INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

On the Importance of Audio-source Separation for Singer Identification in Polyphonic Music

Bidisha Sharma, Rohan Kumar Das, Haizhou Li

Department of Electrical and Computer Engineering,
National University of Singapore, Singapore
{s.bidisha, rohankd, haizhou.li}@nus.edu.sg

Abstract

Singer identification is to automatically identify the singer in a music recording, such as a polyphonic song. A song has two major acoustic components that are singing vocals and background accompaniment. Although identifying singers is similar to speaker identification, it is challenging due to the interference of background accompaniment on the singer-specific information in singing vocals. We believe that separating the background accompaniment from the singing vocal will help us to overcome the interference. In this work, we extract the singing vocals from polyphonic songs using Wave-U-Net based audio-source separation approach. The extracted singing vocals are then used in i-vector based singer identification system. Further, we explore different state-of-the-art audio-source separation methods to establish the role of considered method in application to singer identification. The proposed singer identification framework achieves an absolute accuracy improvement of 5.66% over the baseline without audio-source separation.

Index Terms: Singer identification, audio-source separation, Wave-U-Net, i-vector speaker modeling

1. Introduction

Singer identification is to automatically identify the singer in a music recording. It potentially enables many multimedia applications. The basic modules of a conventional singer identification system are vocal/non-vocal segmentation, feature extraction, modeling of the singers and classification. Although this problem is analogous to speaker identification [1, 2], it is more challenging because of the presence of loud, non-stationary background accompaniment that acts as an additional variable over the singing vocals. Many of the previous works on singer identification have either ignored the influence of background music on singing vocals or attempted to identify singers from speaker-identification standpoint [3–7].

The initial explorations for singer identification consider a speaker recognition system with Gaussian mixture model (GMM), trained using mel frequency cepstral coefficient (MFCC) features to distinguish singers [5]. This is followed by some attempts to improve the feature representation [6]. Others explore the structural knowledge of music for singer identification [7]. The rhythm structure and inter beat time resolution features are used to train classifiers for both vocal segmentation and singer modelling. This study established that instrumental music sections are also unique for each singer apart from the vocal characteristics.

The work presented in [8] uses vocal segmentation using sparse representation based classification to improve the performance of singer identification. The approach reported in [9] estimated the uncertainty from enhanced melody of polyphonic songs in an automatic manner to facilitate more or less importance to the features depending on how reliable they are in different time frames. Unlike the traditional methods, the authors of [10] attempted to extract acoustic features that reflect vibrato information to identify singers.

In another direction of work [11], it is considered that the major issue in automatically identifying singers is the negative influences caused by accompaniment sounds. The problem is attempted in two ways, accompaniment sound reduction and reliable frame selection, which contributed to improve the singer identification performance significantly. Along this direction, study in [12] suggests removing the influence of background accompaniment, with the assumption that substantial similarities exist between the instrumental-only regions and the singing regions with background accompaniment. Similarly, [13] shows that singer identification with singing vocal extraction from polyphonic songs outperforms that without singing vocal extraction. The audio-source separation method employed in their work is a probabilistic approach based on GMMs of the short-time spectra of two sources. Further, they have used adaptation of source model via maximum likelihood linear regression in order to improve the separation quality.

Overall, the prior work in singer identification runs in three major directions that include improving the vocal segmentation [8, 10] , singer-specific acoustic feature extraction [3, 6, 10], and audio-source separation [9, 12, 13]. We note that the audio-source separation methods have advanced greatly in recent time with better accuracy compared to the earlier studies [9, 12, 13]. Nevertheless, most of the existing singer identification methods used GMMs or SVM modeling techniques, whereas i-vector modeling gained popularity for speaker recognition in the current decade [14]. Eghbal et al [15] have used i-vector based singer identification framework using timbral features on Artist20 database [16] and presented a comparative study.

