Automatic Evaluation of Song Intelligibility using Singing Adapted STOI and Vocal-specific Features

Bidisha Sharma and Ye Wang

***Abstract*—An objective machine-driven measure of song intelligibility would be of great utility for various music information retrieval tasks. Song intelligibility mostly depends on two factors, the amount of interference caused by background accompaniment, and the quality of singing vocal. We leverage these two factors to determine the intelligibility of a song. For the first factor, we adapt a well known method for intelligibility prediction of noisy speech, short term objective intelligibility (STOI), to singing. The singing-adapted STOI considers the polyphonic song as a time-frequency weighted noisy version of the extracted singing vocal. We use U-net based audio source separation method to extract singing vocal from a polyphonic song. The singing vocal shares the same underlying physiological mechanism for production as that of speech, with some differences in the pronunciation and prosody of the phonemes. Therefore, for the second factor, we have introduced vocal-specific features to measure the intelligibility of the singing vocal, which are excitation source, spectral, and prosodic singing characteristics. We perform detailed analysis on each of these features to establish their efficacy for quantifying song intelligibility. We train a regression model to derive the intelligibility scores using a combination of the vocal-specific features and singing adapted STOI, obtaining a significant improvement in performance. The correlation between the intelligibility score obtained using proposed framework and human-rated intelligibility score is 0.81, which shows the efficacy of the proposed approach.**

***Index Terms*—Song intelligibility, language learning, song recommendation, music, vocal-specific features, modulation spectrum, excitation source.**

I. Introduction

Song intelligibility is the measure of how comprehensible the sung lyrics are in the presence of background accompaniment in a polyphonic music composition. Estimating this measure of intelligibility could be useful to guide a song recommendation system that specifically aims to recommend intelligible songs. Such a system would not only be effective for entertainment purpose, but also helps to select songs useful for language learning [1]–[3].

However, estimating song intelligibility is a very challenging problem and can even surpass the speech intelligibility prediction problem in difficulty. This is because of two main factors that introduce additional difficulty in song comprehension: the accompaniment's interference with the singing vocal, and singing vocal-specific characteristics such as syllable rate, pitch range, loudness, and voice quality. Moreover, for a given song it is generally unknown how a reference version of that song would sound like, with extremely intelligible lyrics; consequently, intelligibility cannot be simply derived by comparing a given version of a song to a reference. Therefore, in this work, we focus on developing a reference-independent method to automatically evaluate song intelligibility in a way that approximates human judgment.

Several studies in the literature have established that listening to songs can contribute to language development [1], [4]–[8]. However, not all songs are equally suitable for this task. The songs which are challenging to understand can lower student's interest in learning the language. On the other hand, songs which are easily comprehensible can raise the interest of the students, as well as help them to follow and understand the content that they are singing [9]. This is important because, apart from language learning, one of the potential goals of a song for listeners is to gain some level of understanding of the message and mood being communicated [10]. As manually selecting songs for this purpose is onerous and subject to bias, a system which facilitates automatic recommendation of songs that are easy to comprehend would thus be of use. However, to make this possible, a framework for automatically evaluating song intelligibility would be a necessity.

Song intelligibility notion has been analyzed from different perspectives in the literature. The studies in [11], [12] showed that the intelligibility of vowels is significantly reduced, when the sung pitch is very high. Particularly, high-pitched tones were less intelligible when sung with a lowered larynx, as classically trained singers often do. In a comparison between spoken and sung lyrics [13], researchers found that the intelligibility of singing vocal decreased by seven-fold compared with their spoken counterparts. This is because singing adds many constraints to the characteristics of different sounds, such as higher fundamental frequencies (F_0), increased formant frequencies, vowel centralization [12], [14], [15], linguistic and rhythmic factors [16], voice quality [17], change in relative loudness, syllable rate, singing style, the amount of reverberation, and the accompaniment's timbre [18]. Deviations in these factors between singing vocals and spoken utterances lead to incorrect perceptions of phonemes and words in the lyrics. While the authors of [18] analyzed the effects of such features on perceived intelligibility, they did not extract them to quantify intelligibility more explicitly. An investigation of how such extracted acoustic features are useful to measure the intelligibility of a polyphonic music composition would definitely be applicable in a real world scenario.

In addition to the acoustic characteristic of the singing vocal, the lexical characteristics of the lyrics is also a factor related to the song intelligibility [19]. However, the lexical

Bidisha Sharma is with Electrical and Computer Engineering Department, National University of Singapore and Ye Wang is with the School of Computing, National University of Singapore, Singapore 117543. Email:s.bidisha@nus.edu.sg, wangye@comp.nus.edu.sg.

characteristics alone are not enough to represent the intelligibility of songs. The same lyrics can be rendered more or less intelligible depending on, for instance, the speed at which they are sung or the intensity of the background accompaniment. Consequently, any reasonable measure of intelligibility cannot be limited to lexical analysis and must incorporate acoustic content.

A. *Song intelligibility measure*

In [20], Ibrahim et al. performed a pilot study on the intelligibility of sung lyrics. They explored different acoustic features such as vocal-to-accompaniment ratio, harmonics-to-residual ratio, high frequency energy, high frequency components, syllable rate, tempo and event density, and Mel-frequency cepstral coefficients (MFCCs) to classify songs as high, moderate, or low-intelligible. They used these features in a support vector machine (SVM) based classifier to obtain the final classification labels. They reported a 66% average classification accuracy, with poor accuracy (20%) for low-intelligible songs. However, they did not perform a detailed analysis of the significance of each of these features and their complementary aspects on intelligibility. Furthermore, most of the features used in this study are related to the ratio of vocal and background accompaniment energies. However, some vocals with little or no background accompaniment may nonetheless be unintelligible, as can be the case with, for example, opera and classical songs. Furthermore, while classification provides some idea of intelligibility, a regression approach would be better able to provide sufficient granularity for use in music recommendation systems.

B. *Speech intelligibility measure*

While a wide number of methods have been proposed to quantify speech intelligibility, many of them are specifically designed for particular speech applications, which may not be applicable in case of music. The speech intelligibility measures can be broadly classified into two categories: intrusive and non-intrusive methods. Intrusive methods require a reference signal (clean speech) to which the test signal (degraded speech) is compared. However, it is impractical to have a reference song with ideal intelligibility corresponding to each test song. For non-intrusive methods, no reference signal is needed, but existing methods of this type are still designed for particular scenarios that do not necessarily apply to sung utterances. For example, one widely used non-intrusive method for speech intelligibility evaluation is speech to reverberation ratio (SRMR) [21], but this approach is specifically proposed to predict intelligibility of reverberant and dereverberated speech, and it does not include the factors related to song intelligibility.

Although these methods cannot be directly applied to song intelligibility evaluation, they can potentially be adapted for this purpose. For instance, one popular intrusive method of spoken utterance intelligibility estimation is short term objective intelligibility (STOI) [22]. STOI is used to evaluate the intelligibility of time-frequency (TF) weighted noisy speech with respect to clean speech. Our research shows that STOI can be adapted and modified to analyze the intelligibility of sung lyrics as well.

C. *Overview of proposed approach*

We propose a framework to automatically obtain an intelligibility score of a song that correlates with human judgment. Based on the literature, we summarize that the factors related to song intelligibility are mostly from two categories. First, *the interference between the background accompaniment and the singing vocal*. The presence of relatively loud background accompaniment can inhibit comprehension of singing vocals, which is in turn analogous to the relative loudness levels of the vocals and the instrumental accompaniment. Second, the aspects that relate to the *singing vocal-specific features*, including syllable rate, pitch range, loudness, and vocal production specific aspects.

Knowledge of both of these aspects is essential for assessing song intelligibility; neither alone is sufficient. Songs with little or no background accompaniment may nonetheless be unintelligible due to the very high pitch, very fast or slow tempo, or bad voice quality. On the other hand, songs with average pitch, adequate tempo, and good vocal quality may be incomprehensible, if background accompaniment is too high. We hypothesize to extract both aspects independently and combine them to derive the intelligibility information.

We consider the presence of background accompaniment as a noisy component over the singing vocal and the extracted singing vocal as a clean signal. The extracted singing vocal is obtained using the U-Net based audio source separation method [23]. We adapt the STOI measure from speech to singing and incorporate it to determine the correlation between the extracted singing vocal (reference signal) and singing vocal with background accompaniment (test signal).

To quantify the effects of singing vocal-specific characteristics, we leverage on the idea that preservation of speech-like characteristics in the singing vocals contributes to intelligibility [13]. Similar articulatory movements result in the production of phonemes in speech and singing, and these movements also result in similar excitation source and spectral cues. In singing vocals, due to the presence of background accompaniment and modification of pitch, duration, syllable rate, and phoneme pronunciation, their acoustic features deviate from that of speech. These deviations, in turn, cause smearing of the vocal-specific features and increase difficulty in perception, resulting in lower intelligibility. The pronunciation error is a result of the wrong excitation source and vocal-tract configuration for a particular phoneme. Therefore, we extract the vocal-specific features commonly used in the speech literature to characterize the singing vocal.

