Syllable Segmentation of Continuous Speech Using Auditory Attention Cues

Ozlem Kalinli

US R&D, Sony Computer Entertainment, Foster City, California, USA.
e-mail: ozlem_kalinli@playstation.sony.com

Abstract

Segmentation of speech into syllables is beneficial for many spoken language processing applications since it provides information about phonological and rhythmic aspects of speech. Traditional methods usually detect syllable nuclei using features such as energies in critical bands, linear predictive coding spectra, pitch, voicing, etc. Here, a novel system that uses auditory attention cues is proposed for predicting syllable boundaries. The auditory attention cues are biologically inspired and capture changes in sound characteristic by using 2D spectro-temporal receptive filters. When tested on TIMIT, it is shown that the proposed method successfully predicts syllable boundaries and performs as good as or better than the state-of-the art syllable nucleus detection methods.

Index Terms: syllabification, syllable boundary prediction, syllable nuclei detection, auditory attention, auditory gist.

1. Introduction

It has been argued that syllables play a crucial role in human speech perception. There is evidence in psychoacoustic and psycholinguistic studies supporting the hypothesis that the syllable may constitute a natural unit for segmentation and recognition of speech [1]. Hence, segmentation of continuous speech into syllables is beneficial for many applications including speech recognition and speech synthesis. For example, by counting the number of syllables in an utterance, one can estimate a speaking rate, which can be useful for spoken language understanding, selecting appropriate acoustic models for speech recognition, or model adaptation, etc. [2, 3, 4]. Also, syllable nuclei, which can be more reliable source of information in noisy conditions, can be used as anchor points for robust speech processing [1, 5].

Even though, there is no clear definition of a syllable, it's mostly accepted that a syllable contains a central peak of sonority, called syllable nucleus, and consonants surrounding them. Usually the work for syllable segmentation of speech is simplified to the detection of syllable nuclei since they can be located easier and more reliably compared to syllable boundaries [6, 3, 7]. For the detection of syllable nuclei, most of the existing methods rely on estimating a one-dimensional continuous curve from extracted short-time acoustic features and performing a peak search on the curve to locate syllable nuclei [8]. Some of the used features are energy in selected critical bands, linear predictive coding spectra, subband-based correlation, pitch, voicing, etc [9, 2, 6, 3, 7]. The traditional algorithms usually require parameter tuning or they are rule-based. Hence, they are usually less robust on new data or new conditions such as speaking style, noise conditions, etc.

In a speech spectrum, one can usually see edges and local discontinues around syllable boundaries; especially around syllable nucleus boundaries since the syllable nucleus usually exhibit high energy and clear formant structure. For example, in Fig 1, the spectrum of a speech segment that contains three syllables (transcription: "even then", with three syllables: iy, v_ih_n, eh_n) is shown. In the spectrum, one can visually observe three regions/boundaries that correspond to syllable boundaries in the speech segment. Hence, we believe that by detecting the relevant oriented edges and discontinues in the auditory spectrum; i.e. as done visually, syllable segments and/or boundaries in speech can be located.

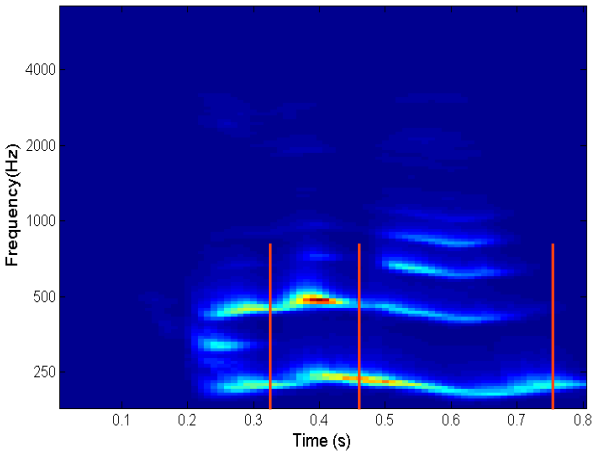

Figure 1: A spectrum of a sample speech segment with transcription "even then" containing three syllables: iy, v_ih_n, eh_n. Vertical orange bars located at 0.31 s, 0.46 s and 0.75 s in the image indicate approximate syllable boundaries.

