Capture-Time Feedback for Recording Scripted Narration

Steve Rubin[1], **Floraine Berthouzoz**[2], **Gautham J. Mysore**[2], **Maneesh Agrawala**[3]
[1]University of California, Berkeley [2]Adobe Research [3]Stanford University
srubin@cs.berkeley.edu {floraine,gmysore}@adobe.com maneesh@cs.stanford.edu

ABSTRACT

Well-performed audio narrations are a hallmark of captivating podcasts, explainer videos, radio stories, and movie trailers. To record these narrations, professional voiceover actors follow guidelines that describe how to use low-level vocal components—volume, pitch, timbre, and tempo—to deliver performances that emphasize important words while maintaining variety, flow and diction. Yet, these techniques are not well-known outside the professional voiceover community, especially among hobbyist producers looking to create their own narrations. We present *Narration Coach*, an interface that assists novice users in recording scripted narrations. As a user records her narration, our system synchronizes the takes to her script, provides text feedback about how well she is meeting the expert voiceover guidelines, and resynthesizes her recordings to help her hear how she can speak better.

Author Keywords

Voiceover; narration; speech emphasis; audio.

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces — Graphical user interfaces

INTRODUCTION

From podcasts and radio stories to explainer videos to commercials and movie trailers, audio narration pervades modern media. High-quality narrations captivate listeners and shape their perceptions of stories. But producing an effective narration, requires good delivery. Professional voiceover actors continuously adjust four primary voice components—volume, pitch, timbre, and tempo—to follow a set of best practices for high-quality audio narrations. These best practices cover four high-level guidelines [11]:

- **Emphasis:** Emphasize important words by adjusting the voice's pitch contour, volume and tempo.
- **Variety:** Vary the tempo and pitch of the delivery to avoid sounding monotonic.
- **Flow:** Control the speed of the narration and avoid pauses between words to allow sentences to flow naturally.
- **Diction:** Articulate words clearly; do not mumble.

Hobbyist content creators also record narrations, but are not always aware of these professional voiceover guidelines. As a result, their voiceovers often sound less captivating, coherent, and polished. Even a novice creator who knows that she should, for example, emphasize the important words in a sentence, may struggle in manipulating the pitch of her voice to achieve that emphasis.

In this paper we present *Narration Coach*, an interface that assists novice users while they record scripted narrations by providing immediate feedback on how well the user emphasizes important words and maintains variety, natural flow and good diction. The user enters her desired script and underlines words that she wants to emphasize. As the user records herself speaking the script, our system segments and aligns the recordings to the corresponding lines in the script. After the user records a set of lines, *Narration Coach* detects the spoken emphasis in those lines, and checks for vocal variety, good flow, and clear diction in the recording. Our system uses these detections in two ways. First, it provides feedback to the user about how successfully she spoke the line and what she can improve. Second, it provides methods for resynthesizing versions of the recording that improve the emphasis and flow. The user can construct her final narration either by re-recording lines with the feedback in mind, or by using one of the resynthesized options as an improved version of the take. This record–feedback–resynthesis pipeline allows users to iteratively improve their narration.

In a pilot study, first-time users of our system successfully created audio narrations that they preferred over narrations they made without *Narration Coach*. Additionally, impartial listeners rated narrations created with our system as higher quality than the narration made without it.

RELATED WORK

Our work builds on research in active capture systems and audio resynthesis methods.

Active Capture. *Narration Coach* follows in a line of research on *active capture* [6], a media-production paradigm that combines capture, interaction, and processing. Chang and Davis [5] present an active capture system to direct users performing and recording a video introduction. Heer et al. [12] describe strategies for active capture systems to deliver feedback to users. Our system differs from these projects by providing an end-to-end system for the broad task of recording narration rather than scene-specific directions like "turn your head" and "scream louder." Carter et al.'s *NudgeCam* [4] helps users record video that follows capture heuristics such as interview guidelines. *Narration Coach* similarly helps users follow capture heuristics, but for audio narration instead of video. Hindenburg Audio Book Creator [13] is

UIST '15, November 08- 11, 2015, Charlotte, NC, USA
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3779-3/15/11$15.00
DOI: http://dx.doi.org/10.1145/2807442.2807464

a digital audio workstation specifically designed for recording audio book narration. It allows users to manually link recordings to a script, but unlike *Narration Coach*, it does not provide feedback and resynthesis tools.

The research most similar to our work is Kurihara et al.'s *Presentation Sensei* [15], a tool that uses speech and image processing to give a speaker feedback on speed, pitch, and eye-contact while he rehearses a slide-based presentation. Like our system, *Presentation Sensei* provides capture-time feedback about speech performance. However, our system focuses on the iterative narration recording process. It organizes recorded audio based on an input script, and it provides automatic, word-level feedback and resynthesis tools.

Resynthesizing Audio Recordings. Existing digital audio workstations (DAWs) like Avid ProTools and Adobe Audition allow users to record audio and improve its quality using low-level signal processing effects. However, these DAWs target professional audio producers and unlike our system, they do not provide high-level tools to help novices record narrations, like associating recordings with lines in a script and providing automated feedback and resynthesis. Rubin et al. [21, 22, 23] present systems for editing audio stories, including tools for auditioning and combining multiple takes of lines. These systems assume that users have already recorded their footage, while *Narration Coach* assists users ini the speech capture process.

Our work draws on techniques from signal processing research to analyze and manipulate speech. Researchers have developed speech prosody manipulation techniques such as the phase vocoder [8], which enables audio adjustments by modifying components in the frequency domain, PSOLA [28], which reconstructs audio by adding, removing, and shifting small windows in the audio, and TANDEM-STRAIGHT [14], a vocoder that allows manipulations by decomposing speech into a smooth spectrogram, pitch contour, and periodicity map. Black and Taylor's Festival system [1, 27] is a complete text-to-speech pipeline, which include concepts relevant to our work such as prosody prediction from text, prosody synthesis, and speech signal analysis. We apply these algorithms in our analysis and resynthesis tools.

