TIME COMPRESSION OF INSTRUCTIONAL VIDEOS VIA MULTIMODAL ANALYSIS

Vinay Melkote, Sonal Patil, S. Mohana Prasad, Ankita Patil, Sumit Negi, and Om Deshmukh

Multimedia Analytics Group, Xerox Research Centre India (XRCI), Bangalore, India

ABSTRACT

The spontaneous and interactive teaching style of the lecturer in audio/video (AV) recordings from real classrooms, while appropriate for the live audience, may seem overly belabored for an internet-only user who lacks the context of the traditional classroom. A time-compressed version of the lecture recording that faithfully replicates the most important segments, speeds through sections that provide little advantage for comprehension of the topic, and reduces the viewing burden may thus be preferred. These classroom recordings are typically distinguished by the presence of a single speaker, the lecturer, who conveys the most important information, presence of disfluencies/filler words in his/her speech, and sections with writing and a corresponding slowing of speech. We propose a technique for non-uniform time-compression that involves condensing information about these distinct characteristics into a saliency score for each frame which in turn drives an overall optimization routine to pick the subset of frames to construct the sped-up signal from, while ensuring AV sync. A subjective test demonstrates a strong user preference for the proposed technique over a widely available compression tool applicable to generic multimedia content. Multimodal analysis techniques to detect these characteristics and their accuracy are described.

Index Terms— temporal compression, classroom recordings, instructional videos, multimodal analysis, variable frame rate

1. INTRODUCTION

The last few years have witnessed an exponential increase in the amount of educational content available online. This is due in part to the success of MOOC (Massive Open Online Courses) platforms such as Coursera, edX, and Udacity. Many universities and K-12 educational institutions are making recordings of their classrooms available on open platforms such as YouTube. Such classroom recordings amount to tens of thousands of hours of instructional video data. While MOOC content is scripted specifically for an internet-based user population, the teaching style of the lecturer in these ad hoc classroom recordings is naturally more spontaneous and interactive as befits a live audience: the lecturer may belabor on specific points, indulge in question-answer style sessions, etc. The former type of content is typically more fast paced and dense in information, especially given the fact that the internet-based user has the option to stop at any point as well as randomly access any part of the recording. In contrast, the unscripted teaching style in classroom recordings may seem excessively belabored and non-engaging to an online audience and it may thus be beneficial to be able to automatically speed up such content without significant loss of information. One may also desire a sped-up version of the recording when in a time crunch, for instance, to simply revise the topic or because one may be in a low-bandwidth region such as is common in developing nations - it may be easier to download a time-compressed version of the content owing to the smaller file size.

Many of these platforms such as YouTube, or video playback software such as Windows Media Player, or VLC media player already possess the capability to speed up multimedia content during playback. However, these generic techniques are agnostic to characteristics unique to instructional videos, employ no content-based differentiation between segments of the lecture, and the resulting time-compression is thus not geared to maximally preserve useful information. Specifically, recorded lectures feature a single main speaker, the instructor, who provides most of the significant information - the students' response or questions during interactions are less important, and may even be reiterated by the teacher. The instructor's speech contains frequent usage of filler words (such as "alright?", "okay", "right", etc.) and disfluency (such as "um", "ah", etc.) when s/he gathers his/her thoughts or is eliciting reaffirmation from the students (the incidence is higher when the instructor is a non-native English speaker). The speaking may be interspersed with writing on the board and the instructor's speech typically slows down while writing, or during a dictation. Thus segments of low information may be identified via multimodal analysis, i.e., via a combination of audio, text, and video features, and either completely excised out of the lecture or suitably sped up with minimal impact on comprehension. Further, note that the content is amenable for offline analysis, given that it is pre-recorded, and the decisions pertaining to how to achieve a target compression rate can in fact be pre-determined.

