Eastern Illinois University
The Keep

Faculty Research & Creative Activity | Technology, School of

January 2009

Algorithms to Automate Estimation of Time Codes for Captioning Digital Media

Daniel P. Harvey II
Eastern Illinois University

Peter Ping Liu
Eastern Illinois University, pliu@eiu.edu

Follow this and additional works at: https://thekeep.eiu.edu/tech_fac

 Part of the Technology and Innovation Commons

Recommended Citation

Harvey, Daniel P. II and Liu, Peter Ping, "Algorithms to Automate Estimation of Time Codes for Captioning Digital Media" (2009). *Faculty Research & Creative Activity*. 12.
https://thekeep.eiu.edu/tech_fac/12

This Article is brought to you for free and open access by the Technology, School of at The Keep. It has been accepted for inclusion in Faculty Research & Creative Activity by an authorized administrator of The Keep. For more information, please contact tabruns@eiu.edu.

Algorithms to Automate Estimation of Time Codes for Captioning Digital Media

Daniel P. Harvey II and Peter Ping Liu

Eastern Illinois University, Charleston, IL 61920, USA

Abstract. Procedures were developed to partially automate the captioning process by estimating caption time codes using plain-text transcripts and audio recordings. Signal analysis is performed on the audio to measure pause location and duration, zero crossing rate (ZCR), and obtain frequency domain data. Algorithms were developed to match pauses in the audio to the ends of sentences in the transcript based on the observation that pause durations are greater at ends of sentences than within sentences. We have observed that ZCR peaks correspond to consonants in speech and that continuous wavelet transforms (CWT) work well for distinguishing between groups of consonants. These measurements will be used to develop algorithms to match selected phonemes in the audio to text in the transcript to supplement the pause matching results.

1 Introduction

The presentation of information on the World Wide Web relies increasingly on multimedia technologies. The audio component of these media presentations on the web remains largely inaccessible to persons who are Deaf or Hard of Hearing [1]. Captioning provides accessibility to media resources for deaf and hearing-impaired persons. However, captioning with currently available captioning software is time and labor intensive because it requires manually generating a text transcript and manually determining time codes. The goal of this study is to develop ways to provide synchronized captions more effectively and efficiently by extracting and analyzing signal data from the audio and using this information to automatically align text captions with audio recordings. In order to simplify implementation as much as possible, we are attempting to accomplish alignment without speech recognition in order to circumvent the need to compile a vocabulary for the system and so that users will not need to train the system for individual speakers.

2 Methods

2.1 Recordings and Transcripts

The recordings used in this study were of professional speakers/readers from American English radio and television broadcasts and of non-professional speakers reading text from a novel. The data set used in this study consisted of seven

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 585–590, 2009.
© Springer-Verlag Berlin Heidelberg 2009

audio files that represented actual cases of media requiring captioning. The text consisted of 7140 words in 341 sentences. The transcripts were manually generated because commercially available speech recognition is currently not accurate enough to automatically generate the transcripts from recorded speech.

2.2 Signal Analysis

WAV audio files were read as binary data using a program written in C. Amplitude was measured. It was found that the RMS amplitude of each audio file performs adequately as a threshold between silence and speech for captioning. The locations and durations of pauses in the audio recordings were compiled by scanning for segments of the recording below the amplitude threshold. Statistically, pause durations at the ends of sentences are significantly greater than those within sentences. This observation was used to match the ends of sentences in the text to pauses in the audio track using the algorithms developed in this study.

Zero Crossing Rate (ZCR) was measured using 20 ms windows. The ZCR was plotted as a function of time. The ZCR peaks were manually matched to what phoneme was occurring at that time.

The times of the ZCR peaks were tabulated. In windows of 20 ms centered on the ZCR peaks, wavelet analysis was performed using the FAWAVE software package. Scalograms were computed using a complex Gabor wavelet [2] with width 0.125 and frequency 16. The magnitudes were graphed using 6 octaves and 16 voices. For comparison purposes, Fourier spectrograms were computed for the 20 ms samples.

