Automatic Word Boundary Detection Using Adaptive Scaling

Ramanathan Subramanian[#1]

Electrical and Computer Engineering Department, Northeastern University, Boston, MA, USA

[#1] subramanian.r@husky.neu.edu

***Abstract*—This project reports the attempt to detect word boundaries from mere statistical descriptions. It is shown that using the moving window based short-time energy feature in the signal together with priors can be used to detect word boundaries satisfactorily.**

I. Introduction

An algorithm for word boundary detection in recorded sentences of English speech is proposed. Prior approaches to this problem include work in speech recognition and caption alignment; hidden Markov models with acoustic features such as mel-scale frequency cepstrum coefficients (MFCCs). This project is to research approaches used in word-boundary detection in other problems and then apply them to our particular problem with all of the relevant priors available to us.

A. Background

In the speech acoustics laboratory called the CADLab (Communication Analysis and Design Laboratory), under the direction of Dr. Rupal Patel of the Speech and Language Pathology Department and the Computer Science Department, NEU, performs novel research on speech acoustics and computer-human interfaces with a focus on prosody in children and disordered speakers.

A common feature of studies done in the CADLab is the collection of large sets of similar sound files. These sound files are typically a single sentence recorded in a low-noise environment such as a sound-proof booth. The speakers are typically reading from a script in many cases are instructed to make specific modifications to the prosody (pitch, loudness, and rate) of their speech.

These sound files are then analyzed for various acoustic characteristics. For any study that's interested in word-level acoustic measures, which is most of them, a necessary preprocessing step for the acoustic analysis is to mark the word boundaries. Currently this is done manually by a speech researcher who uses the waveform, spectrogram, and listening to the sound file to determine the boundaries. Another researcher will independently label a subset of the recordings (typically ~5%) and the boundaries will be compared to establish a measure of inter-labeler reliability. This process can take many hours and even with inter-labeler reliability is error-prone.

Machine learning techniques can be applied to word boundary detection, building on prior approaches used in speech recognition and caption alignment, optimized for the particular set of constraints in CADLab study recordings. The goal is to reduce study analysis time, improve labeling accuracy and consistency, and enable novel computer-human interface applications.

B. Problem Description

Word boundary detection is a complex problem. Silence detection is insufficient as not all words have clearly defined pauses between them (sometimes referred to as coarticulation). In addition, some phonemes (the individual units of speech which are combined to sound words) have characteristics similar to noise, e.g. - fricatives such as 'ff' and sibilants such as 'ss'. Word durations are partly determined by the number and length of the individual phonemes that form the word. The distribution of word durations in a sentence may also be affected by the semantic intent of the speaker.

There is a degree of intra- and inter-speaker variability. Intra-speaker variability is partly due to semantic intent conveyed through prosody (as previously mentioned) but also due to human error. Inter-speaker variability is in part due to varying vocal characteristics (indicated but not fully determined by age, gender, dialect, etc.) and familiarity with the words being spoken.

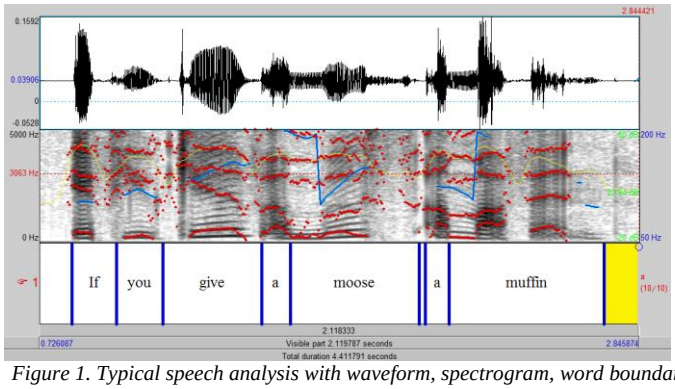

Figure 1. Typical speech analysis with waveform, spectrogram, word boundaries, and phoneme boundaries

This combination of complex, inter-dependent, and not fully specified interactions that make up word boundaries makes this problem suitable for machine learning techniques. In particular, pattern recognition can be used to recognize the duration of an individual word and hidden Markov models can encode the dependencies between words in a sentence.

II. Data Collection and Analysis

The data was primarily collected from two prior studies done in the CADLab. The Read 'N Karaoke study data consists of 14 children, ages 7-9, reading two 50-sentence stories sentence-by-sentence both with regular text and with the presence of visual cues intended to convey prosody. The Contrastive Stress study data consists of 4 norm speakers and 4 disordered speakers reading a set of short sentences multiple times, with each recording placing stress on a different word of the sentence. Only the norm speakers were used. Word boundary detection on disordered speech is out of the scope of this project.