These observations motivate the proposed technique for time compression where segments of the content are first classified into one of the four categories based on audio, visual and transcript information: (A) Filler words and disfluencies (B) Non-speech sounds or ambiance (including where the lecturer writes without talking), (C) Non-lecturer speech (D) everything else (including lecturer speech). Audio segments corresponding to classes (A) and (B) are ranked based on their length and proximity to other similarly labeled segments and marked for excision until the resulting compression meets the target rate. When the required signal length is smaller than what such excision achieves, the audio signal is divided into overlapping frames and a saliency score is derived based on their classification into (C) or (D) type, speaking rate, and energy. An optimization routine then marks some of these frames for removal, while the remaining are overlap added, so that locally the compression rate achieved is inversely proportional to the saliency score. A subset of video frames from the original signal are then chosen to achieve the best synchronization with the retained audio frames in terms of their correspondence on the original uncompressed time-scale (thus enhancing lip sync). However, the original video signal is also analyzed to identify key frames that correspond to when the written content is maximum (e.g., when the blackboard is fully written). The density of frames retained is increased in the vicinity of such key frames in order to maximize visual information in the compressed signal. Note that lip sync is a non-issue when the lecturer is writing as s/he is likely to be facing the board. The superiority of the proposed time-compression approach, tailored to classroom recordings, over a generic solution based on the ffmpeg tool kit[1] is established via a blind subjective test. While we relied on manual labeling of the classes to evaluate the proposed time compression algorithm, methods for automatic labeling and their accuracy are also discussed.

2. RELATED WORK

Prior work on temporal compression techniques use acoustic and semantic cues to either alter the playback rate and/or remove unimportant sections from speech signals. For example, the method proposed in [1] alters the playback rate while maintaining the speaker's pitch contour characteristics. Authors in [2] identify and remove silences or inter-word pauses from recordings using acoustic features. They also use intonations to identify important speech segments allowing users to focus on these. Non-linear time compression techniques that exploit fine-grain structure of human speech to differentially speed-up segments of speech have also been popular [3], [4]. Here the degree of compression varies throughout the playback i.e., different portions of the recordings are played at different speeds. A good example is the algorithm proposed by [3] that tries to mimic the compression strategies used by people when they talk fast in a natural setting. Non-uniform compression techniques have been shown to be superior to linear time compression methods where the speech content in uniformly time compressed [5]. Recent approaches combine acoustic features with semantic features derived from Automatic Speech Recognition (ASR) transcripts for achieving higher compression rate. These methods use text processing techniques on ASR transcripts to identify important regions of the recording [6], [7], [8].

A related topic is that of video summarization, i.e., generation of a short summary of the video either by identifying key still frames or by generating a video skim/preview [2] [9]. In either case the intent is to extract a *gist* or highlights, rather than preserving all useful information as is the focus in our paper.

In contrast to prior work that has largely looked at generic multimedia, sports, news, or movie content, we focus exclusively on spontaneous classroom recordings and exploit characteristics unique to them for time compression. Further, we employ a combination of audio, video, and textual modalities to drive our compression system and leverage the fact that the analysis can be effected offline with a view of the entire signal.

3. PROPOSED COMPRESSION OF THE AUDIO STREAM

The audio stream is divided into overlapping frames that are 40ms each and with 75% overlap. The time-compression mechanism works by selecting a subset of these frames and adding consecutive selected frames back together with the same overlap, after multiplying them with a Hamming window. More sophisticated techniques to put the frames back together such as pitch synchronous overlap-add [10] are expected to provide an additional benefit, however our focus here is primarily on the judicious selection of the subset of frames such that the desired compression is achieved.

3.1. Excisions

Class A: Filler words and disfluency: These are employed by the instructor in order to collect his/her thoughts, or to get feedback from the audience on their comprehension of what is being taught. Segments corresponding to these words convey no information to the student, and can thus be completely excised from the recording. The corresponding video frames are all marked for excision.

Class B: Non-speech sounds: These are produced when the teacher is writing without talking, with signature sounds of chalk or marker on the board, when sliding black boards, when the classroom door shuts, when students are laughing or applauding, etc. All such frames contain no informative speech, and can be marked for removal. Such segments (especially those associated with writing) can form a significant fraction of the recording, depending on the topic. In the test case *Michael Cranston* of Sec. 5, a 2x compression could be achieved simply by excision of the portions of the signal corresponding to chalk sounds. Segments in this category larger than 5s in length typically correspond to a continuous section of writing, and are excised in descending order of their length till they are exhausted or the desired signal length is achieved. Other times the lecture may intermittently speak and write, and non-speech segments smaller than 5s in length are grouped with other non-speech segments within a 1 minute window around them and the groups are excised in descending order of their cumulative lengths. This form of excision ensures that any perceived audible artifacts are localized. Note that we do not include segments of writing where there is no associated audio signature under this label. These could also be exploited, for instance, by techniques that excise silences as in prior work [2].

