Transactions of the International Society for Music Information Retrieval

XX, X., XX, X., and XX, X. (2021). Audiovisual Singing Voice Separation, *Transactions of the International Society for Music Information Retrieval*, V(N), pp. xx–xx, DOI: https://doi.org/xx.xxxx/xxxx.xx

ARTICLE TYPE

Audiovisual Singing Voice Separation

XXXX XX*, XXXXXX XXXX†, and XXXXXX XXXX*

Abstract

Separating a song into vocal and accompaniment components is an active research topic, and recent years witnessed an increased performance from supervised training using deep learning techniques. We propose to apply the visual information corresponding to the singers' vocal activities to further improve the quality of the separated vocal signals. The video frontend model takes the input of mouth movement and fuses it into the feature embeddings of an audio-based separation framework. To facilitate the network to learn audiovisual correlation of singing activities, we add extra vocal signals irrelevant to the mouth movement to the audio mixture during training. We create two audiovisual singing performance datasets for training and evaluation, respectively, one curated from audition recordings on the Internet, and the other recorded in house. The proposed method outperforms audio-based methods in terms of separation quality on most test recordings. This advantage is especially pronounced when there are backing vocals in the accompaniment, which poses a great challenge for audio-only methods.

Keywords: Source separation, audiovisual analysis, singing performance.

1. Introduction

Vocal performance is an important art form of music. The task of singing voice separation is to isolate vocals from the audio mixture, which contains other instrumental sounds that help to define the harmony, rhythm, and genre. Singing voice separation is often the first step towards many application-oriented vocal processing tasks including pitch correction, voice beautification, and style transfer, as implemented in some mobile Apps such as WeSing and Smule. It is also often a pre-processing step for other research tasks such as singer identification (Berenzweig et al., 2002), lyrics alignment (Fujihara et al., 2006), and tone analysis (Fujihara and Goto, 2007).

There are various scenarios when video recordings are available for singing performances, such as operas, music videos (MV), and self-recorded singing activities. In pop music, creative visual performances give artists a substantial competitive advantage. Moreover, due to the rapid growth of Internet bandwidth and smartphone users, videos of singing activities are becoming popular in a number of video sharing platforms such as TikTok and Instagram.

Visual information, e.g., lip movement, has been incorporated and shown its benefits in speech signal processing, such as audiovisual speech separation (Lu et al., 2019), enhancement (Afouras et al., 2018), and recognition (Petridis et al., 2018). Visual information has also been incorporated in music analysis (Duan et al., 2019), such as source association (Li et al., 2019, 2017a,c), source separation (Zhao et al., 2019), multi-pitch analysis (Dinesh et al., 2017), playing technique analysis (Li et al., 2017b), cross-modal retrieval (Li and Kumar, 2019) and generation (Chen et al., 2017; Li et al., 2018). For singing performances, however, little work has been done. It is reasonable to think that visual information would also help to analyze singing activities, and in particular, separate singing voices from background music. This is based on the fact that mouth movements and facial expressions of the singer are often correlated with the singing voice signal fluctuations. The advantages of audiovisual analysis over audio-only analysis can be best reflected on songs with multiple vocal sources while only one source is considered as the separation target, e.g., songs with backing vocals in the accompaniments. However, to what extent does the incorporation of visual information help singing voice separation is still a question. Different from speech signals, singing voices (except for rap music) generally change slower (Mesaros and Virtanen, 2010), showing less frequent matching with mouth movements (Cadalbert et al., 1994). Furthermore, some musically important fluctuations of the singing voice such as pitch modulations show little, if any, correlation with mouth movements (Connell et al., 2013).

Therefore, it is our intention to answer the follow-

*Department of XXX, University of XX, NY, USA.
†XXX company

ing research question in this paper: *Can visual information about the singer improve singing voice separation, and if yes, how much?* It is noted that while traditional singing voice separation tasks (e.g., SiSEC[1], MIREX[2], or AICrowd Music Demixing Challenge[3]) define all vocal components in a song as the singing voice, in this work we define it as separating the solo singing voice from the accompaniments, where the accompaniments may contain backing vocals. We argue that our definition is more rational to the nature of music as it separates solo, typically presenting the main melody, from accompaniment, typically presenting harmony. Separating the solo voice enables many applications such as solo vocal pitch correction (Grell et al., 2009) or vocal effects appliance for the soloist without affecting the backing vocal sources. The solo singing voice separation problem is somewhat similar to speech enhancement with babble noise (Vincent et al., 2018). However, music accompaniment is typically much louder and richer in timbre than background noise in speech enhancement settings. In addition, music accompaniment, especially backing vocal, shows very strong correlations with the solo vocal signal. They make the problem at hand very challenging.

To answer the above-mentioned research question, we design an audiovisual neural network model to separate the solo singing voice from the accompaniments that may contain backing vocals. This network model takes both the audio mixture signal and the mouth region of the singing video as input. The audio processing sub-network is designed based on the MMDenseLSTM (Takahashi et al., 2018b), the champion of SiSEC2018 and the best officially evaluated system by the time of mid 2021. The visual processing sub-network uses convolutional and LSTM layers to encode mouth movements of the singer. The audio and visual encodings are fused before they are used to reconstruct the solo singing magnitude spectrogram. The training target of the proposed audiovisual network is to minimize the Mean-Square-Error (MSE) loss of the magnitude spectrogram reconstruction of the solo singing voice. To facilitate the network to learn audiovisual correlation of singing activities, we add extra vocal signals irrelevant to the solo singer to the audio mixture during training. To investigate the benefits of visual information, we compare the proposed audiovisual model with several state-of-the-art audio-based singing separation methods and an audiovisual speech enhancement method. We further vary the architecture and input of the visual processing sub-network to compare their performances.

One challenge we encounter in this work is the lack of audiovisual datasets of singing. For training, this can be addressed by randomly mixing solo singing videos downloaded from the Internet with irrelevant accompaniment music. We download *a cappella* audition vocal performance videos and randomly mix their audio with accompaniment audio tracks from the MUSDB18 dataset to generate mixtures. We name this the *Audition-RandMix* dataset, and partition it into training, validation and test subsets. For evaluation on real songs, however, we need audiovisual recordings of singing with its relevant accompaniment music in separate tracks. To our best knowledge, no such dataset exists. Therefore, we record a new audiovisual dataset named *URSing*, where singers are recruited to sing along with prepared accompaniment tracks.

We conduct experiments on both the Audition-RandMix test set and the URSing dataset. Results on both sets show that the proposed audiovisual method outperforms baseline methods in most test conditions, no matter if the accompaniment tracks contain the backing vocals or not. We further conduct subjective evaluations on a cappella video performances in the wild to prove the advantages of our proposed method.

The contributions of this paper include:

- The first work to incorporate visual information to the state-of-the-art music source separation framework to address the singing voice separation problem,
- A proposal of solo voice separation where backing vocal components, if exist, are regarded as accompaniment tracks, which better fits many application scenarios, and
- The first audiovisual singing performance dataset, URSing, free for download[4].

