Syncing Pre-Recorded Audio Description to a Live Musical Theater Performance Using a Reference Audio Recording

Dirk Vander Wilt
New York University
New York, NY, USA
dirk.vanderwilt@nyu.edu

Morwaread Mary Farbood
New York University
New York, NY, USA
mfarbood@nyu.edu

ABSTRACT
Audio description, an accessibility service used by blind or visually impaired individuals, provides spoken descriptions of visual content. This accommodation allows those with low or no vision the ability to access information that sighted people obtain visually. A method for deploying pre-recorded audio description in a live musical theater environment is presented here. This method uses a reference audio recording and an online time warping algorithm to automatically align audio description with repeated live performances of a theatrical production. This system is used in two evaluation experiments that show the method successfully aligns multiple recordings of works of musical theater in order to automatically trigger pre-recorded, descriptive audio in real time.

CCS Concepts
•**Human-centered computing** → **Accessibility technologies;** •**Applied computing** → *Performing arts; Sound and music computing;*

Author Keywords
audio description; blind; visually impaired; live theater; time warping

INTRODUCTION
Audio description (AD) is a spoken description of the visual elements of an accompanying work, providing an accessible alternative to obtaining visual information for blind or visually impaired individuals [3, 11]. Like fixed media (television, movies, etc.), AD for live performances must concurrently provide necessary visual information and be timed appropriately so as not to disrupt other aspects of the show (musical numbers, dialogue, etc.). However, since repeated live performances are by design never identical, a fixed AD track cannot be aligned in advance. In live situations, either a professional audio describer narrates the visual elements in real time, or a system is created to allow pre-recorded AD to be triggered [6]. Developing and deploying this type of service is expensive and time-consuming [9, 12].

ASSETS '19, October 28–30, 2019, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-6676-2/19/10.
https://doi.org/10.1145/3308561.3354624

This study proposes an inexpensive and novel method to trigger pre-recorded AD for a live theatrical performance by only using audio obtained from a show's previous performance and an established online time warping algorithm [2]. The contributions of this study include: (1) a novel implementation of online time warping, (2) a unique dataset of annotated musical theater audio recordings, and (3) a proposed software implementation that integrates with existing theatrical technology.

METHOD
Live audio alignment has been successful in a variety of contexts, such as score following [1] and audio-transcription alignment [5]. In the implementation described here, a reference audio recording is aligned to live audio input using an online time warping algorithm [2]. During the live alignment, descriptive audio tracks based on their pre-aligned position in the reference recording are also aligned and played back in real time.

To set up the system, audio from one performance of the entire production is recorded in advance. Features from that audio are extracted and stored in frames of vectors. A second audio track is also created, containing only the AD which aligns to that recorded performance. The descriptive track is broken up into multiple smaller tracks such that one sub-track contains the audio for a single described event, and the points where each sub-track should be triggered are marked based on the reference recording.

The extracted features from both the reference recording and live input are Mel-frequency cepstrum coefficients (MFCCs). MFCCs are a well-established set of features used when analyzing audio signals for human speech recognition [8] and music information retrieval applications [7]. This system uses 13 MFCCs per frame as well as a first-order difference to account for change over time, totaling 26 features per frame. The audio is captured at a low sampling rate of 8 kHz, and features are extracted using a frame length of 100 ms and a hop size of 40 ms. The code implemented to extract MFCCs is based on [4]. Extracted features are stored as two series of vectors, one for the reference and one for the live input.

Dynamic Time Warping (DTW) compares two time series U and V of lengths m and n. The output is an m- by- n matrix D where each element $D(x,y)$ is a cumulative (in this case Euclidean) distance ("path cost") from $D(1,1)$ to $U(x)$ and $V(y)$ [10]. Thus, from any arbitrary frame in time series U, the best match in time series V is determined by examining

the path costs for that frame in D; the lowest cost is the closest match.

Classic DTW requires both time series to be available in advance, which is not suitable for a live performance. An online time warping algorithm [2] requires only one series to be available in advance (the reference), while the other series may be obtained in real time (as the live input). The algorithm outputs a similar matrix D, but it builds as the input is received, one row and/or column at a time. Whenever a new frame of live input is received, the algorithm determines whether to add another row, or column, or both, to D, thus incrementing either the index of the live input (t) or of the reference (j). AD is triggered based on the real-time current value of j.

EVALUATION

To evaluate this method, two different audio recordings of the same theatrical productions were used, with one recording as a reference and the other as live input. Markers were placed manually in both recordings to represent specific moments in the production (such as lines of dialogue or musical cues). The algorithm was then run in real time and the mark locations found during the alignment were compared to the actual mark locations in the live input.

In the first evaluation experiment, two recordings of Gilbert and Sullivan's *H.M.S. Pinafore* were used (recorded in 1949 and 1958). In both recordings 213 specific points were marked. The results showed that 161 marks (76%) were found less than 1 second from the mark's actual location in the reference; 183 marks (86%) were found less than 2 seconds from the actual location; and 200 marks (94%) were less than 5 seconds (SD = 2.68 seconds).