Building on this idea of using *singing adapted STOI* and *vocal-specific features*, we further develop regression models to estimate intelligibility.

The rest of the paper is organized as follows: adaptation of STOI for song intelligibility evaluation is discussed in Section II. In Section III, we provide a detailed discussion on vocal-specific features and their significance in terms of intelligibility. The proposed framework is described in Section IV. The experimental evaluation of the framework is presented in Section V. Finally, Section VI summarizes the contribution, results, and discussion of possible future studies.

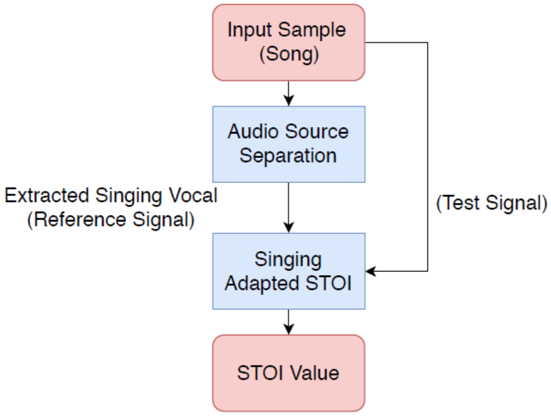

Fig. 1. Block diagram representing singing adapted STOI.

II. Singing Adapted Short-term Objective Intelligibility (STOI)

STOI is one of the most common approaches for automatically measuring intelligibility in the case of noisy and TF-weighted noisy speech [22], and we propose to adapt this method to evaluate song intelligibility. If we have a reference (noisy speech) and a test signal (clean speech), the STOI measure would be the correlation between them.

In this approach, we consider a song as a mixture of two characteristic sources: singing vocal and background accompaniment. The singing vocal combined with background accompaniment can be compared to a speech signal that is contaminated with background noise (test signal), and the extracted singing vocal can be compared to the corresponding clean speech signal (reference signal), as shown in Figure 1. We use the state-of-the-art U-Net based audio source separation method to extract the singing vocal from a polyphonic song [23]. The U-Net architecture (initially developed for medical imaging), builds upon the fully convolutional network [24] with symmetric down-sampling and up-sampling paths. It has the capability to recreate the fine, low-level detail required for high quality audio reproduction. In this work, we have used the pre-trained models corresponding to the iKala dataset and the implementation available in [25], which uses the Chainer framework. From our analysis, we note that the computed singing adapted STOI is not largely dependent on the performance of audio source separation method.

The spectrogram representations for a high-intelligible song excerpt and its extracted singing vocal are shown in Figure 2(a) and (b), respectively. The equivalent spectrograms for a moderate-intelligible excerpt are shown in Figure 2(c) and (d), and those for a low-intelligible excerpt are displayed in Figure 2(e) and (f). It is observed that the correlation between the song's spectrogram and that of the corresponding extracted singing vocal decreases as intelligibility decreases, due to the increase in interference of the background accompaniment. Furthermore, the presence of dominant vocal information in the intelligible songs compared to the less intelligible songs leads to better vocal separation, which is evident from our observation. As STOI captures the correlation between a test signal and its reference signal, it can be effectively used to quantify the interference caused by the background accompaniment.

The method of computing STOI [22] is as follows: both the signals are TF-decomposed in order to obtain a representation that correlates to auditory perception. The lower energy frames with respect to the reference signal are removed from both signals. Next, a one-third octave band analysis is performed by grouping discrete Fourier Transform (DFT) bins. In total 15 one-third octave bands are used. The k^{th} DFT bin of m^{th} frame is denoted by $\hat{x}(k,m)$ of the reference signal. The norm of the j^{th} one-third octave band is defined as,

$$X_j(m) = \sqrt{\sum_{k_1(j)}^{k_2(j)-1} |\hat{x}(k,m)|^2}, \tag{1}$$

where k_1 and k_2 denote the one-third octave band edges, which are rounded to the nearest DFT bin. Similarly, the TF representation of test signal is represented as $Y_j(m)$. Then the TF-unit $Y_j(m)$ is normalized and clipped (Y') to reduce the signal-to-distortion ratio [22]. An intermediate intelligibility measure is then defined as an estimate of the Pearson correlation coefficient $(d_j(m))$ between the clean $(X_j(m))$ and modified/processed $(Y_j'(m))$ TF-units. The average of $d_j(m)$ is measured over all bands and frames, which is defined as the STOI measure,

$$d = \frac{1}{JM}\sum_{j,m} d_j(m), \tag{2}$$

where, J and M represent the number of one-third-octave bands and the total number of frames respectively.

Before computing the singing adapted STOI, we downsample the audio signals to 10 kHz sampling rate, as all the parameters to measure STOI are defined accordingly [22]. In the singing vocal with background accompaniment, the singer's voice cannot be readily heard due to the interference of the frequency characteristics of instruments over the singer's formant, and this interference is specifically present in the higher frequency bands [26]. Due to the higher pitch level in singing relative to speech, the range of the second to the fifth formant in a sung utterance generally lies between 1.25 to 5 kHz [27]. Unlike noisy speech, the interference caused by the background accompaniment is generally found in the higher frequency formants. Our analysis of the STOI measure over different octave bands shows that the interference due to the background accompaniment is more prominent (in the low-intelligible excerpts) over the frequency range of 2-5 kHz. Therefore, unlike STOI in speech, we do not perform averaging over all the frequency bands as shown in (2). Instead, we consider only the upper three bands, denoted as bands number 13, 14, and 15, with center frequencies 2.40 kHz, 3.02 kHz, and 4.30 kHz respectively. This can be expressed as,

$$d_{\text{singing}} = \frac{1}{3M}\sum_{j=13,m=1}^{15,M} d_j(m), \tag{3}$$

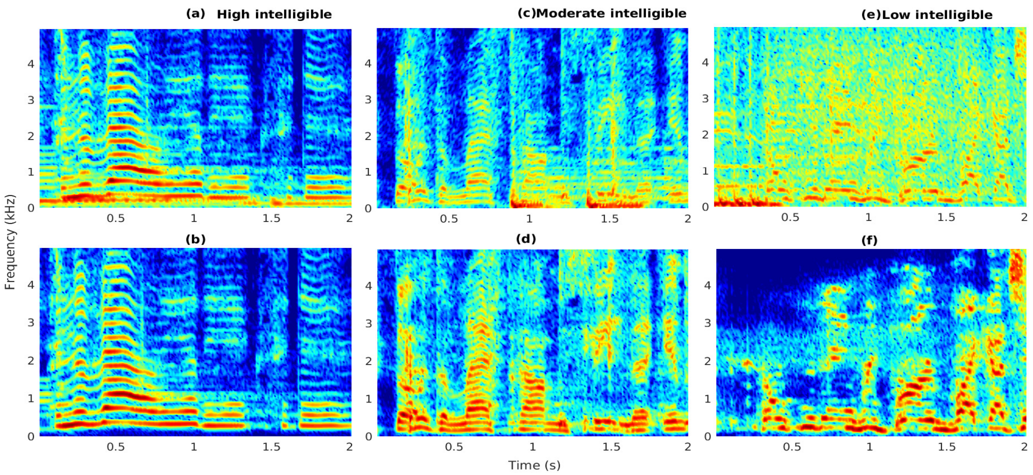

Fig. 2. Spectrogram representations for (a) a high-intelligible song; (b) the extracted vocal from the same high-intelligible song in (a), (c) a moderate-intelligible song; (d) the extracted vocal from the same moderately intelligible song in (c); (e) a low-intelligible song, (f) the extracted vocal from the same low-intelligible song in (e).

where, j and m represent the band number and number of frames in an excerpt, respectively. We have extracted d_{singing} for 200 excerpts. A detailed description of the dataset is presented in Section V-A. The Pearson correlation coefficients obtained between d_{singing} and human-rated scores in 0.39. However, this correlation value is particularly low for 23 excerpts out of 200 excerpts. If we consider these excerpts as aberrations and exclude them, we achieve a Pearson correlation coefficient of 0.60 on the remaining dataset.

A. *Limitations of singing adapted STOI*

Although singing adapted STOI provides a reasonable correlation for most of the excerpts (88%) in the database, we cannot overlook the examples for which the human-rated intelligibility score is not correlated with the obtained singing adapted STOI value. We investigate these 23 (12%) excerpts and observed that independent of the interference caused by the background accompaniment, certain characteristics of singing vocals do contribute to intelligibility and cannot be captured using singing adapted STOI. These factors include:

- Syllable rate: Our articulators move at a certain rate when we produce a vocal signal. Normal hearing listeners can perceive vocal signals as intelligible only if the amplitude fluctuations of those signals is limited to a certain frequency range [28]. If the rate of production of syllables is very fast or slow while singing, the song will be less intelligible.
- Pitch: It is evident from earlier studies that difficulty in comprehending singing vocals increases with increase in pitch [29].
- Loudness and pronunciation: Independent of the interference caused by the background accompaniment, the loudness of the sung lyrics and the correctness of the pronunciation of the words, syllables and phonemes are important aspects of song intelligibility.
- Voice quality: Another factor related to song intelligibility is vocal quality. If the singer's voice contains attributes such as breathiness, roughness, or hoarseness, the song is likely to be less intelligible.