Singing vocal shares the same underlying physiological mechanism for production as that of speech. However, singing requires a higher level of vocal effort and a larger range of variation in loudness than those of natural speech [17]. There are two ways to combat these differences. Firstly, one can reduce the influence of background accompaniment and secondly, introduce features to represent uniqueness in each singer's voice, despite of the presence of background accompaniment. The timbre features are well established in the literature to represent singer identity [15]. Therefore, in this work we attempt to follow the former direction by extracting singing vocals from polyphonic songs, to reduce the interference on extraction of singer-specific information from songs. We hypothesize that the interference of background accompaniment over the singing vocals affects the extracted features that propagate to the singer identification models. In this regard, an end-to-end audio-source separation

http://dx.doi.org/10.21437/Interspeech.2019-1925

method over i-vector based system is considered for singer identification. The studies presented in this work are conducted on Artist20 database.

The rest of the paper is organized as follows. Section 2 discusses an end-to-end audio-source separation technique that could be suitable for our study. The proposed framework of singer identification with audio-source separation is described in Section 3. In Section 4, the details of the experiments are presented. Section 5 reports the results and discussion. The work is finally concluded in Section 6.

2. Audio-source Separation

To overcome the background accompaniment on extracting singer-specific information, we incorporate an audio-source separation module to extract the singing vocals from polyphonic songs. We study the effect of three different audio-source separation methods on our singer identification: harmonic/percussive, convolutional neural network (CNN) based, and Wave-U-Net based approach.

Percussion component in the background accompaniment introduces vertical lines in the spectrogram, which makes it noisy. Therefore, we first attempt to remove these using the traditional harmonic/percussive method [18], which is reported to be simple and effective. This method uses median filters individually in the horizontal and vertical directions to separate the harmonic and the percussive events. This separation method is integrated in the widely used audio and music analysis library librosa[1] [19].

As an alternative we are also interested a CNN based solution to the audio-source separation [20]. This method achieves the same performance as that of multi-layer perceptron based audio-source separation with less time complexity and compact representation. In this case, we use the model trained on iKala[2] dataset, for voice, bass, and drums separation [21].

Wave-U-Net[3] represents a recent success in audio-source separation [22]. It is an adaptation of the U-Net architecture [23, 24] into one-dimensional time domain to perform end-to-end audio-source separation, which repeatedly resamples feature maps to compute and combine features at different time scales. This architecture extracts an increasing number of higher-level features on coarser time scales using down-sampling blocks. These features are combined with the earlier obtained local, high-resolution features using up-sampling blocks, yielding multi-scale features, which are used for making predictions [22]. We have used the pre-trained best vocal separation model (M5-HighSR), whose implementation is available in [25].

In Figure 1, we show a comparison of the vocals obtained from the three different audio-source separation methods. Figure 1 (a) shows the waveform corresponding to a segment of a polyphonic song from Artist20 dataset. Figure 1 (b), (c), (d), (e) show spectrogram (with 20 ms frame-size, 10 ms frame-shift, sampling rate 10 kHz) corresponding to original polyphonic song and extracted vocals for the same audio segment using harmonic/percussive, CNN and Wave-U-Net based audio-source separation, respectively. If we compare each of Figure 1 (c), (d), (e) with Figure 1 (b), we can observe that using harmonic/percussive method (Figure 1 (c)), the percussive component is removed, however the other components are preserved

[1]https://librosa.github.io/librosa/

[2]http://mac.citi.sinica.edu.tw/ikala/

[3]https://github.com/f90/Wave-U-Net

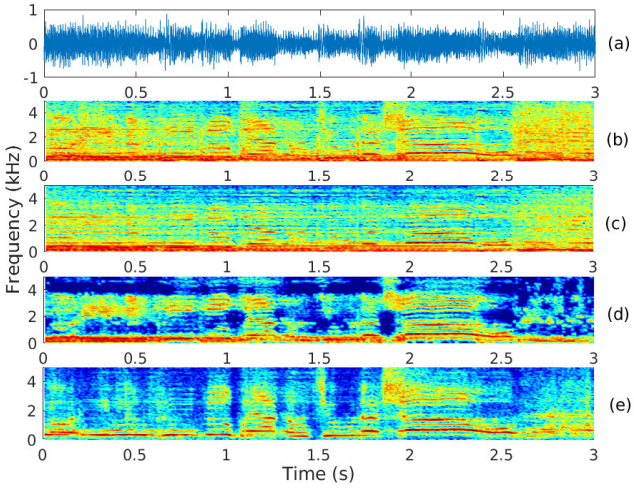

Figure 1: *Comparison of spectrograms for different audio-source separation methods for a segment of polyphonic song from Artist20 dataset, (a) original mixed audio waveform; spectrogram of (b) polyphonic song, extracted vocal using (c) harmonic/percussive, (d) CNN based, (e) Wave-U-Net based audio-source separation methods.*