2. Related Work

2.1 Singing Voice Separation

Early methods for singing voice separation include non-negative matrix factorization (Vembu and Baumann, 2005), adaptive Bayesian modeling (Ozerov et al., 2005, 2007), robust principal component analysis (Huang et al., 2012; Chan et al., 2015), and autocorrelation (Rafii and Pardo, 2011). Some methods address the singing separation problem using extra information such as vocal pitches (Hsu et al., 2012) or voice activities (Chan et al., 2015). Recently, deep learning based methods are proposed to model convolutional (Chandna et al., 2017) or recurrent structures (Huang et al., 2014; Uhlich et al., 2017) of magnitude spectral representations of music signals. Some works also learn to reconstruct spectral phases in addition to magnitudes (Takahashi et al., 2018a; Choi et al., 2019), while others directly work on time-domain waveforms with an end-to-end training strategy (Lluis et al., 2019; Stoller et al., 2018). Official blind evaluations and comparisons of these meth-

[1] A community-based signal separation evaluation campaign. https://sisec18.unmix.app/#/

[2] Music Information Retrieval Evaluation eXchange. https://www.music-ir.org/mirex/wiki/MIREX_HOME

[3] https://www.aicrowd.com/challenges/music-demixing-challenge-ismir-2021

[4] https://sites.google.com/view/ursing/home

ods can be referred in the SiSEC2018 (Stöter et al., 2018), where the best performing method MMDenseLSTM (Takahashi et al., 2018b) uses a DenseNet structure with a recurrent structure to process magnitude spectrograms. Later more systems are proposed and open-sourced with comparable or better results, such as Open-Unmix (Stöter et al., 2019), Spleeter (Hennequin et al., 2019), D3Net (Takahashi and Mitsufuji, 2021), DEMUCS (Défossez et al., 2019), LaSAFT (Choi et al., 2021), where DEMUCS is ranked best (referring to "Browse State-of-the-Art"[5], an unofficial platform to collect and compare all music separation results). More recently proposed music separation systems can be referred in the AICrowd Music Demixing Challenge, another official contest to conduct blind evaluations following SiSEC2018.

2.2 Audiovisual Source Separation

Most audiovisual separation works are proposed for speech signals. For speech separation, one challenge is the permutation problem where the separated components need to be assigned to the correct talkers. Lu et al. (2018) specifically address the problem by applying the visual information as a post-processing step to adjust the separation mask. Later the same group proposes to fuse the visual information to an audio-based deep clustering framework to propose an audiovisual deep clustering model for speech separation (Lu et al., 2019). Another work is described in (Ephrat et al., 2018), where the input is the mixture spectrogram and the face embeddings of all the appeared speakers in the audio sample. The training target is the complex mask that can be applied to the original spectrogram to recover the complex spectrogram of each speaker. It is noted that speech separation algorithms typically assume a noiseless or less noisy environment in which speech signals are mixed. In addition, speech signals to be separated are typically assumed to be from different speakers. Both assumptions are not true in solo singing separation, as the background music is often quite strong and the backing vocal often comes from the same singer as the soloist (?).

Speech enhancement aims at separating speech signal from background noise. It is more relevant to singing voice separation from background music considering the foreground-background relations of sources. Hou et al. (2018) address the speech enhancement problem using a two-stream structure that takes both noisy speech and frames of the cropped mouth regions as inputs to compute their features. These features are then concatenated by a fusion network which also outputs corresponding clean speech and reconstructed mouth regions. Another audiovisual speech enhancement work proposed in (Afouras et al., 2018) uses 1D convolutional layers to reconstruct the magnitude spectrogram of the clean speech and uses it to further estimate its phase spectrogram. The input of the visual branch is the feature embeddings on the lip region that are pre-trained on lip reading tasks.

[5] https://paperswithcode.com/sota/music-source-separation-on-musdb18

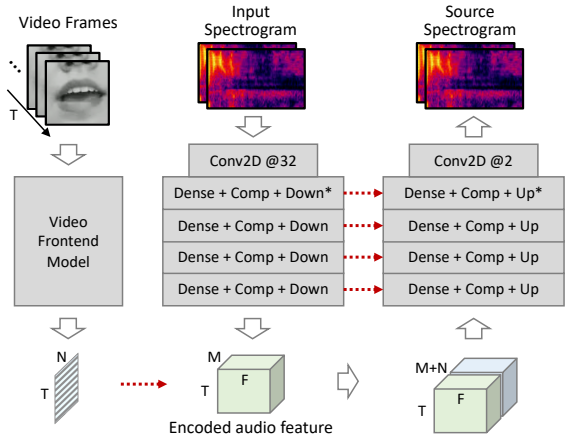

Figure 1: The proposed model structure. Dashed arrows denote the concatenation operation. Downsample/upsample are applied to both time and frequency dimensions in the outer layers (marked by *), while they are only applied to the frequency dimension in the inner layers.

Less work has been proposed for audiovisual music separation. Parekh et al. (2017) apply non-negative matrix factorization (NMF) to separate string ensembles, where the bowing motions are used to derive additional constraints on the activation of audio dictionary elements. This method, however, is only evaluated on randomly assembled video scenes of string instruments where distinct bowing motions of each player are clearly captured. Zhao et al. (2018) propose to learn static audiovisual correspondence with cross-modal source localization; The correlation between each pixel in a given video frame and the sound component can be constructed. Followup works for separating music sources include recognizing the audiovisual correspondence from visual motions (Zhao et al., 2019) and gestures (Gan et al., 2020) in music instrument performances. Similar works have been proposed in (Gao and Grauman, 2019) and (Tzinis et al., 2021), where correspondence between audio and video are learned in unsupervised manner to guide source separation. This line of research achieves promising results in audiovisual music separation, but have not addressed singing voice separation.

3. Method

3.1 Network Architecture

The proposed model takes the input of the magnitude spectrogram of the original audio mixture which contains both the solo vocal and background music, and the mouth region of the video frames corresponding to the vocals. The output is the magnitude spectrogram of the source audio of vocals. It builds upon a state-

of-the-art audio separation model named MMDenseLSTM (Takahashi et al., 2018b) with a video front-end model. The MMDenseLSTM model consists of convolutional layers stacked into dense blocks, which alternates downsample/upsample layers to form a multi-scale structure. It first embeds an input magnitude spectrogram into an encoded feature space and decodes it to recover the separated magnitude spectrogram. Skip connections as U-Net structure are applied. This "encoder-decoder" structure with skip connections is widely applied in several music separation models (Jansson et al., 2017; Stoller et al., 2018; Zhao et al., 2019; Liu and Yang, 2018). The video front-end model extracts visual features from mouth movements, which are fused with the encoded audio feature. The network structure is illustrated in Figure 1. We explain each part of the model in detail as follows.

3.1.1 *Audio Separation Model*

Following MMDenseLSTM, our audio separation model consists of:

- Dense Block. It applies 2D convolutional layers[6] and the output feature maps of all layers are concatenated with each other along the channel dimension. This structure reuses the feature maps from previous layers and greatly reduces the model size.
- Compression layer. It is a convolutional layer with 1×1 kernels applied after each dense block. We use a compression ratio of 0.2, which means that the number of feature maps (channels) is reduced by 80% after each compression layer. We apply a compression layer right after each dense block, which improves the model compactness.
- Downsample/Upsample. These layers are applied after compression layers to resize the feature maps without changing the the number of channels. Downsample layers are average pooling with 2×2 kernels after the first compression layer, and 1×2 kernels in the following layers. In other words, downsampling is performed along both the time and freuqncy dimensions in the first layer, but only to the frequency dimension in other layers. Symmetrically, upsample layers apply transposed convolutional layers with 2×2 kernels and strides at the last upsample layer but 1×2 for the other layers. Different from (Takahashi et al., 2018b) where downsample/upsample always addresses both time and frequency dimensions in multiple scales, our proposed strategy downsamples/upsamples the time dimension only once, making the audio stream have the same frame rate as the video stream. The encoded audio spectrogram feature is denoted as $\mathbf{S}_A \in \mathbb{R}^{M\times T\times F}$, with the channel ($M$), downsampled time ($T$), and frequency ($F$) dimensions. Skip connections are added as concatenations on the corresponding layers with the same feature map size, as the U-Net structure.
- Multi-Band. Following (Takahashi and Mitsufuji, 2017), we also equally divide the spectrogram into a low-frequency band and a high-frequency band and apply the above-mentioned U-Net encoder-decoder structure on each sub-band. The dense blocks of low-frequency band have a higher channel number. Detailed parameters can be referred to (Takahashi and Mitsufuji, 2017).