To take into account the above factors related to song intelligibility, we must incorporate each of them into the song intelligibility evaluation framework. The literature supports the contention that singing sounds can be essentially regarded as modified speech sounds [30], and this finding inspires us to analyze different features widely used to characterize speech signals. We thus study the vocal-specific features with respect to intelligibility and incorporate them into the song intelligibility evaluation.

III. Vocal-specific Features

TABLE I
Vocal-specific features and their significance.

Features	Significance
Pitch	Fundamental frequency of singing vocal
Smoothed Hilbert envelope (HE)	Excitation source energy
Peak-to-sidelobe ratio	Strength of excitation
Slope of peaks of HE of LP residual	Strength of excitation
Spectral peak energy	Energy of spectrum
Modulation spectrum energy	Rate of articulation
Sub-band correlation	Formant structure
Spectral slope	Spectral energy dustribution
Normalized autocorrelation peaks	Periodicity
Suprasegmental feature	Periodicity
Jitter	Variation in pitch
Shimmer	Variation in amplitude

In this work, we analyze three categories of vocal-specific features: excitation source, spectral, and prosodic aspects. The excitation source characteristics used are pitch, smoothed Hilbert envelope, peak-to-sidelobe ratio and slope of peaks of Hilbert envelope (HE) of linear prediction (LP) residual. The spectral features used are the sum of spectral peak values, modulation spectrum energy, sub-band correlation, and spectral slope. Periodicity, jitter and shimmer are the prosodic aspects discussed in this section. The features used and their significance are briefly summarized in Table I.

We note that except for pitch, all of these features are derived from the polyphonic song and not the extracted singing vocal. The extracted singing vocal may be distorted by the audio source separation process and smearing or preservation of vocal-specific characteristics due to the interference of background accompaniment cannot be captured by the extracted singing vocal in any event. To extract the vocal-specific features, the audio signals are downsampled to 16 kHz sampling rate and analyzed for 25 ms frame-size with a shift of 5 ms. We perform vocal segmentation as a pre-processing step to remove the segments with only background accompaniment. The dataset used consists of 200 excerpts with a human-rated intelligibility score corresponding to each excerpt, which is described in Section V-A.

A. *Vocal segmentation*

We are interested in measuring intelligibility, which applies only to the vocal sections of a piece of music and not to purely instrumental segments. We note that in the extracted singing vocal, sections that contain only instrumental accompaniments are suppressed and their energy is very low. We divide the spectrum of each frame of the extracted singing vocal into 4 equal sub-bands. On account of the observation that the 2$^{\text{nd}}$ sub-band energy shows a prominent difference between segments with vocals and those without vocals, we set thresholds based on the average 2$^{\text{nd}}$ sub-band energy and segment length, then checks frames to see if they fall below the thresholds and thus likely contain no lyrics. The detected non-vocal sections are removed from both the extracted vocal signal and the polyphonic song in this pre-processing step.

B. *Excitation source features*

1) Pitch: Pitch is considered to be the most fundamental aspect of singing voice, and it is related to both the intelligibility and the quality of songs [17], [29]. Studies have shown increase in F_0 results in lowering intelligibility [13]. We use the librosa library [31] to perform pitch extraction from the extracted singing vocal (Unvoiced segments are assigned a pitch value of 0) for the entire dataset.

2) Smoothed Hilbert envelope: LP analysis is a useful tool to deconvolve the speech signal into vocal-tract and excitation source information [32]. LP analysis is performed to derive the LP coefficients which are then inverse filtered to obtain the residual signal. The LP residual is a good approximation of the excitation source signal. One issue with this approach is that glottal closure instants (GCIs) in the production of the vocal are manifested as large amplitude fluctuations of either positive or negative polarity in the LP residual. This difficulty can be overcome by using the HE of LP residual [33].

The HE of LP residual for 100 ms vocal segment corresponding to low- and high-intelligible excerpts are shown in Figure 3(b) and Figure 3(d). They are obtained from the audio signals shown in Figure 3(a) and Figure 3(c) respectively. It is clear that the peaks of the HE of LP residual are far less prominent in the case of the low-intelligible excerpt (Figure 3(b)) as compared to that of the high-intelligible excerpt (Figure 3(d)).

In this representation, both fine and gross level changes in excitation characteristics are present. For instance, the fine level change may be from closed to open phase in a pitch period, and the gross level change may be in the energy level. To capture only gross level information, small changes are smoothed away by convolving the HE of the LP residual using a Hamming window of 25 ms [34].

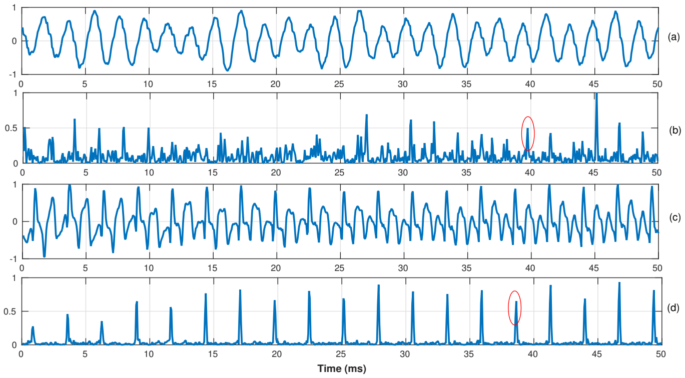

Fig. 3. Comparison of Hilbert envelope of LP residuals for high-intelligible and low-intelligible songs; (a) singing vocal with background accompaniment from a low-intelligible song, (b) HE of LP residual extracted from the signal in (a), (c) singing vocal with background accompaniment from a high-intelligible song, (d) HE of LP residual extracted from the signal in (c).

The smoothed Hilbert envelope for vocal segments corresponding to low-intelligible excerpt (Figure 4(a)) and high-intelligible excerpt (Figure 4(g)) are shown in Figure 4(b) and Figure 4(h) respectively.

3) Peak-to-sidelobe ratio: The peaks in the HE of LP residual signal correspond to GCIs, which is indicated with the red circle in Figure 3(b) and Figure 3(d). The difference between the nature of the HE of LP residual signals in Figure 3(b) and Figure 3(d) is that the sidelobes around each peak have higher values in case of the low-intelligible excerpt compared to that of the high-intelligible excerpt. This reflects the impulse-like nature of the excitation source during production of vocals [35], which is an indicator of the strength of excitation. Distinct peaks with suppressed sidelobes indicate loudness in the produced vocal and the presence of less interference from instrumental sounds.

To quantify this observation, we consider a short segment of 1.5 ms towards the right of each peak of each HE of LP residual. The peak-to-sidelobe ratio is obtained by dividing the peak value by the mean of the samples from 0.5 ms to 1 ms of the 1.5 ms segment [36]. We thus obtain the feature value for each peak of a given HE of LP residual signal, then average the feature values over a 25 ms frame.

4) Slope of peaks of HE of LP residual: The slope of the peaks of HE of LP residual also signifies the excitation source strength [35]. As shown in Figure 3, the peaks of HE of LP residual are sharper for high-intelligible singing vocal segments compared to low-intelligible segments. Both the left and right side slopes of each peak are calculated and their average is treated as the slope of the peak. The slope of peaks within each frame is averaged to determine the average slope of the frame.

Although all of the above features are extracted from the same HE of LP residual signal, they represent different information. The smoothed Hilbert envelope gives us gross level information, while the other two provide fine level information.

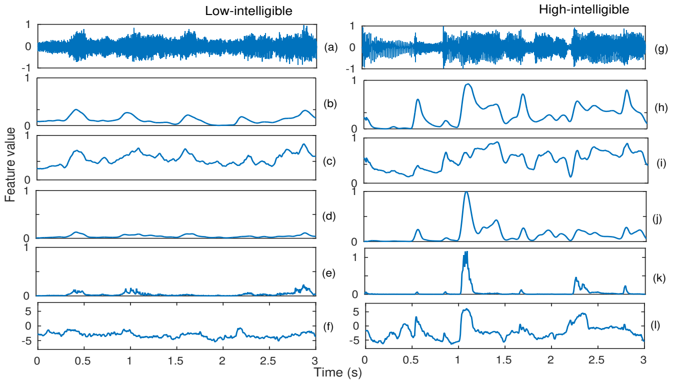

Fig. 4. (a) Singing vocal with background accompaniment from a low-intelligible song, (b) smoothed Hilbert envelope, (c) sum of spectral peaks, (d) modulation spectrum energy, (e) spectral correlation, (f) spectral slope extract from the signal shown in (a); (g) singing vocal with background accompaniment from a high-intelligible song, (h) smoothed Hilbert envelope, (i) sum of spectral peaks, (j) modulation spectrum energy, (k) spectral correlation, (l) spectral slope extracted from the signal shown in (g).