in the spectrogram, similar to the polyphonic song shown in Figure 1(b). After applying CNN based source separation, although the vocal specific characteristics are preserved in the spectrogram, as shown in Figure 1(d), there are some glitches at the boundaries of the phonemes and some of the harmonic components are removed. This type distortion is also evident for the songs with high intensity background accompaniment during informal listening. Figure 1 (e) shows that the Wave-U-Net based audio-source separation not only removes the background accompaniment, but also preserves the vocal part.

3. Singer Separation and Identification Framework

In this work, we propose a framework for automatic singer identification, which considers the audio-source separation method to eliminate the interference introduced by the background accompaniment on singer identity cues. Figure 2 shows the block diagram of the proposed framework for singer identification. Instead of directly extracting the features, we first extract the singing vocals from polyphonic songs using end-to-end Wave-U-Net architecture proposed in [22]. The audio-source separated vocals are then used for singer modeling and testing to identify the singers.

We use i-vector based system for modeling the singer characteristics. The i-vector is a factor analysis approach that represents the dominant speaker information in terms of a low dimensional vector. This compact representation is learned using a total variability space that is trained using background data. The total variability space captures all the variability like channel/session information, and hence necessary compensation techniques have to be applied on i-vectors to compensate them [26, 27]. The train i-vectors of singers are computed using features extracted after audio-source separation. Similarly, the test i-vectors from songs of test data are obtained after audio-source separation. Given a test song i-vector, its similarity with all the train i-vector models is computed, to identify the singer with the highest similarity.

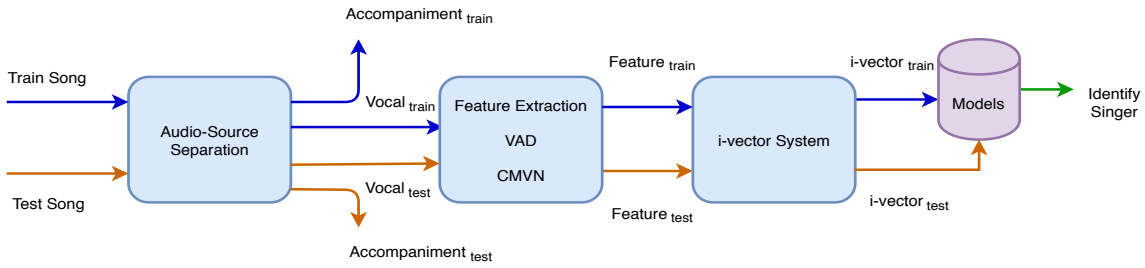

Figure 2: *Proposed framework of singer identification with audio-source separation module.*

4. Experiments

This section describes the experiments related to the singer identification framework proposed in this paper. We also discuss the details of the corpus and experimental setup.

4.1. Database

To validate our assumption and experiments, we have used Artist20[4] dataset, which is a standard corpus for singer/artist identification [16]. It contains six albums of English popular songs from 20 artists that include a total of 1,413 tracks. The database also defines a canonical 6-fold train/test scheme, where each fold consists of training on five albums per artist, and testing with the remaining one. The audio files are provided in the form of 32 kbps mono MP3 with 16 kHz sampling frequency. The previous recent work in singer identification [15] used the same corpus and reported the results for the predefined 6-fold train/test scheme.

4.2. Experimental setup

As mentioned in Section 3, firstly we use Wave-U-Net to extract singing-vocals from polyphonic audio, for all the 1,413 tracks available in the Artist20 database. The Wave-U-Net architecture computes an increasing number of higher-level features on coarser time scales using down-sampling blocks. These features are combined with the earlier computed local, high-resolution features using up-sampling blocks, yielding multi-scale features which are used for making predictions.