Here, a novel method that uses auditory attention cues is proposed for syllabification of continuous speech. The model shares the same front-end processing with the bottom-up (BU) saliency-driven auditory attention model in [10], which emulates saliency driven human auditory attention. For example, it was shown in [10, 11] that the BU attention model could successfully detect salient audio events/sounds in an acoustic scene by capturing changes that make such salient events perceptually different than their neighbours. Hence, the model can be used for change point detection such as the boundary between a syllable nucleus and consonants surrounding it.

The auditory attention model is biologically inspired and mimics the processing stages in the human auditory system. The 2D auditory spectrum of input sound is analyzed by extracting multi-scale features using 2D spectro-temporal receptive filters which are based on the processing stages in the central auditory system. The features consist of *intensity, frequency contrast, temporal contrast* and *orientation*. Here, the auditory spectrum is analogous to an image of a scene in vision and some

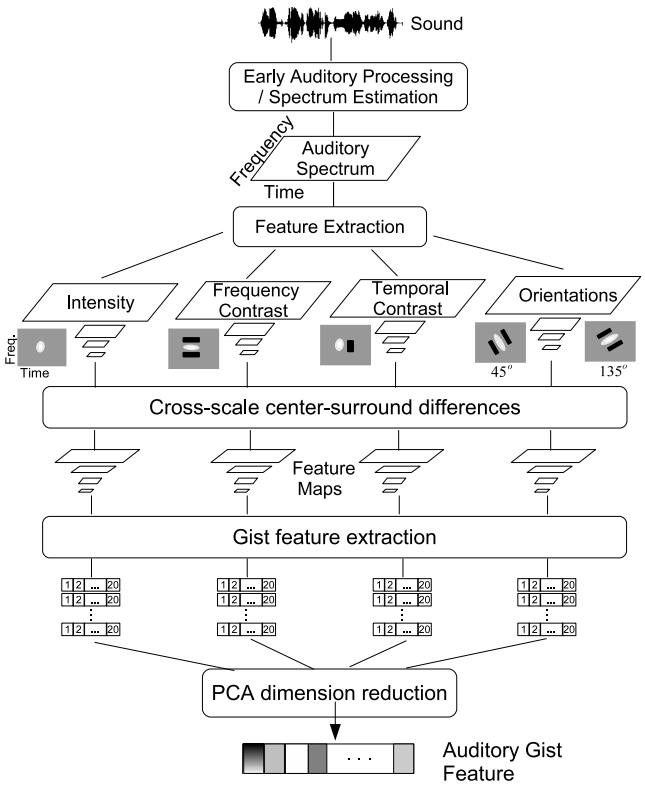

Figure 2: Auditory Attention Model and Gist Extraction

of these features are tuned to different local oriented edges; i.e. frequency contrast features are tuned to local horizontally oriented edges, which are good for detecting and capturing formants and their changes as discussed later. Next, the multi-scale feature maps are converted to low level auditory gist features, which capture and summarize the overall statistics and contextual information of the acoustic scene. Finally, a neural network is used to discover the relevant oriented edges and to learn the mapping between the gist features and syllable boundaries.

The rest of the paper is organized as follows. The auditory attention model together with gist extraction is explained in Section 2, which is followed by experimental results in Section 3. The concluding remarks are presented in Section 4.

2. Auditory Attention Model

The block diagram of the auditory attention model is shown in Fig 2. As stated earlier, the model is biologically inspired and hence mimics the processing stages in the human auditory system. First, the auditory spectrum of the input sound is computed based on early stages of the human auditory system. The early auditory system model used here consists of cochlear filtering, inner hair cell, and lateral inhibitory stages mimicking the process from basilar membrane to the cochlear nucleus in the auditory system [10]. The cochlear filtering is implemented using a bank of 128 overlapping constant-Q asymmetric band-pass filters with center frequencies that are uniformly distributed along a logarithmic frequency axis. For analysis, audio frames of 20 milliseconds (ms) with 10 ms shift are used, i.e. each 10 ms audio frame is represented by a 128 dimensional vector.