The audio separation model described in this section is the same as the method proposed in (Takahashi et al., 2018b), except the downsample/upsample parameters which are adjusted for audiovisual fusion when visual inputs are applied.

Note that since SiSEC2018, new methods are proposed to advance the state of the art of the music separation tasks. However, we still take MMDenseLSTM as an important reference. The reasons are twofold. First, after SiSEC2018 there are no public music separation contest running a blind evaluation for direct comparison of different methods. MMDenseLSTM is the most reliable framework to refer as an audio sub-network to build our audiovisual separation model, especially when we aim to prototype the first audiovisual vocal separation work instead of yielding a high rank in the traditional music separation task. Second, MMDenseLSTM has small model size, which is especially beneficial for our audiovisual fusion and experiments when the audiovisual singing performance dataset is not in a large scale.

3.1.2 *Video Front-End Model*

We propose to apply a visual branch to parse the input video stream and fuse it with the encoded audio features. The video stream is a sequence of mouth region RGB images in consecutive video frames. The video front-end model has four convolutional layers, followed by a fully connected layer, an LSTM layer, then a final fully-connected layer, with the parameters of Conv2D@16 (channel number is 16), Conv2D@16, Conv2D@32, Conv2D@32, FC@256, LSTM@128, and FC@N. N is the dimension of the encoded feature vector for each video frame. The input video stream with T frames results in a feature map $\mathbf{S}_V \in \mathbb{R}^{N\times T\times 1}$. There is no pooling operation along the time dimension thus the temporal information is preserved. Raw RGB values are normalized to zero mean and unit variance.

3.1.3 *Audiovisual Fusion*

The extracted visual feature map from the video branch is fused with the encoded audio spectrogram feature map $\mathbf{S}_A \in \mathbb{R}^{M\times T\times F}$. To do so, the visual feature map $\mathbf{S}_V \in \mathbb{R}^{N\times T\times 1}$ is broadcast along the third dimension and then concatenated with the audio feature

[6] A convolutional layer includes BatchNormalization+ReLU+Conv2D throughout the paper.

to obtain the audiovisual feature $\mathbf{S}_{AV} \in \mathbb{R}^{L \times T \times F}$, where $L = M + N$ is the concatenated channel dimension. Note that the temporal information from both the audio and video branches is correlated during this fusion; This is different from some works where audiovisual fusion is performed on feature maps that aggregate information along time.

In addition to minor structural changes, we also drop the LSTM structure of the original MMDenseLSTM model (Takahashi et al., 2018b) when we design the audio branch of our proposed model. This follows the observation that the addition of the LSTM structure does not achieve substantial improvement in SiSEC2018 yet the number of parameters would be increased significantly for audiovisual fusion.

3.2 Training

We train the model to predict the magnitude spectrogram of the source signal and use the original mixture's phase to recover the time-domain waveform. Many spectral-domain source separation methods, especially those for speech signals, use a spectrogram mask as the training target; This mask is then multiplied elementwise with the mixture signal's magnitude spectrogram to recover the source magnitude spectrogram. For music separation, some recent works train networks to directly output the source magnitude spectrogram (Uhlich et al., 2017; Takahashi et al., 2018b) using a Mean-Squared-Error (MSE) loss. We follow the same way to take the source magnitude spectrogram as the training target. However, we have a mask operation as one layer of the model that regularizes the feature maps into the range of [0, 1] using a Sigmoid function and multiplies the mask layer with the input spectrogram to get the model output. We find that this regularization step is beneficial for for our audiovisual separation model. We have a comparative experiment in Section 5.4.

Compared to the audio mixture input, the visual input provides much less information about the source signals, therefore, the training loss may not be propagated back sufficiently into the visual branch, making the audiovisual network difficult to train. One way to address this is to explicitly learn audiovisual matching, either through pre-training (Lu et al., 2018) or early audiovisual fusion (Lu et al., 2019). Another way might be to add visual reconstruction as another training target, leading to a chimera-like network structure (Hou et al., 2018).

In this work, we address this problem by adding some extra vocal components to the original mixture, which are not related to the mouth movements and thus are not included in the target vocal spectrogram. This is similar to adding an additional speaker in the training data in the case of audio-visual speech separation (Ephrat et al., 2018), which forces the model to learn audiovisual correlations after the fusion and only separate the vocal components that are related to the visual input. Note that in the training samples all of the vocal and accompaniment components are randomly mixed, so neither the extra vocal components or the solo vocal components have harmonic relations with the accompaniment tracks. In the experiments, we show that the strategy of training with randomly generated vocal-accompaniment pairs performs decently on real songs.

4. Dataset

Since there is no publicly available audiovisual singing voice dataset containing isolated vocal tracks, we collect our own data for training and evaluating the proposed method.

4.1 A Cappella Audition Vocals (AAV)

We curated 491 YouTube videos of solo singing performances by querying the YouTube search API with the keyword "Academic Acappella Audition". We only selected video excerpts where the singer faces the camera and sings without accompaniment. The total length of these excerpts is about 8 hours. As it is difficult to find relevant and appropriate accompaniment tracks, in our experiments we simply randomly chose instrumental accompaniment tracks (from the "accompaniments" track in the MUSDB18 dataset) and mixed them with the solo singing excerpts to create singing-accompaniment mixtures. To prepare the extra vocal components, we also download 2 hours of chorus recordings from YouTube.

The randomly mixed samples are used for training, validation, and evaluation. Before the mixing process, vocals in AAV are divided into training/validation/evaluation sets roughly as 8:1:1 (50 tracks for evaluation). Accompaniment tracks from MUSDB18 (which contains a wide range of music genres and instrument types) are also divided into the three sets following the official way (also 50 tracks for evaluation). Then mixing is applied on each split independently to form the training/validation/evaluation sets. Volume of each track is normalized using the root-mean-square (RMS) value. For training and validation sets, each track is split into short samples (around 2.5 seconds) for random mixing, resulting in a massive amount of mixed samples. We do not balance the volume of each individual sample so the mixing may have different SNRs. During training, for half of the training/validation samples we add extra vocal components that are not related to the mouth movements to encourage the model to learn audiovisual correlations. Half of the extra vocal components are solo vcoals from other irrelevant singers in the AAV dataset, and the other half are samples from the chorus recordings. We apply a random gain between -6dB to 0dB for the extra vocal components, considering that in real singing performances, the solo vocal usually still takes the lead position. For evaluation, mixing is

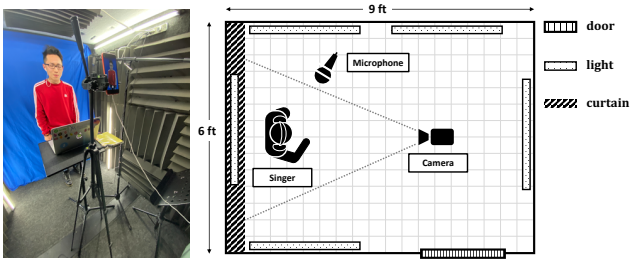

Figure 2: A sample photo and floor plan of the sound booth for the recording process of the URSing dataset.

performed on a random bijection between the 50 vocals and 50 accompaniments. For each mixture, we pick a 30-second excerpt (with both vocal and accompaniments present) for evaluation, following the same strategy as the MUSDB18 dataset. This set is referred to as "Audition-RandMix" in the following experiments. For the same 50 mixtures, we randomly add extra vocals following the same strategy as preparing the training set, which is referred as "Audition-RandMix (v+)", in order to explore the model performance in more challenging cases.

Note that all the samples in this condition are artificial mixtures that cannot represent real songs, since vocals and accompaniments are not correlated in harmony. However, training music separation model on artificial mixtures can be still helpful to separate real songs (Song et al., 2021), and artificial mixtures have been also used as evaluate data for music separation tasks (Luo et al., 2017).

