END-TO-END AUTOMATIC SINGING SKILL EVALUATION USING CROSS-ATTENTION AND DATA AUGMENTATION FOR SOLO SINGING AND SINGING WITH ACCOMPANIMENT

Yaolong Ju **Chun Yat Wu** **Betty Cortiñas Lorenzo**
Jing Yang **Jiajun Deng** **Fan Fan** **Simon Lui**
Huawei Technologies Co., Ltd., China
{yaolongju, wu.chun.yat, cortinas.lorenzo.betty
yangjing201, deng.jiajun, fanfan1, luisiuhang}@huawei.com

ABSTRACT

Automatic singing skill evaluation (ASSE) systems are predominantly designed for solo singing, and the scenario of singing with accompaniment is largely unaddressed. In this paper, we propose an end-to-end ASSE system that effectively processes both solo singing and singing with accompaniment using data augmentation, where a comparative study is conducted on four different data augmentation approaches. Additionally, we incorporate bi-directional cross-attention (BiCA) for feature fusion which, compared to simple concatenation, can better exploit the inter-relationships between different features. Results on the 10KSinging dataset show that data augmentation and BiCA boost performance individually. When combined, they contribute to further improvements, with a Pearson correlation coefficient of 0.769 for solo singing and 0.709 for singing with accompaniment. This represents relative improvements of 36.8% and 26.2% compared to the baseline model score of 0.562, respectively.

1. INTRODUCTION

In recent years, the widespread use of digital media has changed the way users interact with music, giving rise to new applications like streaming services and online karaoke platforms [1, 2]. As numerous singing content is published daily by these applications, it becomes very expensive and practically unscalable to retrieve high-quality content manually. One such scenario is the discovery of vocal talent in the vast online platforms, where automatic singing skill evaluation (ASSE) systems can be used to examine and rate all the singing content, so that the top-tier can be distributed for more views, subscribers, and ultimately more profits.

Despite the potential commercial values, ASSE is a difficult task that encompasses both subjective preferences and multi-dimensional objective features (e.g., intonation accuracy, rhythm accuracy, range, and dynamics) that professional judges also consider when evaluating vocal performances [3]. Over the years, different ASSE systems have been proposed. Depending on whether a reference melody is taken as the ground truth, these ASSE systems can be classified as reference-dependent [4–9] or reference-independent approaches [10–18]. Recent research on ASSE has been mainly focused on reference-independent deep learning-based approaches, where CNN-based architectures are often used to extract useful patterns from input spectrograms [11, 14, 15, 17, 18]. Other features including pitch histograms [11,14,15,17] and singer timbre embeddings [17, 18] are also used, and these features are usually fused via concatenation. Although this is a simple way of feature fusion, the more advanced techniques that could uncover deeper relationships between these features are still unexplored in ASSE.

Another limitation of the current ASSE research stems from the lack of open-source datasets and high-quality annotations. For example, among the three recent datasets: neither the Smule DAMP dataset [14] nor the YJ-16K dataset [18] is open-sourced, and although Lyra-SA [19] is available after filling out an application form [1], the authors claimed that singing skill annotations are still immature and therefore not sufficiently curated for research purposes yet. Other ASSE datasets including self-made recordings [9,20] or collections from singing platforms [7,12] are also non-public. The lack of publicly available datasets is one of the major impediments that significantly hinder the advancement of ASSE research.

Finally, most ASSE systems require solo singing as input, leaving the scenario of singing with accompaniment largely unexplored [7–9, 11, 12, 14, 15]. On the other hand, [17] proposed an ASSE system that can process singing with accompaniment, but it is achieved by employing a singing voice separation tool [21] as a pre-processing step to remove the accompaniment, which not only results in a more complicated and computationally expensive system, also the model input is still essentially solo

Attribution: Y. Ju, C. Y. Wu, B. C. Lorenzo, J. Yang, J. Deng, F. Fan, and S. Lui, "End-to-end automatic singing skill evaluation using cross-attention and data augmentation for solo singing and singing with accompaniment", in *Proc. of the 25th Int. Society for Music Information Retrieval Conf.,* San Francisco, United States, 2024.