The two-dimensional auditory spectrum with time and frequency axes is analogous to an image of a scene in vision. In the next stage, multi-scale features, which consist of *intensity* (I), *frequency contrast* (F), *temporal contrast* (T), and *orientation* (O_θ) with $\theta = \{45^o, 135^o\}$, are extracted from the auditory spectrum based on the processing stages in the central auditory system [10, 12].

These features are extracted using 2D spectro-temporal receptive filters mimicking the analysis stages in the primary auditory cortex. Each of the receptive filters (RF) simulated for feature extraction is illustrated with gray scaled images in Fig 2 next to its corresponding feature. The excitation phase and inhibition phase are shown with white and black color, respectively. For example, the frequency contrast filter corresponds to receptive fields in the primary auditory cortex with an excitatory phase and simultaneous symmetric inhibitory side bands. Each of these filters is capable of detecting and capturing certain changes in signal characteristics. For example, the frequency contrast features are capable of detecting and capturing changes along the spectral axis, whereas the orientation features are capable of capturing and detecting moving ripples (i.e. raising and falling curves). One important point is that in the attention model feature contrast is computed rather than the absolute feature strength, which is also crucial for change point detection and segmentation.

The RF for intensity feature has only an excitation phase and is implemented using a 2D Gaussian kernel. The RF for generating frequency contrast, temporal contrast and orientation features are implemented using 2D Gabor filters with angles 0^o, 90^o, $\{45^o, 135^o\}$, respectively. The multi-scale features are obtained using a dyadic pyramid: the input spectrum is filtered and decimated by a factor of two, and this is repeated. Finally, eight scales are created (if the scene duration is larger than 1.28 s; otherwise there are fewer scales), yielding size reduction factors ranging from 1:1 (scale 1) to 1:128 (scale 8). For details of the feature extraction and filters, one may refer to [10, 12].

After multi-scale features are obtained, the model computes "center-surround" differences by comparing "center" fine scales with "surround" coarser scales yielding feature maps. The center-surround operation mimics the properties of local cortical inhibition and detects local temporal and spatial discontinuities. It is simulated by across scale subtraction ($\ominus$) between a center scale c and a surround scale s yielding a feature map $\mathcal{M}(c, s)$:

$$\mathcal{M}(c, s) = |\mathcal{M}(c) \ominus \mathcal{M}(s)|, \quad \mathcal{M}\epsilon\{I, F, T, O_\theta\} \qquad (1)$$

The across scale subtraction between two scales is computed by interpolation to the finer scale and point-wise subtraction. Here, $c = \{2, 3, 4\}$, $s = c + \delta$ with $\delta\epsilon\{3, 4\}$ are used, which results in 30 feature maps when there are eight scales.

Next, an "auditory gist" vector is extracted from the feature maps of I, F, T,O_θ such that it covers the whole scene at low resolution. To do that, each feature map is divided into m-by-n grid of sub-regions and mean of each sub-region is computed to capture the overall properties of the map. For a feature map $\mathcal{M}_i$ with height h and width w, the computation of feature can be written as:

$$G_i^{k,l} = \frac{mn}{wh} \sum_{u=\frac{kw}{n}}^{\frac{(k+1)w}{n}-1} \sum_{v=\frac{lh}{m}}^{\frac{(l+1)h}{m}-1} \mathcal{M}_i(u, v), \qquad (2)$$

where $k = \{0, \cdots, n-1\}, l = \{0, \cdots, m-1\}$, and feature map index $i = \{1, \cdots, 30\}$. An example of gist feature extraction with $m = 4, n = 5$ is shown in Fig 2, where a $4 \times 5 = 20$ dimensional vector is shown to represent a feature map. After extracting a gist vector from each feature map, we obtain