4.2 URSing

To evaluate the proposed method in more realistic singing performances, we create the XXX (*hidden for anonymous review*) Multi-Modal Singing Performance Dataset (URSing). In this paper, we only use the URSing dataset for evaluation. A brief description of the creation process is described below.

4.2.1 Singer Recruiting

Singers are students at the University of XXX (*hidden for anonymous review*). Audition is performed to filter out unqualified singers who could not sing in tune. Each participant receives $5 for recording each song, and is allowed to record up to 5 songs. Each singer has signed a consent form about ethical approval, which authorizes the release of the dataset for research purpose.

4.2.2 Piece Selection

To ensure high recording efficiency, the singers pick their own songs and their favorite accompaniment tracks to sing along. We do not put constraints on song genres, but filter out songs of which the accompaniment tracks are of low sound quality.

Figure 3: Examples of video frames of the URSing dataset and cropped mouth region pictures as the input to the video branch of the proposed method.

4.2.3 Recording

To ensure synchronization, the singers listen to the accompaniment track through earphones while recording their singing voice. Their voices are recorded using an AT2020 condenser microphone hosted by Logic Pro X, and their videos are recorded using iPhone 11. The recording is conducted in a semi-anechoic sound booth. A sample photo and the floor plan of the sound booth are shown in Figure 2.

4.2.4 Post-processing

For each solo vocal recording we use the following plug-ins to simulate the typical audio production procedure in commercial recordings: a) static noise reduction (*Klevgrand Brusfri* and *Waves X-noise*), b) pitch refinement (*Melodyne*), c) sound compression (*Fabfilter Pro-C 2*), and d) reverberation (*Fabfilter Pro-R*). We also adjust the vocal volume to balance it with the accompaniment track. Beyond this, we do not perform any other editing on the audio recording (e.g., time warping or rhythmic refinement) to preserve the synchronization with the visual performance. To synchronize the audio recording captured by the AT2020 microphone with the video recording captured by the smartphone, we use the audio recording captured by the built-in microphone of the smartphone as the bridge, through cross correlation.

4.2.5 Annotation

Since the mouth movements are mostly relevant to the singing performance, we provide the annotations of the mouth regions in the dataset. This is performed using the Dlib library (King, 2009), an automatic tool for facial landmark detection, followed by manual check. The mouth region is represented as a square bounding box with the side length equal to 1.2 times of the maximum horizontal distance for all mouth landmarks.

This results in 65 songs, totaling 4 hours of audiovisual recordings of singing performance. For each song, we provide:

- The audio recording of the solo singing voice (in WAV, 44.1 KHz, 16 bits, mono).
- The corresponding accompaniment audio track (in WAV, 44.1 KHz, 16 bits, mono or stereo).
- The video recording of the soloist's upper body

(in MP4, 1080P portrait, 29.97 FPS).
- The annotations of mouth regions for each video frame.

Note that when we prepare the accompaniment tracks, we do not avoid the tracks containing backing vocals, as they are the challenging and useful cases to study in this paper. Example video frames and the cropped mouth region pictures using the provided mouth region annotations are provided in Figure 3.

We also choose a set of 30-sec excerpts where both solo vocal and accompaniment tracks are prominent to form a benchmark evaluation set. Specifically, for each of the 65 songs, we choose one 30-sec excerpt without backing vocal and one with back vocal, if such excerpts are available. We provide this information in the metadata. This results in 54 excerpts with accompaniment tracks that only contain instrumental components (referred as "URSing" in the following experiments) and 26 excerpts with accompaniment tracks that also contain backing vocals (referred as "URSing (v+)". The latter, presumably, are more challenging for solo vocal separation and more useful for showing advantages of audiovisual methods. In this paper, since we do not use any songs from URSing for training, we only use these 30-sec excerpts for evaluation.

5. Experiments

5.1 Implementation Details

For audiovisual singing videos, audio is downsampled to 32 KHz. We use a frame length of 1024 and a hop size of 640 (20 ms) for spectrogram calculation. Magnitude spectrogram has been converted to logarithm scale followed by normalization along each frequency axis, which better weighs the contribution of high frequency bins. Video data is converted to 25 FPS (equivalent to 40 ms frame hop size). For the origianl singing performance videos, the mouth regions are cropped as square bounding box using the Dlib library (King, 2009) and then interpolated with the size of 64 × 64. RGB videos have been converted to grayscale. Each training sample is 2.56 seconds long, containing 128 audio frames and 64 video frames. The input/output audio spectrogram has the shape of 2×128×513 (channels × frames × frequency bins), and each input video stream has the shape of 64×64×64 (frames × width × height). We use RMSProp optimization with a learning rate of 0.01. The learning rate decays every 5 epochs by multiplying with 0.8. We use batch size of 8 for training on a TITAN X GPU with 11.9 GB graphic memory. It takes about 40 hours to train for 50 epochs. We adopt early stopping when the validation loss does not decrease for 10 consecutive epochs.

For evaluations, we calculate the signal-to-distortion ratio (SDR) between the separated vocal waveforms and the ground-truth ones using the BSS Eval toolbox V4, same as the evaluation measure applied in SiSEC2018. Specifically, for each 30-sec evaluation excerpts, we calculate the median SDR over all 1-sec audio segments.

5.2 Baselines

We first use the original mixture recording (referred as "MIX" in the experiments) as the separated vocal for evaluation on our dataset. This sets lower bounds of separation results without any separation techniques. Then we apply two oracle filtering techniques that utilize ground-truth source signals: The ideal binary mask (IBM) assigns each time-frequency bin to the predominant source. The ideal ratio mask (IRM) distributes the power of each time-frequency bin into different sources according to the power ratio of the ground-truth sources. The IBM and IRM set upper bounds for time-frequency masking-based source separation methods.

We then compare our proposed method with several audio-based music separation methods as baselines.

- RX7. A commercial software developed by iZotope[7]. We apply batch processing of the "music rebalance" function with the preset "isolate vocals" on "medium" level. Training data for the model inside this software is unknown to us.
- UMX (Stöter et al., 2019). An open-sourced separation tool known as "Open-unmix" . The model employs the BLSTM structure and is trained on the MUSDB18 dataset.
- Spleeter (Hennequin et al., 2019). An open-sourced music separation method with a CNN+Unet model trained on their in-house dataset of 24,097 songs.
- Spleeter-train. Same model as "Spleeter" but trained on our Audition-RandMix dataset using the same conditions as those for our proposed audiovisual method as a direct comparison.
- Demucs. An open-sourced music separation method with U-Net and LSTM structure to process the signal in waveform domain. It achieved the best separation performance among all open-sourced tools up to date.
- MMDenseLSTM (Takahashi et al., 2018b). The method that achieved the best results in SiSEC2018, even without training on extra data. We implemented this method from scratch. Our implementation has been validated by achieving 7.44dB of SDR of vocal separation results on the MUSDB18 test set. We then trained this model on our Audition-RandMix dataset as a direct comparison.

We also implement an audiovisual speech enhancement method named AVDCNN proposed in (Hou et al., 2018). This method applies 2D CNNs to take noisy speech and the mouth region visual recording as inputs, fuses encoded audio and visual features to output

[7] https://www.izotope.com

Method	UMX	Spleeter	MMDenseLSTM	AVDCNN	Proposed
Parameter ($\times 10^6$)	8.5	19.7	1.22	11.3	2.05

Table 1: Comparison of model size of different methods.

the enhanced speech signal as well as reconstructed video frames of mouth movements. After the fusion layers, we used LSTM instead of fully-connected layers as used in (Hou et al., 2018), which shows higher performance in our experiment scenarios.

We choose audiovisual speech enhancement instead of audiovisual speech separation as the baseline, because we believe that speech enhancement is more relevant to singing voice separation from background music in terms of foreground-background relations of sources, as explained in Section 2.2. In addition, audiovisual speech separation usually assumes the availability of all talkers, while in our setting, only the video of the solo singing voice is used.