To train the audio-source separation model available in [25], 75 tracks from the training partition of the MUSDB[5] [28] multi-track database are randomly assigned. In this work, we use the best performing pre-trained model M5-HighSR, whose implementation is available in [25]. The M5-HighSR model has several modifications over the baseline model, which are difference output layer, input context and re-sampling, stereo channels and learned up-sampling with 44.1 kHz sampling rate [23].

In ideal case, the extracted singing vocals should contain only singer's voice and silence segments corresponding to instrumental sections of the songs. However, due to the errors in the singing vocal separation method, the instrumental accompaniments are suppressed only to some extent, and these non-vocal segments do not contain any useful information. To detect the non-vocal segments, we divided the spectrum of each frame (framesize 25 ms, frameshift 5 ms, sampling frequency 16 kHz) into four equal subbands to detect these suppressed non-vocal segments. The energy corresponding to the 2nd subband shows a prominent difference between the segments with vocals and without vocals. A threshold based on the average 2nd subband energy is set to classify the frames into vocal and non-vocal categories. The non-vocal segments with very long duration are removed from the audio as a pre-processing step.

The singing vocals after audio-source separation and vocal/non-vocal segmentation are then used for singer identification studies. The previous work on Artist20 dataset using i-vector framework presented a comparative study on the use of different feature dimensions, channel/session compensation and scoring techniques [15]. It is found that the best performance is achieved by using 20-dimensional MFCC features along with their first and second derivatives over i-vector with linear discriminate analysis (LDA) [29] followed by cosine distance for scoring. Therefore, we use the exact best experimental setup to compare our proposed framework to that without audio-source separation presented in [15].

We perform short-term processing and extract 20-dimensional MFCC features along with their first and second derivatives (60-dimensional) for every frame [30]. The features are then subjected to cepstral mean variance normalization (CMVN) before developing the models [31].

The studies in this work follow a 6-fold validation for performance evaluation as discussed earlier. In each fold, five albums are used for training and the remaining one is used for testing. The universal background model (UBM) with 1024 Gaussian components and a total variability matrix (T-matrix) with 400 factors are used for i-vector extraction [14, 32]. This set up is similar to the previous work in [15]. No other corpus is used to train the UBM and T-matrix. Instead, in each fold, one third of the training set is used to train UBM and the entire set for T-matrix learning. The 150-dimensional LDA is computed with same data by which T-matrix is trained.

5. Results and Discussion

In this section, we discuss the results for singer identification studies with the proposed framework. Further, we extend the studies to compare Wave-U-Net based audio-source separation to other methods and short song segment scenario.

5.1. Studies with proposed framework

Table 1 shows the performance of the proposed framework with audio-source separation and its comparison to baseline without audio-source separation. The results of the baseline are cited from [15] that shows the average identification accuracy across the 6-fold studies. We report the results of every fold for the proposed framework, followed by the the average identification accuracy. Each fold result also represents performance of different albums used for identifying the singers from Artist20 corpus. It is observed from Table 1 that the proposed framework with audio-source separation helps to improve the performance by a large margin, which is 89.97% compared to 84.31% for baseline method. This indicates the importance of audio-source separation for singer identification that confirms our hypothesis.

[4]https://labrosa.ee.columbia.edu/projects/artistid/
[5]https://sigsep.github.io/datasets/musdb.html

Table 1: *Performance comparison of proposed framework to that without audio-source separation on Artist20 corpus.*

Train/test	Accuracy (%)
Baseline	
Avg. [15]	84.31
Proposed	
Fold 1	85.17
Fold 2	90.41
Fold 3	89.96
Fold 4	95.63
Fold 5	90.21
Fold 6	88.24
Avg.	**89.97**

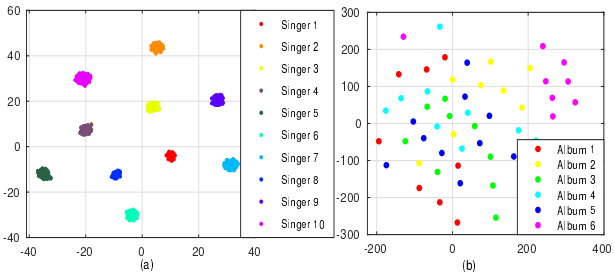

Figure 3: *t-SNE visualization of i-vectors for (a) 10 different singers, (b) different albums corresponding to one singer.*

