LANGUAGE-RESOURCE INDEPENDENT SPEECH SEGMENTATION USING CUES FROM A SPECTROGRAM IMAGE

Su Jun Leow[1]*, Eng Siong Chng*[1] *and Chin-Hui Lee*[2]

[1] School of Computer Engineering, Nanyang Technological University, Singapore
[2] School of ECE, Georgia Institute of Technology, Atlanta, GA. USA
sleow001@e.ntu.edu.sg, aseschng@ntu.edu.sg, chl@ece.gatech.edu

ABSTRACT

In this paper, we use image processing techniques on the speech spectrogram to perform speech phoneme segmentation. The proposed method relies solely on visual cues on the spectrogram, without the need for language-specific training data. The results are evaluated on the TIMIT corpus, and compared to other unsupervised speech segmentation techniques, with comparable results obtained. We also fuse the results with those obtained by *hidden Markov models (HMM)* and HMM-based forced alignment to investigate if image features can provide an additional feature representation for speech processing tasks. With the fusion, up to 10% absolute improvement in segmentation accuracy over the HMM baselines can be obtained. Results are promising and suggests a strong potential for image-based features applying to speech processing.

Index Terms— speech processing, spectrogram processing, image processing, speech segmentation, low-resource languages

1. INTRODUCTION

Speech parameters currently being used are mostly spectral features, such as *mel frequency cepstral coefficients (MFCC)* and *perceptual linear predictive (PLP)* coefficients. These features are proven to be effective as a form of speech representation as they are derived from modeling of the human vocal tract.

In this paper, we explore an alternative view of representing speech, by investigating the use of visual features to perform speech processing. This work is motivated by the fact that human spectrogram readers are able to perform phoneme labeling from spectrogram images as proven by experiments done in [1], [2], [3] and [4]. These studies suggest that visual cues can provide additional information if they are properly used in current speech processing applications. More often than not, speech researchers visualize the correctness of their algorithms using spectrograms. Such visual inspections of the cues on the spectrogram provide an effective way of speech analysis, which further confirms that visual cues exist on the spectrograms to enable such analysis to be done.

If such image features can be reliably extracted, it can provide an additional feature representation to improve the accuracy of speech processing tasks. More importantly, if such features can be encoded as heuristics without the need for data-driven training, they will deem valuable for the processing of low-resource languages. One might argue that human spectrogram readers need to go through adequate training in order to recognize phonemes from spectrograms. However, many early research studies ([5], [6], [7], [8] and [9]) have suggested that these visual cues can indeed be encoded by rules and used to perform phoneme classification. These experiments however were done on small test sets, and the features were either manually extracted, or by using simple peak selection on the spectrograms.

To automate the extraction of visual features from image spectrograms, image processing techniques for phoneme recognition have been used as described in [10], [11], [12], [13], [14] and [15]. The effectiveness of these methods has been shown on restricted tasks, such as stops or vowels identification, and digit classification. However we wish to use image processing techniques to perform phoneme segmentation which is essential for many tasks in speech processing, such as model training and speech synthesis. Ironically, the state-of-the-art segmentation algorithms often involve the use of hidden Markov models (HMMs) [16] to perform phoneme recognition as discussed in [17] and [18]. This implies that manually transcribed and/or time-labeled data needs to be available to build the HMMs in order to perform segmentation. Obtaining manually transcribed training data is extremely labor intensive, and may not be available for low-resource languages.

The proposed framework in this study overcomes the need to train speech models to perform speech segmentation similar in essence to those described in [19], [20], [21], [22] and [23]. By performing image processing on the spectrogram, phoneme boundaries in the TIMIT corpus are detected and compared to the manually transcribed ground truth and compared with the phoneme boundaries obtained by other unsupervised phoneme boundary detection methods. Finally, we fuse the results of the proposed algorithm with that achieved by HMM and show that image features on the spectrogram provides additional information and can be harnessed to improve current speech technologies.

The rest of the paper is organized as follows. In Section 2, details of the algorithm are provided with discussion of the experimental results provided in Section 3. Finally, the conclusion and future work are presented in Section 4.

2. SEGMENTATION ALGORITHM

The algorithm starts with the generation of an image representation of the speech signal, which is the time-frequency speech spectrogram. With the intuition that locations on the spectrogram with significant intensity changes along the time-axis indicates a possible phoneme change, we detect for such differences to locate possible phonetic boundaries. As there could be spurious detections, the algorithm performs a post processing to select the best set of boundaries by using a minimum distortion reduction criterion.

2.1. Spectrogram Generation

The time-frequency spectrogram is obtained by performing *short-time Fourier transform (STFT)* [24] on the speech signal using a window size of 128 samples with a 64-sample overlap size. 128 samples correspond to an 8 ms window when STFT is applied to TIMIT sequences sampled at 16000 Hz. In order to preserve the frequency details, the number of FFT bins is set to 256. The logarithm is then applied on the *power spectral density (PSD)* matrix, where each column corresponds to the PSD for each window, from which we obtain the image representation of the signal.