[1] Available at: https://lyracobar.y.qq.com/singvoicedataset.html

singing. In this paper, we propose a new ASSE system capable of processing both solo singing and singing with accompaniment in an end-to-end manner, thereby eliminating the need for a singing voice separation tool. This is achieved through data augmentation during training, where we present the same singing clip in three distinct versions: solo singing, singing with its original accompaniment, and singing remixed with a different accompaniment. For the remixed version, we explore and compare four different approaches, which will be detailed in Section 2.2.2. Furthermore, we explore feature fusion techniques beyond simple concatenation, since such methods can better integrate and amalgamate diverse data sources with greater efficacy [22, 23]. In particular, we adopt a Bi-directional Cross-Attention (BiCA) mechanism during feature fusion, given its effectiveness in capturing the reciprocal knowledge exchange between the source and target features in both directions [24]. The contributions of this paper are:

- We propose an ASSE system that processes both solo singing and singing with accompaniment. The system functions in an efficient end-to-end manner, thereby eliminating the need for a singing voice separation tool required by the baseline model [17].
- We adopt a BiCA mechanism during feature fusion, which better exploits the inter-relationships between different features and facilitates their reciprocal knowledge exchange, compared to simple feature concatenation (see Section 2.2.1).
- We explore data augmentation for ASSE and compare four different approaches: the existing Shuffle-And-Remix [25], the proposed Same-Song Remix, the proposed Key-Match Remix, and All Remix that combines the data augmented from the above three methods (see Section 2.2.2).
- Results show that BiCA and data augmentation boost performance individually (see Section 3.4). The combination of both results in further improvements, with a Pearson correlation coefficient of 0.769 for solo singing and 0.709 for singing with accompaniment on the 10KSinging dataset. This represents relative improvements of 36.8% and 26.2% compared to the baseline score of 0.562 [17], respectively.

2. METHODOLOGY

2.1 The Baseline Model

In the ASSE literature, singing skills can be presented as a ranking [11], a category [8, 9, 12, 18] (e.g., awesome, mediocre, or inferior), or a numerical score [14, 15, 17, 20] (e.g., 60 out of 100). We consider numerical scores for singing skills since they can be mapped into discrete categories or sorted as a ranking, which can be used in different scenarios. Within this range, the existing literature on ASSE is quite limited, and we consider [17] the baseline model for our study, due to its superior performances to the recent ASSE system [14].

The pipeline of the baseline model is shown in Fig. 1(a): it begins by extracting solo singing from the input using an existing singing voice separation tool [21], then the Constant-Q Transform (CQT) is computed and processed by a Convolutional Recurrent Neural Network (CRNN) with an attention mechanism. Following this, the 200-dimensional output from the CRNN and attention is fused with the 120-dimensional pitch histogram [2] and the 512-dimensional X-vector [3] using concatenation. Finally, the combined features are subsequently fed into a streamlined pair of dense layers to output the predicted singing rating. [4]

Compared to [14], three improvements were made in [17]: (1) the attention mechanism was added to the CRNN structure to further explore the useful, long-term relationships in the feature space; (2) X-vector [26] was added as additional features to depict the singing voice timbre, representing the control, resonance, and power that can be essential in singing skill evaluations; (3) the network structure was also finetuned to accommodate the first two improvements, where an extra dense layer was added to optimize the performance. Furthermore, they presented the 10KSinging dataset, which includes the singing skill ratings for 9,756 songs from 93 Chinese male singers and 97 Chinese female singers, and it was further divided into training, validation, and testing sets with 8,000, 756, and 1,000 songs, respectively. Each song from 10KSinging has two versions: the original singing with accompaniment version and the solo singing version, where the accompaniment of the latter was removed using a singing voice separation tool [21]. They used both versions to train their proposed ASSE model and found the solo singing version achieved better performances. Therefore, they considered singing voice separation an integral part of the pipeline, extracting solo singing as the input to their ASSE model shown in Fig. 1(a). As a result, a 62.4% relative improvement was achieved on Pearson correlation coefficient compared to [14] (0.562 VS. 0.346) on the 10KSinging dataset, serving as a solid baseline in this paper.