To test the algorithm in a more realistic situation, a second experiment using two recordings with notable, audible differences were obtained: the Broadway (2007) and London (2010) cast recordings of *Legally Blonde, The Musical*. The London version was recorded in front of a live audience and contains audience noise, laughter, ambience, etc. that is not present in the Broadway recording. One track was removed from the London version because it was out of order. In both versions of *Legally Blonde*, 169 locations were manually marked. The results showed that 117 marks (69%) were found within 1 second of the reference, 133 (79%) were within 2 seconds, and 147 (87%) were within 5 seconds (SD = 3.32 seconds).

In both experiments, the total duration of each recording was over an hour, and the system was able to keep up with the long live input, and automatically correct itself after moments of difference. Within the context of a live theatrical environment, these results show that most (79-86%) AD will be triggered within two seconds of the actual event occurring. These metrics indicate that theatergoers would be able to follow the visual elements of the production in a timely way.

DETECTING INTERRUPTIONS

Performance interruptions are ubiquitous in live theater and must be handled accordingly. If an intermission, audience applause, or other such break is not included in the reference recording, subsequent marks will mis-fire until sufficient time has passed for the system to adjust. To account for this, the system listens for specific audio identifiers that are representative of typical interruptions—audience chatter, applause, and others. If the ongoing live input is closer to this training data, then the system will stop incrementing j until the performance has resumed. This is done by calculating the path costs of c previous input frames against c frames of training data. If the live input is closer (Euclidean distance-wise) to the training data than to the next frames of the reference, then the system will stop incrementing j until the input data more closely aligns with the reference.

To evaluate pause detection, two 3-minute audio clips were obtained to represent two theatrical environments that could indicate a cue to pause (a small empty theater and audience chatter in a large theater during intermission). The first 30 seconds of each recording were used as training data, and the remaining 2.5 minutes were inserted after the overture of *H.M.S. Pinafore* in the live input only. This emulates an unexpected interruption not present in the reference audio.

The results show that the mark accuracy after the interruption greatly increased when the algorithm was paused. For the first 40 marks, the average accuracy with interruption-detection enabled was 1.32 seconds, whereas without the detection the accuracy was 16.32 seconds.

CONCLUSION AND FUTURE WORK

One goal of automating AD deployment is to allow every performance of a repeating production to be accessible, without needing a live describer. Another goal is to provide easy-to-use tools for production companies which may help increase the availability of AD where it may not have been available before. Our ongoing research in developing this novel approach for triggering AD for live theater shows promise in addressing both of these goals.

This proposed system only requires audio, which can be obtained either from the theater's existing setup, or from a separate microphone. The computer must have one mono output channel and two mono input channels: one input to capture the reference recording and live input, and the other to capture the live audio describer. The simplicity of the method's technical setup and its overall flexibility provide a promising new way to make theater experiences for visually impaired audience members more inclusive and accessible.

ACKNOWLEDGEMENTS

We thank Juan Pablo Bello (NYU MARL) and Amy Hurst (NYU Ability Project) for their assistance and feedback.

REFERENCES

[1] Andreas Arzt. 2016. *Flexible and robust music tracking*. Ph.D. Dissertation. Johannes Kepler University, Linz.

[2] Simon Dixon. 2005. Live tracking of musical performances using on-line time warping. In *Proceedings of the 8th International Conference on Digital Audio Effects*. 92–97.

[3] Louise Fryer. 2016. *An Introduction to Audio Description: A Practical Guide*. Routledge.

[4] D. S. Pavan Kumar. 2016. A simple MFCC extractor using C++ STL and C++11. (2016). https://github.com/dspavankumar/compute-mfcc

[5] N. Lertwongkhanakool, N. Kertkeidkachorn, P. Punyabukkana, and A. Suchato. 2015. An Automatic Real-time Synchronization of Live Speech with Its Transcription Approach. *Engineering Journal-Thailand* 19, 5 (2015), 81–99.

[6] Elena Litsyn and Hagai Pipko. 2019. System and method for distribution and synchronized presentation of content. (May 2 2019). US Patent App. 16/092,775.

[7] Beth Logan. 2000. Mel Frequency Cepstral Coefficients for Music Modeling. In *ISMIR*, Vol. 270. 1–11.

[8] Lindasalwa Muda, Mumtaj Begam, and I. Elamvazuthi. 2010. Voice Recognition Algorithms using Mel Frequency Cepstral Coeffient (MFCC) and Dynamic Time Warping (DTW) Techniques. *Journal of Computing* 2, 3 (2010), 138–143.

[9] Malgorzata Plaza. 2017. Cost-effectiveness of audio description production process: comparative analysis of outsourcing and 'in-house' methods. *International Journal of Production Research* 55, 12 (2017), 3480–3496.

[10] H. Sakoe and S. Chiba. 1978. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics Speech and Signal Processing* 26, 1 (1978), 43–49.

[11] Joel Snyder. 2014. *The Visual Made Verbal: A Comprehensive Training Manual and Guide to the History and Applications of Audio Description*. American Council of the Blind, Inc.

[12] Agnieszka Szarkowska. 2011. Text-to-speech audio description: towards wider availability of AD. *Journal of Specialised Translation* 15 (2011), 142–162.