C. Spectral features

1) Spectral peak energy: The amplitudes of the formants obtained from the spectrum of the vocal signal represent the vocal-tract shape. These spectral peaks can be estimated by selecting some of the largest peaks in the LP spectrum. Since the objective is just to have gross information about the vocal-tract shape, we obtain the sum of the ten largest peaks [34]. The normalized spectral peak energy contours for vocal segment corresponding to low-intelligible (Figure 4(a)) and high-intelligible (Figure 4(g)) excerpts are shown in Figure 4(c) and Figure 4(i), respectively.

2) Modulation spectrum energy: The modulation spectrum represents the evolution of the amplitude content of various frequency bands in the short time Fourier transform (STFT) spectrum over time [37]. Normal hearing listeners can perceive speech or other vocal signals as intelligible only if the amplitude fluctuation is limited to a certain frequency range [28]. Therefore, capturing modulation spectrum energy over specific bands of the vocal signal can be useful. We use the method of Greenberg et al. [37], [38] to extract the modulation spectrum [37], [38]. When extracting the modulation spectrum for speech processing, a sampling frequency of 8 kHz is usually used. However, maximum frequency of singing vocal may exceed 4 kHz, therefore, we use a sampling frequency of 16 kHz for this purpose.

The modulation spectrum energy for 3 sec vocal segments corresponding to low-intelligible and high-intelligible excerpts are shown in Figure 4(d) and Figure 4(j), respectively. We observe that the average modulation spectrum energy in Figure 4(d) is higher than that shown in Figure 4(j).

3) Sub-band correlation: It is evident from the spectrograms shown in Figure 2 that the formant structure is more preserved in high-intelligible compared to a low-intelligible song. This is partially because the introduction of more noise-like components in the spectrum smears out the song's harmonic nature while making the audio unintelligible. The sub-band correlation measure has been used in speech processing in [39] to detect the vowel formant structure. To capture this information, we divided the spectrum into four bands and determined the correlation between compressed energy envelopes of these bands. As evident from Figure 4(e) and Figure 4(k), high-intelligible excerpts have higher values in their sub-band energy contours than low-intelligible excerpts.

4) Spectral slope: The spectral slope represents the prominence of higher and lower frequency energy in a spectrum. Because the suppression of higher harmonics make sounds muffled and less intelligible in the presence of noise or any other interfering signal, the spectral slope can be treated as an indication of intelligibility [40]. Moreover, it has been used for this purpose in tasks such as voice quality analysis and Lombard speech analysis [40], [41]. Motivated by these previous studies, we examine the difference between the spectral slopes of high-intelligible and low-intelligible songs. Figure 4(f) and Figure 4(l) show the spectral slope contours for the excerpts shown in Figure 4(a) and Figure 4(g) respectively. It can be observed that high-intelligible segments have a more negative spectral slope. In the case of loud or hyper-articulated speech, there tend to be strong high frequency components which result in a comparatively flatter spectrum.

D. Prosodic features

Prosodic features are long-term features, as the segments affected (syllables, words and phrases) are larger than phonetic units. These features are mainly manifested as sound duration, tone, and intensity variation.

1) Normalized autocorrelation peaks: From the low-intelligible and high-intelligible song segments shown in Figure 3(a) and Figure 3(c) respectively, it is evident that over a small frame of about 25 ms the periodic nature is more intact in the case of the high-intelligible compared to the low-intelligible. Periodicity is thus a potentially helpful feature for intelligibility prediction, and we estimate it by using short-term autocorrelation analysis. The value of the first peak (after the central peak) in the autocorrelation sequence is an indication of the periodicity. The central peak is the peak of the autocorrelation sequence at the origin. The value of the first maximum peak is normalized with respect to the central peak, which gives the normalized autocorrelation peak value. In Figure 5 (a) and Figure 5 (b), the autocorrelation sequence for the frames of low-intelligible and high-intelligible excerpts are respectively shown. The peak values in both are marked with rectangles. We can observe that the normalized autocorrelation peak value is significantly less in the case of the low-intelligible, compared to the high-intelligible excerpt.

2) Suprasegmental feature: To capture this periodic structure, we also introduce a suprasegmental feature which has previously been used to capture the tendency of an acoustic signal to repeat the same structure over a longer duration [36]. Firstly, the GCIs are determined from the extracted vocal signal using the zero frequency filtering method [42]. From these GCIs, we can obtain the pitch intervals in the polyphonic song. The correlation between 10 successive pitch periods from the song is computed to obtain an estimate of how repetitive its structure is. The periodicity values are certainly

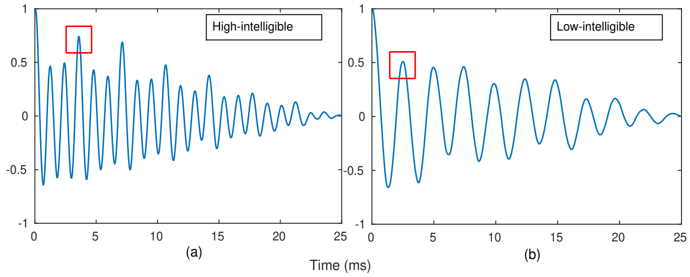

Fig. 5. Normalized autocorrelation plot for a selected portion of singing vocals with background accompaniment corresponding to (a) high-intelligible and (b) low-intelligible excerpts.

lower for low-intelligible excerpts compared to that of high-intelligible excerpts.

3) Jitter and shimmer: Jitter and shimmer are the measures of the cycle-to-cycle variation of F_0 and amplitude, respectively. They have previously been used for the study of voice quality [43]. Jitter is affected mainly because of lack of control of vocal fold vibration, more variation of which leads to unintelligibility in the voice. Shimmer, for its part, is related to the reduction of glottic resistance and mass lesions in the vocal folds, which are in turn correlated with the presence of noise at emission and breathiness [43]. Very high values of jitter and shimmer are an indication of breathy, rough or hoarse voice quality. From our analysis we found, these two prosodic aspects have higher values for low-intelligible excerpts compared to high-intelligible excerpts.

After exploring these features with respect to the intelligibility of a polyphonic song, we combine them into a 12-dimensional feature vector, which is used to derive the song intelligibility score. To check whether these features are correlated to the human-rated intelligibility scores, we average each feature over the entire excerpt. The correlation of the average feature values with intelligibility score for each evidence is shown in Table II. A higher magnitude of correlation value indicates a stronger potential of the feature to quantify intelligibility. The correlation value corresponding to the combined 12-dimensional feature is 0.50, which indicates the efficacy of these features in representing song intelligibility. We observe that suprasegmental feature has the lowest correlation, whereas the slope of peaks of HE of LP residual and modulation spectrum energy show the highest correlations. The negative signs indicate an inverse relationship between the feature and the intelligibility score; for instance, intelligibility decreases as pitch increases. As mentioned in Section II-A, for 23 (12%) excerpts we observe particularly lower correlation of STOI and intelligibility score. The correlation between the proposed vocal-specific features and intelligibility score for these 23 excerpts is 0.54, which shows the complementary aspects of vocal-specific features and STOI. We expect better performance with statistical models using a combination of all the features.

To further observe the redundancy among different vocal-specific features, we have also included the canonical correlation analysis [44] in Table II. We find canonical correlation value of each vocal-specific feature with all other vocal-specific features. The higher canonical correlation value (closer to 1) for a feature represents more similarity with other features. From table II, we observe that although correlation exists between the vocal-specific features there is some extra information captured by each feature, as the correlation value is less than 1 in each case. For instance, the suprasegmental feature has less correlation with human rated intelligibility score, however it has contrasting information compared to all other features.

TABLE II
CORRELATION BETWEEN DIFFERENT VOCAL-SPECIFIC FEATURES AND HUMAN-RATED INTELLIGIBILITY SCORES, AND CANONICAL CORRELATION ANALYSIS OF EACH VOCAL-SPECIFIC FEATURES WITH OTHER VOCAL-SPECIFIC FEATURES, FOR 200 EXCERPTS.