2.3 Global Averaging Method

In the Global Averaging Method, the number of characters in each sentence and in the entire transcript are counted. The algorithm differentiates between punctuation points at the ends of sentences (periods, question marks, and exclamation points) and punctuation points within sentences (commas, semi-colons, colons, and dashes). All characters, except for punctuation marks, are assumed to represent equal amounts of time. This algorithm has an additional assumption that the pauses with the longest durations correspond to punctuation marks in the transcript. The total duration of each caption ($t_{caption}$) is the sum of the duration of the pauses (t_{pauses}) and the duration of the speech segments, or articulation time, (t_{speech}) in that caption as expressed in Equation 1.

$$t_{caption} = t_{pauses} + t_{speech} \tag{1}$$

The duration of the articulation time for each sentence was estimated from the number of characters in each sentence as a proportion of the total recording time. Character Count Weighting (CCW) assumes that each non-punctuation character, including blank spaces, in the text transcript corresponds to an equal amount of time. This assumption is based on the observation that one of the

factors in syllable timing is the number of phonemes it includes [3]. For every caption, the number of spaces, the number of punctuation points within sentences, the number of punctuation points at the ends of sentences, and the number of alphanumeric characters are tabulated. The articulation time of each caption (t_{speech}) is estimated by calculating the ratio of the number of non-punctuation characters in each caption ($cc_{caption}$) to the number of non-punctuation characters in the entire transcript (cc_{total}) and multiplying that ratio by the clip time (t_{total}) as shown in Equation 2.

$$t_{speech,caption} = t_{speech,total}(cc_{caption}/cc_{total}) \tag{2}$$

To improve upon the accuracy of the algorithm, factors are added to account for pauses corresponding to punctuation. This is done by measuring the longest pauses detected in the audio track, calculating an average pause duration associated with punctuation marks within sentences and at the ends of sentences, and adding a pause time based on the number of punctuation marks in each caption.

The estimated timeline of the captions is constructed by:

1) estimate articulation time for a caption using CCW
2) add averaged pause time based on number of punctuation marks
3) repeat process to the end of the transcript.

2.4 Local Maxima Method

In the Local Maxima Method, the text is parsed into captions using end-of-sentence punctuation as break points. The number of end-of-sentence punctuation points (n) equals the number of captions. A total articulation time is estimated by subtracting the total duration of the n longest pauses from the total cliptime. The duration of each caption articulation time is estimated with Character Count Weighting (CCW). The estimated timeline of the captions is then constructed by:

1) moving out the estimated caption duration for sentence 1($t_{speech,caption1}$)
2) querying pauses in an area centered on $t_{speech,caption1}$
3) using a range that is defined as a percentage of the estimated duration of the preceding caption with a set minimum range
4) finding the longest pause starting within that range
5) if there are multiple candidate pauses in the search range, obtaining manual feedback from user to locate end of sentence (optional)
6) adding the chosen pause duration to the estimated caption duration
7) moving out the estimated caption duration for sentence 2 ($t_{speech,caption2}$)
8) and then repeating the process to the end of the transcript

The Local Maxima Method constructs a timeline without using average values for pauses at ends of sentences, but rather uses actual pause duration values.

When matching estimated times from text analysis to the location of pauses measured from the audio file, the local maxima method limits the pause pool by limiting the time range searched; whereas the global threshold method limits the pause pool by a pause duration criterion.

3 Results

3.1 Pause Matching

Table 1 shows the maximum and the average errors for the three types of speech tested in this study. The maximum error and the average error are much less for the Local Maxima Method than for the Global Averaging Method. In order to successfully distinguish between within-sentence pauses and end-of-sentence pauses on the basis of duration, it is advantageous to utilize data over localized portions of a file rather than over the entire audio file. The incorporation of manual feedback with the alignment algorithm resulted in a greater reduction in error. The overall rate of manual feedback requests for the files tested was approximately one request for every ten captions. For the files tested in this study, the Local Maxima Method with Manual Feedback accurately estimated the timing of 96% of the captions within 0.5 seconds. Comparable accuracy has been achieved in ongoing testing with actual media captioning projects using a web application that incorporates the Local Maxima Method with Manual Feedback.

3.2 Zero Crossing Rate and Frequency Analysis

ZCR peaks were distinctive and easily detected. We found a total of 14 consonants that corresponded to ZCR peaks. These consonants are listed in Table 2. The number of peaks within a given amount of time (90/min) is too large to be helpful in matching them to the transcript. We then analyzed the audio data near the ZCR peaks using spectrograms from Fourier analysis and continuous wavelet transforms (CWT).

