Report from Dagstuhl Seminar 19052

Computational Methods for Melody and Voice Processing in Music Recordings

Edited by Meinard Müller, Emilia Gómez, and Yi-Hsuan Yang

3.29 Singing Voice Modelling for Language Learning

Ye Wang (National University of Singapore, SG)

Singing is a popular form of entertainment, as evidenced by the millions of active users of Karaoke apps like Smule's Sing! and Tencent's Quanmin K Ge. Singing is presumed to be the oldest form of music making and can be found in human cultures around the world. However, singing can be more than just a source of entertainment: parents sing nursery rhymes to their young children to help them learn their first language, music therapists use singing to help aphasia patients speak again, and medical studies have revealed that singing, in general, has many health benefits. Consequently, computational methods for singing analysis have emerged as an active research topic in the music information retrieval community.

Pedagogical research has shown that actively singing in a foreign language helps with pronunciation, vocabulary acquisition, retention, fluency, and cultural appreciation [1]. Inspired by this scientific discovery, we have developed a novel multi-language karaoke application called SLIONS (Singing and Listening to Improve Our Natural Speaking), designed to foster engaging and joyful language learning [2]. We followed a user-centered design process that was informed by conducting interviews with domain experts and by conducting usability tests among students. The key feature of SLIONS is an automatic speech recognition (ASR) tool used for objective assessment of sung lyrics, which provides students with personalized, granular feedback based on their singing pronunciation.

During its proof of concept phase, SLIONS employed Google's ASR technology to evaluate sung lyrics. However, this solution lacks technical depth and has several critical limitations. First, Google ASR is proprietary technology and is effectively a black box. As a result, it is impossible for us to understand precisely why it succeeds in evaluating certain sung lyrics but fails in others. This not only prevents us from gaining insights into the underlying models but also affects SLIONS' value in real-world applications. It is also impossible for us to modify Google's ASR technology even though we wish to use SLIONS for widely varying applications that range from language learning to melodic intonation therapy. Google ASR technology is designed for speech recognition and is suboptimal for analyzing singing voice, as the characteristics of sung utterances differ from those of spoken utterances. Therefore it is desirable to investigate better and more versatile computational methods for objective assessment of sung lyrics. Furthermore, it is also useful to address some important human–computer interaction (HCI) questions. For example, while previous studies have shown that singing can help with learning pronunciation, the critical question of which factors are essential for not only improving pronunciation but also maintaining engagement during singing exercises remains. It is vital to design interface/interaction features that support both learning and engagement aspects.

Although speech and singing share a common voice production organ and mechanism, singing differs from speech in terms of pitch variations, possible extended vowels, vibrato, and more. It is interesting to exploit the similarities between speech and singing in order to employ existing methods/tools and datasets in the relatively mature ASR field while also developing new methods to address the differences. To this end, we have created and published the NUS Sung and Spoken Lyrics Corpus, a small phonetically annotated dataset of voice utterances [3]. Furthermore, we have also attempted to address the problem of lyrics and singing alignment [4, 5, 6, 7], evaluation of sung lyrics [8, 9], and intelligibility of sung lyrics [10]. While many challenges remain as to adequately modeling and analyzing singing voice for real-world applications such as language learning, our efforts are already pointing the way towards a robust, versatile model that can enable the automatic evaluation of sung utterance pronunciation.

References

[1] Arla J. Good, Frank A. Russo, and Jennifer Sullivan. The Efficacy of Singing in Foreign-Language Learning. Psychology of Music, 43(5), 2015, pp. 627–640.

[2] Murad, D., Wang, R., Turnbull, D., Wang, Y.. SLIONS: A Karaoke Application to En- hance Foreign Language Learning. Proceedings of the ACM International Conference on Multimedia, Seoul, Korea, 2018, pp. 1679–1687.

[3] Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim, and Ye Wang. The NUS Sung and Spoken Lyrics Corpus: A Quantitative Comparison of Singing and Speech. Proceedings of the IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013, pp. 1–9.

[4] Chitralekha Gupta, Rong Tong, Haizhou Li, and Ye Wang. Semi-Supervised Lyrics and Solo-Singing Alignment. Proceedings of International Society for Music Information Re- trieval Conference (ISMIR), Paris, France, 2018, pp. 600–607.

[5] Min-Yen Kan, Ye Wang, Denny Iskandar, Tin Lay Nwe, and Arun Shenoy. LyricAlly: Automatic Synchronization of Textual Lyrics to Acoustic Music Signals. IEEE Transactions on Audio, Speech, and Language Processing, 16(2), 2008, pp. 338–349.

[6] Denny Iskandar, Ye Wang, Min-Yen Kan, and Haizhou Li. Syllabic Level Automatic Synchronization of Music Signals and Text Lyrics. Proceedings of the ACM International Conference on Multimedia, 2006, pp. 659–662.

[7] Ye Wang, Min-Yen Kan, Tin Lay Nwe, Arun Shenoy, and Jun Yin. LyricAlly: Automatic Synchronization of Acoustic Musical Signals and Textual Lyrics. Proceedings of the ACM International Conference on Multimedia, 2004, pp. 212–219.

[8] Chitralekha Gupta, Haizhou Li, and Ye Wang. Automatic Pronunciation Evaluation of Singing. Proceedings of the International Conference on Spoken Language Processing (In- terspeech), Hyderabad, India, 2018, pp. 1507–1511.

[9] Chitralekha Gupta, David Grunberg, Preeti Rao, and Ye Wang. Towards Automatic Mispronunciation Detection in Singing. Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 2017, 390–396.

[10] Karim M. Ibrahim, David Grunberg, Kat Agres, Chitralekha Gupta, and Ye Wang. Intelligibility of Sung Lyrics: A Pilot Study. Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 2017, pp. 686–693.