3.2. Speed-up

If higher compression than what can be achieved by excision is desired, the remaining signal is sped up by intermittent dropping of frames (as opposed to continuous dropping characteristic of excision). A positive saliency score with a unit maximum value is assigned for each frame as follows and regions of lower saliency are sped up more.

Class C: Non-lecturer speech: Segments of the video corresponding to where a student asks a question or responds to one tend to contain information of lower value than where the instructor speaks. Further, students may not be within the range of the recording microphone and the instructor very likely repeats the response for the benefit of the online audience. Rather than completely excising such sections we reduce the saliency to a number in $[0.25, 0.5]$, with segments of lower energy (corresponding to farther students and lower information) getting a lower score. All frames within the same non-lecturer segment get the same saliency.

Class D: All other segments: The remaining frames are assumed to correspond to the speech of the lecturer and assigned a saliency based on speaking rate. We identify the location of syllable nuclei in the audio stream [11], and apply a moving average filter of 10s duration to compute the speaking rate (number of syllables per second) for each frame. Fig. 1 compares the distribution of speaking rate for an ad hoc class room recording and a scripted course. Ad hoc classroom recordings tend to have significant variations in speaking rate of the lecturer, especially when the lecturer simultaneously writes and talks, and thus the corresponding distribution is more spread out. In contrast, the pace of the lecturer's speech is far more consistent in the scripted course, and the speaking rate distribution is peaky. The saliency s_k for frame k is computed as:

$$s_k = 1 - \alpha(1 - \frac{min(\bar{r}, r_k)}{\bar{r}}) \tag{1}$$

where r_k is the speaking rate for the frame, $\bar{r}$ is the mean speaking rate over the entire signal, and $\alpha = 0.5$ (based on tuning for

[1] https://www.ffmpeg.org/

[2] This is also called a moving-image abstract, moving story board, or summary sequence.

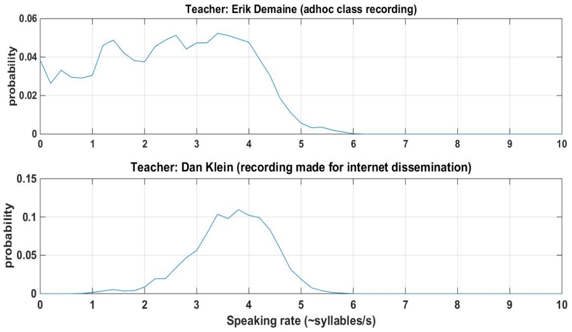


Fig. 1. Distribution of speaking rate for an ad hoc class room recording (top) and a scripted course (bottom)

best compression quality). Thus frames with lower speaking rate are assigned a lower saliency. The saliency is lower bounded to 0.25.

Offline Optimization for Audio Frame Selection The instantaneous compression rate for frame k is defined as $R_k = \frac{\beta}{s_k}$. Let $B = \{b_0, \cdots, b_{N-1}\}$ denote the sequence of decisions b_k for the N frames of the signal: $b_k = 0$ if the frame is dropped and 1 if retained. Initially $b_k = 0$ for all excised frames (Sec. 3.1) and 1 otherwise. A sliding window is used to compute the average number of retained frames around the current frame k, and if the corresponding *achieved* compression rate is less than R_k, b_k is set to 0. The window is then slid over to $k+1$ and the process repeated. If the length of the compressed signal on reaching the last frame is not as desired, the value of β is changed and another pass through the signal is effected.

3.3. Segmentation and Labeling

The compression routine above requires segmentation of the content into sections and labeling each section as one of the four different classes A - D. The experiments described in Sec. 5 utilize manually segmented and labeled data, in order to quantify an upper bound on the quality of the compressed signal one could obtain via the proposed system. Nevertheless, we describe below an autolabeling technique as well, to establish the feasibility of the proposed approach.