GUIDELINES FOR HIGH-QUALITY NARRATION

We designed *Narration Coach* to address common problems in recording narration [7, 9, 11, 18]. Voiceover actors use four high-level guidelines to improve the quality of their delivery [11, 19]. They emphasize words that help users follow the story (e.g. descriptive words, proper nouns, action verbs), add vocal variations to the delivery, adjust the speed of the narration and the location of pauses to control the flow of the speech, and enunciate words clearly. To achieve these high-level goals, voiceover actors continuously adjust four components of their voice, i.e. the volume, pitch, timbre, and tempo.

Emphasis

When speaking, voiceover actors adjust their pitch, tempo and volume to emphasize or *hit* a word. Emphasizing a word makes it stand out or signals the end of a thought; it helps the listener follow the story and identify the most important messages. The speaker needs to understand the intended emotion and meaning of the narration to determine which words to emphasize. However, general-purpose voiceover guidelines suggest: emphasizing one of the first two words in a sentence can help listeners focus attention on the remainder of the sentence; emphasizing the end of a sentence signals the end of a thought; emphasizing action words and descriptive words helps listeners focus on the action and the subject; and emphasizing reference words (e.g. subject names) helps listeners focus on the subject, location and objects in the story [11].

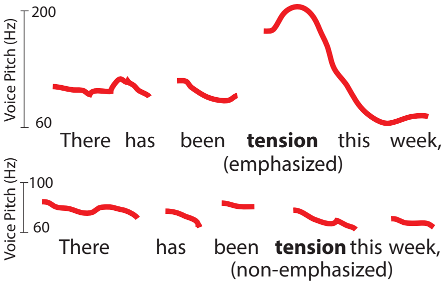


Figure 1. In two spoken versions of the same phrase, the pitch contour for the word "tension" features a steep pitch increase when the speaker emphasizes the word (top) versus when she does not (bottom).

Voiceover acting books further suggest that the most common way to emphasize a word is to increase the pitch on the stressed syllable of the word while slightly increasing the volume and duration of the word [2, 11, 29]. A common emphasis pitch contour of a word starts at a low pitch, rises to a greater pitch at the stressed syllable in the word, and then falls back down to a lower pitch by the end of the word (Figure 1). We can provide feedback on whether the speaker emphasized each word by analyzing its pitch contour, volume, and duration. To resynthesize a sentence to emphasize a non-emphasized word, our system adjusts the pitch contour of the word to mimic a target emphasis contour, and increases the duration and volume to draw further attention to the word.

Variety

If a speaker uses a small range of pitches and always speaks at the same tempo, the resulting speech can sound robotic and listeners may lose focus. Instead, to keep the voiceover interesting, speakers add variety to their delivery. They use pitch variation to expand their range, and elongate words to change the tempo of the sentence. We analyze the pitch and tempo to provide feedback about the variety in the recording. Our system focuses on helping users speak with more variety. Although some speakers have *too much* vocal variety, novice speakers are less likely to have this issue.

Flow

To control the flow of a sentence, voice actors continuously adjust the tempo of the narration, so as not to speak too fast or too slow. They are also careful about where they insert pauses to avoid unnecessarily disrupting a sentence. Less experienced speakers often read in a disconnected way; they insert extraneous pauses between words or speak so fast that they forget to pause entirely. Speakers can improve the flow of a narration by speaking at a natural tempo and minimizing unintentional pauses. We provide feedback about the flow by analyzing the speed of the narration, i.e. by comparing

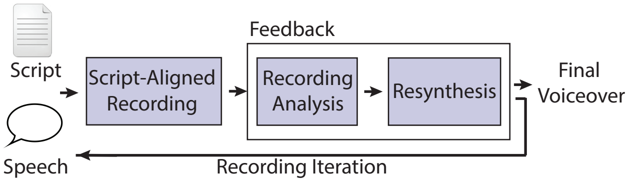


Figure 2. A user records part of a narration script in our system. After she records part of the script, our system displays text-based feedback and provides audio resyntheses that correct problems in the narration. This text and audio feedback helps her understand how to iterate and improve the recording.

the duration of each spoken word to the typical duration of that word (learned from data). We also analyze the duration and location of pauses within a sentence. To resynthesize the recording with improved flow, we can slow down or speed up the tempo, as well as insert or remove pauses as necessary.

Diction

To make voiceovers easy to understand and follow, speakers must articulate their words clearly. Clear enunciation correlates with exaggerated mouth movements [19]. We thus provide feedback about the quality of the diction by using a face tracker to analyze the speaker's mouth movements throughout a recording session. We do not resynthesize speech to correct for poor diction. Increasing or decreasing enunciation in recorded speech is an open problem for future research.

INTERFACE

This section describes our system's interface, while the ALGORITHMIC METHODS Section explains the techniques that enable these interactions. At a high-level, the user records part of the script and *Narration Coach* continuously time-aligns the recording with the script (Figure 2). When the user stops recording, the interface provides feedback for each recorded take based on the four high-level narration guidelines. *Narration Coach* also resynthesizes the recorded speech to automatically improve it. Users can incorporate the resynthesized version in their final narration, or use it as a suggestion of how to improve their narration. In our system, we call a sentence in the script text a ***line***, and we call an audio recording of one line a ***take*** of that line.

Transcript-guided recording

The user launches *Narration Coach* after she writes her script. She imports the script text file and our interface shows a line-by-line view of the script in the main window (Figure 3, left), where each line corresponds to a sentence. She can edit, add, and remove lines from the script in this window. She can also select words and mark them as targets for emphasis (⌘+U). Our interface underlines such emphasis words.

If the user is unsure about which words to emphasize, she can select a line and click on the "Annotate line" button in the toolbar, which underlines suggested words to emphasize. *Narration Coach* does not annotate the entire script automatically by default because desired emphasis can be context-dependent; a speaker may emphasize the same sentence in different ways to convey different meanings. Nevertheless, if the user does not have a specific interpretation in mind, our system applies a general set of rules as described in the GUIDELINES Section to pick words to emphasize.