Table 1: Syllable Boundary Detection Results at Frame-Level for Varying Window Duration

W(s)	D	Ac	Pr	Re	Fs
0.2	24	89.1	78.9	96.6	86.9
0.3	33	90.6	82.6	95.1	88.4
0.4	**72**	**92.1**	**85.3**	**95.2**	**90.0**
0.6	95	90.9	82.8	94.7	88.3

Table 2: Syllable Boundary Detection Results at Frame-Level for Individual Feature with $W = 0.4$ s

Feat.	D	Ac	Pr	Re	Fs
I	24	87.6	77.6	91.8	84.1
F	25	91.2	84.1	93.9	88.7
T	30	88.2	78.2	93.9	85.3
O	28	87.2	78.0	91.4	84.1
IFTO	**72**	**92.1**	**85.3**	**95.2**	**90.0**

Table 3: Syllable Boundary Detection Results at Frame-Level with NTIMIT3m

Condition	Ac	Pr	Re	Fs
Matched	88.5	80.2	89.9	84.7
Mis-matched	82.1	68.7	94.2	79. 5

the cumulative gist vector by augmenting them. Then, principal component analysis (PCA) is used to remove redundancy and to reduce the dimension to make machine learning more practical.

3. Experiments and Results

TIMIT database is used in the syllable segmentation experiments. TIMIT doesn't have syllable annotation; hence the syllabification software *tsyllb2* [13] is used for syllabifying words using their phoneme transcription. Then, the timing information for syllables is automatically extracted using phoneme timing information provided with TIMIT and the phoneme sequence constituting each syllable. The official TIMIT train and test split is used in the experiments. The test set contains 1344 utterances, which contains approximately 17190 syllables. In addition, to test the method in a noisy condition, TIMIT database is re-recorded in real office noise environment using a PlayStation Eye camera which has a built-in microphone array. The sound source is located 3 meter away from the camera in the recordings. The TIMIT far-field recordings, referred as NTIMIT3m here after, contain linear and non-linear distortions such as additive noise, channel and non-linear DSP distortion.

In the experiments, a 3-layer neural network is used for learning the mapping between the auditory gist features and the syllable boundaries. The neural network has D inputs, $(D + N)/2$ hidden nodes and N output nodes, where D is the length of auditory gist vector after PCA dimension reduction when 95% of the variance is retained, and N is the number of classes, which is two; i.e. boundary vs. non-boundary.

The exact syllable boundaries for multi-syllabic words can be ambiguous in English; i.e. it is hard to decide which consonants belong to the first or the second syllable. Hence, the experiments are conducted such that the goal was to estimate the end of syllable nucleus where usually there is a vowel-to-consonant transition. In the rest of the paper, the term syllable boundary will be used to mean the end of syllable nucleus.

The auditory gist features are estimated every 50 ms using a window that centers on the current frame to capture the context. A 50 ms error margin is allowed for the syllable boundary detection. For example, if there is a boundary at 130 ms, the auditory gist features corresponding to the frames at 100 ms and 150 ms are both labeled as a boundary in the training. During evaluation, frame/frames detected as a boundary within 50 ms window of a reference boundary is/are accepted correct. For the above example, detecting a boundary for either frames located at 100 ms or 150 ms is accepted correct, when there is a reference syllable boundary at 130 ms. The excessive detected boundaries are counted as insertions and having no detected boundary for a reference one is counted as deletions.

The role of window duration W is investigated in the experiments by varying duration from 0.2 s, which is mean syllable duration, up to 0.6 s to analyze the effect of neighbouring left and right context on the performance. The grid size determines the temporal and spectral resolution. Different grid sizes are tested for auditory gist extraction for varying the temporal and spectral resolution. It was found that a grid size of 4-by-10 is sufficient and performs well in this task with a reasonable feature dimension. In Table 1, the frame-level syllable boundary detection results, Accuracy (Ac), Precision (Pe), Recall (Re), and F-score (Fs), for varying window duration are presented together with the auditory gist dimension D. The boundary detection performance is lower for shorter window duration; i.e. when $W = 0.2$ s (mean syllable duration). These results indicate that contextual information helps syllable boundary detection. Increasing the window duration improves the performance and the best performance achieved is 92.1% syllable boundary detection accuracy at frame-level with $W = 0.4$ s.