The magnitude peaks of sections of CWT scalograms taken at ZCR peaks are distinctive and occur at different octaves for different phonemes. The results of these analyses are summarized in Table 2. Based on these preliminary observations, the octave at which the maximum peak in the scalogram slice can be used to distinguish four phonemes from the set of 14 consonants that correspond to ZCR peaks. The rate of occurrence for those four phonemes in the recordings tested was seven per minute. This should facilitate the process of matching these phonemes to their occurrence in the transcript. The approach of combining ZCR and wavelet analysis has been used to classify segments of speech signals into broad phonetic categories including silence, voiced, unvoiced, and plosive release [4]. The results of the CWT were more straightforward and easier to interpret than the spectrogram results. This is consistent with a previous study using wavelet analysis to distinguish between classes of phonemes [5].

Table 1. Error in Pause Matching Algorithms

Audio Type	Cases	Algorithm	Max. Error (s)	Average Error (s)
novel reading	2	GAM	24.83	3.66
		LMM	15.14	3.26
		LMM with FB	5.58	0.10
radio broadcasts	2	GAM	10.56	3.24
		LMM	10.11	1.36
		LMM with FB	5.06	0.42
scripted narrations	3	GAM	6.84	0.53
		LMM	0.00	0.00
		LMM with FB	0.00	0.00

Table 2. Octave and Magnitude of maximum peak in CWT scalogram of phonemes corresponding to ZCR peaks

Phoneme	Octave	Magnitude
ch	4.83	2.69E+10
j	4.94	1.17E+10
sh	4.84	2.04E+10
zh	4.63	4.15E+09
d	0.69	4.58E+10
f	3.19	3.64E+08
g	0.63	2.18E+11
h	3.81	6.78E+08
k	3.00	1.21E+10
p	2.81	2.70E+08
s	2.41	1.07E+10
t	2.69	6.89E+09
th	2.00	3.08E+08
z	1.88	1.74E+09

4 Conclusions and Future Work

The level of accuracy achieved in this study indicates that the method proposed for aligning text to audio using pauses is sufficient for the task of estimating the timing of captions for media. However, future work will include testing the algorithms on a larger data set taken from a standard corpus such as the TIMIT speech database [4, 5] in order to verify these results and to facilitate comparison to standard methods of text alignment. It is also planned to test the pause matching algorithms with other languages such as Hindi and Spanish.

The use of ZCR and wavelet analysis appears to be a promising approach for distinguishing a relatively small subset of phonemes. Because the feature matching process relies primarily on time domain parameters and because the wavelet analysis is distinguishing consonants that have a component of turbulent airflow, we are avoiding the use of signal features that are strongly influenced by the size and shape of the vocal tract and thereby again avoiding the need for training the system to individual speakers. However, testing on a larger data set will be necessary to verify these preliminary results. Based on the results of the pause matching algorithms, the approach that will be taken to match these four phonemes is to use character count weighting to approximate the temporal location of the candidate phonemes found in the transcript and then match them to the appropriate audio events based on the location of the ZCR peaks. These results will be used to supplement the pause matching results in order to improve the accuracy of the alignment process for a wider variety of speech recordings and to reduce the need for manual feedback in the alignment process.

References

1. Canadian Network for Inclusive Cultural Exchange: Online enhanced captioning guidelines (2004), `http://cnice.utoronto.ca/guidelines/caption.pdf` (retrieved September 22, 2004)
2. Walker, J.S.: Primer on wavelets and their scientific applications. CRC Publishing, Boca Raton (1999)
3. Campbell, W.N.: Syllable-based segmental duration. In: Bailly, G., Benoit, C., Sawallis, T.R. (eds.) Talking machines: Theories, models, and designs, pp. 211–224. Elsevier Science, Amsterdam (1992)
4. Pernkopf, F., Van Pham, T., Bilmes, J.A.: Broad phonetic classification using discriminative Bayesian networks. Speech Communication 51, 151–166 (2009)
5. Tan, B.T., Fu, M., Spray, A., Dermody, P.: The Use of Wavelet Transforms in Phoneme Recognition. In: Proceedings of ICSLP 1996 (1996)