To begin recording the script, the user clicks the "Record" button on the toolbar and starts reading the script line by line (Figure 3a). She does not need to read the lines exactly as they appear in the script; the speaker can modify lines, e.g., by re-wording phrases.

When the user presses the "stop recording" button, *Narration Coach* analyzes the speech. It first transcribes the speech and then automatically segments and aligns the speech transcription to the best matching set of consecutive lines in the script. The main script view displays these recorded lines in a light blue font to indicate to the user that those lines have recorded takes associated with them (Figure 3b). If after the speech analysis, our system determines that a recorded line follows all four high-level narration guidelines, the script view instead colors the line in dark blue. The font colors let the user quickly see which parts of the script she has yet to record (black lines), and which parts to record again to improve the quality of the narration (light blue lines). When the user clicks a blue-colored line in the script, the *take inspector* appears and shows all of the recorded takes for that line (Figure 3, right). The take inspector also provides feedback about the delivery and allows users to resynthesize takes.

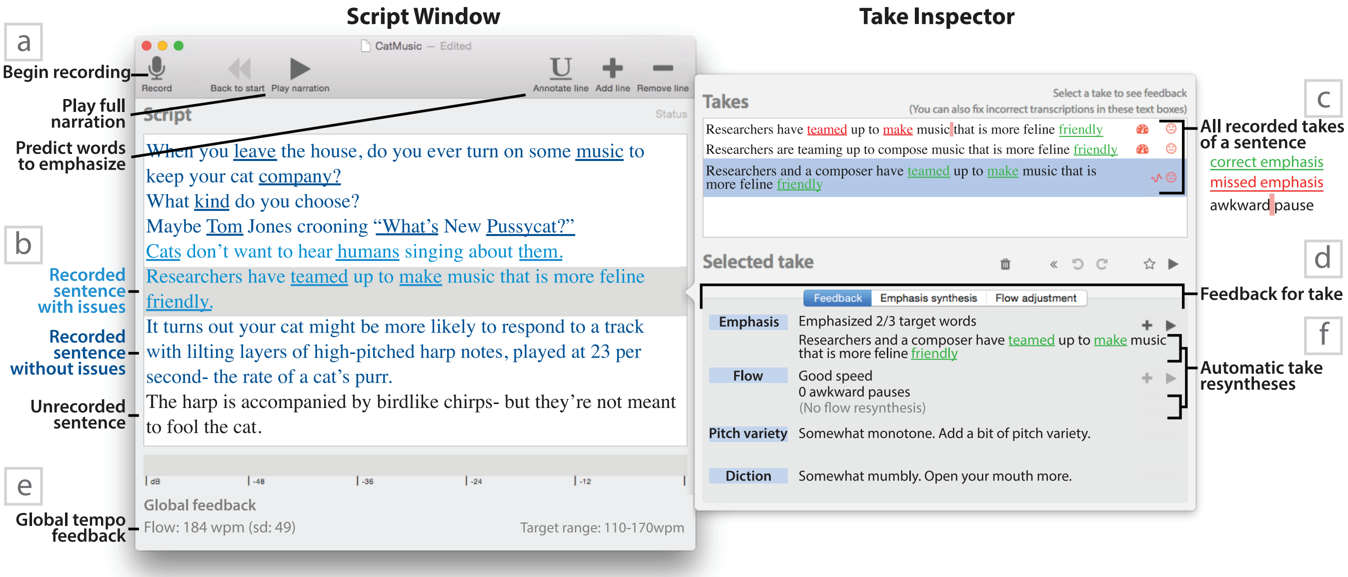
Speech feedback

In the take inspector (Figure 3, right), *Narration Coach* provides feedback about the four high-level guidelines: emphasis, variety, flow, and diction. *Narration Coach* detects the emphasized words in each recorded take. As in the script view, the take inspector underlines each word that the user intended to emphasize. For each of these words, the take inspector renders the word in green if the user successfully emphasized the word and red if the user did not emphasize the word (Figure 3c). This visual encoding allows the user to see where she needs to place more emphasis in subsequent recordings. If the user disagrees with the detected emphasis, she can select words in the recorded take and add or remove the emphasis (⌘+B).

For other guidelines, we aim to give the user actionable advice rather than have her focus on specific quantitative properties of her speech. To do this, we provide text-based feedback rather than solely giving numerical or visualization-based feedback.

Narration Coach provides textual feedback describing the variety of the performance. Feedback lines indicate the variety in the pitch and tempo of the speech and suggest that the user add more or less variety in further recordings if needed. Our system reports pitch variety feedback for each recorded take (Figure 3d). While excessive *tempo* variety within a single line sounds unnatural, tempo variety over multiple sentences helps to hold the listener's attention. Our system gives tempo variety feedback about the full narration in the script window (Figure 3e).

To provide feedback about the flow of the speech, our system displays the speech tempo and the number of mid-sentence



When you leave the house, do you ever turn on some music to keep your cat company?
What kind do you choose?
Maybe Tom Jones crooning "What's New Pussycat?"
Cats don't want to hear humans singing about them.
Researchers have teamed up to make music that is more feline friendly.
It turns out your cat might be more likely to respond to a track with lilting layers of high-pitched harp notes, played at 23 per second- the rate of a cat's purr.
The harp is accompanied by birdlike chirps- but they're not meant to fool the cat.

Global feedback
Flow: 184 wpm (sd: 49)
Target range: 110-170wpm

Researchers have teamed up to make music that is more feline friendly
Researchers are teaming up to compose music that is more feline friendly
Researchers and a composer have teamed up to make music that is more feline friendly

Emphasis	Emphasized 2/3 target words Researchers and a composer have teamed up to make music that is more feline friendly
Flow	Good speed 0 awkward pauses (No flow resynthesis)
Pitch variety	Somewhat monotone. Add a bit of pitch variety.
Diction	Somewhat mumbly. Open your mouth more.

