Method and System for Aligning Audio Description to a Live Musical Theater Performance

Dirk Vander Wilt and Morwaread Mary Farbood

New York University
[dirk.vanderwilt,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. In this paper a method for deploying pre-recorded audio description in a live musical theater environment is presented. This method uses a reference recording and an online time warping algorithm to align audio descriptions with live performances. A software implementation that is integrated into an existing theatrical workflow is also described. 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.

Keywords: audio description, blind, visually impaired, accessibility, disability, musical theater, time warping

1 Introduction

Audio description (AD) is an accommodation used by people who are visually impaired or blind. It is a spoken description of the visual elements of an accompanying work, providing an accessible alternative to obtaining information that sighted individuals may obtain visually. Users of AD should be able to ascertain with audio what a sighted person at the same time and place may ascertain with vision. At live theatrical events that provide AD, such as some Broadway musicals, patrons are either provided with a wireless audio receiving device or are asked to download a software application onto their smartphone, so the transmission of the AD will not disrupt other theatergoers. For an overview on the art and practice of AD, see Fryer [6] and Snyder [14].

Live theatrical events pose an interesting problem for AD services. Like fixed media, the descriptions must be timed appropriately so as not to disrupt other aural aspects of the show (musical numbers, dialogue, etc.) [6, 14]. 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 describes the visual elements in real time, or a system is created to allow pre-recorded AD to be triggered [9]. Developing and deploying this type of service is expensive and time-consuming. A recent study showed that media producers view AD as “a costly service with no revenue potential.” Creating audio description for a 30-minute television show with 24 cues may cost between $698 and $1,462, depending on how the description is produced [12]. According to Szarkowska [15], “A lengthy preparation process and high production costs are among the greatest obstacles to the wider availability of audio description.”

This paper proposes an inexpensive and novel method to trigger AD for a live theatrical performance by only using audio obtained from a show’s previous performance. The process described in this paper warps the audio from the live show in real time to a reference recording using an established online time warping algorithm [5]. This method is a step towards being able to reduce the cost of deploying live theatrical audio description, thus making it more available to visually impaired people.

2 Method

Live audio-to-audio alignment has been used successfully in music score following [1, 2] and audio-transcription alignment [8] tasks. In this implementation, a reference recording is aligned to live input using an online time warping algorithm [5]. During the live alignment, descriptive audio tracks based on their pre-aligned position in the reference recording are also aligned and played back for blind and visually-impaired audience members.

First, a performance of the entire production is recorded in advance. Relevant features from that audio are then 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 within the recorded performance, and the points where each sub-track should be triggered are marked. Once the marks, descriptive track, and extracted features of the reference recording are obtained, the live alignment may begin.

2.1 Audio Features

In the system described here, Mel-frequency cepstrum coefficients (MFCCs) are extracted from both the live input and reference recording. MFCCs are a well-established set of features used when analyzing audio signals for human speech recognition and music information retrieval applications. Both uses are applicable here since theatrical productions often contain both speech (dialogue) and music (showtunes). The coefficients are derived from the Mel scale, which captures patterns audible to the human ear. Although the system does not recognize speech explicitly, it uses MFCCs to compare different (but similar) samples of speech and music patterns. The code implemented to extract MFCCs here was based on [7].

Starting with audio at a sampling of 8 kHz, the MFCC extraction process begins with the application of a pre-emphasis filter so that the higher frequencies of the signal have greater energy. The signal is then segmented into frames, and a Hamming window and FFT are applied to each frame. The results are filtered through a Mel filterbank with 40 filters, which is where the raw frequency data gets "refined" to frequencies based on the Mel scale. A discrete cosine transform (DCT) is performed on the log-amplitudes of the filter output, resulting in 13 coefficients which are the MFCCs for that frame. To account for change over time, a first-order difference of each coefficient from the previous frame is appended to the 13 current coefficients, making the total number of coefficients used for analysis 26 per frame. Given the real-time nature of this system, the features must be extractable in less time than it takes the corresponding audio to play out in real time. In this case, the system extracts one MFCC feature vector at a frame length of 100 ms and a hop size of 40 ms. For a description of MFCC feature extraction, see [10, 8, 11].

