Acoustic Stress Detection for Improved Navigation of Educational Videos

Sonal Patil, Harish Arsikere, Om Deshmukh

Data Analytics Lab, Xerox Research Center India (XRCI), Bangalore, Karnataka, India

{Sonal.Patil, Harish.Arsikere, Om.Deshmukh}@xerox.com

Abstract

This paper presents a system that uses acoustic stress detection to identify important concepts in educational videos. The proposed system is part of a non-linear navigation system that contains additional features like dynamic word cloud and 2-D timeline. An important feature of the word cloud is that the color used to represent the word depicts its spoken emphasis. This emphasis is estimated by quantifying the acoustic stress of each word. Stressed instances of a given word are also highlighted on the 2-D timeline using different colors. The primary focus of this paper is to detect words spoken with higher acoustic stress and provide an efficient means to navigate to corresponding instances. In the training phase, words are labeled manually as 'stressed' or 'unstressed' by speech experts. An SVM classifier is trained using three types of acoustic features: intensity-based, pitch-based and duration-based. Considering the data imbalance in terms of the ratio of 'stressed' to 'unstressed' words, the performance achieved (70% correct detection at a false-alarm rate of 19%) is satisfactory. The usability studies show that the time taken to detect and navigate to stressed instances of words is significantly less ($p < 0.01$) than that using a youtube-type baseline system.

Index Terms: Video Navigation, Acoustic Stress, Education, Word Cloud

1. Introduction

The growth of online courses and Open Educational Resources (OER) is considered as one of the remarkable achievements in education in recent times. As online educational content, especially video content, is increasing rapidly it is becoming important to develop methods for its efficient consumption. Various studies have shown that learners use non-linear navigation to browse through a lecture before actually watching it [1].

We have proposed a multimodal non-linear video navigation system [2]. There are three major components of the proposed navigation system: (1) identification and placement of important keywords on a word cloud such that the relative temporal order and significance of these keywords is depicted, (2) identification of video frames that correspond to maximum written material (e.g., a full blackboard) to be used for visual summarization, and (3) identification of acoustically stressed words which indicate important acoustic events in the lecture. The word cloud is created using the audio transcription and utilizes four dimensions to visualize the different keywords: The x-coordinate of the word cloud signifies average temporal occurrence of the word in the video. The y-coordinate signifies spread of the word in the lecture. The font size of the word is proportional to its frequency of occurrence in the video. The color of the word signifies the proportion of instances that were acoustically stressed. If a word has been emphasized majority of times in the given time window, it is displayed in dark red color the word-cloud (Red, one of the psychological primary colours, is associated with intensity and excitement[1]). At the other extreme, if a word is never stressed, it is displayed in blue color which is associated with coolness.

Figure 1 shows the snapshot of the system with different navigation components. If the user clicks on a word in word cloud, all occurrences of that word are highlighted on the 2-D timeline. A word occurrence is highlighted in red color on this 2-D timeline if it is stressed otherwise it is shown in white. Thus, the word cloud and the timeline enable user to search for the desired concepts that were emphasized by the speaker. The main focus of this paper is to describe the acoustic analysis of the spoken content to detect words emphasized by the speaker.

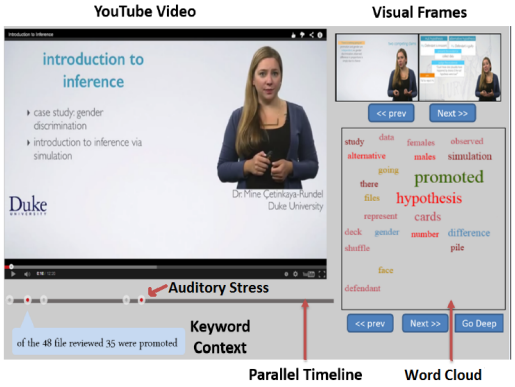

Figure 1: *Snapshot of the system prototype.*

In speech and audio processing, automatic detection of spoken stress is an active area of research [3]. It has potential applications in spoken language learning and evaluation. In an educational setting, given that speakers often emphasize certain words to convey important points [4], stress analysis can also be used to detect the key concepts conveyed in the lecture. But so far, to our knowledge, none of the previous studies has explored that. Thus, our main contribution is to identify acoustic events using acoustic stress detection and to utilize them for efficient browsing of educational videos.

2. System Description

The proposed technique for automatic stress detection is discussed using Figure 2.

2.1. Data Preprocessing

First, all the video lectures selected for the study are processed to extract audio. Using the audio and text transcriptions, forced alignment is performed using the time-stamps provided with

[1] http://www.colour-affects.co.uk/psychological-properties-of-colours

these transcriptions. We are currently developing in-house ASR capabilities to deal with videos without transcriptions. In the current setting, HVite (HTK tool) is used to get phoneme- and word-boundary estimation. The triphone acoustic models, used for forced alignment, were trained on Wall Street Journal data with dictionary of size 15k words. To obtain better time alignments for inter-word silences, each word in the dictionary had two entries: one with and the other without "sil" appended to its phonetic expansion.