Figure 3. The main window of our interface shows the narration script (left). The user (a) underlines words that she intends to emphasize and records part of the script. The script window (b) uses font color to show which lines the user has not recorded yet, which lines have problematic recordings, and which lines have good recordings. When the user clicks on a blue line in the script, our system displays the take inspector (right) for recordings of that line. The take inspector (c) shows the user words she correctly emphasized and words she failed to emphasize. By selecting a specific take of the line, she can (d) see feedback about that take's emphasis, flow, variety, and diction. Our system also shows (e) global tempo variety. She can (f) listen to resynthesized versions of the recording that address emphasis and flow issues and replace the take with a resynthesized version.

silences. These displays suggest that the user slows down or speeds up, and reduces unnecessary mid-sentence pauses if necessary. *Narration Coach* renders awkward pauses as red boxes in the text of the take. We also use text to provide feedback about the quality of the diction, suggesting that a user open her mouth more to improve mumbled speech (see ALGORITHMIC METHODS for details on text feedback).

Speech resynthesis

Resynthesizing audio—splicing multiple takes together or manipulating the underlying pitch, volume, and duration contours (time-varying functions) of individual takes—allows the user to create and hear modified versions of her performance. Our system automatically generates resynthesized takes to address problems with emphasis and flow. As our system resynthesizes audio with contours further from the original take's contours, the resulting audio becomes more audibly different from the original recording, but also gains more audio artifacts. The user can use these resynthesized takes to replace her original take, or she can use them as audio feedback to guide her performance next time she records a take of that line.

Narration Coach provides two forms of speech resynthesis: emphasis and flow (Figure 3f). When the user selects a sentence in the take inspector, our system automatically generates these two resynthesized versions of the sentence. If the user fails to emphasize words that she underlined in the script, our system generates a version of the sentence that adds emphasis to those words. The user can click on the "plus" button to replace the original take with the resynthesized take.

If the user spoke too quickly or too slowly in the recording, *Narration Coach* resynthesizes a slower or faster version of the take, respectively. This resynthesized take adjusts the speed while preserving the other characteristics of the recording. Our system also automatically adds pauses between sentences and reduces unnecessary pauses within a sentence.

Constructing the final narration

As the user records and re-records the lines in her script, *Narration Coach* captures a complete recording of the script. The user can listen to this recording at any point by clicking the "play narration" button in the main script window. Our system analyzes the problems in each line to automatically select and play the best take of each line for the final version of the voiceover. The user can override this default "best-take" selection behavior by clicking the "star" button on her favorite take in the take inspector. When the user finishes recording the narration, she selects "Export" from the file menu to export her full narration as a high quality, uncompressed WAV file.

ALGORITHMIC METHODS

The features in *Narration Coach* rely on text and audio processing algorithms.

Script Analysis

Narration Coach provides suggestions for words in the script to emphasize by applying rules proposed by voiceover experts [11]. Our system suggests users emphasize: descriptive words (adjectives), proper nouns, action verbs, and words at the beginnings and ends of sentences. *Narration Coach* uses the Mac OSX built-in `NSLinguisticTagger` API to tag

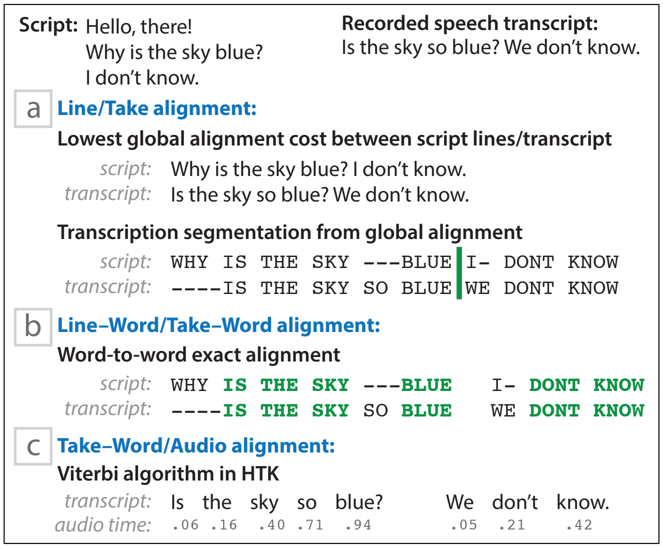

Script: Hello, there!
Why is the sky blue?
I don't know.
Recorded speech transcript:
Is the sky so blue? We don't know.
a Line/Take alignment:
Lowest global alignment cost between script lines/transcript
script: Why is the sky blue? I don't know.
transcript: Is the sky so blue? We don't know.
Transcription segmentation from global alignment
script: WHY IS THE SKY ---BLUE I- DONT KNOW
transcript: ----IS THE SKY SO BLUE WE DONT KNOW
b Line–Word/Take–Word alignment:
Word-to-word exact alignment
script: WHY IS THE SKY ---BLUE I- DONT KNOW
transcript: ----IS THE SKY SO BLUE WE DONT KNOW
c Take–Word/Audio alignment:
Viterbi algorithm in HTK
transcript: Is the sky so blue? We don't know.
audio time: .06 .16 .40 .71 .94 .05 .21 .42

Figure 4. ***Narration Coach*** **aligns recorded speech with the script. First, (a) it finds the set of script lines that correspond to the recording. Then, it segments the transcript based on a global alignment with the script lines. It (b) finds word matches between the script lines and the transcript. Finally, (c) it aligns the words in the take to the speech audio.**

the part-of-speech of each word, and then applies the rules listed in the GUIDELINES Section to label emphasis. The tagger does not differentiate between action verbs and non-action verbs, so we supply a list of non-action verbs to ignore during annotation.