In Figure 3 (a), we have shown the visualization of i-vector based singer representation for randomly selected 10 singers. We use t-Distributed Stochastic Neighbor Embedding (t-SNE) technique [33] for this, which is widely used for the visualization of high-dimensional data. It is observed that the proposed framework is efficient to represent uniqueness in each singer. To establish that our singer models are album independent, we also show the same t-SNE representation in Figure 3 (b), for 6 albums of a randomly chosen singer. As the i-vectors are overlapping for different albums, we can conclude that the models are album independent.

5.2. Comparison to various audio-source separation methods for proposed framework

We also compare different audio-source separation methods discussed in Section 2 for singer identification. The training and test samples from fold 2 are randomly chosen for this comparative study. Table 2 shows the comparison of three different methods of audio-source separation for singer identification. It is observed that the Wave-U-Net based audio-source separation method outperforms the other two by a large margin.

We recall our observations from Figure 1 that presented spectrograms of singing vocals obtained using different audio-source separation methods. It showed better singing vocals with CNN based method than the harmonic/percussive approach. However, the performance achieved with the CNN based method is much poorer for singer identification as observed. The work done in [34] on singing-to-lyrics alignment showed that the effectiveness of CNN based method is closer to Wave-U-Net based approach, than the harmonic/percussive audio-source separation. This shows that the CNN based method can be effective for audio-source separation in tasks, where there is no importance of speaker information. However, the results is this work depict that it may not be useful for singer identification application. Thus, the consideration of Wave-U-Net for audio-source separation to have an improved speaker characterization is justified in this work.

Table 2: *Performance comparison of different audio-source separation methods under Fold 2 on Artist20 corpus.*

System	Accuracy (%)
Harmonic/percussive	85.84
CNN	78.54
Wave-U-Net	90.41

Table 3: *Performance comparison of baseline and proposed system under 10 seconds singing speech scenario.*

System	Accuracy (%)
Baseline	47.03
Proposed	68.04

5.3. Studies using short song segments

With the rising importance of practical applications in every field, the short utterance scenario in speaker recognition has gained attention in the recent years [35–39]. The same can be applicable in this case of identifying singers, when there is a short segment of song available. Therefore, we investigate our proposed framework with audio-source separation for identifying singers with short song segments. Instead of 5 minutes of song, in this case 10 seconds of singer's voice is considered to test with our proposed approach and the baseline without audio-source separation. We note that the same trained models described in Section 5.1 are used in this scenario. Table 3 shows the results for singer identification under short test song segment. It is visible that the proposed method achieves an accuracy of 68.04% against the baseline method accuracy 47.03%. This shows that the gain is more significant with the proposed framework when there is limited amount of data present to identify a singer. This further showcases that audio-source separation plays a major role for singer identification that is more evident for short song segments.

The studies presented here shows the importance of audio-source separation for singer identification. Consideration of Wave-U-Net as the most suitable one is validated by comparing to other available methods. Further, the gain achieved with short song segments strengthens the use proposed framework for singer identification.

6. Conclusion

This work focuses on proposal of a novel framework for singer identification using polyphonic songs, by incorporating the audio-source separation module. An audio-source separation method based on Wave-U-Net is used for separating the background accompaniment from the songs. The signing vocals thus obtained are then used to build the singer models with i-vector based approach. The studies are conducted on Artist20 database reveals that the proposed framework outperforms the baseline without audio-source separation by a large margin. Further, we demonstrate the effectiveness of considering Wave-U-Net over other audio-source separation methods by performing their studies for singer identification. The future work will focus to extend this work under singer verification application.

7. Acknowledgement

This research work is supported by Programmatic Grant No. A18A2b0046 and A1687b0033 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain).

8. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, pp. 12 – 40, 2010.

[2] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, Nov 2015.

[3] Y. E. Kim and B. Whitman, "Singer identification in popular music recordings using voice coding features," in *Proceedings of the 3rd international conference on music information retrieval*, 2002, pp. 164–169.