The audio signal is segmented based on a combination of frame-level energy and fundamental frequency features, computed every 10ms using the Wavesurfer software [12]. Contiguous regions of 250ms or more with energy below 5% of the overall median energy are identified as silence regions and are automatic candidates for segment boundaries. Among the non-silence regions, locations of energy minima that satisfy the following three conditions are marked as segment boundaries: (a) the energy at the minima is less than or equal to 10% of the overall median, (b) the maximum frame energy in the segment is at least 75% of the median energy, and (c) at least 25% of the frames within the region have a non-zero fundamental frequency, indicating voicing regions. Our initial analysis shows that this strategy leads to a near-perfect segmentation, i.e., the boundaries align well with manually determined boundaries between segments of different classes. The instructor's speech gets segmented at what we refer to as "phrasal boundaries", where s/he pauses to either catch his/her breath, to change topics or to write on the board.

Class A and Class D are initially considered to be of one single class as the disfluencies are really part of the instructor's speech. To estimate the labels of these segments, a 3-class J48 decision tree classifier is trained using Weka [13]. Each segment instance is parameterized using 38x features as described in [14]. These features consist of (a) loudness, defined as normalized intensity raised to a power of 0.3, (b) 12 Mel Frequency Cepstral Coefficients (MFCCs) along with the log energy ($MFCC_0$) and their first and second order delta values, and (c) voicing related features fundamental frequency (F0), voicing probability, Harmonic Noise Ratio and Zero Crossing Rate, and their functionals based on percentiles and quartiles.

Once the segments corresponding to the lecturer's speech are obtained the subset containing disfluencies are to be identified to separate out Class A from Class D. Disfluency detection in spontaneous speech is a well studied area in the speech literature [15], [16], [17]. Any existing technique based purely on acoustic analysis or a combination of acoustic and speech transcript can be used to automatically detect disfluency in the audio.

4. PROPOSED COMPRESSION OF THE VIDEO STREAM

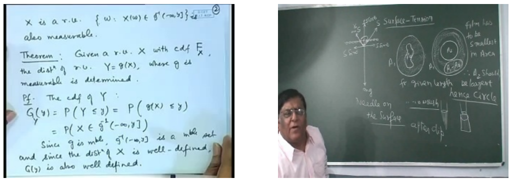
The audio frame selection determines the audio stream and the length of the compressed multimedia signal. Given a fixed video frame rate the timestamp for video frames in the compressed signal is known. For each such timestamp, the closest audio frames in the compressed stream (generally 3 - 4) and their timestamps in the original audio stream are determined. The video frames in the original video within the extent of these audio timestamp are all candidate frames to be considered for positioning at the current timestamp in the compressed signal. The best candidate video frame is chosen to maximize the number of the audio frames it is closest to within a threshold, as well as minimizing the total distance from all the audio frames under consideration. Such an AV sync approach ensures that the lip sync present in the original video is preserved in the compressed version (to within limits of perception).

4.1. Frame Selection to Maximize Visual Information

Key frames within the original video sequence are identified where the instructor has just stopped writing. These key frames are assumed to contain dense text regions. During the writing period the instructor typically faces the board, and lip-sync is a non issue. We exploit this fact by overriding the decisions of the AV sync approach in the 5s period prior to the key frame. In that region, the visual frames dictated by the AV sync approach are replaced by an equal number of consecutive video frames leading up to the key frame. This maximizes the written information in the compressed time signal. The other advantage is that retaining such key frames helps compensate for any audio artifacts that the speed up may produce.

4.2. Detection of Key Video Frames

As mentioned above, the key frames referred above contain dense text regions. In a typical class-room lecture (where an instructor uses a white-board or black-board) we expect the content to monotonically increase across successive frames as the lecture progresses and then drastically drop when the instructor erases the board. Detection of video frames with high text density is accomplished via a four-stage process:
(1) Identification of frames containing written text: A two-class linear SVM classifier which uses the Histogram of Oriented Gradients (HOG) features is trained to distinguish "textual" and "non-textual" frames. Example non-textual frames are where the camera is zoomed in on the lecturer or when the camera pans the classroom.
(2) Identification of the frame region containing text content: Once the textual frames are identified as above the next step is to identify



x is a r.v. $\{\omega : x(\omega) \in g^{-1}(-\infty, y]\}$ is
also measurable.

Theorem: Given a r.v. X with cdf F_X, the dist^n of r.v. $Y = g(X)$, where g is measurable is determined.