Transcript-guided recording

While the user is speaking, our system records the audio and runs Mac OSX's Enhanced Dictation tool in the background to compute a text transcript. In order for our interface to organize, analyze, and manipulate the recorded audio, it requires alignment meta-data (Figure 4). When the user stops recording, our system performs a multi-staged alignment process to generate the following meta-data:

- The **line/take alignment** is the link between each recorded take and a line in the original script, so our system can show users all takes of a given script line.
- The **line-word/take-word alignment** links words in each recorded take and words in the line corresponding to that take. If the user did not read the script verbatim, some words in the recorded take may not have links to the original line, and some words in the original line may not have links to the recorded take. We need this meta-data to determine if the take includes the words that the user intended to emphasize.
- Finally, the **take-word/audio alignment** is a set of timestamps in a take's recorded audio for the beginning and end of each word and phoneme, so our system can analyze and manipulate the vocal components of the recording at the word and phoneme levels.

Our alignment process makes no assumption about where the user starts speaking within the script, but does assume that she speaks lines sequentially from the starting point. The user also does not have to speak lines exactly as written—she can change wordings while she is recording. Because the user can record multiple script lines at once, our system first segments the transcript t into its associated script lines. The transcript corresponds to a consecutive interval $[i, j]$ of script lines $\mathbf{s}$. Our system finds the interval with the best match to the transcript, as measured by the minimum cost global alignment $c(t, s_{[i,j]})$ between the transcript and the concatenated script lines $s_{[i,j]} = s_i + \cdots + s_j$. We find this interval by first applying Needleman-Wunsch [17] to compute the global alignment between the transcript t and each script line s_i. The cost of this alignment corresponds to a score for interval $[i, i]$. Then, our system repeatedly extends each of these intervals by one line as long as the alignment cost decreases, i.e., $c(t, s_{[i,j+1]}) < c(t, s_{[i,j]})$. From these final intervals, our system selects the one with the smallest cost. Our system then segments t into $j - i$ sentences according to this minimum cost alignment, which gives us the line/take alignment (Figure 4a).

Our system finds the line-word/take-word alignment by searching the global alignment from the previous step for words in the transcript that are exactly aligned to words in the script (Figure 4b). Finally, *Narration Coach* computes word and phoneme timestamps for the recorded audio using the HVite component of HTK, a hidden Markov model toolkit [30] (Figure 4c).

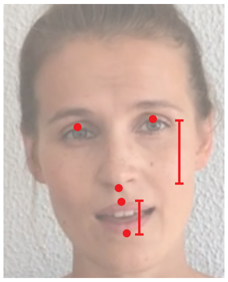
Speech feedback

To detect emphasized words, *Narration Coach* applies a two-phased process. AuToBI [20]—a tool for predicting prosodic annotations in speech—reliably detects words that have pitch accents, which are pitch "configurations that lend prominence to their associated word" [25]. While pitch accents are necessary for emphasis, they are not sufficient. For example, a word preceding a pitch accented word may contain a much larger pitch accent, overpowering any emphasis that a listener may otherwise hear in the second word. We apply the AuToBI pitch accent detector to find a subset of words that have pitch accents. In order to find the pitch-accented words that sound emphasized, our system searches for typical emphasis contours (Figure 1): our system runs a second pass over the pitch-accented words, finding those that are louder (> 1.25 decibels) or higher pitched (> 1.25 times) than the preceding word, or those that have more pitch variation (> 1 standard deviation) than the sentence as whole. We set these thresholds by analyzing contours before, during, and after emphasized words in speech recordings. *Narration Coach* underlines the user's target emphasized words, rendering successfully hit target words in green and missed target words in red.

Our system finds pitch variety by first applying Mauch and Dixon's [16] probabilistic YIN smooth pitch estimation algorithm and computing the standard deviation of the log of the pitch in the take. *Narration Coach* displays this standard deviation in the take inspector, mapping the value to a level of text feedback— $[0, .2)$: "Monotone. Add more pitch variety," $[.2, .3)$: "Somewhat monotone, Add a bit of pitch variety," $[.3, .4)$: "Good pitch variety," and $[.4, \infty)$: "Excellent pitch variety." We set these mappings empirically by listening to and qualitatively rating narration audio clips. If the user finds the mapping to be inaccurate, she can adjust the mapping through our system's preferences window.

We provide feedback on tempo variety for the narration as a whole. *Narration Coach* computes the words per minute (WPM) and the standard deviation of a moving average of WPM of the full narration and displays these values below the script in the main script window.

Our system uses a different tempo metric to provide flow feedback at the take-level. Sentences have large differences in word length distribution, so WPM is less useful as a metric for take-level speed analysis. Instead, our system computes the speed of each take by comparing the duration of each spoken word to the expected duration of that word. We use the transition probabilities in a monophonic hidden Markov model [31] to model the duration of each English phoneme. The expected duration of a word is the sum of the expected durations of the phonemes in that word. Our system computes the cumulative distribution function (CDF) for the duration of the words in each take to report a value between zero—the speaker said the sentence quickly—and 1—the speaker said the sentence slowly. *Narration Coach* maps the probability to different levels of text feedback— [0, .25): "Very slow! Speed up," [.25, .3): "Somewhat slow. Speed up," [.3, .4): "Good speed," [.4, .45): "Somewhat fast. Slow down," and [.45, 1]: "Very fast! Slow down." As with the pitch variety mappings, we set these empirically and they are user-customizable.



When people articulate clearly, they open their mouths wider than when they mumble [19]. In addition to recording audio, *Narration Coach* captures video from the computer's webcam as the user speaks. Our system detects mumbling by analyzing the speaker's facial movement. *Narration Coach* runs Saragih's face tracker [24] on the captured video and records four points for each captured frame: the topmost mouth point, the bottommost mouth point, the topmost eye point, and the bottommost nose point (see inset). Our system computes the ratio $\frac{mouth_{top} - mouth_{bottom}}{eyes_{top} - nose_{bottom}}$ for each frame and then computes the standard deviation of these ratios. The higher the variance of the $mouth_{top} - mouth_{bottom}$ difference, the wider the speaker is opening her mouth while speaking. The denominator acts as a normalizing term to preserve scale-invariance in this ratio. *Narration Coach* maps the standard deviation of these ratios to text feedback in the take inspector— $[0, .02)$: "Very mumbly! Open your mouth more," $[.02, .04)$: "Somewhat mumbly. Open your mouth more," and $[.04, \infty)$: "Well-articulated." Before the user starts recording, we calibrate these mappings based on the size of the user's closed mouth and the size of the user's wide open mouth.