Features	**Correlation value**	**Canonical correlation**
Pitch	-0.22	0.71
Smoothed Hilbert envelope	0.43	0.89
Peak-to-sidelobe ratio	0.18	0.79
Slope of peaks of HE of LP residual	0.47	0.69
Spectral peak energy	0.33	0.89
Modulation spectrum energy	0.46	0.82
Sub-band correlation	0.42	0.79
Spectral slope	0.35	0.85
Normalized ACR peaks	0.18	0.74
Suprasegmental feature	0.09	0.54
Jitter	-0.19	0.53
Shimmer	-0.26	0.66
Combined features	**0.50**	–

IV. PROPOSED FRAMEWORK

Based on the features discussed in the previous sections, we outline the proposed framework to automatically evaluate song intelligibility, as shown in Figure 6. A given input song is first passed through an audio source separation module to extract the singing vocal. Both the polyphonic song and the extracted vocal are passed through the vocal segmentation block, where sections with only instrumental music are identified and removed. As per Section II, we then passes the extracted and segmented singing vocal, as well as the segmented polyphonic song through the singing adapted STOI module, which gives an STOI value for each frame of the song. The 12-dimensional vocal-specific features are also extracted from the audio signal corresponding to the segmented polyphonic song. We combine the STOI values and vocal-specific features to create a 13-dimensional feature vector, which is in turn used to train an SVM based regression model, with the human-rated intelligibility scores treated as reference scores. During testing, we apply the 13-dimensional feature vector to the regression model to predict an intelligibility score for all the frames of a given song. The final intelligibility score for a song is obtained by averaging the scores over all the frames of that song. An implementation of this proposed automatic evaluation of song intelligibility is made available[1] for use.

V. EXPERIMENTAL EVALUATION

In this section we first describe the dataset used for the analysis and experiments. Then we discuss our experimental set-up and results.

[1] https://github.com/bidishasharma/Automatic-Song-Intelligibility

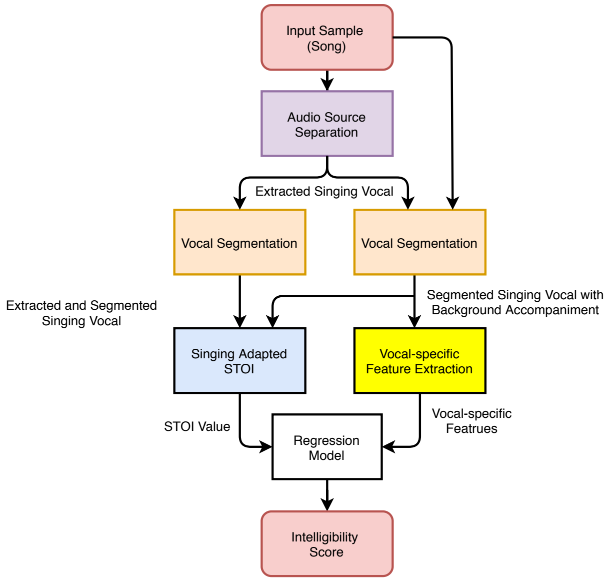

Fig. 6. Framework for automatic evaluation of song intelligibility using singing adapted STOI and vocal-specific features.

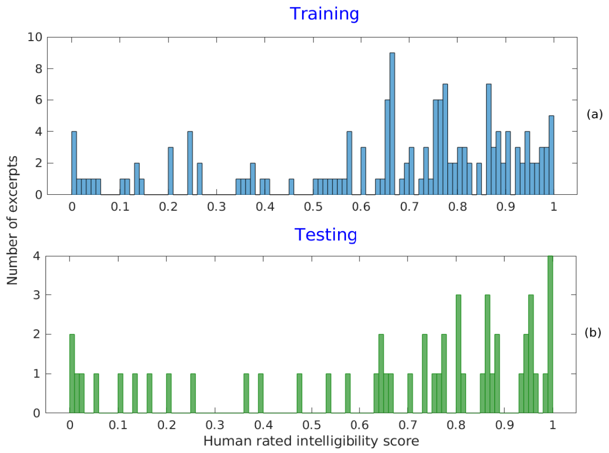

Fig. 7. Histogram plot representing the distribution of human-rated intelligibility scores for (a) train and, (b) test dataset.

A. Dataset

For the analysis and evaluation presented in this work, we use the database of English songs collected in Ibrahim et al. [20]. There are 10 genres in the database: classical, electronic, folk, jazz, metal, pop/rock, punk, rap, reggae, and RnB. To create this database, songs were randomly selected from the Rate Your Music database, with the only restrictions being that songs had to be in English and have few ratings with low popularity scores; this latter requirement was in order to reduce the chances that a user had previously heard a song and thus knew its lyrics from some other context. From each song, two excerpts were selected; these excerpts included a complete utterance and had an average duration of 6.5 seconds. A total of 200 such excerpts were thus created. Each of the 200 excerpts is then transcribed by 17 participants and the transcription was compared with the original lyrics to obtain the word accuracy rate. The average word accuracy rate for each excerpt over 17 participant is the *human-rated intelligibility score* used as the ground-truth measure.

The dataset used has a limited number of excerpts for each genre and an unbalanced distribution over the entire range of human-rated intelligibility score (0-1). As each genre has different acoustic characteristics, to analyze the performance of our algorithm over all 10 genres for both high- and low-intelligible excerpts, we have to include at least some examples of each genre in the test dataset. With this strategy, we divide the dataset of 200 excerpts into training and testing sets. As there are a total of 10 genres which each has 20 excerpts in the dataset, the training set consists of 140 excerpts (14 excerpts per genre) and testing set consists of 60 excerpts (6 excerpts per genre). The histogram representations of training and testing excerpts with respect to intelligibility score are shown in Figure 7. It is evident that both the training and testing sets have an identical distribution over the entire range of intelligibility scores. Similar representations for each genre are depicted in Figure 8, which shows that folk and jazz have the highest human-rated intelligibility scores in this dataset, whereas punk and metal have the lowest.

B. Experimental setup

We perform three types of experiments to validate the proposed framework. Although we are using a regression model to obtain the intelligibility scores, to validate the efficacy of the proposed features for quantifying intelligibility, we also perform 2-class (low- and high-intelligible) and 3-class (low-, moderate- and high-intelligible) classification tasks. To effectively use the limited amount of data, we employ SVM based models for all the above mentioned experiments.

C. Two-class and three-class classification

To develop the 2-class classifier, the excerpts with human-rated intelligibility scores $(<\ 0.5)$ are labeled as *low-intelligible* and (> 0.5) are labeled as *high-intelligible*. In case of the 3-class classifier, excerpts with human-rated intelligibility score (> 0.66), $(< 0.66\ \&\ \ > 0.33)$ and (< 0.33) are labeled as *high-*, *moderate-* and *low-intelligible*, respectively. We develop 6 different types of systems (for each of the 2- and 3-class classifiers) using the features listed below,

- **Acoustic feat**: Acoustic feature vector proposed in Ibrahim et al. (6-dimensional) [20]
- **MFCC**: MFCCs along with its first derivative (34-dimensional)
- **Vocal feat**: Proposed vocal-specific features (12-dimensional)
- **Vocal feat+STOI**: vocal-specific features fused with singing adapted STOI (13-dimensional)
- **Vocal feat+MFCC**: vocal-specific features fused with MFCCs along with its first derivative (46-dimensional)
- **STOI+MFCC**: Singing adapted STOI fused with MFCCs along with its first derivative (35-dimensional)
- **Acoustic feat+MFCC**: Acoustic features used in Ibrahim et al. [20] fused with MFCCs along with its first derivative (40-dimensional)
- **Vocal feat+STOI+MFCC**: Vocal feat+STOI fused with MFCCs (47-dimensional)

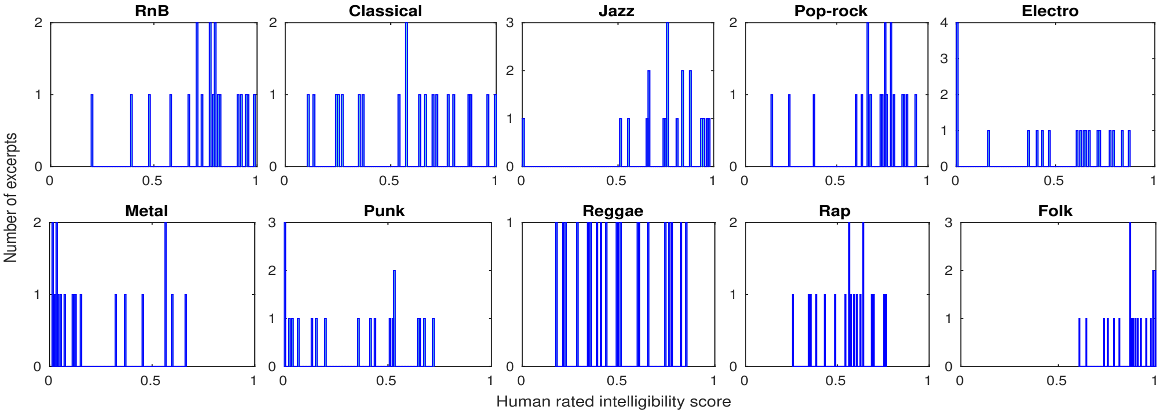

Fig. 8. Histogram plot representing the distribution of human-rated intelligibility scores for different genres.

TABLE III
CLASSIFICATION ACCURACY (%) FOR TWO-CLASS CLASSIFIERS FOR DIFFERENT FEATURE SETS CORRESPONDING TO VARIOUS GENRES.