2.2 Online Time Warping

Dynamic Time Warping (DTW) uses dynamic programming to recursively compare 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 between $U(x)$ and $V(y)$. The value of each cell is the cumulative "path cost" from $D(1, 1)$ up to that point [13]:

$$D(i,j) = d(i,j) + min \left\{ \begin{array}{l} D(i-1,j) \\ D(i-1,j-1) \\ D(i,j-1) \end{array} \right\} \tag{1}$$

Every cell in the matrix is calculated by obtaining the distance between $U(i)$ and $V(j)$, and adding it to the least of one of three adjacent (previous) cells. In this way, the cumulative path cost between $D(1, 1)$ and the current cell is determined. The smaller the cumulative cost, the better the match up to that point. When the whole matrix is calculated through $D(m, n)$, backtracking the smallest distance back to $D(1, 1)$ will be the warp path which relates the closest points of U to V. Unfortunately, this algorithm requires both series to be fully available in advance and has a running time of $O(N^2)$, making it unsuitable for live performance tracking.

This online time warping algorithm, developed by Dixon [5], requires only one of the series to be available in advance, while the other series may be obtained in real time (V is known fully in advance, and U is only partially known, but increases as new live input is received). The algorithm outputs a similar matrix D, but it builds as the input is received, one row and/or column at a time, and only in the forward direction. Plus, it is only able to estimate the total size of the resulting matrix, so it is instead bounded by a constant, which is determined in advance. Thus, it does not have the advantage of being able to backtrack from future input.

In online time warping, whenever a new frame of input is received in real time as $U(t)$, where t is the current live input frame, the system must determine whether to add another row, or column, or both, to matrix D. It does this by checking all the path cost's previous c elements of the current row t and column j of the matrix. If the lowest cost is in a row, it increments the row. If the lowest path is in the column, it increments the column. If the lowest cost is $D(t, j)$, the current cell, it increments both. Also, if a row or column has been incremented $MaxRunCount$ times, it then increments the other, thus preventing the system from running away in one direction. This implementation sets $c = 500$ and $MaxRunCount = 3$ as described in [5].

Indices t and j are pointers to the current real-time position in U and V. At any point $U(t)$ (the current frame in the real-time input), the value of j is the current estimated location in V. Since index t is the current live input frame, it will always increment steadily. Index j, however, will increment based on where the online time warping algorithm estimates the current temporal location to be in the reference recording. AD is inserted based on the real-time current value of j.

3 The Alignment Process

Three inputs are needed to trigger AD: the reference recording V, one or more frames of ongoing live input U, and an array of frame numbers F, where $F(1...x)$ represents the frame at which AD number x should be triggered. U and V are arrays of feature vectors. Both $U(n)$ and $V(n)$ are a single feature vector at frame n. Prior to the live performance commencing, all features of the reference recording are extracted and placed in V. U is extracted in real time during the live performance.

When the show begins, $t = 1$ and $j = 1$, which are references to the indices of the first frames of U and V, respectively. Each time t increases (meaning the live recording has progressed by one frame), the new value of j is determined, based on the online time warping algorithm. If the algorithm determines that U is progressing faster than V at that moment, then $t+ = 1$. If U is slower than V, then $j+ = 1$. If they are both moving at the same speed at that moment, then both t and j are incremented. Index j will keep increasing until it matches t's estimated location, and a new t (live input frame) is obtained (or, alternately, j will not increase while t catches up). The AD number x is triggered when $j = F(x)$. In this way, the descriptive tracks are able to align with the live performance based on the online time warping's estimation of the current index j of the reference.

Since the actual size of the matrix is unknown, an empty square matrix of 40,000-by-40,000 was created in the current implementation, which holds approximately 25 minutes of data on either t or j given the feature extraction parameters presented earlier. During the alignment, when one index is incremented up to the size of the matrix, the matrix is reset and the path cost returns to 0. In this manner, the alignment can run indefinitely, and the calculated path cost does not increase indefinitely. During the feature extraction phase, the MFCCs for each minute of audio was calculated and extracted in less than 2 seconds while running on a 2.7 GHz MacBook Pro using a C/C++ implementation. Offline tests of the online algorithm were able to process one hour of alignment (including extraction and matrix calculation) in about 4 minutes. This process therefore runs comfortably in a real-time scenario.

4 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 spoken words 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: a D'Oyly Carte recording from 1949 and a Malcolm Sargent recording from 1957. In both recordings 213 specific points were marked; these points were meant to simulate where AD may be triggered. This experiment used the D'Oyly Carte version as the reference and the Malcolm Sargent version as the live input.

After completing the alignment, the ground truth was compared with the marks automatically located by the algorithm. A total of 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. The mean difference between the marks indicated by the algorithm and the ground truth was 1.2 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. The only alteration made to the recordings was the removal of one track from the London version because it was out of order in the Broadway recording.

In both versions of *Legally Blonde*, 169 locations were manually marked, and the alignment was run in real time using the London recording as the reference, and the Broadway recording as the live input. 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. The mean difference between the generated marks and ground truth was 1.79 seconds (SD = 3.32 seconds).