The contribution of each feature in the attention model is presented in Table 2 for $W = 0.4$ s. All of the features individually perform well above the chance level, which is 56.5% (obtained by labeling all frames with the majority class). The most informative feature about syllable boundary detection is *frequency contrast*, which achieves 91.2% syllable boundary detection accuracy at frame-level. However, the best performance is achieved with the combined features $IFTO$. The high performance achieved with frequency contrast features can be attributed to that they are tuned to local horizontally oriented edges, which can be good for detecting and capturing formants and their changes.

To test the proposed method with noisy data, a set of experiments is conducted on NTIMIT3m using $IFTO$ features with $W = 0.4$ s window. The experiments are conducted for matched and mis-matched conditions. The neural network is trained using the train split of the re-recorded NTIMIT3m and clean TIMIT for matched and mis-matched conditions, respectively. The syllable boundary detection results at frame-level are presented in Table 3. For the matched condition, 88.5% syllable boundary detection accuracy is achieved at frame-level. As one can observe, the performance is lower than the one achieved in clean conditions which indicates that NTIMIT3m contains distortions which make the boundary detection difficult even in the matched condition. One can think 88.5% accuracy as an upper-bound for NTIMIT3m since it is achieved with models trained with matched data. The experiments are repeated for the mis-matched condition and 82.1% syllable boundary detection accuracy is achieved when clean-trained neural network is used. As one can expect, the mismatch between noisy and clean data causes some performance degradation; however, the syllable boundary detection accuracy, 82.1%, in mis-matched condi-

Table 4: Comparison of Syllable Nuclei Detection Results

Method	Re	Pr	Fs
TCSSC[3]	86.06	99.69	90.21
nRG[7]	79.97	99.84	88.58
RG [7]	86.54	98.86	92.07
Attention Cues	**92.23**	**94.09**	**93.15**

tion is still well above the chance level. It can be concluded that the proposed method for syllable boundary detection is robust to noise.

We cannot directly compare our results with the results reported in the literature due to differences in the definition of the problem, evaluation metrics, data sets used in the experiments, etc. In the literature, the work on syllable segmentation is focused on the syllable nucleus detection. Most of the existing methods rely on estimating a one-dimensional continuous curve from extracted acoustic features and performing a peak search on the curve to locate syllable nuclei [3, 8, 7]. Then, if there is peak located within ϵ (usually $\epsilon = 50$ ms) window of a reference syllable nucleus, it's counted as correct. However, here more detailed syllable boundary information is extracted at frame-level; i.e. for each frame a probability score of the frame being a syllable boundary can be estimated. By definition of the problem, the syllable boundary estimation has a more strict time constraint, and hence allows a smaller error margin compared to syllable nucleus detection methods.

For comparison with the work in the literature, syllable nuclei detection experiments are also conducted and syllable level results are obtained. First, a neural network is trained such that frames corresponding to the middle of syllable nuclei are labeled as targets to be detected. For each frame, the neural network returns a value between $[0, 1]$ which can be thought as the posterior probability of a frame being the middle of a syllable nucleus given auditory gist features. Next, the neural network output score is used for generating a one-dimensional curve as a function of time and a peak search is performed on the curve to locate local maxima. Finally, peaks that are larger than 0.5, which is the standard threshold used in binary classification, are used to locate syllable nuclei. For scoring, a time-alignment between the detected syllable nuclei and the reference ones is used. As done in [3, 7], if a peak is detected within 50 ms window of a reference syllable nucleus, it is accepted as correct. Here, no peak could validate more than one reference syllable nucleus, the excessive detected peaks are counted as insertions, and having no detected peak for a reference syllable nucleus is counted as a deletion.