Observing the spectrogram image generated, as illustrated by the top image in Fig. 1, the contrast of the formants with the background is not very distinct, which can make boundary detection challenging at a later stage. Therefore, we perform a contrast stretching and histogram equalization [25] of the whole spectrogram. This is commonly done in image processing to improve the contrast of images.

In addition to enhancing the contrast, we also perform median filtering on the image, which is usually used in image processing to remove salt-and-pepper noise [26]. As seen from the top image in Fig. 1, regions of silence does not appear as smooth regions, but as grainy patches of gray regions. If not properly dealt with, these regions will generate spurious detections creating high false alarms. Therefore, a median filter with a window size of B is applied to each column of the image to remove such noises, with the effect of B discussed in Section 3. The filtering is not done row-wise so as not to blur the phoneme boundaries which we eventually wish to detect. Fig. 1 shows an example of the spectrogram before and after enhancement. Comparing the top and bottom images, regions of silence now appear as white smooth regions, and the formants appears stronger after enhancement.

2.2. Boundary Detection

To detect for possible phoneme boundaries, the algorithm searches for significant intensity changes along the time-axis in three sub-bands. To avoid getting many spurious detections, a window of 5 columns is used to do frame differencing. We compute $\mathbf{d}(s, c)$ which is the difference of each sub-band s at column c as follows,

$$\mathbf{d}(s,c) = \frac{\sum_{r=r_1}^{r_2} \sum_{w=1}^{5} |\mathbf{I}(r, c-5+w) - \mathbf{I}(r, c+w)|}{(r_2 - r_1 + 1) \times 5} \quad (1)$$

where $\mathbf{I}$ refers to the 2-D spectrogram, and r_1 to r_2 defines the rows in sub-band s of interest.

The first sub-band, s_1, is defined as 1 to $(0.1 \times h)$, where h refers to the image height. The voicing bars [28] are usually detected in this sub-band. The second sub-band, s_2, is defined as $(0.1 \times h + 1)$ to $(0.6 \times h)$, and is where formants of vowels can be detected. The last sub-band, s_3, refers to $(0.6 \times h + 1)$ to h. The high frequencies of the fricatives are frequently found in this region. This is illustrated by the bottom image in Fig. 1.

Finally, we compute the image difference profile, $\mathbf{D}$, in Eq. (2).

$$\mathbf{D}(c) = \max(\mathbf{d}(s_1, c), \mathbf{d}(s_2, c), \mathbf{d}(s_3, c)) \quad (2)$$

$\mathbf{D}$ only records the sharpest intensity change that happens in any of the sub-bands. The peaks in $\mathbf{D}$ are found and peak locations whose values are above $f\times$ mean($\mathbf{D}$) are chosen as boundary locations, where the effect of f is discussed in Section 3.

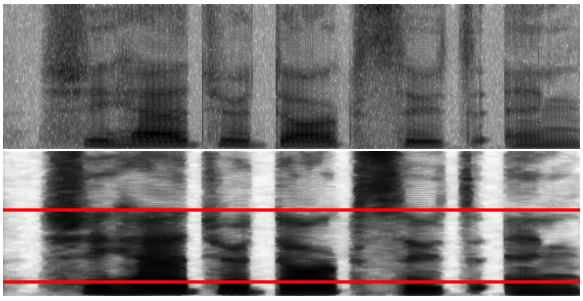


Fig. 1. (Top) Spectrogram before enhancement. (Bottom) Spectrogram after enhancement and divided into the three sub-bands.

2.3. Boundary Selection

Spurious boundaries detected above due to noise are removed here by using a minimum distortion reduction criterion. Given a segment on the image, the distortion E, of this segment is given by Eq. (3),

$$E = \frac{\sum_{r=1}^{m} \sum_{c=2}^{n} |\mathbf{I}(r,c) - \mathbf{I}(r,c-1)|}{m \times (n-1)} \quad (3)$$

where m and n refer to the number of rows and columns in spectrogram $\mathbf{I}$. Essentially, E gives the amount of intensity change accumulated in the segment. Intuitively, if the segment does indeed contain a boundary, selection of this boundary would split the image into two portions, left and right of the boundary. The accumulated distortion of the two resultant segments would be lesser than the original segment since the boundary, which is the biggest contributor to the distortion measure, is removed and does not contribute to E.

With this intuition in mind, the boundary selection process starts with the entire image being one segment. The best boundary from the set of detected boundaries is selected and the image is split into two segments. The algorithm is repeated on the newly split segments and the entire process iterates until no more boundaries can be selected from the current set of segments.

There are three cases whereby no more boundaries can be selected from a segment:

- no more detected boundaries on the segment, or
- the resultant distortion reduction is less than the specified threshold, set to 0.0025 in our algorithm, or
- the selected boundary would result in a segment smaller than the minimum width, set to 5 here to coincide with the window length used in Section 2.2.