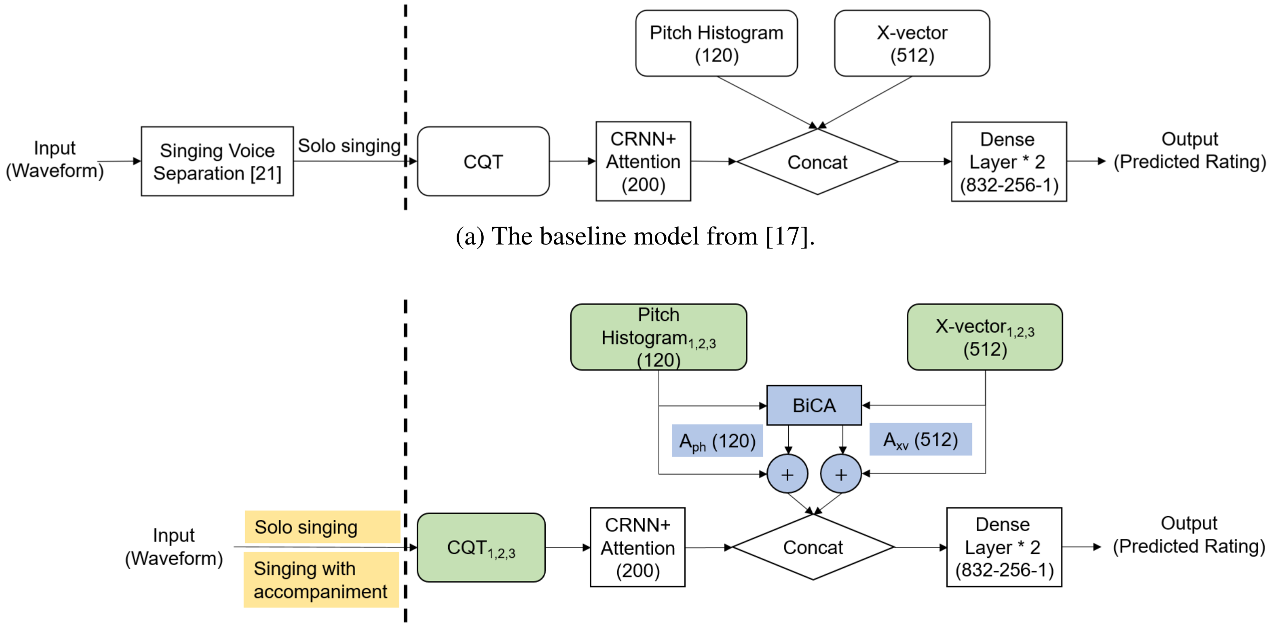
2.2 The proposed Model

Our proposed ASSE model is shown in Fig. 1(b), which highlights the three improvements compared to the baseline model Fig. 1(a) in different colours. The yellow part indicates that the proposed model can process both solo singing and singing with accompaniment, while the baseline model requires singing voice separation to extract solo singing from the input; the blue part represents the use of bi-directional cross-attention between the pitch histogram

[2] Originally proposed in [14], pitch histogram is a global representation of pitch distribution for music, where all octave-equivalent pitches are folded, resulting a range of 12 pitch classes. The distance between two adjacent pitch classes is represented with 10 bins.

[3] According to [26], X-vector distinguishes different kinds of voice timbre. It has been applied to areas including speaker/emotion recognition [27], and singer identification [28].

[4] The reader can refer to [14] for more specifics regarding the CRNN structure and the pitch histogram generation, and [17] for details on the attention mechanism integrated into the CRNN and X-vector extraction.



(a) The baseline model from [17].

(b) The proposed model, with three key distinctions compared to (a) highlighted in different colours.

Figure 1. Illustration of the baseline model (a) and the proposed model (b). The diagram on the right of the dashed line indicates the architecture of the two ASSE systems, where (a) requires singing voice separation as pre-processing [21] to extract solo singing while (b) can process both solo singing and singing with accompaniment (yellow) in a more efficient, end-to-end manner. The proposed model features Bi-directional Cross-Attention (BiCA, blue) and data augmentation (green) using the three distinct training sets of 10KSinging, namely: subindex 1 as singing with accompaniment, subindex 2 as solo singing, and subindex 3 as singing and accompaniment remix discussed in Section 2.2.2. The number in parentheses represents the number of dimensions, while A_{ph} and A_{xv} denote the attention output for the pitch histogram and X-vector, respectively. Both A_{ph} and A_{xv} enter a sum operation with their respective input feature via residual connections.

and X-vector features (see Section 2.2.1); the green part represents data augmentation, where three distinct sets of 10KSinging: singing with accompaniment, solo singing, and singing and accompaniment remix are used during the training process (see Section 2.2.2).