Pf. The cdf of Y:
$G_Y(y) = P(Y \le y) = P(g(X) \le y)$
$= P(X \in g^{-1}(-\infty, y])$
Since g is mble, $g^{-1}(-\infty, y]$ is a mble set and since the dist^n of X is well-defined, G(y) is also well defined.

Fig. 2. Detected key frames with maximum text in instructional videos

the region within this frame that contains text (i.e., the boundary of the blackboard or the slide projection). A Mean Shift Segmentation (MSS)-based [18] nonparametric clustering approach is applied on each of these textual frames to segment it into multiple regions. We assume that the maximum segment in the frame is the region containing written text.
(3) Calculation of text density: The chalk (or marker) color is typically in contrast from the background board color. We thus use the Sobel operator [3] to estimate these sharp transitions which act as a proxy for the written content. The proportion of these edge pixels to board pixels is the estimate of the text region density for the frame.
(4) Identification of key frames: The text density scores across the frames can be thought of a time signal. Frames corresponding to the local maxima of this temporal sequence are the likely key frames. We notice that frame occlusion due to the lecturer movement or camera movement leads to a few duplicate key frames. A nearest neighbor based approach is applied on the Scale Invariant Feature Transform (SIFT) representation of these likely key frames to weed out the spurious frames and retain the final set of key frames.

This key frame detection method was trained on 4 videos of 30 minutes each from different lecturers where the full-written key frames were manually labeled. On 6 test videos, [4] the proposed method gave an average precision and recall of 79% and 77.6%, respectively. A detected key frame is marked as correct if it occurs within ±2.5s of a labeled key frame. Example key frames detected by this method for two different videos are shown in Figure 2.

5. RESULTS

We evaluate the utility of the proposed approach against a competitor obtained by time-compression of the audio and video streams via the ffmpeg utility. Informal evaluation indicates that the subjective quality of the resulting compressed multimedia signals are similar to what is observed due to the native compression in YouTube or Windows Media Player. The chosen approach to create the competitor allowed us to obtain the compressed signal as a file, which could then be uploaded to the Dropbox based test portal for access by the test subjects over the internet. Four different 20 minute long instructional videos (from 2 American male tutors, 1 American female, and 1 Indian male) are compressed with the proposed and competing approaches, two of the videos at a target rate of 2x and the other two at 2.5x as indicated in the x-axis labels of Fig. 3. A double blind subjective test required each of the 12 test subjects (graduate students in engineering colleges in India) to pairwise compare the first 5 minutes of compressed signal from the two competing methods. Each test subject performed this comparison for two instructional videos, one compressed at 2x and the other at 2.5x. The ordering of the two videos and that of the two compression methods was alternated across the test subjects to minimize any conditioning bias.

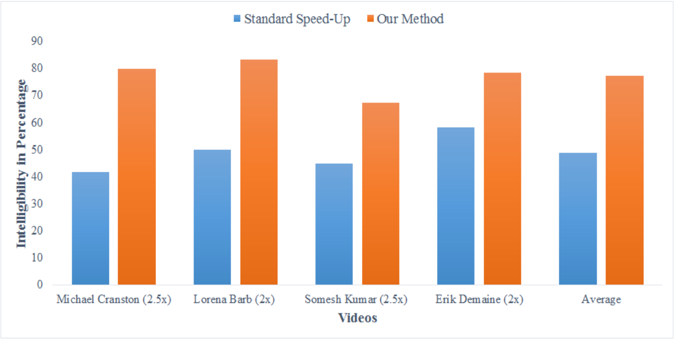


Fig. 3. Comparison of the proposed approach with the generic competitor for four instructional videos (instructor name on x-axis)

The subjects were required to answer three questions after viewing each pair of video sequences:

Q1 How much of the content in video 1 was intelligible? $(0 - 100\%)$

Q2 How much of the content in video 2 was intelligible? $(0 - 100\%)$

Q3 Which video would you prefer to learn from? (1 or 2)

The test subjects overwhelmingly preferred the proposed approach over the competitor (the response to **Q3** was 100% in favour of the proposed method). Fig. 3 provides a comparison of the two compression techniques in terms of their comprehensibility, i.e., the averaged responsens for questions **Q1** and **Q2** for the four instructional videos. The improved intelligibitly due to the proposed method is obvious.