We present the model sizes of baseline models that are open-source or implemented by us in Table 1, together with that of the proposed model.

5.3 Objective Evaluation on Synthetic Mixtures

We evaluate the comparison methods on the four test sets described in Section 4: Audition-RandMix, Audition-RandMix (v+), URSing, and URSing (v+). "v+" means that the accompaniments contain vocal components. Note that all these songs are synthetic mixtures, e.g., Audition-RandMix is random mixed samples and URSing is recorded in controlled environment. Boxplots of SDR results are shown in Figures 4, where each data point in the boxplots is the median SDR of the separated vocal of all 1-sec segments of a 30-sec excerpt. The horizontal line inside each box indicates the median value across all excerpts. Several interesting observations can be made from the results.

5.3.1 Benefits of Visual Information

The proposed method outperforms all audio-based separation baselines in most of the evaluation sets. This shows the advantage of incorporating visual information about the singer's mouth movement for solo singing voice separation. Among the audio-based baseline methods, MMDenseLSTM is much stronger than RX7, because MMDenseLSTM is our own implementation and is trained on our dataset while RX7 is not. However, Spleeter slightly outperforms our proposed system on the URSing set. We believe that this is because Spleeter is trained on a much larger in-house dataset that contains 24,097 songs totalling 79 hours. This is verified by the fact that, Spleeter-train, the same model as Spleeter but trained on our dataset as a fair comparison, does not outperform MMDenseLSTM nor the proposed method. We suggest that this is because our proposed model (and MMDenseLSTM) has a much smaller model size than Spleeter, making it less prone to overfitting given a small training set.

Comparing songs with backing vocals (Audition-RandMix (v+) and URSing (v+)) to songs without backing vocals (Audition-RandMix and URSing), we can see that the outperformance of the proposed method is better pronounced on songs with backing vocals. Wilconxon signed-rank tests show that the improvement of the proposed method over MMDenseLSTM on Audition-RandMix (v+) and URSing (v+) are both significant, with p values of 6.2×10^{-3} and 4.3×10^{-2}, respectively. We argue that this is because audio-only methods tend to assign all the vocal components to the separated singing voice, while the proposed audiovisual method learns to only separate the vocal signals that are correlated to the solo singer's mouth movements.

The reason that the improvement is more pronounced on Audition-RandMix (v+) than on URSing (v+), we argue, are twofold: 1) backing vocals in URSing (v+) are not as strong as the intentionally added backing vocals in Audition-RandMix (v+), and 2) backing vocals in URSing (v+) often overlap with solo vocals and share the same lyrics, showing high correlations with the mouth movements of the solo singer, while the added backing vocals in Audition-RandMix (v+) are irrelevant to the solo vocal.

Figure 5 shows one 10-sec sample as an extreme case to compare the spectrograms of audio-based MMDenseLSTM method and the proposed audiovisual method when backing vocal components are strong (e.g., the middle part of the sample). We also show the mouth movement in several frames throughout this excerpt. It can be seen that MMDenseLSTM recognizes the backing vocal components in the middle frames as the solo vocal, while the audiovisual method suppresses those components significantly.

On songs without backing vocals, the outperformance of the proposed method can still be observed. Subjective listening by the authors suggests that the visual information helps to reduce high-frequency percussive sounds from the solo vocal, as the former do not correlate with mouth movements well.

5.3.2 Superiority of Proposed Audiovisual Architecture

The proposed method outperforms the audiovisual speech enhancement baseline significantly in all evaluation sets. Note that the baseline is trained and evaluated on the same dataset as the proposed method. This shows the superiority of the proposed network architecture on the solo singing voice separation task. In particular, we argue two main reasons. First, the proposed model utilizes the commonly used U-net structure with skip connections, which generally achieves good results in music separation (Jansson et al., 2017; Stoller et al., 2018; Takahashi and Mitsufuji, 2017). Second, in our audiovisual fusion scheme we preserve the temporal correspondence, which prevents a sub-

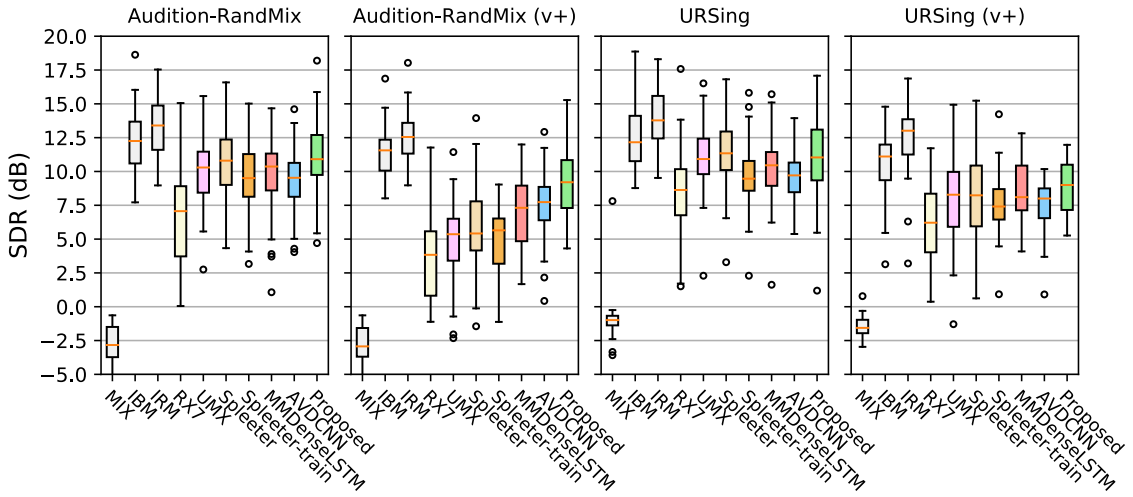

Figure 4: The SDR (dB) comparison on separated solo vocals with different methods on different evaluation sets. ("v+" denotes for songs where accompaniments contain vocal components.)

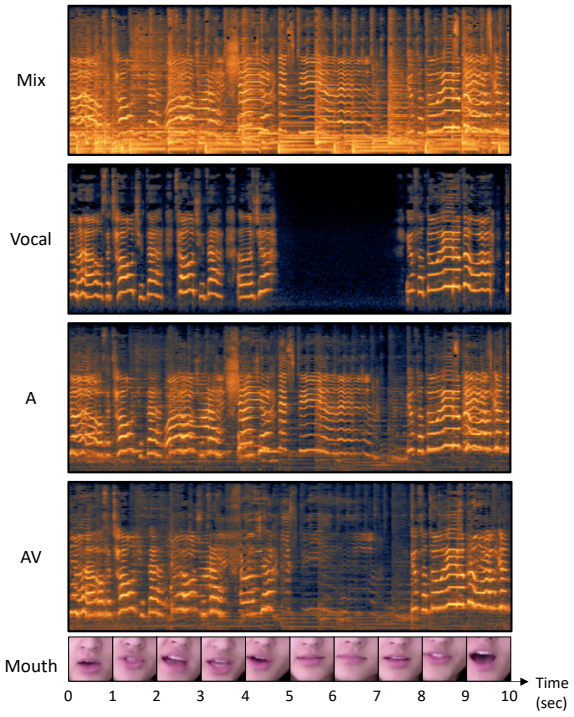

Figure 5: One 10-sec example comparing audio-based separation (MMDenseLSTM) with audiovisual separation (proposed) on a song excerpt with strong backing vocals. The four spectrograms from top to bottom are original mixture, ground-truth vocal, audio-based vocal separation result from (Takahashi et al., 2018b), and audiovisual vocal separation result from the proposed method. This sample result has 10-sec long, and one mouth frame of each second is attached.)

stantial increase of the number of trainable parameters in the fusion layer. This is important when the DenseNet-based audio sub-network has a small model size. The variations of different video sub-networks, however, does not make much difference on the separation performance, as we analyzed in Section 5.4.