Speech resynthesis

Narration Coach provides automatic speech resynthesis methods for emphasis and for flow. If the user fails to emphasize words she underlined in the script, *Narration Coach* resynthesizes the take with emphasis added to those words in a two-phased procedure. First, for each missed target word, our system checks if the user properly emphasized the word in another take of the same script line. If she did, *Narration Coach* creates a new version of the current take with the target words replaced with the audio of the properly emphasized words. Second, if the user did not emphasize the target word in another take, our system applies a parameter-based resynthesis method. This method uses PSOLA [3, 28], a speech processing technique that modifies the pitch and duration of speech by manipulating small pitch-aligned overlapping windows of the audio signal— moving them closer together or further apart to adjust pitch, and repeating or eliminating windows to adjust duration. Our system constructs new pitch, volume, and duration contours to match typical emphasis contours:

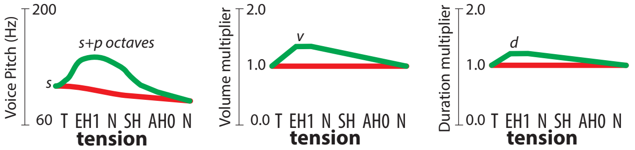


Figure 5. Our system adds emphasis to a word by modifying words' vocal parameter contours (red: before; green: after) and creating resynthesized audio using PSOLA.

Pitch. Every word contains a stressed vowel phoneme. Our system creates the new pitch contour by setting a pitch peak at the start of that phoneme. The new pitch peak is the maximum pitch of the word plus a parameter p octaves. We construct the contour according to the TILT [10] pitch contour generation model. In this model, the contour follows a piecewise quadratic function ascending from the word's original starting pitch to the new maximum pitch, followed by another piecewise quadratic function descending to the word's original ending pitch (Figure 5, left).

Volume. Our system manipulates the volume of the target word by setting a volume multiplier of v from the start of the stressed vowel phoneme through the end of the stressed vowel phoneme. It sets the volume multiplier to 1.0 elsewhere and adds a linear transitions to and from the stressed vowel phoneme, so that the volume variations are not perceived as too abrupt (Figure 5, center).

Duration. Speakers can extend the typical duration of a word to give it more weight. Our system sets the new duration contour akin to the volume contour, substituting a duration multiplier d for v (Figure 5, right). This extends and adds emphasis to the vowel phonemes, which sounds more natural than extending all of the word's phonemes.

Our system parameterizes the contours by the values p, v, and d. We set these defaults to $.25$ octaves, 1.25, and 1.1 respectively, which represents an audible but not extreme increase in these parameters. If the user wants more control, she can modify the defaults for p, v, and d and fine-tune the synthesized emphasis for each word.

The first phase of emphasis resynthesis has the advantage of using emphasized words that the user said in same context as the target words, and avoids adding artifacts that the

		Recording comparison				*Narration Coach* usage								
Name	Script lines	Takes per line (Aud.)	Takes per line (*NC*)	Final dur. (Aud.)	Final dur. (*NC*)	Start rec.	Open take insp.	Edit dict.	Edit detected emph.	Play take	Play emph. resynth.	Play flow resynth.	Favorite take	Play full narr.
ballotselfies	8	1.375	2	0:56	0:54	10	14	9	2	17	4	3	8	1
blue	8	1	2.125	0:42	0:48	7	32	10	1	26	14	10	8	2
coloradosv	8	2	3.125	0:47	0:57	14	25	4	0	14	2	0	9	1
mosquitos	10	1.8	1.6	0:46	0:55	17	38	14	1	15	2	11	6	3
voting	11	1.36	2	0:58	0:59	12	26	0	1	13	1	3	4	4

Table 1. Five participants each recorded two narrations for a script—one using a traditional DAW (Adobe Audition) and one using *Narration Coach*. Users recorded more takes per line and produced slower-paced narrations using our system. We instrumented *Narration Coach* to record usage statistics. "Start rec." is the number of times the user initiated a recording session, and "open take insp." is the count of times the user clicked a line in the script to view the take inspector. "Edit dict". refers to the number of times the user corrected the speech-to-text dictation, and "edit detected emph." tallies times when the user disagreed with and changed the emphasis detection. "Play take," "play emph. resynth.," and "play flow resynth." refer to the user playing different versions of the take within the take inspector. "Star take" tracks the user selecting a favorite take, while "play full narr." tracks when the user listened to the entire narration.

parameter-based resynthesis method may introduce; however, it requires that she emphasized the target word in another take and it can introduce artifacts if the speaker had different tones between the takes.

We can apply similar techniques to remove emphasis from words that the speaker over-emphasizes. However, our system focuses on fixing non-emphasized words because novice speakers tend to forget to emphasize words.

Narration Coach modifies the flow of a sentence by creating a faster or slower version of the sentence as needed. It lengthens or shortens the durations of words using PSOLA time-stretching. *Narration Coach* also removes unneeded and awkward pauses in the take, i.e., pauses over a quarter of a second long that do not occur directly after commas or other punctuation.

Constructing the final narration

Narration Coach creates a final narration by playing the best take of each line in succession. Our system defines the best line as the line that follows the most guidelines. A take follows the emphasis guideline if the speaker emphasized each target word. A take follows the variety, flow, and diction guidelines if the respective computed feedback values fall within the "good" range according to the text feedback mappings. If multiple takes follow the same number of guidelines, our system favors the most recent take, and if the user clicks the "star" to mark a take of a line as a favorite, our system uses that take instead. *Narration Coach* mixes room tone into the final narration to prevent audible, unnatural dips to silence in the transitions between sentences.