2.2.1 *Bi-directional Cross-Attention*

As discussed above, the baseline model [17] includes a self-attention mechanism in CRNN to capture the long-term relationships from the input CQT spectral representation. This is based on the scaled dot-product attention layer proposed by Vaswani et al [29]:

$$\begin{aligned} Attn(Q,K,V) &= softmax\left(\frac{Q \times K^T}{\sqrt{D}}\right) \times V \\ &= softmax(S) \times V, \end{aligned}$$

where Q, K, V, S, and D denote the query, key, value, similarity matrix, and dimension of the attention layer, respectively.

In addition to the self-attention mechanism adopted by the baseline model, we further improve our approach by applying cross-attention to the remaining two features: pitch histogram and X-vector, since there can be correlations between singers' pitch accuracy and timbre quality that are beneficial for ASSE. In the cross-attention mechanism, the query Q_t is derived from the target t, with the key K_s and the value V_s derived from the source s. The attention output $Attn_t(Q_t, K_s, V_s)$ is then added to the target t via a residual connection, leaving the source s unmodified. This means if we aim to apply cross-attention to both pitch histogram and X-vector features, we need to do it twice: one using pitch histogram as target (t), X-vector as source (s) and vice versa for the other one. To reduce the excessive computational demands in this case, we adopted Bi-directional Cross-Attention (BiCA) [24] that contains a reciprocal attention mechanism, where a shared query-key (QK) matrix [30] is applied to update both the target t and the source s in parallel. Concretely, the similarity matrices of S_t and S_s in BiCA can be calculated as:

$$S_t = \frac{(QK)_t \times (QK)_s^T}{\sqrt{D}} = S_s^T,$$

where $(QK)_t$ and $(QK)_s$ are the shared query-key matrices projected from t and s, respectively. As a result, the attention features of t and s can be respectively obtained by multiplying the corresponding similarity matrix with the value matrix projected from both t and s:

$$Attn_t = softmax(S_t) \times V_s; \; Attn_s = softmax(S_s) \times V_t$$

Finally, we perform a residual connection in both t and s to add the corresponding attention features $Attn_t$ and $Attn_s$. Overall, we enhance the learning of inter-relationships between pitch histogram and X-vector by implementing the cross-attention mechanism, specifically BiCA[5] to our approach illustrated in Fig. 1(b). This way, our proposed model is capable of maintaining the effectiveness of cross-attention while being computationally efficient [24].

2.2.2 *Data Augmentation for ASSE*

As discussed in Section 1, the lack of data can hinder the research and development of ASSE models, and we aim to mitigate this problem by adopting data augmentation. For this purpose, we use the 10KSinging dataset from [17], which contains 9,756 songs in two versions: singing with accompaniment and solo singing. It is further divided into training, validation, and testing sets with 8,000, 756, and 1,000 songs, respectively. We combine the two versions (singing with accompaniment and solo singing) of the training sets and develop a third one called "singing and accompaniment remix", with an additional 8,000 songs generated by remixing the solo singing with a different accompaniment to create more data. For this purpose, we compare three different remixing approaches:

- **Shuffle-And-Remix** [25]: this existing approach remixes each solo singing with a randomly selected accompaniment from another song. Note that with this approach, the singing and accompaniment may not be in the same musical key, and combining the two will introduce differences in musical key irrelevant to ASSE and may interfere with the training process. Therefore, we propose two new remixing techniques that ensure the same key between singing and accompaniment as follows.

- **Same-Song Remix**: instead of using a different song, we can shift the accompaniment track of the same song by a random duration between 5 to 15 seconds ahead or behind the singing track and remix both. This creates a unique alignment where the vocals and music are out of their original synchronization but still ensures both are in the same musical key.

- **Key-Match Remix**: we use the Madmom key detection algorithm [31] to iterate all the accompaniments and eliminate the ones that are in a different key than the solo singing. Among the remaining candidates, we randomly pick one accompaniment, and remix it with the solo singing.

As a result, we have an augmented training set of 24,000 songs in total, where singing with accompaniment, solo singing, and singing and accompaniment remix all contribute 8,000 songs, indicated respectively as subindex 1, 2, 3 in Fig. 1(b). Furthermore, we can extend the set of subindex 3 by combining the augmented data from all three remixing approaches above (8000 $\times$ 3 songs) and propose a fourth approach: **All Remix**, with 40,000 songs in total.