5.3.3 *Limitations and Room for Improvement*

Compared with reported SDR values in SiSEC2018, the SDR values in Figure 4 are much higher. For example, MMDenseLSTM reaches over 10 dB on URSing but only less than 7 dB in SiSEC2018 (method "TAK1" in (Stöter et al., 2018)). We argue that the songs used in SiSEC2018 (i.e., the MUSDB18 dataset) are professionally recorded, mastered and mixed vocals. They often contain complex components such as polyphonic vocals, background humming, and strong reverberation. They are mastered and mixed by professional music producers to intentionally make them better fused into the background music. In contrast, the ground-truth vocals in our datasets are solo vocals recorded in controlled environments with limited vocal effects added. It is reasonable to believe that the benefits of visual information can be further demonstrated on more professionally produced songs. In addition, the performance difference between the Audition-RandMix test sets and the URSing test sets seems to be small for all methods, including the oracle results. This shows that randomly mixed songs, although lacking harmonic and rhythmic coherence, are not easier to separate than the more realistically mixed songs, suggesting that it may be reasonable to use randomly mixed songs to train the methods (Luo et al., 2017). However, whether this is still true for professionally produced songs is still a question.

On the other hand, there is still some gap between the proposed method and the oracle results on the SDR metric in our evaluation sets. It is likely that this gap

will be even bigger on professionally produced songs. This suggests that much work can be done to improve the separation performance. For example, time-domain separation for the audio branch may further improve the performance (Luo and Mesgarani, 2018).

5.4 Different Video Front-End Models

To investigate the key factors of the audiovisual separation framework and the robustness, we replace the proposed Conv2D+LSTM video front-end with several other widely-used visual feature extraction frameworks:

- No-mask. This experiment has the same video branch, but without a mask layer after the audio-visual fusion.
- Conv3D (Tran et al., 2015). The Conv3D model takes all the video frames from each sample as a feature map and a 4-th dimension is added as the channel dimension set as 64. We then apply 2 Conv2D layers (with the channel dimension 128 and 256) on each frame to share the channel dimension with Conv3D. Followed by pool operation and fully-connected layers, we obtain the video feature with the same dimension as $\mathbf{V}_{\text{Conv3D}} \in \mathbb{R}^{N\times T}$. Note that in this structure, the temporal information is only parsed at the very first Conv3D structure, since no recurrent network is applied.
- Dense+LSTM (Huang et al., 2017). Different from the proposed model, we replace the Conv2D layers with a dense block from the DenseNet structure. Each dense block has 2 layers with growth rate of 12. Then a Conv2D layer with 1×1 kernels is applied to compress the channel number to 32, resulting in the same feature dimension as the proposed CNN+LSTM model before feeding into the FC@256.
- Lip-reading. This variation uses a pre-trained model proposed in (Petridis et al., 2018) on the lip reading task on the LRW dataset (Chung and Zisserman, 2016). The original model structure consists of Conv3D, ResNet-34, and GRU. We only use the pre-trained model to extract the visual feature to integrate into our proposed audio-visual source separation model.

A comparison of different video front-end models is shown in Figure 6. It can be seen that the proposed (Conv2D+LSTM) model achieves the highest SDR values for most cases, but some video front-end models do not make much difference. Applying a mask layer is critical, as otherwise audiovisual method even degrades from the audio-based method. Note that for audio-based baseline method (MMDenseLSTM), we have also experimented models with a mask layer or not, but it does not make difference on the separation results. The Conv3D framework slightly degrades the performance, but still outperforms the audio-based

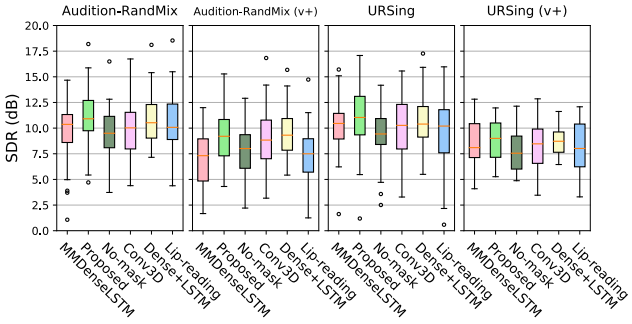

Figure 6: The SDR (dB) comparison on the separated solo vocal from the audiovisual method using different video front-end models.

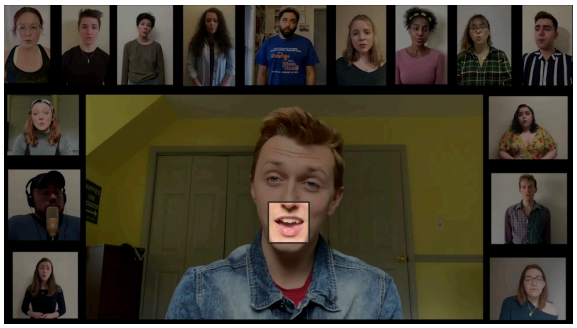

Figure 7: One sample frame of an a cappella song for subjective evaluation.

baseline method (MMDenseLSTM). One reason for this performance drop may be that in this framework, there is no recurrent structure, and the temporal evolution of visual information is only processed by the Conv3D structure. As the Conv3D structure takes the raw input of mouth frames, it may be sensitive to mouth position changes due to landmark detection errors. The model pre-trained on lip reading ranks the worst among the audiovisual models. This is because the lip reading model was trained on the LRW dataset where for each sample containing several words, only one word around the center frames is annotated as the training target. This makes the model only attend to the middle frames of a video excerpt, leading to limited guidance for the singing voice separation and even degradation from audio-based methods. We have also conducted experiments using the pre-trained lip reading model but finetuned on our separation task, but it does not boost the separation performance from our proposed video frontend model. It is possibly because lip movements in speech and singing are different.

5.5 Subjective Evaluation on Professional A Cappella Songs

In this section, We further evaluate the benefits of visual information incorporated in our proposed method on real a cappella songs in the wild. We collect 35 audiovisual a cappella recordings from YouTube. These collections represent the extreme cases where all the accompaniment components are vocals (except for sev-

Figure 8: Statistics of the 26 subjects' musical background related to the subjective evaluation.

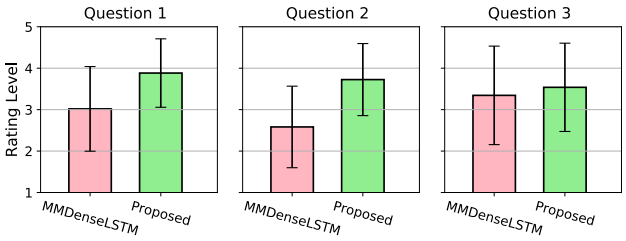

Figure 9: The subjective ratings of the separation quality in response to the three questions. Each error bar shows mean ± standard deviation.

eral cases where additional percussive instruments are also present), to study how much the proposed audiovisual method is advantageous while the audio-based method is very likely to fail. Here we use the MMDenseLSTM baseline as the audio-based method for comparison, which yields the best separation results among audio-based baselines. Most of these songs are chorus performance with a solo singer accompanied by harmonic vocals and/or vocal beatbox, while some are performance with multiple solo singers. We only keep the videos where the solo singer's mouth is visible and clear, without video shot transition for at least 10 seconds. A sample frame of one song is shown in Figure 7 with the mouth region of the targeted solo singer highlighted.