In both experiments, the total duration of each recording was over an hour, and the algorithm was able to keep up with the long live input, and automatically correct itself after moments of difference between the reference and the live input. If there is a "mistake" between productions and the AD becomes misaligned,

Table 1. Results of the two evaluation experiments. Marks refers to the total number of annotated marks for each show; values in the < 1, 2, 5 sec columns indicate the percentage of marks found within 1, 2, and 5 seconds of the ground truth; St. dev refers to standard deviation of the differences between the found marks and the ground truth.

	Marks	**< 1 sec**	**< 2 sec**	**< 5 sec**	**St. Dev**
Pinafore	213	75.57%	85.92%	93.92%	2.68 sec
Blonde	169	69.23%	78.70%	86.98%	3.32 sec

the algorithm may correct itself as the production progresses. For example, the longest difference between reference and live for all experiments was about 21 seconds, which occurred during a significant period of divergence between the two recordings of *Legally Blonde*. However, the algorithm was back to correctly aligning once again less than 2 minutes later, with the next marks falling 120 ms from the reference.

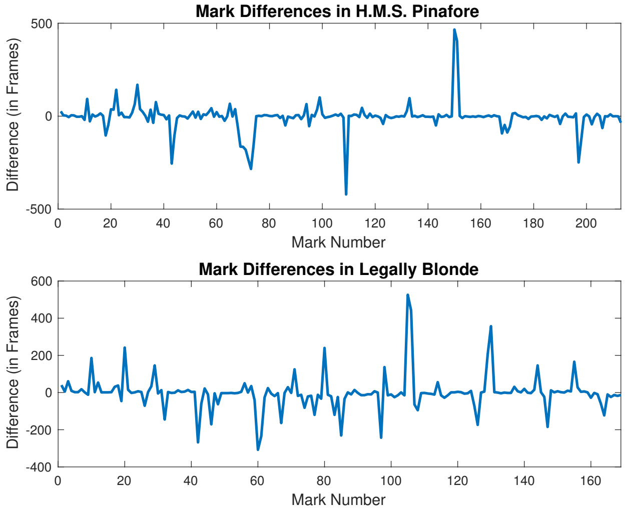

Fig. 1. Accuracy of all marks shown for *H.M.S. Pinafore* and *Legally Blonde*, shown as deviations from ground truth in frames. X axis indicates mark number; Y axis indicates difference in number of frames (25 frames = 1 second).

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.

5 Implementation

Software implementations to increase the availability of audio description generally take the form of automating some task of AD creation or deployment, thus decreasing cost and complexity, and ultimately increasing availability. For example, the *CineAD* system [4] uses information from existing closed captions and a teleplay or screenplay of a fixed film or television program to generate a descriptive script automatically, which may then be read by a synthetic voice or by a human. Alternately, *LiveDescribe* [3] seeks to recruit amateur describers to describe videos; it does this in part by analyzing a video and allowing describers to record descriptive audio only during breaks in dialogue.

In a live setting, and in particular when a live performance is imperfectly repeated, automated AD must accommodate for variations, but must still be able to follow some static representation of the performance in order to correctly align the description. Thus, we propose both a software implementation and workflow that takes into account existing theatrical technology (audio recording) as well as established music information retrieval techniques (feature extraction and online time warping) to align AD in real time. The previous sections of this paper have discussed the algorithm and parameters of the software; this section describes how the software may be ideally used in a live setting.

Computer and Theatrical Requirements. The software in this implementation runs comfortably on a 2012-era MacBook Pro with a solid-state hard drive and 16 gigabytes of memory. The computer must be able to obtain two channels of mono input simultaneously—one input to capture the reference recording and (later) the live input, and the other to capture the live audio describer. Importantly, both channels must be isolated from each other such that they do not capture audio from each other, meaning the describer must be listening with headphones so that the reference recording is not captured on the descriptive audio track. The computer must also have a single mono audio output channel, though this will not run simultaneously with the input. This output is transmitted (by either wireless or some other mechanism) to audience members using the descriptive service.

Software Interface. The software's core functions are *Record* and *Begin Show* buttons. The *Record* button activates the two mono inputs simultaneously to capture both the reference recording and the descriptive audio. The software records the reference's alignment to the AD by detecting when the speaker begins to talk, and the corresponding frame number is retained for the particular mark. Alternately, a third function, *Add Mark*, can be manually activated each time the describer wishes to add a new mark.

The software may have supplemental editing features. The ability to replace portions of the reference recording may be helpful if a long-running show has scene changes. Additionally, the ability to minutely correct the trigger timing of specific descriptive audio may help with fine tuning the AD between performances.