The syllable nucleus detection results are given in Table 4 along with the state-of-the art (to the best of our knowledge) results reported in [3, 7] on TIMIT. The auditory attention features can detect 92.2% of syllable nuclei with 94.1% precision. [3, 7] optimized algorithm parameters to obtain the best Recall, the best Precision, and the best F-score. One should refer to F-score in Table 4 for comparison since the best Recall and the best Precision cannot be obtained simultaneously at a given time; whereas, F-score, by definition, considers both Precision and Recall at a given time. The results from Table 4 show that the proposed method with auditory attention features performs better than the state-of-the art on syllable nucleus detection based on F-score. In addition, the method does not require parameter/threshold tuning or need any information about the nature of the problem. Hence, it can be easily used for different sets of conditions, speaking styles, and languages.

4. Conclusion and Future Work

In this paper, biologically inspired auditory attention cues are proposed for syllable segmentation of continuous speech. First, an auditory spectrum is computed from input sound using an early auditory system model. A set of multi-scale auditory features is extracted in parallel from the auditory spectrum using 2D spectro-temporal receptive filters. It can be thought that the auditory spectrum is analogous to an image of a scene in vision and the extracted multi-scale features capture/detect different oriented edges and discontinues in the scene. Next, multi-scale features are converted into low-level auditory gist features that capture the essence of sound. Using a neural network, the model learns the mapping between the gist features and syllable boundaries.

The proposed method achieves 92.1% syllable boundary detection accuracy at frame-level when tested on TIMIT. The method is tested with noisy TIMIT recorded in real office environment and it's shown that the method can detect syllable boundaries robustly in a noisy condition. Finally, for comparison, it is shown that the method can successfully detect syllable nuclei with high accuracy and perform better than the state-of-the art (to the best of our knowledge) in [3, 7] on TIMIT.

As part of our future work, we plan to conduct experiments in other languages, speaking styles, and noise conditions.

5. Acknowledgement

The author would like to thank Dr. Ruxin Chen and Dr. Gustavo Hernandez-Abrego at SCEA for valuable discussions and for providing TIMIT far-field recordings.

6. References

[1] S. Wu, M. L. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *Proc. of ICASSP*, 1997.

[2] N. Morgan and E. Fosler-Lussier, "Combining multiple estimators of speaking rate," in *Proc. of ICASSP*, 1998.

[3] D. Wang and S. Narayanan, "Robust speech rate estimation for spontaneous speech," vol. 15, no. 8, pp. 2190–2201, 2007.

[4] J. Zheng, H. Franco, and A. Stolcke, "Rate-dependent acoustic modeling for large vocabulary conversational speech recognition," in *Proc. of the NIST Speech Transcription Workshop*, 2000.

[5] C. D. Bartels and J. A. Bilmes, "Use of syllable nuclei locations to improve asr," in *Proc. of ASRU*, 2007.

[6] A. Howitt, *Automatic syllable detection for vowel landmarks*. PhD Thesis, MIT, 2000.

[7] Y. Zhang and R. Glass, "Speech rhythm guided syllable nucleis detection," in *Proc. of ICASSP*, 2009.

[8] T. Dekens, M. Demol, W. Verhelst, and P. Verhoeve, "A comparative study of speech rate estimation techniques," in *Proc. of Interspeech*, 2007.

[9] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *J. Acoust. Soc. Am*, pp. 880–883, 1975.

[10] O. Kalinli and S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *Proc. of Interspeech*, 2007.

[11] O. Kalinli, S. Sundaram, and S. Narayanan, "Saliency-driven unstructured acoustic scene classification using latent perceptual indexing," in *Proc. of MMSP*, 2009.

[12] O. Kalinli and S. Narayanan, "Prominence Detection Using Auditory Attention Cues and Task-Dependent High Level Information," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, pp. 1009–1024, 2009.

[13] B. Fisher, "Syllabification software," National Institute of Standards and Technology,http://www.itl.nist.gov/iad/mig//tools/, 1997.