Genre	**% Accuracy for 2-class classification**							
	Acoustic feat	**MFCC**	**Vocal feat**	**Vocal feat+STOI**	**Vocal feat+MFCC**	**STOI+MFCC**	**Acoustic feat+MFCC**	**Vocal feat+STOI+MFCC**
Feature dimension→	6-dimensional	34-dimensional	12-dimensional	13-dimensional	46-dimensional	35-dimensional	40-dimensional	47-dimensional
All	65.00	81.66	65.51	77.10	83.34	83.33	81.36	**88.13**
Classical	50.00	83.33	83.33	**83.33**	83.33	66.67	83.33	66.67
Electro	66.67	66.67	83.33	83.33	83.33	100.00	66.67	**100.00**
Folk	100.00	100.00	60.00	100.00	100.00	100.00	100.00	**100.00**
Jazz	100.00	100.00	100.00	83.33	100.00	83.33	**100.00**	83.33
Metal	50.00	66.67	50.00	75.00	100.00	100.00	83.33	**100.00**
Pop-rock	50.00	83.33	50.00	50.00	66.67	66.67	83.33	**83.33**
Punk	66.67	83.33	50.00	83.33	66.67	83.33	66.67	**83.33**
Rap	50	83.33	83.33	83.33	66.67	83.33	83.33	**100.00**
Reggae	50.00	66.67	50.00	50.00	66.67	66.67	66.67	**66.67**
RnB	66.67	83.33	40.00	80.00	100.00	83.33	83.33	**100.00**

TABLE IV
CLASSIFICATION ACCURACY (%) FOR THREE-CLASS CLASSIFIERS FOR DIFFERENT FEATURE SETS CORRESPONDING TO VARIOUS GENRES.

Genre	**% Accuracy for 3-class classification**							
	Acoustic feat	**MFCC**	**Vocal feat**	**Vocal feat+STOI**	**Vocal feat+MFCC**	**STOI+MFCC**	**Acoustic feat+MFCC**	**Vocal feat+STOI+MFCC**
Feature dimension→	6-dimensional	34-dimensional	12-dimensional	13-dimensional	46-dimensional	35-dimensional	40-dimensional	47-dimensional
All	50.00	67.21	53.22	54.80	66.84	68.00	54.83	**69.23**
Classical	66.67	**66.67**	33.33	33.33	50.00	66.67	50.00	50.00
Electro	42.86	50.00	42.85	57.14	71.42	57.14	28.57	**85.71**
Folk	83.33	100.00	100.00	100.00	83.33	83.33	100.00	**100.00**
Jazz	60.00	80.00	80.00	80.00	100.00	80.00	80.00	**80.00**
Metal	75.00	**100.00**	87.50	87.50	87.50	100.00	75.00	87.50
Pop-rock	0.00	16.67	33.33	**33.33**	16.67	16.67	16.67	33.33
Punk	66.67	83.33	50.00	50.00	83.33	100.00	50.00	**83.33**
Rap	33.33	**50.00**	16.67	16.67	33.33	33.33	33.33	33.33
Reggae	28.57	28.57	14.28	14.28	42.86	42.86	28.57	**40.00**
RnB	40.00	100.00	80.00	80.00	100.00	100.00	100.00	**100.00**

All the features mentioned above are derived from the excerpts using a sampling frequency of 16 kHz, frame-size of 25 ms and a frame-shift of 5 ms. To match the dimension during fusion with other features, we obtain the STOI values for each frame of an excerpt by removing the averaging over all frames as shown in (2). Each dimension of the feature vector is normalized to obtain a minimum value of 0 and a maximum of 1. We use SVM models with a radial basis function (RBF) kernel, where the values of c and γ are set using 5-fold cross validation over the entire training dataset. During testing, the same features are extracted for each of the test excerpts and normalized using the minimum and maximum values obtained during training. The frame level normalized features are fed to the classifier, which in turn provides predicted class information for each frame. We perform majority voting over the frames of an excerpt to derive it's predicted class. The overall classification accuracy along with the genre specific classification accuracies for the 2-class and 3-class classifiers are depicted in Table III and Table IV respectively.

For the 2-class classifier, the average classification accuracy using proposed vocal-specific features along with singing adapted STOI is 77.10%, and the accuracy in its weakest genres, reggae and pop-rock is 50% (Table III). From the genre-wise distribution of the intelligibility score shown in Figure 8, we can observe that the reggae genre has many excerpts with an intelligibility score near the threshold of 0.5; these excerpts fit almost equally well in either category and are thus likely to be misclassified. In all the classifiers we achieve good accuracy for the folk genre, where most of the excerpts are in the high-intelligible class. We also achieve excellent accuracy (83.33%) for the classical excerpts, which are evenly distributed between the high- and low-intelligible classes. This shows that the proposed features (Vocal feat+STOI) work reasonably well for this task. The addition of MFCC features provides an absolute improvement of 11% in accuracy, and the accuracy of pop-rock genre improves the most. This validates the efficacy of the proposed vocal-specific features and singing adapted STOI to represent song intelligibility.

In the 3-class classification task shown in Table IV, the accuracy obtained using proposed features is 54.80%, which improves to 69.23% after fusion with MFCC. According to the distribution of training and testing data shown in Figure 7, many excerpts have intelligibility scores very close to 0.66, which is the threshold between the high- and moderate-intelligible classes. From our investigation, we found that most of the wrongly classified instances have an intelligibility score value near this threshold. The accuracy obtained using MFCCs alone is 67.21%. Similar to the 2-class classification task, the lowest accuracy using the proposed features is for the reggae genre and highest for folk. In this case, the performance of the classical genre is not favorable. We observe improvement in performance of both the classifiers when each of the vocal-specific features and STOI is individually combined with MFCC. This also validates the presence of useful information in the proposed features to represent intelligibility.

D. Regression model

In this work, we are particularly fascinated to automatically obtain an intelligibility score for a song that correlates with human perception. To obtain the intelligibility score corresponding to each song, we use SVM based regression model. We develop different models using the different combination of features as listed for the classification task; in total, 6 different types of systems are developed and compared.

The features are processed in an identical manner as discussed in the classification task. Similarly, we use the RBF kernel for the regression model, where c and γ values are set using 5-fold cross validation for the entire training database. While testing, each dimension of the feature vector is normalized with respect to minimum and maximum values obtained from the training feature set. The predicted intelligibility scores corresponding to all the frames of an excerpt are averaged to derive a single intelligibility score for each excerpt.

The correlation and mean absolute error (MAE) values between the intelligibility scores obtained from the regression model and human-rated intelligibility scores are measures of the efficacy of the proposed method. These correlation and MAE values are depicted in Table V for different sets of features. Using only MFCCs and delta MFCCs (34-dimensional), we obtain a correlation of 0.78. For the proposed feature set (Vocal feat+STOI), the average correlation value is 0.75. For metal, this correlation is the lowest (0.19) and it is the highest for jazz (0.87). After appending the proposed features, with MFCCs we achieve an average accuracy of 0.81 over all the genres. The MAE using only MFCCs is 0.15, which improves after appending vocal-specific features (0.14) and STOI (0.13). The MAE drops down to 0.10 using the proposed feature set with MFCCs. We achieve the highest correlation of 0.95 for the RnB genre in this case. The MAE is 0.10 using the proposed vocal-specific features and singing adapted STOI in combination with MFCCs, which is 0.15 using only MFCCs.

The scatterplots in Figure 9 show the derived intelligibility scores for different feature sets with respect to the human-rated intelligibility scores for the entire test set. We can observe that the points converge towards a linear line as we append the proposed features with MFCCs. The genre-specific correlation values for different regression models from Table V can be compared in Figure 10. These experiments establish the proposed framework for automatic evaluation of song intelligibility.

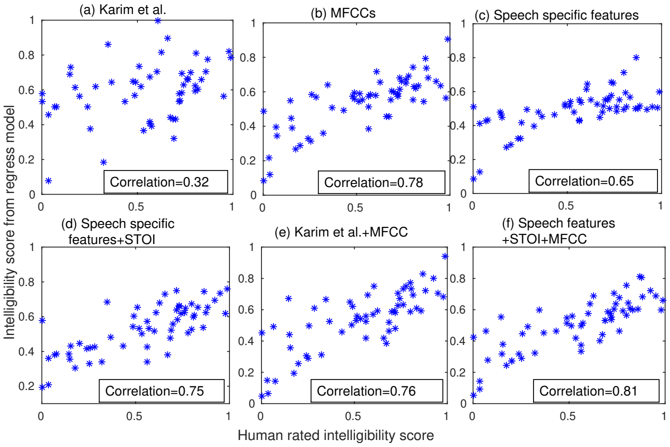

Fig. 9. Scatter plot representing correlation between intelligibility scores obtained from regression models and human-rated intelligibility scores, for different feature sets

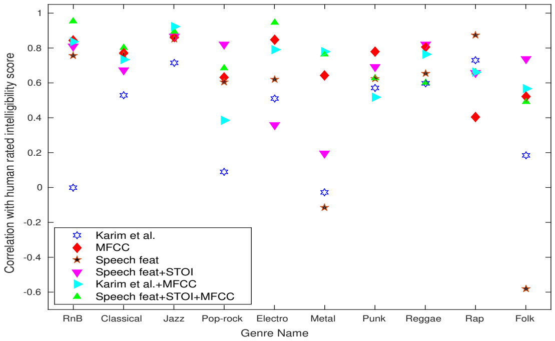

Fig. 10. Comparison of the correlation values from Table V, for different feature sets corresponding to various genres.