3. EXPERIMENTS

As indicated in [17], each song of the 10KSinging dataset is associated with an overall, normalized rating between 0 and 1, and the goal of our ASSE model is to predict a regressed value close to the ground truth rating. Although the work presented in this paper is not open-source for proprietary restrictions, most of the essential components are open-source as follows: the code base and the fundamental structure of CRNN, including the pitch histogram calculation can be found at Github[6]; the annotations for 10KSinging, the attention machism appended after CRNN, and the X-vector calculation can be found at [17], and we will respectively explain our experimental settings and relevant implementation details in Section 3.1 and Section 3.2 for the ease of reproducing our work.

3.1 Experimental Settings

We first investigate the effect of data augmentation in five settings: no data augmentation, Shuffle-And-Remix (SAR), Same-Song Remix (SSR), Key-Match Remix (KMR), and All Remix (ALL) proposed in Section 2.2.2. In each data augmentation setting, we can either use the baseline architecture (Fig. 1(a)) or adopt BiCA (Fig. 1(b)), resulting in a total of 10 experiments. In each experiment, we present the performance on the 1000-song test set from the two versions of 10KSinging: singing with accompaniment ("w/ acc") and solo singing ("w/o acc"). As shown in Table 1, the first two experiments involve no data augmentation and each has two distinct ASSE models that are trained using the two versions of 10KSinging ("w/ acc" and "w/o acc"), same as [17].

For the remaining eight experiments involving data augmentation, each uses an augmented training set of 10KSinging, which contains songs from the following three sets: singing with accompaniment, solo singing, and singing and accompaniment remix introduced in Section 2.2.2. Unlike the first two experiments, each of the eight experiments has only one ASSE model, which is evaluated in both "w/ acc" and "w/o acc" test sets. Altogether, 12 models are trained in total to explore the effects of data augmentation and BiCA.

3.2 Implementation Details

We adopt the same parameters for generating CQT and pitch histograms as described in [17], namely 96-bin CQT and 120-bin pitch histogram. For training, we use Mean Squared Error (MSE) as the loss function, where the epoch with the lowest MSE on the validation set is chosen as the best-performing model, both in the "w/ acc" and "w/o acc" settings. We use the Adam optimizer and a learning rate of 0.0001. The number of epochs is set to 250 with a batch size of 4. All other parameters remained consistent with those outlined in [17], except for a few adjustments, which are detailed below.

[5] We use an open-source implementation of BiCA available at: https://github.com/lucidrains/bidirectional-cross-attention.

[6] Implementation can be found at: https://github.com/AME430/Towards-Training-Explainable-Singing-Quality-Assessment-Network-with-Augmented-Data.

Data Aug	BiCA	MSE(↓)		MAE(↓)		Bad P%(↓)		Pearson(↑)	
		w/o acc	w/ acc	w/o acc	w/ acc	w/o acc	w/ acc	w/o acc	w/ acc
No (8,000)	No [17]	0.0042	0.0046	0.0495	0.0524	10.1%	11.8%	0.562	0.497
No (8,000)	Yes	0.0038	0.0043	0.0459	0.0499	8.7%	10.5%	0.623	0.539
SAR (24,000)	No	0.0041	0.0044	0.0495	0.0519	9.3%	9.4%	0.561	0.522
SAR (24,000)	Yes	0.0031	0.0033	0.0386	0.0415	6.5%	7.3%	0.697	0.670
SSR (24,000)	No	0.0041	0.0044	0.0497	0.0515	9.6%	10.2%	0.555	0.514
SSR (24,000)	Yes	0.0029	0.0033	0.0385	0.0416	5.8%	7.0%	0.714	0.673
KMR (24,000)	No	0.0041	0.0044	0.0496	0.0521	8.9%	9.8%	0.562	0.517
KMR (24,000)	Yes	0.0028	0.0031	0.0375	0.0413	6.2%	6.9%	0.730	0.687
ALL (40,000)	No	0.0039	0.0040	0.0471	0.0478	9.4%	10.1%	0.593	0.576
ALL (40,000)	Yes	**0.0025**	**0.0030**	**0.0351**	**0.0387**	**4.7%**	**6.1%**	**0.769**	**0.709**

Table 1. The ASSE results of Mean Squared Error (MSE), Mean Absolute Error (MAE), Bad Case Proportion (Bad P %), and Pearson correlation coefficient (Pearson) on the singing with accompaniment test set (w/ acc, 1,000 songs) and the solo singing test set (w/o acc, 1,000 songs) from the 10KSinging dataset [17]. SAR, SSR, KMR, and ALL refer to the four different data augmentation methods introduced in Section 2.2.2: Shuffle-and-Remix, Same-Song Remix, Key-Match Remix, and All Remix, respectively, where the number in parenthesis indicates the number of songs used as training data. For experimental purposes, no data augmentation, SAR, SSR, KMR, and ALL is respectively applied to the model architecture without and with Bi-Directional Cross-Attention (BiCA, illustrated in Fig. 1(b)) to demonstrate the individual and reciprocal effects of data augmentation and BiCA. The downward and upward arrows on the evaluation metrics respectively represent the desirable lower or higher values for better performances. The best results are highlighted in bold, which concentrate on the ASSE model employing both All Remix data augmentation and BiCA (ALL-Yes), is therefore our proposed method in this paper.