As we do not have access to the source tracks, we cannot evaluate the separation performance using common objective evaluation metrics. Instead, we conduct a subjective evaluation on the source separation quality ((Cartwright et al., 2016) and (Cartwright et al., 2018)) over 51 people. Some subjects are students or faculty from the University of XXX (*hidden for anonymous review*), others are subscribers from the International Society for Music Information Retrieval (ISMIR) community. Statistics of the subjects' music background is shown in Figure 8. Each survey asks a subject to rate 7 of the 35 songs, and each subject may take more than one surveys. For ratings from the same subject, we take the average to avoid bias. For each song, the subjects first watch a 10-sec excerpt of the original performance and then watch the same video twice with the solo singing voice separated by two different singing voice separation methods in a random order to rate the separation quality. Due to the variations across these songs, the original recording serves as a reference for a consistent scoring scheme. For each video we also highlight the mouth region of the target solo singer (see Figure 7) to help subjects focus on the corresponding solo voice. The specific evaluation questions are:

- Question 1: *What do you think about the overall separation quality for the targeted singer?*
- Question 2: *What do you think about the separation quality in terms of removing backing vocal accompaniments in the separated solo voice?*
- Question 3: *What do you think about the separation quality in terms of not introducing artifacts into the separated solo voice?*

The subjects need to answer each question using a scale from 1 to 5, where "1" represents *Very bad* and "5" represents *Very good*. The three questions are related to the common definitions of the three objective source separation evaluation metrics, SDR, SIR, and SAR, respectively.

The results of the subjective evaluations are presented in Figure 9. According to the collected responses for Question 1, the proposed audiovisual method is rated significant higher than the baseline audio-based method (Wilconxon signed-rank test shows a p value of 3.5×10^{-31}); The average rating is raised from 3.1 to 3.9. For Question 2, the difference is even more significant, as the average rating is increased from 2.6 to 3.8 (with a p value of 3.1×10^{-45}), showing that the proposed method is especially beneficial for removing accompaniments from the mixture. Regarding the artifacts introduced into the separated solo vocals in Question 3, both methods achieve a rating between "neutral" and "good", and the difference is not statistically significant (with a p value of 0.46).

5.6 Ablation Studies on Non-informative Visual Input

To further study how the visual information helps with the separation performance, we design several complementary experiments as ablation studies. We first modify the network structure by replacing the video front-end model with other existing widely-applied visual feature extraction framework to explore the key factor of the audiovisual separation framework and the robustness. Then we feed the visual branch with non-informative or even misleading inputs to observe how the separation quality degrades.

5.6.1 Non-Informative Visual Input

To further investigate how the incorporation of visual information affects the separation performance, in this section, we substitute the visual input (i.e., mouth region of the solo singer) with some irrelevant content.

- Constant. We feed the visual branch with constant zero values all the time.
- White-noise. We feed the visual branch with white noise that is normalized to the same range as the videos of mouth regions.
- White-noise*. The white noise is directly fused

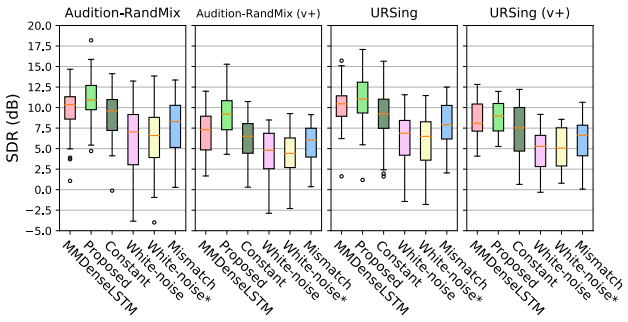

Figure 10: The SDR (dB) comparison on the separated solo vocal of the proposed audiovisual method with non-informative visual inputs.

with the audio embedding, replacing the whole visual branch.

- Mismatch. The input of the visual branch is the mouth region video of an irrelevant singer to provide misleading information about the singing activity.

Figure 10 shows the separation results on different experimental settings. The model performance always degrades from the audio-based baseline MMDenseLSTM when feeding with irrelevant or misleading information. This suggests that a non-informative visual input is harmful for separation. The performance degradation by feeding white noise or a mismatched singer is more noticeable than a constant input. This may be because the model is more likely to overfit irrelevant visual fluctuations in the training data, while for a constant visual input the model is more likely to ignore it. Nonetheless, in all of these circumstances, the separation performance still achieves a median SDR over 5dB for most cases. This suggests that the audio branch is dominant in the model inference. Comparing with the "No-mask" results in Figure 9, this also confirms our claim in Section 5.4 that the mask layer helps to improve the model robustness, even when the visual input is less informative.

6. Discussion

Since we are the first work to address audiovisual separation for singing performance, there are still much room to improve and many areas to explore. First, we are not building our model upon the most state-of-the-art audio separation methods due to the reasons described in Section 3.1.1. Other techniques like attention or transformer-based models may further improve the performance. Second, in this paper we crawled the Audition-RandMix data for training and created the URSing dataset for evaluation. While it is a challenging process to record the audiovisual singing performance with ground-truth tracks, collecting randomly mixed data for training is an easier process, since there are many solo singing performance videos in the Internet. Since using randomly mixed data has been proved beneficial for training music separation (Song et al., 2021), one could potentially improve the audiovisual vocal separation results by collecting more data. Third, it is worth investigating how other kinds of visual performances could help with the analyses of singing voice, such as facial expressions, body gestures, and body movements, etc

7. Conclusion

In this paper, we proposed an audiovisual approach to address the solo singing voice separation problem by analyzing both the auditory signal and mouth movement of the solo singer in the visual signal. To evaluate our proposed method, we created the URSing dataset, the first publicly available dataset of audiovisual singing performances recorded in isolation for singing voice separation research. We also curated a solo singing voice dataset from YouTube for training. Both objective evaluations on artificially mixed singing music and subjective evaluation on professionally produced a cappella songs showed that the proposed method significantly outperforms state-of-the-art audio-based methods. The advantages of the proposed method is especially pronounced when the accompaniment track contains backing vocals, which have been difficult to separate from solo vocals by audio-based methods.

Acknowledgment

We thank all of the singers who participated in our dataset recording process, and XXX for post-processing the audio recordings. We also acknowledge the funding agency.

References

Afouras, T., Chung, J. S., and Zisserman, A. (2018). The conversation: Deep audio-visual speech enhancement. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*.

Berenzweig, A., Ellis, D. P., and Lawrence, S. (2002). Using voice segments to improve artist classification of music. In *Proceedings of the AES 22nd International Conference: Virtual Synthetic and Entertainment Audio*.

Cadalbert, A., Landis, T., Regard, M., and Graves, R. E. (1994). Singing with and without words: Hemispheric asymmetries in motor control. *Journal of Clinical and Experimental Neuropsychology*, 16(5):664–670.

Cartwright, M., Pardo, B., and Mysore, G. J. (2018). Crowdsourced pairwise-comparison for source separation evaluation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 606–610. IEEE.

Cartwright, M., Pardo, B., Mysore, G. J., and Hoffman, M. (2016). Fast and easy crowdsourced percep-

tual audio evaluation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 619–623. IEEE.

Chan, T.-S., Yeh, T.-C., Fan, Z.-C., Chen, H.-W., Su, L., Yang, Y.-H., and Jang, R. (2015). Vocal activity informed singing voice separation with the iKala dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 718–722.

Chandna, P., Miron, M., Janer, J., and Gómez, E. (2017). Monoaural audio source separation using deep convolutional neural networks. In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation*, pages 258–266. Springer.

Chen, L., Srivastava, S., Duan, Z., and Xu, C. (2017). Deep cross-modal audio-visual generation. In *Proceedings of the ACM Thematic Workshops of Multimedia*, pages 349–357.

Choi, W., Kim, M., Chung, J., and Jung, D. L. S. (2019). Investigating deep neural transformations for spectrogram-based musical source separation. *arXiv preprint arXiv:1912.02591*.

Choi, W., Kim, M., Chung, J., and Jung, S. (2021). Lasaft: Latent source attentive frequency transformation for conditioned source separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 171–175. IEEE.

Chung, J. S. and Zisserman, A. (2016). Lip reading in the wild. In *Proceedings of the Asian Conference on Computer Vision*, pages 87–103. Springer.

