Retrieval*, pages 43–57. Springer, 2011.

[29] Martın Rocamora and Perfecto Herrera. Comparing audio descriptors for singing voice detection in music audio files. In *Brazilian symposium on computer music, 11th. san pablo, brazil*, volume 26, page 27, 2007.

[30] Daniele Schon, Sylvain Moreno, Mireille Besson, Isabelle Peretz, and Regine Kolinsky. Songs as an aid for language acquisition. *Cognition*, 106(2):975–983, 2008.

[31] Robert Speer, Joshua Chin, Andrew Lin, Lance Nathan, and Sara Jewett. wordfreq: v1.5.1. `https://doi.org/10.5281/zenodo.61937`, September 2016.

[32] Valerie Trollinger. The brain in singing and language. *General Music Today*, 23(2), 2010.

[33] Xinxi Wang and Ye Wang. Improving content-based and hybrid music recommendation using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 627–636. ACM, 2014.

[34] Christopher Watts, Kathryn Barnes-Burroughs, Julie Estis, and Debra Blanton. The singing power ratio as an objective measure of singing voice quality in untrained talented and nontalented singers. *Journal of Voice*, 20(1):82–88, 2006.

[35] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.