We use the sigmoid activation function following the final dense layer to constrain the output range between 0 to 1. Also, the Exponential Linear Unit (ELU) activation function is introduced within the dense layer. These adjustments can facilitate the model's ability to learn a more accurate distribution of the output score.

3.3 Evaluation Metrics

Although correlation coefficients are often used as the evaluation metric in ASSE [5, 13–15, 17, 32], we aim to incorporate additional metrics to demonstrate the performances of ASSE models more comprehensively. Overall, four evaluation metrics are considered:

- Mean squared error (MSE) (↓): same as the loss function introduced in Section 3.2.
- Mean absolute error (MAE) (↓): it shows how much the predicted rating deviates from the ground truth in the linear scale.
- Bad case proportion (↓): same as [17], the predicted rating will be considered a bad case if its MAE is no less than 0.1.
- Pearson correlation coefficient (↑): it demonstrates the degree of correlation between the predicted rating and the ground truth, within the range of $[-1, 1]$.

3.4 Results and Discussions

The results are shown in Table 1, where we use acronyms to represent each experiment. For example, No-No indicates the experiment without data augmentation nor BiCA, and KMR-Yes indicates the experiment using both KMR augmentation and BiCA, etc.

3.4.1 *Results on BiCA*

We first investigate the effects of BiCA by comparing the models with and without BiCA under five different data augmentation settings (No-No VS. No-Yes; SAR-No VS. SAR-Yes; SSR-No VS. SSR-Yes; KMR-No VS. KMR-Yes; ALL-No VS. ALL-Yes), finding that using BiCA results in consistent performance improvements in all cases. This demonstrates that the employment of BiCA effectively helps the ASSE models capture useful interrelationships between pitch histogram and X-vector and facilitate their reciprocal knowledge exchange, leading to better results. This is reasonable since there can be correlations between singers' pitch accuracy and timbre quality that are beneficial for ASSE. For example, singers with excellent singing skills tend to have great pitch accuracy (indicated by pitch histogram) and timbre quality (indicated by X-vector), and vice versa for mediocre or inferior singers.

3.4.2 *Results on Data Augmentation*

We then compare the four data augmentation approaches: SAR, SSR, KMR, and ALL to no data augmentation. When using the baseline architecture (Fig. 1(a) without BiCA), results show overall marginal improvements in almost all cases (SAR-No VS. No-No; SSR-No VS. No-No; KMR-No VS. No-No; ALL-No VS. No-No). When

BiCA is applied, the improvement is much more apparent (SAR-Yes VS. No-Yes; SSR-Yes VS. No-Yes; KMR-Yes VS. No-Yes; ALL-Yes VS. No-Yes). These results demonstrate the effectiveness of data augmentation and its reciprocal advantage with BiCA. As discussed in Section 3.4.1, it seems that data augmentation provides more samples for BiCA to further exploit correlations between pitch histogram and X-vector, which can be beneficial for evaluating singing skills in ASSE.

Of particular interest, we notice that KMR achieves superior results among all three data augmentation approaches (KMR-Yes VS. SSR-Yes VS. SAR-Yes). This could be that KMR combines the advantages of SAR and SSR, where the former mixes the singing with a different accompaniment and the latter ensures the same key between singing and accompaniment, leading to better performances. Despite their differences, we can combine the augmented data from SAR, SSR, and KMR as All Remix (see Section 2.2.2) for even more training data, resulting in the best results overall (SAR-Yes VS. SSR-Yes VS. KMR-Yes VS. ALL-Yes).