E. Genre independent training

In the previous experiments, we include excerpts from all the genres in both the training and testing data, which do not demonstrate whether the model is genre-independent, or equivalently, whether the model will work accurately on a genre that is not represented in the training set. Therefore, we perform another experiment in which all the excerpts corresponding to 9 genres are used to train the SVM based regression model, and the remaining genre is used for testing. Effectively, this results in 10 regression models, each of them to be tested against the one genre excluded from training that particular model. We use the 47-dimensional feature vector (*Vocal feat+STOI+MFCC*). For each genre, the correlation value obtained between the predicted and human-rated intelligibility score is depicted in the bar plot shown in Figure 11. We observe that the correlation values drop relatively steeply for the folk and punk genre when they are not included in the training data. However, other genres have only a relatively small drop in correlation value when compared to the correlations in Table V. This shows that the proposed

TABLE V
THE CORRELATION (CORR) AND MEAN ABSOLUTE ERROR (MAE) BETWEEN THE INTELLIGIBILITY SCORES OBTAINED FROM REGRESSION MODELS AND HUMAN RATED INTELLIGIBILITY SCORES, FOR DIFFERENT FEATURE SETS CORRESPONDING TO VARIOUS GENRES.

Genre↓	**Acoustic feat**		**MFCC**		**Vocal feat**		**Vocal feat+STOI**		**Vocal feat+MFCC**		**STOI+MFCC**		**Acoustic feat+MFCC**		**Vocal feat+STOI+MFCC**	
Feature dimension→	6-dimensional		34-dimensional		12-dimensional		13-dimensional		46-dimensional		35-dimensional		40-dimensional		47-dimensional	
Evaluation parameter→	Corr	MAE	Corr	MAE	Corr	MAE	Corr	MAE	Corr	MAE	Corr	MAE	Corr	MAE	Corr	MAE
All	0.32	0.33	0.78	0.15	0.65	0.19	0.75	0.17	0.78	0.14	0.79	0.13	0.76	0.14	**0.81**	0.10
Classical	0.53	0.96	0.77	0.13	0.76	0.14	0.67	0.15	0.78	0.13	0.80	0.13	0.73	0.14	**0.80**	0.11
Electro	0.50	0.22	0.85	0.18	0.67	0.22	0.36	0.20	0.86	0.09	0.88	0.12	0.79	0.17	**0.94**	0.08
Folk	0.18	0.36	0.52	0.14	-0.58	0.28	**0.73**	0.16	0.52	0.14	0.50	0.14	0.56	0.13	0.49	0.13
Jazz	0.71	0.19	0.86	0.17	0.85	0.17	0.87	0.16	0.91	0.12	0.88	0.14	**0.92**	0.13	0.89	0.09
Metal	-0.03	0.30	0.64	0.17	-0.11	0.23	0.19	0.20	0.58	0.15	0.62	0.19	**0.78**	0.11	0.77	0.10
Pop-rock	0.09	0.25	0.63	0.19	0.60	0.22	**0.82**	0.20	0.67	0.19	0.64	0.13	0.38	0.21	**0.68**	0.12
Punk	**0.78**	0.23	0.57	0.17	0.62	0.21	0.69	0.20	0.60	0.17	0.58	0.17	0.52	0.20	0.63	0.15
Rap	0.73	0.28	0.41	0.10	**0.87**	0.10	0.66	0.10	0.52	0.10	0.60	0.10	0.66	0.10	0.66	0.10
Reggae	0.60	0.23	0.80	0.10	0.65	0.11	**0.82**	0.13	0.62	0.10	0.58	0.10	0.77	0.09	0.60	0.11
RnB	0.00	0.24	0.84	0.17	0.76	0.24	0.73	0.20	0.83	0.17	0.86	0.12	0.83	0.14	**0.95**	0.08

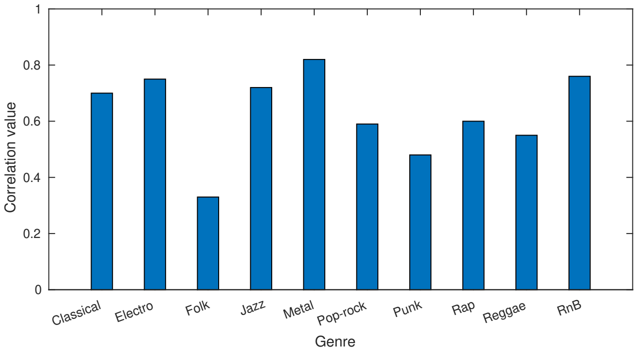

Fig. 11. Correlation of the intelligibility score for different genres obtained from the regression models trained using all other genres.

features also work relatively well independent of the genres used in training the model.

F. Comparison of STOI and vocal-specific features

To test whether the proposed singing adapted STOI and the vocal-specific features carry complementary information, we develop another SVM based regression model in a similar manner as described in Section V-D, but using only the singing adapted STOI values. We obtain a correlation value of 0.41 between the predicted and human-rated intelligibility scores. We observe that for a particular number of excerpts this correlation is relatively low. Therefore, we divided the testing excerpts into two groups, A and B, based on whether singing adapted STOI works reasonably well for the excerpt or not. As shown in Table VI, the correlation value for the excerpts corresponding to group A is 0.58 and that of group B is 0.25. It is evident that the intelligibility of the excerpts corresponding to group A depends on the interference caused by the background accompaniment, whereas that of group B depends on vocal-specific characteristics. We also observe that the model obtained using only vocal-specific features performs in a reverse way for the two groups of excerpts. However, after combining both STOI and vocal-specific features, the performance for both the groups is enhanced. This experiment supports that both the evidence are complementary to each other, and both are necessary for calculating intelligibility.

TABLE VI
COMPARISON OF STOI AND VOCAL-SPECIFIC FEATURES.

Group	STOI	Vocal feat	Vocal feat+STOI
A	0.58	0.61	0.79
B	0.25	0.69	0.71

G. Features derived from the extracted vocal

We also consider the vocal-specific features obtained from the extracted vocal instead of the polyphonic song. However, based on our experiments, we found that the correlation of the combined vocal-specific features derived from the extracted singing vocal with human-rated intelligibility score is 0.39, while it is 0.50 for the same features extracted from the polyphonic song. This may be due to some distortion introduced in the audio source separation method. In this case, the regression model also shows a poor correlation (0.60) with the human-rated intelligibility score. We also use MFCCs derived from extracted vocal and develop a regression model, that results in a correlation of 0.60 (Table VII). By combining all the features obtained from the extracted vocal, we obtained a correlation of 0.66. Whereas, MFCCs from the polyphonic song and vocal-specific features from the extracted vocal, gives a correlation of 0.73. These results are weaker than the ones in which the features are calculated from the polyphonic song. More prominent characteristics of the vocal in presence of the background accompaniment leads to better intelligibility.

TABLE VII
CORRELATION OF INTELLIGIBILITY SCORES OBTAINED FROM REGRESSION MODELS TRAINED USING THE FEATURES DERIVED FROM THE EXTRACTED SINGING VOCALS.

Regression correlation			**Classification accuracy(%)**	
Vocal feat +STOI	MFCCs	Vocal feat+ STOI+MFCCs (47-dim)	Two-class 47-dim	Three-class 47-dim
0.60	0.60	0.66	72.41	54.83

VI. SUMMARY

In this work, we focus on developing a framework to automatically estimate intelligibility of polyphonic song. Our proposed strategy uses acoustic cues extracted from the song to derive its intelligibility, which include singing adapted STOI and vocal-specific features. The proposed features are validated with 2-class and 3-class intelligibility classification tasks. Finally, to derive the intelligibility score against each excerpt of a song, we used regression models trained using the proposed features. The 2-class classification accuracy for the proposed feature set is 88.13%, which is 69.23% for the 3-class classification. We use correlation and MAE measures between intelligibility score obtained from the regression model and human-rated intelligibility score, to establish the efficacy of the proposed method. This correlation is 0.81 and MAE is 0.10 for the proposed framework. Among 10 different genres, our method achieve the best performance for Electro and RnB.

We also analyze the performance of the proposed framework, by excluding one genre from training. We find that the proposed method is genre independent to a certain extent. We also explore that the singing adapted STOI and vocal-specific features perform in a complementary way to each other. Although we use extracted singing vocal in the singing adapted STOI, we could not achieve good performance with the vocal-specific features derived from the extracted singing vocal. Rather, we observed degraded performance using the features derived from the extracted singing vocals.