Figure 2. A typical snapshot of the TextGrid corresponding to an utterance

There are 3,266 single-sentence recordings from these two studies. There are 18 different speakers, 415 unique words, and 30,926 word/duration pairs. There are an average of 9.5 words per sentence with an average sentence duration of .3.78 seconds. The average word duration is 0.384 seconds. There are an average of 74 instances of each individual word, but the distribution is exponential. Only 76 words have more than 50 instances. The most common word is 'the' with 857 instances.

Many of the recordings were labeled with numbers (e.g. - s1, s2, etc.) instead of the actual words spoken. The transcripts had to be identified and matched up with the marked word boundaries. Several recordings contained labeled errors such as microphone pops and speaker disfluencies (mispronounced words).

III. Algorithm

For an initial baseline, the current functionality of the analysis suite used in the CADLab was duplicated and the error was computed. The suite takes as an input the number of words in each recording and then places that number of word boundaries distributed uniformly in the recording. A speech analyst then reviews each recording and adjusts the boundaries and adds others for additional pauses between words.

The approach is to use our prior knowledge of word durations to build a statistical estimator. Please find in the folder the Excel sheet carrying the word statistics. The corpus was analyzed and a table was built with the mean and variance of each word in the training set and the mean onset time (delay from start of recording to start of first word) and ending time (delay from end of last word to the end of the recording) as well as the average sentence length.

Using that information to predict a new sentence, the onset is first estimated and ending delays by scaling the mean onset and ending delay by the ratio of the new sentence to the average sentence. This gives an estimate for how much of the recording is silence and how much is sounded. Then for the sounded portion of the sentence, the mean duration of each word scaled by the ratio of the length of the sounded portion of the new sentence to the sum of the mean durations of each word is used, distributed proportionately to the variance of each word.

For sentence *i*,
$Onset_i = Onset_\mu * Sentence\ Duration_i / Sentence\ Duration_\mu$
$Endingi = Ending_\mu * Sentence\ Duration_i / Sentence\ Duration_\mu$
$C = (Sentence\ Duration_i - Onset_i - Ending_i - \Sigma\ Word_\mu) / \Sigma\ Word_\sigma$
$Word_i = Word_\mu + C * Word_\sigma$

For sentences with new words for which the mean or variance is unknown, instead of distributing the extra time to each word according to its variance, the extra time is uniformly distributed to the unknown words. If the sentence is shorter than the sum of mean durations of the known words, we can scale them smaller by double the difference and uniformly split that time between the unknown words.

IV. Results

Different values of the window length and window shift was tried and the ones having a nice temporal resolution versus smoothness trade-off was retained. The short-time energy waveform was smoothed to manage its lingering around the energy threshold and was further thresholded based on peak value in the signal to give the word boundaries. The performance is general satisfactory for initial marking of the boundaries. The algorithm successfully handles less complex sentences.

This implementation might be suitable for reducing the total distance word boundaries have to be moved, but it is unsuitable for doing word-level acoustic analysis. The sensitivity of results to word distribution is a clear indication that our handling of sentences with unknown words is poor.

However, the results are nowhere close to human labeller accuracy.

A. *Relevant Plots*

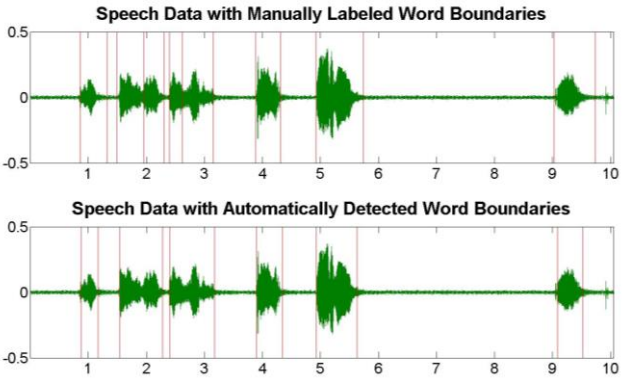

Figure 3. Manually/Automatically detected word boundaries

The top figure shows a waveform corresponding to an utterance with manually labelled word boundaries. The bottom shows the automated detection of the word boundaries for the same utterance.