RESULTS

We conducted a pilot study where we introduced *Narration Coach* to five users (3 female, 2 male), none of whom has had formal voiceover or audio recording/editing training. We began each session by providing the user with a script and asking her to read it out loud to learn its words and phrasing. We then had her record a narration of that script using Adobe Audition, a traditional audio recording and editing tool. We allowed the user to record additional takes—of the entire script or of specific lines—until she was satisfied with the narration, and we helped her edit the takes together so she did not have to learn how to use the software. After they finished recording the narration with Audition, we loaded the script in *Narration Coach*. We then gave the user a 10-minute demonstration of our system's recording, feedback, and resynthesis tools, and asked her to use *Narration Coach* to create a narration. At the end of each session, we solicited the user's feedback on our system's features, and asked her to select whether she preferred her first narration or her narration created in *Narration Coach*. In total, each pilot study session lasted one hour and we compensated each participant with a $15 gift card.

We had the participants record their narrations in Audition before using our system to prevent them from transferring specific feedback from *Narration Coach* to Audition. However, a different type of learning effect is possible in our non-counterbalanced pilot: users may get better at the narration task over time, which could lead to users always producing better narrations in the second task. We tried to avoid this bias by having our users read the script out loud before recording, and by allowing them to re-record the script in Audition as many times as they wanted in order to construct a narration that they liked.

Overall, the users were enthusiastic about our narration tool. Each of the five users preferred the narration they created with *Narration Coach* over the narration they created with Audition, and every user noted that they would use our tool to record narrations. Table 1 summarizes the participants' usage of Audition and *Narration Coach* in the recording session. The supplemental material[1] for this paper includes the final audio from these sessions. Users demonstrated different usage patterns in *Narration Coach*; for example, some perfected one line at a time while others listened to all available resyntheses for the entire script. They spent more time recording narrations using our system than with Audition, noting that this was because they tried to improve on their takes based on *Narration Coach*'s feedback. They recorded more takes per line in our system than when using Audition.

The recording tools in *Narration Coach* excited the users. Notably, they liked how it organized audio by script line into different takes, how they could start recording from anywhere in the script, and how they could star a favorite take to play in the final narration instead of manually splicing in the best take as in a timeline-based editor. One participant noted that

[1] `http://vis.berkeley.edu/papers/narrationcoach/`

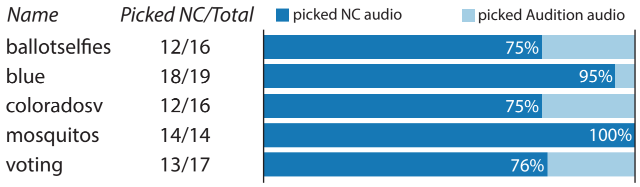


Figure 6. Listeners on Mechanical Turk thought the narrations made with Narration Coach were higher quality than the narrations made without it for all five of the audio pairs created by participants in our pilot study.

"Organizing audio line by line into takes was super helpful because it didn't invalidate an entire recording session if I stumbled over one word." Another appreciated the "star" feature, as it enabled *"[re-recording] several times without the hassle of splicing the new takes in."* Our system functions well even if the speech transcription has minor errors (e.g., with one or two errors in a take, *Narration Coach* can still compute accurate alignment meta-data and thus provide take feedback and resyntheses). However, the automatic speech-to-text transcription was highly inaccurate for one user who had an accent. That user had to manually correct each automatically generated transcript. We expect that off-the-shelf dictation tools will improve as speech recognition research progresses.

Users found the feedback tools useful, as well, with the feedback on speed and awkward pauses being the most useful. One user described that speed detection was useful because *"I usually speak very fast without realizing it,"* and added that emphasis detection was useful because it *"allowed me to make a mental note of whether or not I emphasized certain words."* The participants referenced the pitch variety and diction feedback less often, in part because users did not trust the feedback: *"I... didn't feel like I was very monotone."* Even if the user *was* being monotone or mumbling, our system needs to better communicate feedback that users may find incorrect or even insulting. *Narration Coach* addresses this problem with emphasis detection by allowing users to toggle the feedback if they disagree with the analysis.

Our participants repeatedly used and complimented the automatic flow resynthesis feature that altered a take's speed and removed awkward pauses. However, they found the automatic emphasis resynthesis tool to be less useful. Instead of using the tool as a way to hear what proper emphasis might sound like to guide their future takes, the users treated the emphasis resynthesis as a candidate for a final recording. As such, the users found audio artifacts caused by the parameter-based resynthesis to be off-putting. In a longer recording session, users would likely record more takes of each line, so the emphasis resyntheses would be more likely to use replacement-based emphasis rather than contour-based emphasis, which tends to produce fewer audio artifacts.

Listening evaluation. While all of our participants preferred their narration created with *Narration Coach* to their narration created with Audition, we sought additional input from crowd workers. For each script, 20 US workers with over 95% approval rates on Amazon Mechanical Turk listened to both versions of the narration. We asked them to select which clip had the higher overall narration quality. We rejected workers that did not spend enough time on the task to listen to both clips in their entirety, which left an average of 16 workers per narration pair. Figure 6 shows the proportion of listeners that selected each narration. For each narration pair, we performed a binomial test to see if the proportion of listeners that selected the *Narration Coach* recording was greater than chance (50%). In all five cases, more listeners preferred the *Narration Coach* audio clip to the other clip ($p < .05$).

CONCLUSION AND FUTURE WORK

High-quality narrations are integral to creating captivating digital content, but novice users are not aware of the guidelines necessary to produce such voiceovers. We have presented *Narration Coach*, an interface to help users create narrations by providing text and audio feedback throughout the recording process. Using our system, novice users can better create high-quality narrations that more closely adhere to professional voiceover guidelines.

Our system focuses on constructing a better audio narration but does not consider other production constraints such as syncing a narration to a video. We could integrate these per-line alignment constraints into the system an add additional feedback and resyntheses that help the user hit exact times. The system could additionally alert the user when lines in the script are too verbose to fit within their video constraints.