The proposed framework is uncomplicated and oriented towards the application of song recommendation for language learning. It is true that the SVM-based statistical model does retain the limitations of having a small dataset and possibility of train-test data mismatch. To get around this drawback, a larger database would be of use. Fortunately, the proposed idea can be easily replicated on a new larger dataset. Moreover, we also derive an intelligibility score without any statistical model and using only average of raw vocal-specific features. By combining the proposed vocal-specific features, we achieve a correlation of 0.50 as shown in Table II. In combination with the singing adapted STOI, a correlation of 0.59 is obtained. This demonstrates the efficacy of the proposed feature set to represent song intelligibility.

In the future, using the proposed features, we would like to derive an intelligibility measure without using any statistical models, evaluated on a larger dataset.

VII. Acknowledgements

This research is supported in part by an unrestricted gift from Smule Inc.

References

[1] D. Schön, M. Boyer, S. Moreno, M. Besson, I. Peretz, and R. Kolinsky, "Songs as an aid for language acquisition," *Cognition*, vol. 106, no. 2, pp. 975–983, 2008.

[2] M. Schwantes, "The use of music therapy with children who speak english as a second language: An exploratory study," *Music Therapy Perspectives*, vol. 27, no. 2, pp. 80–87, 2009.

[3] D. Fisher, "Early language learning with and without music. reading horizons," *Reading Horizons*, vol. 42, no. 1, pp. 40–49, 2001.

[4] A. Tierney and N. Kraus, "Music training for the development of reading skills," vol. 207, pp. 209–241, 2013.

[5] A. D. Patel, "Language, music, syntax and the brain," *Nature neuroscience*, vol. 6, no. 7, pp. 674–681, 2003.

[6] V. L. Trollinger, "The brain in singing and language," *General Music Today*, vol. 23, no. 2, pp. 20–23, 2010.

[7] A. Kultti, "Singing as language learning activity in multilingual toddler groups in preschool," *Early Child Development and Care*, vol. 183, no. 12, pp. 1955–1969, 2013.

[8] C. F. Mora, "Foreign language acquisition and melody singing," *ELT journal*, vol. 54, no. 2, pp. 146–152, 2000.

[9] T.-a. Kao and R. L. Oxford, "Learning language through music: A strategy for building inspiration and motivation," *System*, vol. 43, pp. 114–120, 2014.

[10] K. Mori and M. Iwanaga, "Pleasure generated by sadness: Effect of sad lyrics on the emotions induced by happy music," *Psychology of Music*, vol. 42, no. 5, pp. 643–652, 2014.

[11] L. A. Smith and B. L. Scott, "Increasing the intelligibility of sung vowels," *The Journal of the Acoustical Society of America*, vol. 67, no. 5, pp. 1795–1797, 1980.

[12] M. S. Benolken and C. E. Swanson, "The effect of pitch-related changes on the perception of sung vowels," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1781–1785, 1990.

[13] L. B. Collister and D. Huron, "Comparison of word intelligibility in spoken and sung phrases," *Empirical Musicology Review*, vol. 3, no. 3, pp. 109–125, 2008.

[14] J. W. Gregg and R. C. Scherer, "Vowel intelligibility in classical singing," *Journal of Voice*, vol. 20, no. 2, pp. 198–210, 2006.

[15] H. Hollien, A. P. Mendes-Schwartz, and K. Nielsen, "Perceptual confusions of high-pitched sung vowels," *Journal of Voice*, vol. 14, no. 2, pp. 287–298, 2000.

[16] R. B. Johnson, D. Huron, and L. Collister, "Music and lyrics interactions and their influence on recognition of sung words: an investigation of word frequency, rhyme, metric stress, vocal timbre, melisma, and repetition priming," *Empirical Musicology Review*, vol. 9, no. 1, pp. 2–20, 2013.

[17] P. Fine and J. Ginsborg, "Perceived factors affecting the intelligibility of sung text," in *Proceedings of the Third Conference on Interdisciplinary Musicology (CIM07)*, 2007, pp. 15–19.

[18] N. Condit-Schultz and D. Huron, "Catching the lyrics: intelligibility in twelve song genres," *Music Perception: An Interdisciplinary Journal*, vol. 32, no. 5, pp. 470–483, 2015.

[19] J. Ginsborg, "The influence of interactions between music and lyrics: what factors underlie the intelligibility of sung text?" *Empirical Musicology Review*, vol. 9, no. 1, pp. 21–24, 2013.

[20] K. M. Ibrahim, D. Grunberg, K. Agres, C. Gupta, and Y. Wang, "Intelligibility of sung lyrics: A pilot study," in *ISMIR*, 2017, pp. 686–693.

[21] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[23] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *18th International Society for Music Information Retrieval Conference*, 2017, pp. 745–751.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[25] "A chainer implementation of u-net singing voice separation model," https://github.com/Xiao-Ming/UNet-VocalSeparation-Chainer, [Online; accessed 28-October-2018].

[26] D. Z. Borch and J. Sundberg, "Spectral distribution of solo voice and accompaniment in pop music," *Logopedics Phoniatrics Vocology*, vol. 27, no. 1, pp. 37–41, 2002.

[27] S. O. Ternstrom, "Hi-fi voice: observations on the distribution of energy in the singing voice spectrum above 5 khz," *Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3379–3379, 2008.

[28] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.

[29] N. S. Di Carlo and A. Germain, "A perceptual study of the influence of pitch on the intelligibility of sung vowels," *Phonetica*, vol. 42, no. 4, pp. 188–197, 1985.

[30] J. Sundberg and T. D. Rossing, "The science of singing voice," *The Journal of the Acoustical Society of America*, vol. 87, no. 1, pp. 462–463, 1990.

[31] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[32] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[33] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Transactions on Acoustics, Speech and Signal Processing,*, vol. 27, no. 4, pp. 309–319, 1979.

[34] S. M. Prasanna, B. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 556–565, 2009.

[35] G. Seshadri and B. Yegnanarayana, “Perceived loudness of speech based on the characteristics of glottal excitation source,” *The Journal of the Acoustical Society of America*, vol. 126, no. 4, pp. 2061–2071, 2009.

[36] B. Sharma and S. Mahadeva Prasanna, “Sonority measurement using system, source, and suprasegmental information,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 3, pp. 505–518, 2017.

[37] S. Greenberg and B. Kingsbury, “The modulation spectrogram: In pursuit of an invariant representation of speech,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings*. IEEE, 1997, p. 1647.

[38] H. Dudley, “Remaking speech,” *The Journal of the Acoustical Society of America*, vol. 11, no. 2, pp. 169–177, 1939.

[39] D. Wang and S. S. Narayanan, “Robust speech rate estimation for spontaneous speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2190–2201, 2007.

[40] V. C. Tartter, H. Gomes, and E. Litwin, “Some acoustic effects of listening to noise on speech production,” *The Journal of the Acoustical Society of America*, vol. 94, no. 4, pp. 2437–2440, 1993.

[41] B. Sharma and S. R. M. Prasanna, “Enhancement of spectral tilt in synthesized speech,” *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 382–386, 2017.

[42] K. Murty and B. Yegnanarayana, “Epoch extraction from speech signals,” *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 16, no. 8, pp. 1602–1613, 2008.

[43] H. F. Wertzner, S. Schreiber, and L. Amaro, “Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders,” *Brazilian journal of otorhinolaryngology*, vol. 71, no. 5, pp. 582–588, 2005.

[44] W. Krzanowski, *Principles of multivariate analysis*. OUP Oxford, 2000, vol. 23.

Bidisha Sharma is a Postdoctoral Research Fellow in the Electrical and Computer Engineering Department at the National University of Singapore (NUS). She received Ph.D. degree from Indian Institute of Technology (IIT) Guwahati in India in 2018, B.E. degree in Electronics and Telecommunication Engineering from Girijananda Chowdhury Institute of Management and Technology, Gauhati University, Guwahati, India, in 2012. Her research interests are in speech signal processing, speech synthesis, automatic speech recognition and singing voice analysis.

Ye Wang is an Associate Professor in the Computer Science Department at the National University of Singapore (NUS) and NUS Graduate School for Integrative Sciences and Engineering (NGS). He received his Ph.D. degree from Tampere University of Technology in Finland in 2002, M.Sc. degree from Braunschweig University of Technology in Germany in 1993, and B.Sc. degree from South China University of Technology in China in 1983. He established and directed the sound and music computing (SMC) Lab (www.smcnus.org). Before joining NUS, he was a member of the technical staff at Nokia Research Center in Tampere, Finland for 9 years. His research interests include sound analysis and music information retrieval (MIR), mobile computing, and cloud computing, and their applications in music edutainment and e-Health, as well as determining their effectiveness via subjective and objective evaluations. His most recent projects involve the design and evaluation of systems to support 1) therapeutic gait training using Rhythmic Auditory Stimulation (RAS), 2) second language learning, and 3) motivating exercise via music-based systems.