Connell, L., Cai, Z. G., and Holler, J. (2013). Do you see what i'm singing? visuospatial movement biases pitch perception. *Brain and cognition*, 81(1):124–130.

Défossez, A., Usunier, N., Bottou, L., and Bach, F. (2019). Demucs: Deep extractor for music sources with extra unlabeled data remixed. *arXiv preprint arXiv:1909.01174*.

Dinesh, K., Li, B., Liu, X., Duan, Z., and Sharma, G. (2017). Visually informed multi-pitch analysis of string ensembles. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3021–3025. DOI: https://doi.org/10.1109/ICASSP.2017.7952711.

Duan, Z., Essid, S., Liem, C., Richard, G., and Sharma, G. (2019). Audiovisual analysis of music performances: Overview of an emerging field. *IEEE Signal Processing Magazine*, 36(1):63–73. DOI: https://doi.org/10.1109/MSP.2018.2875511.

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., and Rubinstein, M. (2018). Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (TOG)*, 37(4). DOI: https://doi.org/10.1145/3197517.3201357.

Fujihara, H. and Goto, M. (2007). A music information retrieval system based on singing voice timbre. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, pages 467–470.

Fujihara, H., Goto, M., Ogata, J., Komatani, K., Ogata, T., and Okuno, H. G. (2006). Automatic synchronization between lyrics and music cd recordings based on viterbi alignment of segregated vocal signals. In *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, pages 257–264.

Gan, C., Huang, D., Zhao, H., Tenenbaum, J. B., and Torralba, A. (2020). Music gesture for visual sound separation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10478–10487.

Gao, R. and Grauman, K. (2019). Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 3879–3888.

Grell, A., Sundberg, J., Ternström, S., Ptok, M., and Altenmüller, E. (2009). Rapid pitch correction in choir singers. *The Journal of the Acoustical Society of America*, 126(1):407–413.

Hennequin, R., Khlif, A., Voituret, F., and Moussallam, M. (2019). Spleeter: A fast and state-of-the-art music source separation tool with pre-trained models.

Hou, J.-C., Wang, S.-S., Lai, Y.-H., Tsao, Y., Chang, H.-W., and Wang, H.-M. (2018). Audio-visual speech enhancement using multimodal deep convolutional neural network. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Hsu, C.-L., Wang, D., Jang, J.-S. R., and Hu, K. (2012). A tandem algorithm for singing pitch extraction and voice separation from music accompaniment. *IEEE Transactions on audio, speech, and language processing*, 20(5):1482–1491.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708.

Huang, P.-S., Chen, S. D., Smaragdis, P., and Hasegawa-Johnson, M. (2012). Singing-voice separation from monaural recordings using robust principal component analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 57–60.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. (2014). Singing-voice separation from monaural recordings using deep recurrent neural networks. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, pages 477–482.

Jansson, A., Humphrey, E., Montecchio, N., Bittner,

R., Kumar, A., and Weyde, T. (2017). Singing voice separation with deep u-net convolutional networks. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.

King, D. E. (2009). Dlib–ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul.):1755–1758.

Li, B., Dinesh, K., Duan, Z., and Sharma, G. (2017a). See and listen: score-informed association of sound tracks to players in chamber music performance videos. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2906–2910. DOI: https://doi.org/10.1109/ICASSP.2017.7952688.

Li, B., Dinesh, K., Sharma, G., and Duan, Z. (2017b). Video-based vibrato detection and analysis for polyphonic string music. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, pages 123–130.

Li, B., Dinesh, K., Xu, C., Sharma, G., and Duan, Z. (2019). Online audio-visual source association for chamber music performances. *Transactions of the International Society for Music Information Retrieval*, 2(1).

Li, B. and Kumar, A. (2019). Query by video: Cross-modal music retrieval. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, pages 604–611.

Li, B., Maezawa, A., and Duan, Z. (2018). Skeleton plays piano: online generation of pianist body movements from MIDI performance. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.

Li, B., Xu, C., and Duan, Z. (2017c). Audiovisual source association for string ensembles through multi-modal vibrato analysis. In *Proceedings of the Sound and Music Computing (SMC) Conference*, pages 159–166.

Liu, J.-Y. and Yang, Y.-H. (2018). Denoising auto-encoder with recurrent skip connections and residual regression for music source separation. In *Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 773–778.

Lluis, F., Pons, J., and Serra, X. (2019). End-to-end music source separation: is it possible in the waveform domain? In *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*.

Lu, R., Duan, Z., and Zhang, C. (2018). Listen and look: audio–visual matching assisted speech source separation. *IEEE Signal Processing Letters*, 25(9):1315–1319.

Lu, R., Duan, Z., and Zhang, C. (2019). Audio–visual deep clustering for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11):1697–1712.

Luo, Y., Chen, Z., Hershey, J. R., Le Roux, J., and Mesgarani, N. (2017). Deep clustering and conventional networks for music separation: Stronger together. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 61–65.

Luo, Y. and Mesgarani, N. (2018). Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700.

Mesaros, A. and Virtanen, T. (2010). Recognition of phonemes and words in singing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2146–2149. IEEE.

Ozerov, A., Philippe, P., Bimbot, F., and Gribonval, R. (2007). Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1564–1578.

Ozerov, A., Philippe, P., Gribonval, R., and Bimbot, F. (2005). One microphone singing voice separation using source-adapted models. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 90–93. IEEE.

Parekh, S., Essid, S., Ozerov, A., Duong, N., Perez, P., and Richard, G. (2017). Motion informed audio source separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6–10. DOI: https://doi.org/10.1109/ICASSP.2017.7951787.

Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., and Pantic, M. (2018). End-to-end audiovisual speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6548–6552.

Rafii, Z. and Pardo, B. (2011). A simple music/voice separation method based on the extraction of the repeating musical structure. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 221–224.

Song, X., Kong, Q., Du, X., and Wang, Y. (2021). Catnet: music source separation system with mix-audio augmentation. *arXiv preprint arXiv:2102.09966*.

Stoller, D., Ewert, S., and Dixon, S. (2018). Wave-u-net: A multi-scale neural network for end-to-end audio source separation. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, pages 334–340.

Stöter, F.-R., Liutkus, A., and Ito, N. (2018). The 2018

signal separation evaluation campaign. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 293–305. Springer.

Stöter, F.-R., Uhlich, S., Liutkus, A., and Mitsufuji, Y. (2019). Open-unmix - a reference implementation for music source separation. *Journal of Open Source Software*.

Takahashi, N., Agrawal, P., Goswami, N., and Mitsufuji, Y. (2018a). Phasenet: Discretized phase modeling with deep neural networks for audio source separation. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*, pages 2713–2717.

Takahashi, N., Goswami, N., and Mitsufuji, Y. (2018b). Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation. In *Proceedings of the International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 106–110. IEEE.

Takahashi, N. and Mitsufuji, Y. (2017). Multi-scale multi-band densenets for audio source separation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 21–25.

Takahashi, N. and Mitsufuji, Y. (2021). D3net: Densely connected multidilated densenet for music source separation. *arXiv preprint arXiv:2010.01733*.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4489–4497.

Tzinis, E., Wisdom, S., Jansen, A., Hershey, S., Remez, T., Ellis, D. P., and Hershey, J. R. (2021). Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N., and Mitsufuji, Y. (2017). Improving music source separation based on deep neural networks through data augmentation and network blending. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 261–265.

Vembu, S. and Baumann, S. (2005). Separation of vocals from polyphonic audio recordings. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, pages 337–344. Citeseer.

Vincent, E., Virtanen, T., and Gannot, S. (2018). *Audio source separation and speech enhancement*. John Wiley & Sons.

Zhao, H., Gan, C., Ma, W.-C., and Torralba, A. (2019). The sound of motions. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1735–1744.

Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., and Torralba, A. (2018). The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 1, pages 587–604. DOI: https://doi.org/10.1007/978-3-030-01246-5_35.