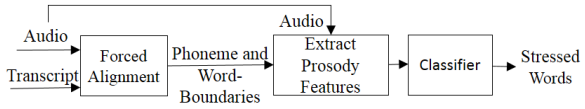

Figure 2: *Block diagram for Acoustic Stress Detection.*

2.2. Prosodic Features Extraction and Classification

Three types of prosodic features were used for stress detection namely, pitch-based, intensity-based and duration-based. Intuitively, stressed words exhibit rise in pitch frequency and modulation as compared to unstressed words. The SNACK sound toolkit [2] was used to get pitch contours. Pitch estimates at 10ms intervals were achieved with the help of pitch contours. Two pitch features: (1) 95th percentile of word's pitch (to capture rise in pitch frequency due to stress), and (2) standard deviation of word's pitch estimates (to capture rise in pitch modulation due to stress) were computed for a given word using the contours. The pitch median of the corresponding speech segment was used for feature normalization.

Intensity is also associated with stressed words. 30ms frames separated by intervals of 10ms were used to obtain intensity contours. A 5-tap FIR filter was used to smooth these contours. The 95th percentile of word's intensity estimate was computed as an intensity feature. The intensity median of the corresponding speech segment was used for normalization of the feature.

Stressed words also demonstrate relation with phoneme elongation. Duration-based feature was captured using duration of the longest phoneme in the word. These durations were calculated using time-stamps obtained through forced alignment.

Further these features were passed to the linear SVM classifier with RBF-kernel with kernel parameter 0.25. The objective of the study is to make a binary decision i.e. decide whether a given word is stressed or not.

Figure 3 compares distributions for stressed versus unstressed instances for the four features discussed above.

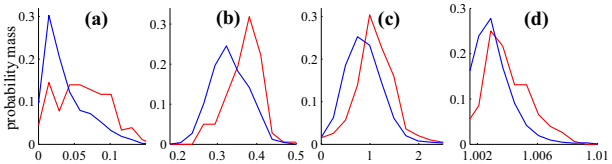

Figure 3: *Distributions of the four prosodic features, (a) Standard deviation of pitch (b) 95th percentile of pitch (c) 95th percentile of intensity (d) duration of longest phoneme, for stressed (red curve) and unstressed (blue curve) instances.*

3. Experimental Evaluation

Five video lectures were selected for this study with one male and two female speakers who had different intonation patterns. Classifier was trained using four videos and evaluated on fifth one. For training, each word was manually labeled as "stressed" or "unstressed" purely by listening to audio files. The labeling was done by the second author. Overall 345 words were labeled "stressed" and 5441 words were labeled "unstressed". Since the stress labeling is found to be subjective [3], we had the third author also to label one of the video. For that video, the inter human agreement for stressed labels of 53.3% was achieved, confirming stress labeling is indeed subjective.

Since the number of unstressed words were around 20 times more than stressed words, we penalized positive class more while training. The test audio file had 50 stressed and 2822 unstressed words. Figure 4 shows ROC curve for test audio along with the highlighted operating point which corresponds to 70% correct-detection rate at false-alarm rate of 19%.

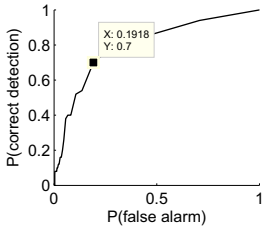

Figure 4: *ROC curve for the proposed acoustic stress detector.*

We performed a few sessions of usability study for proposed system versus baseline system. The baseline system consists of transcripts displayed alongside the video and each sentence in the transcript is hyperlinked to its corresponding time in the video. The participants were asked to navigate to the time at which lecturer emphasizes certain words. We observed significant ($p < 0.001$) reduction in time while navigating to instances of spoken terms which were acoustically stressed. The color code of a word in the word cloud and the word's instances on the timeline for stressed words helped user to visually perceive the audio cue.

4. Demo Plan

In this paper, we present a system that identifies important concepts in educational video using acoustic stress detection and provides a mechanism to easily navigate to these acoustic events in the video. Our demo will consists of a walk-through of all features of the proposed system followed by a hands-on demonstration of how easy it is to browse a lecture using the proposed non-linear navigation. We will also show how easy it is to perceive variations in acoustic stress when the system is accompanied with visual cues.

5. References

[1] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, and R. C. Miller, "Understanding in-video dropouts and interaction peaks in online lecture videos," in *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 2014, pp. 31–40.

[2] K. Yadav, K. Shrivastava, S. Mohana Prasad, H. Arsikere, S. Patil, R. Kumar, and O. Deshmukh, "Content-driven multi-modal techniques for non-linear video navigation," in *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 2015, pp. 333–344.

[3] O. D. Deshmukh and A. Verma, "Nucleus-level clustering for word-independent syllable stress classification," *Speech Communication*, vol. 51, no. 12, pp. 1224–1233, 2009.

[4] G. Brown and G. Yule, *Teaching the spoken language*. Cambridge University Press, 1983.

[2] http://www.speech.kth.se/snack/