6. CONCLUSION

A novel technique for time-compression of instructional video content that are ad hoc classroom recordings is proposed. Characteristics unique to educational multimedia are exploited in a multimodal analysis dependent paradigm, so that regions of low information are excised or suitably sped up to allow efficient consumption of the content. Given that the analysis can be performed offline, a non-myopic optimization routine picks frames to be retained from the entire signal based on a per-frame saliency calculation so that a target compression rate is achieved. On detection of key video frames, the approach deviates from its tendency to maintain AV sync and instead focuses on preserving maximum visual information. A subjective test demonstrates a significant benefit for students compared to a competitor designed for generic multimedia content.

7. REFERENCES

[1] K. Sreenivasa Rao and Anil Kumar Vuppala, "Non-uniform time scale modification using instants of significant excitation and vowel onset points.," *Speech Communication*, vol. 55, no. 6, pp. 745–756, 2013.

[2] Barry Arons, "Speechskimmer: A system for interactively skimming recorded speech.," *ACM Trans. Comput.-Hum. Interact.*, vol. 4, no. 1, pp. 3–38, 1997.

[3] Liwei He and Anoop Gupta, "Exploring benefits of non-linear time compression," in *Proceedings of the Ninth ACM International Conference on Multimedia*, New York, NY, USA, 2001, MULTIMEDIA '01, pp. 382–391, ACM.

[3] http://en.wikipedia.org/wiki/Sobel_operator

[4] Average length of test videos was 42 minutes

[4] Michele Covell, Margaret Withgott, and Malcolm Slaney, “MACH1: nonuniform time-scale modification of speech,” in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP ’98, Seattle, Washington, USA, May 12-15, 1998*, 1998, pp. 349–352.

[5] Francis C. Li, Anoop Gupta, Elizabeth Sanocki, Li wei He, and Yong Rui, “Browsing digital video.,” in *CHI*, Thea Turner and Gerd Szwillus, Eds. 2000, pp. 169–176, ACM.

[6] Chiori Hori and Sadaoki Furui, “A new approach to automatic speech summarization.,” *IEEE Transactions on Multimedia*, vol. 5, no. 3, pp. 368–378, 2003.

[7] Kathleen McKeown, Julia Hirschberg, Michel Galley, and Sameer Maskey, “From text to speech summarization.,” in *ICASSP (5)*. 2005, pp. 997–1000, IEEE.

[8] Simon Tucker and Steve Whittaker, “Novel techniques for time-compressing speech: An exploratory study.,” in *ICASSP (1)*. 2005, pp. 477–480, IEEE.

[9] Y Li, Shih-Hung Lee, Chiah-Hung Yeh, and C.-C.J Kuo, “Techniques for movie content analysis and skimming: tutotial and overview on video abstraction techniques,” *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 79–89, 2006.

[10] F. Charpentier and M. Stella, “Diphone synthesis using an overlap-add technique for speech waveforms concatenation.,” in *ICASSP (5)*. 2005, pp. 2015–2018, IEEE.

[11] Nivja de Jong and Ton G Wempe, “Praat script to detect syllable nuclei and measure speech rate automatically,” *Behavior Research Methods*, vol. 41, no. 2, pp. 385–390, 2009.

[12] Kåre Sjölander and Jonas Beskow, “Wavesurfer-an open source speech tool.,” in *INTERSPEECH*, 2000, pp. 464–467.

[13] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten, “The weka data mining software: An update,” *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[14] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Müller, and Shrikanth S Narayanan, “The interspeech 2010 paralinguistic challenge.,” 2010.

[15] Simon Zwarts and Mark Johnson, “The impact of language models and loss functions on repair disfluency detection.,” in *ACL*, Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, Eds. 2011, pp. 703–711, The Association for Computer Linguistics.

[16] Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary P. Harper, “Enriching speech recognition with automatic detection of sentence boundaries and disfluencies.,” *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[17] Om Deshmukh, Harish Doddala, Ashish Verma, and Karthik Visweswariah, “Role of language models in spoken fluency evaluation,” in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2010, pp. 2866–2869.

[18] D. Comaniciu and P. Meer, “Mean shift: a robust approach toward feature space analysis,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603 – 619, May 2002.