Capturing the Reference and Audio Description. With the software running, the setup configured as above, and the show commencing, the live describer presses *Record* on the computer and the system begins to capture both the performance and describer's voice. The system records the entirety of both mono inputs, capturing the sample numbers for each audio described event. By the end, the system will have captured all relevant data needed to play back the description to future audiences. After this point, the human describer is no longer needed, and the system may be automated.

Providing Audio Description to Audiences. A theatrical technician activates the *Begin Show* button as the production commences. At this point, the mono output (to the wireless receivers of audience members) and the mono input for the live recording (which is the same input as for the reference) is activated. The online time warping process is activated, and the AD is triggered at the correct moment.

6 Conclusion and Future Work

In this paper we presented an automated approach to triggering audio description for a live musical theater performance using audio from a previous performance and an online time warping algorithm. The method is able to correct itself and adapt to variations between the reference and live performance, which is necessary for an effective real-time method. Although the method could be further refined in the future by taking into account very large variations due to intermissions and audience applause that are typical in live performances, the evaluation experiments showed that significant differences such as changes in casting, script, and instrumentation, are already handled robustly.

The software implementation of the system described here can be integrated into a theater's existing setup with minimal interference. For example, the system is able to capture, process, and deploy AD independent from the existing software or hardware controls, since it only uses an audio signal which is readily available in the performance space. Other than an initial reference recording and the descriptive tracks themselves, no other setup or configurations were required. The simplicity of the method's technical setup and its overall flexibility provide a new way to make theater experiences for visually impaired audience members more inclusive and accessible.

Given the proliferation of accessibility on personal computing devices, using a smartphone to align and deliver the description would improve the overall success of the system. Being able to track a live performance from a mobile device without having to be connected to a wireless transmission mechanism would allow AD to be completely in control of the user, not reliant on the setup of the theater.

Audio description is a relatively new but quickly expanding accommodation for those with visual impairments. While it is becoming more common in film and television, AD for live theatrical performances remains rare. Decreasing the cost and complexity of creating and deploying AD would increase its availability, thus making the enjoyment of live theater more accessible to blind and visually impaired individuals.

References

1. Arzt, A.: Flexible and Robust Music Tracking, Ph.D. dissertation, Johannes Kepler University, Linz (2016)
2. Arzt, A., Widmer, G., Dixon, S.: Automatic Page Turning for Musicians via Real-Time Machine Listening. Proceedings of the European Conference on Artificial Intelligence, pp. 241-245 (2008)
3. Branje, C. J., Fels, D. I.: Livedescribe: can amateur describers create high-quality audio description? Journal of Visual Impairment & Blindness. Vol 106, No. 3, pp. 154-165 (2012)
4. Campos, V. P., de Araujo, T, M, U., de Souza Filho G.L., Goncalves, L M. G.: CineAD: a system for automated audio description script generation for the visually impaired. Universal Access in the Information Society, pp. 1-13 (2018)
5. Dixon, S.: Live tracking of musical performances using on-line time warping. Proceedings of the 8th International Conference on Digital Audio Effects, pp. 92-97 (2005)
6. Fryer, L: An Introduction to Audio Description: A Practical Guide. Routledge (2016)
7. Kumar, D. S. P.: A simple MFCC extractor using C++ STL and C++11. Source code at `http://www.github.com/dspavankumar/compute-mfcc` (2016)
8. Lertwongkhanakool, N., Kertkeidkachorn, N., Punyabukkana, P., Suchato, A.: An Automatic Real-time Synchronization of Live Speech with Its Transcription Approach. Engineering Journal, Vol 19, No. 5, pp. 81-99 (2015)
9. Litsyn, E., Pipko, H.: System and method for distribution and synchronized presentation of content. U.S. Patent Application 16/092,775 (filed May 2, 2019)
10. Logan, B.: Mel Frequency Cepstral Coefficients for Music Modeling. ISMIR. Vol. 270. (2000)
11. Muda, L., Begam, M., Elamvazuthi, I.: Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. Journal of Computing, Vol 2, Issue 3. (March 2010)
12. Plaza, M.: Cost-effectiveness of audio description process: a comparative analysis of outsourcing and "in-house" methods. International Journal of Production Research, pp. 3480-3496 (2017)
13. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimisation for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 26, pp. 43?49 (1978)
14. Snyder, J.: The visual made verbal: A comprehensive training manual and guide to the history and applications of audio description. American Council of the Blind. (2014)
15. Szarkowska, A.: Text-to-speech audio description: towards a wider availability of AD. Journal of Specialised Translation. Volume 15, pp. 142-162 (2011)