3.4.3 Results on Solo Singing and Singing with Accompaniment

Additionally, we compare the performances presented in solo singing (w/o acc) and singing with accompaniment (w/ acc) scenarios. Results show that the ASSE models consistently perform better in solo singing, which is understandable considering singing with accompaniment contains irrelevant accompaniment information that can interfere with the training of ASSE models. Although we can follow [17] to add a singing voice separation step (see Fig. 1(a)) to remove accompaniment for better performances, we choose to keep the end-to-end nature of our ASSE system (see Fig. 1(b)) and consider the performance gap between solo singing and singing with accompaniment for our proposed ASSE model (ALL-Yes) non-essential, since both are performing better than the baseline No-No in the solo singing (w/o acc) condition.

3.4.4 Overall Results

Finally, once we combine data augmentation with BiCA, our proposed ALL-Yes model yields notably better results than the baseline [17] (No-No) across all metrics: reaching relative improvements of 40.5% on MSE (0.0025 VS. 0.0042, likewise for the following ones), 29.1% on MAE, 53.5% on Bad P %, and 36.8% on Pearson in solo singing (w/o acc); 34.8% on MSE (0.0030 VS. 0.0046, likewise for the following ones), 26.1% on MAE, 48.3% on Bad P %, and 42.7% on Pearson in singing with accompaniment (w/ acc) scenario.

As discussed in Section 3.1, there are two models under No-No, one trained for w/o acc and the other trained for w/ acc conditions. [17] then proposed the former in the paper due to its superior performance. However, it can only process solo singing data and requires a singing voice separation tool to remove accompaniment from the input. In comparison, our proposed ALL-Yes model can process both solo training and singing with accompaniment inputs, and this is what we refer to as an end-to-end ASSE model, which does not require singing voice separation and also yields notably better performances, with Pearson correlation coefficients of 0.769 in w/o acc and 0.709 in w/ acc, compared to the baseline model of 0.562.

4. CONCLUSIONS

In this paper, we introduce a new ASSE system using data augmentation and compare four specific augmentation approaches: the existing Shuffle-And-Remix [25], the novel Same-Song Remix, Key-Match Remix, and All Remix we propose. Results show that our All Remix approach achieves the best performances, and our system can process both solo singing and singing with accompaniment in an end-to-end manner, thereby eliminating the need for a singing voice separation tool required by the baseline model [17]. We also introduce a Bi-directional Cross-Attention mechanism (BiCA) as a feature fusion method to ASSE for the first time, which discovers useful interrelationships between pitch histogram and X-vector and results in consistent performance improvements in our experiments.

With the combination of BiCA and All Remix data augmentation approach, we not only achieve notable improvements in ASSE performances compared to the baseline [17], we also develop a versatile model capable of processing both solo and instrumentally accompanied vocal performances. To the best of our knowledge, such encompassing ASSE models have not been proposed in existing literature before.

5. FUTURE WORK

Looking ahead, we will continue this research by incorporating future open-source ASSE datasets proposed in the literature. Indeed, we could only make use of the 10KSinging dataset due to the lack of open-source datasets in this domain. Our future work also includes exploring alternative features that could potentially improve the performances of our ASSE models. For instance, we will consider large-scale music models that employ self-supervised learning, since features extracted by those models such as Jukebox [33] and MERT [34] have recently been proven effective and even established new SOTA performances in various music-related tasks. Therefore, these features will be incorporated into our model to exert their potential. Finally, since data augmentation combining solo singing with different versions of accompaniment results in consistent performance improvements, we will explore more data augmentation methods for solo singing by, for example, adding noise, adjusting gain, and applying high/low-pass filters that have been employed in other MIR-related tasks [35] for better performances.

6. REFERENCES

[1] S. Jones, "Music and the internet," *The handbook of internet studies*, pp. 440–451, 2011.

[2] C. Shin and Y. Lim, "Design and implementation of impromptu mobile social karaoke for digital cultural spaces in the new normal era," *Applied Sciences*, vol. 13, no. 22, 2023.

[3] A. M. Studiorum, "Subjective evaluation of common singing skills using the rank ordering method," in *Proceedings of the Ninth International Conference on Music Perception and Cognition*, 2006, pp. 1507–1512.

[4] W.-H. Tsai and H.-C. Lee, "An automated singing evaluation method for karaoke systems," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 2428–2431.

[5] C. Gupta, H. Li, and Y. Wang, "A technical framework for automatic perceptual evaluation of singing quality," *APSIPA Transactions on Signal and Information Processing*, vol. 7, no. e10, 2018.

[6] ——, "Perceptual evaluation of singing quality," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 577–586.

[7] H. Zhang, Y. Jiang, T. Jiang, and P. Hu, "Learn by referencing: Towards deep metric learning for singing assessment," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021, pp. 825–832.