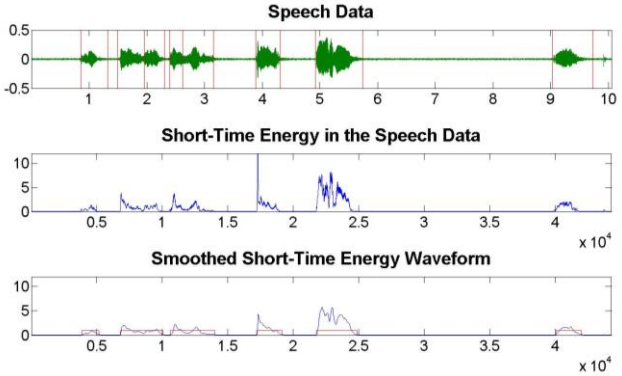

Figure 4. Layered extraction of the word boundaries

V. Discussion

Even though statistical estimating is an improvement, it's nowhere close to human labeler accuracy. This implementation might be suitable for reducing the total distance word boundaries have to be moved, but it is unsuitable for doing word-level acoustic analysis. The sensitivity of results to word distribution is a clear indication that our handling of sentences with unknown words is poor.

VI. Future Directions

We could improve the statistical estimator in a number of ways. We could improve the handling of unknown words by using a similarity measure to estimate the mean of an unknown word from the words in our table. The best choice for this would be a phonetic similarity measure, as word duration is more determined by number and type of phoneme than letters. We could build on the phonemic approach by building a table of mean durations and variance by phoneme instead of by word. There are anywhere from 200,000 to over a million English words depending on how one counts, but only approximately 40 phonemes. This would have difficulty with homographs, which are words with the same spelling but different pronunciation (e.g. desert as in to leave and desert as in a place).

The big improvement made to the statistical estimator is the silence detection step instead of doing a statistical estimation of sounding to silence ratio. A modification of the simple silence detector might use a threshold value and minimum durations to distinguish silence from sounding (e.g. 25 dB difference and 0.1s minimum duration). Further we can use statistical methods to find the best fit of the words to the sounding intervals using their means and variances.

One difficulty with silence detection is that one set of parameters will not accurately detect silences in every recording. One method to mitigate this is to use the ratio of the mean word durations to the current sentence duration to find a reasonable estimate for the sounding to silence ratio of the sentence and tune the silence detection parameters until that ratio is reached. Then then those silence detection parameters can be used to in turn fit the words to the sounding intervals of the sentence.

Other approaches to the problem would be to set a new baseline by taking apart an open source speech recognition library (e.g. CMUSphinx) to find what sort of estimate it makes for word boundary information and compare its error to the uniform and statistical models. An ASR might also serve as a reasonable baseline for measuring performance improvements made by incorporating our prior information into a similar algorithm.

Including acoustic features such as Fourier transform coefficients or mel-scale frequency cepstrum coefficients (MFCCs) would allow use to use pattern recognition to detect when a word deviates from its mean duration. Mel-scale refers to a logarithmic perceptual frequency scale. Cepstrum is a signals engineering term for the Fourier transform of a Fourier transform. This representation has some theoretically important benefits for machine learning methods on acoustic data, such as linearity and correspondence to human perception.

A hidden Markov model can be trained to capture the subtle dependencies between words in a sentence. The work in this area would be heavily influenced by what is found in the ASR systems' operation and theory.

VII. Acknowledgment

Sincere thanks to Prof. Deniz Erdogmus for approving and getting me started with this topic. I would also like to thank my friend William Furr for introducing me to this problem.

VIII. REFERENCES

[1] M. S. Lewicki, "*Information theory: A signal take on speech*," Nature, vol. 466, no. 7308, pp. 821–822, August 2010.

[2] Sahar E. Bou-Ghazale and Khaled Assaleh, "*A Robust Endpoint Detection of Speech for Noisy Environments with Application to Automatic Speech Recognition*", IEEE 2002.

[3] Jean-Claude Junqua, Brian Mak and Ben Reaves, "*A Robust Algorithm for Word Boundary Detection in the Presence of Noise*," IEEE Transactions of Speech and Audio Processing, Vol 2, NO.3, July 1994.

[4] Chin-Teng Lin, Jiann-Yow Lin, and Gin-Der Wu, "*A Robust Word Boundary Detection Algorithm for Variable Noise-Level Environment in Car*," IEEE Transactions on Intelligent Transportation Systems, Vol. 3, No. 1, March 2002.

[5] H. Agaiby, T. J. Moir, "*A Robust Word Boundary Detection Algorithm with Application to Speech Recognition*,"

[6] Lori F. Lamel, Lawrence R. Rabiner, "*An Improved Endpoint Detector for Isolated Word Recognition*," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-29, No. 4, August 1981.

[7] Colin W. Wightman and Mari Ostendorf, "*Automatic Labeling of Prosodic Patterns*", IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 4. October 1994.