Narration Coach includes feedback on emphasis, variety, flow, and diction, but it does not compute feedback on the *tone*— the timbre and emotion of the voice. A user may aim to record an excited or sullen narration. We would need to investigate the low-level characteristics of these kinds of speech to understand how to give users feedback about how to change their tone. In order to keep the user focused on recording her lines, our system provides feedback *after* the user records lines, but not *during*. If the user wanted to use our tool while delivering a public speech or presentation, we would need to add live feedback that was actionable without being distracting. We could also add feedback tools that focus on lower-level audio quality issues like compression, microphone noise and plosives. Manual correction of speech-to-text errors adds undesired usage time to our tool. We need to study how much transcription inaccuracy our algorithms can tolerate before the user must manually correct the entire transcript.

While our system focuses mainly on feedback, another narration assistance system could use the speaker's voice as input and apply voice transformation techniques [26] to generate a narration in a different voice entirely. For example, if the user does not like her own voice, she could generate a version of her narration that sounded like a famous voice actress.

ACKNOWLEDGMENTS

We thank Amy Pavel and Julian Suhr for helping us test *Narration Coach*, and Marissa Pilger for narrating the accompanying video.

REFERENCES

1. Black, A. W., and Taylor, P. A. The Festival Speech Synthesis System: System documentation. Tech. Rep. HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Scotland, UK, 1997. Avaliable at http://www.cstr.ed.ac.uk/projects/festival/.

2. Blu, S., Mullin, M. A., and Songé, C. *Word of Mouth: A Guide to Commercial and Animation Voice-over Excellence*. Silman-James Press, 2006.

3. Boersma, P., and Weenink, D. Praat, a system for doing phonetics by computer.

4. Carter, S., Adcock, J., Doherty, J., and Branham, S. Nudgecam: Toward targeted, higher quality media capture. In *Proceedings of the International Conference on Multimedia*, ACM (New York, NY, USA, 2010), 615–618.

5. Chang, A. R., and Davis, M. Designing systems that direct human action. In *Proceedings of the SIGCHI Extended Abstracts on Human Factors in Computing Systems*, ACM (2005), 1260–1263.

6. Davis, M. Active capture: integrating human-computer interaction and computer vision/audition to automate media capture. In *Proceedings of the International Conference Multimedia and Expo*, vol. 2, IEEE (2003), II–185—II–188.

7. Design, S. S. Voice lesson 2: Marking a script. `https://www.youtube.com/watch?v=LwS7WD7WQ3Y`. Accessed: 2015-03-31.

8. Dolson, M. The phase vocoder: A tutorial. *Computer Music Journal 10*, 4 (1986), 14–27.

9. Drew, P. Take your copy to the woodshed. `http://www.peterdrewvo.com/html/analyze_the_copy_first.html`. Accessed: 2015-03-31.

10. Dusterhoff, K., and Black, A. W. Generating f0 contours for speech synthesis using the tilt intonation theory. In *Intonation: Theory, Models and Applications* (1997), 107–110.

11. Goldberg, D. *The Voice Over Technique Guidebook with Industry Overview*. Edge Studio, 2010.

12. Heer, J., Good, N. S., Ramirez, A., Davis, M., and Mankoff, J. Presiding over accidents: system direction of human action. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2004), 463–470.

13. Hindenburg ABC. `http://hindenburg.com/products/hindenburg-abc/`, Apr. 2015.

14. Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H. Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE (2008), 3933–3936.

15. Kurihara, K., Goto, M., Ogata, J., Matsusaka, Y., and Igarashi, T. Presentation sensei: A presentation training system using speech and image processing. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, ACM (New York, NY, USA, 2007), 358–365.

16. Mauch, M., and Dixon, S. pYIN: a fundamental frequency estimator using probabilistic threshold distributions. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (2014), 659–663.

17. Needleman, S. B., and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology 48*, 3 (1970), 443–453.

18. Robertson, H. 10 ways to build your voice-over skills. `http://www.videomaker.com/article/15804-10-ways-to-build-your-voice-over-skills`. Accessed: 2015-03-31.

19. Rodgers, J. *The Complete Voice & Speech Workout: 75 Exercises for Classroom and Studio Use*. Hal Leonard Corporation, 2002.

20. Rosenberg, A. Autobi-a tool for automatic tobi annotation. In *Proceedings of INTERSPEECH* (2010), 146–149.

21. Rubin, S., and Agrawala, M. Generating emotionally relevant musical scores for audio stories. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, ACM (2014), 439–448.

22. Rubin, S., Berthouzoz, F., Mysore, G., Li, W., and Agrawala, M. Underscore: musical underlays for audio stories. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, ACM (2012), 359–366.

23. Rubin, S., Berthouzoz, F., Mysore, G. J., Li, W., and Agrawala, M. Content-based tools for editing audio stories. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, ACM (2013), 113–122.

24. Saragih, J. M., Lucey, S., and Cohn, J. F. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision 91*, 2 (2011), 200–215.

25. Silverman, K. E., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., Pierrehumbert, J. B., and Hirschberg, J. Tobi: a standard for labeling english prosody. In *Proceedings of The Second International Conference on Spoken Language Processing* (1992), 867–870.

26. Stylianou, Y. Voice transformation: a survey. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.*, IEEE (2009), 3585–3588.

27. Taylor, P. *Text-to-speech synthesis*. Cambridge university press, 2009.

28. Valbret, H., Moulines, E., and Tubach, J. Voice transformation using psola technique. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1 (Mar 1992), 145–148 vol.1.

29. Wilcox, J. *Voiceovers: Techniques and Tactics for Success*. Allworth Press, 2007.

30. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. *The HTK book*, vol. 2. Entropic Cambridge Research Laboratory Cambridge, 1997.

31. Yuan, J., and Liberman, M. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America 123*, 5 (2008), 3878.