[8] J. Böhm, F. Eyben, M. Schmitt, H. Kosch, and B. Schuller, "Seeking the superstar: Automatic assessment of perceived singing quality," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 1560–1569.

[9] Barı, Bozkurt, O. Baysal, and D. Yüret, "A dataset and baseline system for singing voice assessment," in *Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR)*, 2017, pp. 25–28.

[10] T. Nakano, M. Goto, and Y. Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH)*, vol. 4, 2006, pp. 1706–1709.

[11] C. Gupta, H. Li, and Y. Wang, "Automatic leaderboard: Evaluation of singing quality without a standard reference," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 13–26, 2020.

[12] N. Zhang, T. Jiang, F. Deng, and Y. Li, "Automatic singing evaluation without reference melody using bi-dense neural network," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 466–470.

[13] C. Gupta, L. Huang, and H. Li, "Automatic rank-ordering of singing vocals with twin-neural network." in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, 2020, pp. 416–423.

[14] L. Huang, C. Gupta, and H. Li, "Spectral features and pitch histogram for automatic singing quality evaluation with crnn," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 492–499.

[15] J. Li, C. Gupta, and H. Li, "Training explainable singing quality assessment network with augmented data," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 904–911.

[16] C. Gupta, J. Li, and H. Li, "Towards reference-independent rhythm assessment of solo singing," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 912–919.

[17] Y. Ju, C. Xu, Y. Guo, J. Li, and S. Lui, "Improving automatic singing skill evaluation with timbral features, attention, and singing voice separation," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2023, pp. 612–617.

[18] X. Sun, Y. Gao, H. Lin, and H. Liu, "Tg-critic: A timbre-guided model for reference-independent singing evaluation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[19] T. M. L. L. team from Tencent Music Entertainment Group, "Lyra-SA Dataset," https://lyracobar.y.qq.com/singvoicedataset.html, 2023, [Online; accessed 21-March-2023].

[20] W.-H. Tsai and H.-C. Lee, "Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1233–1243, 2012.

[21] C. Li, Y. Li, X. Du, Y. Ju, S. Hu, and Z. Wu, "VocEmb4SVS: Improving singing voice separation with vocal embeddings," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 234–239.

[22] Y. Yang, J. Song, Z. Huang, Z. Ma, N. Sebe, and A. G. Hauptmann, "Multi-feature fusion via hierarchical regression for multimedia analysis," *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 572–581, 2013.

[23] J. Chen, Y. He, and J. Wang, “Multi-feature fusion based fast video flame detection,” *Building and Environment*, vol. 45, no. 5, pp. 1113–1122, 2010.

[24] M. Hiller, K. A. Ehinger, and T. Drummond, “Perceiving longer sequences with bi-directional cross-attention transformers,” *arXiv preprint arXiv:2402.12138*, 2024.

[25] T.-H. Hsieh, K.-H. Cheng, Z.-C. Fan, Y.-C. Yang, and Y.-H. Yang, “Addressing the confounds of accompaniments in singer identification,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1–5.

[26] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[27] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, “X-vectors meet emotions: A study on dependencies between emotion and speaker recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7169–7173.

[28] X. Zhang, J. Wang, N. Cheng, and J. Xiao, “Singer identification for metaverse with timbral and middle-level perceptual features,” in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–7.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, pp. 6000–6010.

[30] N. Kitaev, Łukasz Kaiser, and A. Levskaya, “Reformer: The efficient transformer,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[31] F. Korzeniowski and G. Widmer, “Genre-agnostic key classification with convolutional neural networks,” in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 264–270.

[32] C. Gupta, H. Li, and Y. Wang, “Automatic evaluation of singing quality without a reference,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 990–997.

[33] R. Castellon, C. Donahue, and P. Liang, “Codified audio language modeling learns useful representations for music information retrieval,” in *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021, pp. 88–96.

[34] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Lin, A. Ragni, E. Benetos, N. Gyenge, R. Dannenberg, R. Liu, W. Chen, G. Xia, Y. Shi, W. Huang, Y. Guo, and J. Fu, “MERT: Acoustic music understanding model with large-scale self-supervised training,” *arXiv preprint arXiv:2306.00107*, 2023.

[35] J.-C. Wang, Y.-N. Hung, and J. B. Smith, “To catch a chorus, verse, intro, or anything else: Analyzing a song with structural functions,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 416–420.