[4] C.-C. Liu and C.-S. Huang, "A singer identification technique for content-based classification of MP3 music objects," in *Proceedings of the eleventh international conference on Information and knowledge management*, 2002, pp. 438–445.

[5] T. Zhang, "Automatic singer identification," in *International Conference on Multimedia and Expo. ICME'03.*, vol. 1, 2003, pp. I–33.

[6] M. A. Bartsch and G. H. Wakefield, "Singing voice identification using spectral envelope estimation," *IEEE Transactions on speech and audio processing*, vol. 12, no. 2, pp. 100–109, 2004.

[7] N. C. Maddage, C. Xu, and Y. Wang, "Singer identification based on vocal and instrumental models," in *Proceedings of the 17th International Conference on Pattern Recognition*, vol. 2, 2004, pp. 375–378.

[8] W. Cai, Q. Li, and X. Guan, "Automatic singer identification based on auditory features," in *IEEE Seventh International Conference on Natural Computation*, vol. 3, 2011, pp. 1624–1628.

[9] M. Lagrange, A. Ozerov, and E. Vincent, "Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning," in *ISMIR*, 2012.

[10] T. L. Nwe and H. Li, "Exploring vibrato-motivated acoustic features for singer identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 519–530, 2007.

[11] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Singer identification based on accompaniment sound reduction and reliable frame selection," in *ISMIR*, 2005, pp. 329–336.

[12] W.-H. Tsai and H.-M. Wang, "Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 330–341, 2006.

[13] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," in *ISMIR*, 2007, pp. 375–378.

[14] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[15] H. Eghbal-Zadeh, B. Lehner, M. Schedl, and G. Widmer, "I-vectors for timbre-based music similarity and music artist classification," in *ISMIR*, 2015, pp. 554–560.

[16] D. P. Ellis, "Classifying music audio with timbral and chroma features." in *ISMIR*, 2007, pp. 339–340.

[17] K. Vijayan, H. Li, and T. Toda, "Speech-to-singing voice conversion: The challenges and strategies for improving vocal conversion processes," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 95–102, Jan 2019.

[18] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *13th International Conference on Digital Audio Effects*, 2010.

[19] 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.

[20] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *International Conference on Latent Variable Analysis and Signal Separation*, 2017, pp. 258–266.

[21] T. Chan, T. Yeh, Z. Fan, H. Chen, L. Su, Y. Yang, and R. Jang, "Vocal activity informed singing voice separation with the ikala dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015*, April 2015, pp. 718–722.

[22] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *ISMIR*, 2018.

[23] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *ISMIR*, 2017.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

[25] "Wave-U-Net," [Online; accessed 23-November-2018]. [Online]. Available: https://github.com/f90/Wave-U-Net

[26] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Mar 2005, pp. 629–632.

[27] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "A comparison of session variability compensation approaches for speaker verification," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 802–809, Dec 2010.

[28] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[29] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley and Sons, 2000.

[30] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[31] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr 1981.

[32] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[33] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[34] B. Sharma, C. Gupta, H. Li, and Y. Wang, "Automatic lyrics-to-audio alignment on polyphonic music using singing-adapted acoustic models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 396–400.

[35] R. K. Das and S. R. M. Prasanna, "Speaker verification from short utterance perspective: A review," *IETE Technical Review*, vol. 35, no. 6, pp. 599–617, 2018.

[36] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, , and M. Mason, "i-vector based speaker recognition on short utterances," in *Interspeech*, 2011, pp. 2341–2344.

[37] R. K. Das and S. R. M. Prasanna, *Speaker Verification for Variable Duration Segments and the Effect of Session Variability*. Lecture Notes in Electrical Engineering: Springer, 2015, ch. 16, pp. 193–200.

[38] A. Poddar, M. Sahidullah, and G. Saha, "Speaker verification with short utterances: a review of challenges, trends and opportunities," *IET Biometrics*, vol. 7, no. 2, pp. 91–101, 2018.

[39] R. K. Das and S. R. M. Prasanna, "Investigating text-independent speaker verification systems under varied data conditions," *Circuits, Systems, and Signal Processing*, Jan 2019.