Once the boundary selection process terminates, the selected set of boundaries will give the final segmentation of the spectrogram into its phoneme segments.

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

3.1. Image-based Phoneme Segmentation

The proposed algorithm, which we term *ISeg*, was used to segment the TIMIT testing set consisting of 1344 utterances. The results were compared against the results of [20] using the same set of evaluation metrics, *correct detection rate (CDR)*, *over-segmentation (OS)* and *false alarm (FA)*, defined in Eqs. (4) to (6) below:

$$\text{CDR} = \frac{\text{Number of correct boundaries}}{\text{Number of true boundaries}} \times 100\%, \quad (4)$$

$$\text{OS} = \left(\frac{\text{Number of boundaries found}}{\text{Number of true boundaries}} - 1 \right) \times 100\%, \quad (5)$$

$$\text{FA} = \left(1 - \frac{\text{Number of true boundaries found}}{\text{Number of boundaries found}} \right) \times 100\%, \quad (6)$$

where a correct boundary was defined as being at most 20 ms from a true boundary. Performance of the segmentation algorithm was given in Fig. 2. The curves were obtained by first arbitrarily fixing the peak selection threshold f described in Section 2.2 to 0.5 and varying the median filter window size B, which affected the level of details preserved and therefore the number of boundaries detected. The best window size was found to be and fixed at 8. We then vary f.

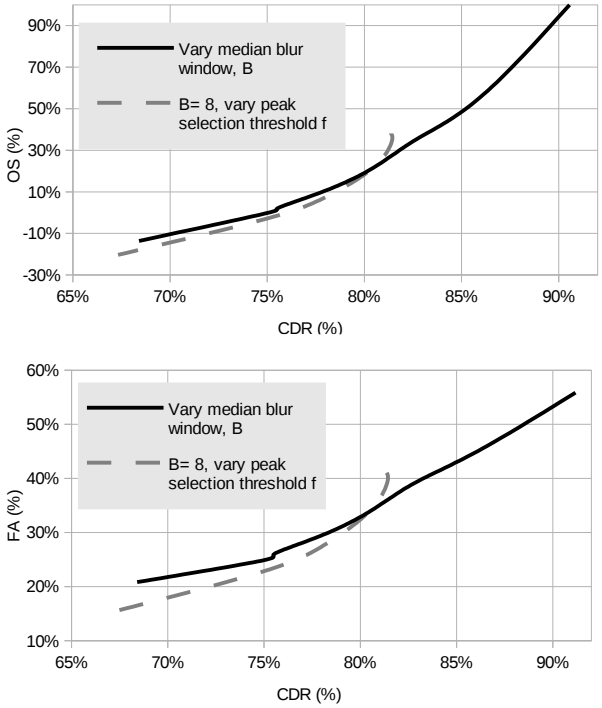


Fig. 2. (Top) OS plotted against CDR for *ISeg*. (Bottom) FA plotted against CDR for *ISeg*.

The results obtained were very similar to those found in [20] and [22], with slight degradation, and were tabulated in Table 1. As only the CDRs obtained at 0% OS were provided in both [20] and [22], we could only compare our results based on that.

The best operating point was found when $B = 8$ and $f = 0.625$, with CDR at 78.07%, OS at 7.74% and FA at 27.54%, which would be used for subsequent comparisons. Although the proposed method did not outperform the existing unsupervised methods, we were still encouraged by the results as it shows the potential for image features on spectrograms to be applied to speech processing tasks. With simple image processing techniques applied to the spectrogram, the algorithm proved to attain a similar performance to other unsupervised segmentation methods.

Method	**CDR**
MMC [20]	76.0%
RD-IDA [22]	77.5%
Proposed Method - ISeg	75.0%

Table 1. Comparison of *ISeg* with other unsupervised segmentation algorithm at OS = 0%.

3.2. Fusion with HMM-based Segmentation

Next, we fused the results of *ISeg* with those obtained by HMM phoneme recognition, *HMM-rec*, and those obtained by HMM forced alignment, *HMM-fa*, to investigate if fusing our results with existing methods applied on spectral features could give rise to extra performance gains. If an accuracy gain was indeed observed, it would imply that visual cues could provide an additional feature to improve the performance of speech processing tasks.

To obtain HMM-based segmentation, the HTK toolkit and the TIMIT training set consisting of 4120 utterances were used to train context-independent HMMs for the 61 phonemes in the TIMIT corpus. It was believed that context-independent HMM gives better segmentation results as investigated in [17]. Each HMM contained 5 states with left-to-right transitions, and each state contained 16 Gaussian mixture components. The HMMs were then used to obtain HMM segmentations for *HMM-rec* and *HMM-fa*.

We subsequently performed a simple fusion of the results by first comparing all the boundaries found in *ISeg* with those obtained by *HMM-rec* or *HMM-fa*. Identical boundaries were ignored and the remaining boundaries discovered by *ISeg* were fused into the boundary set discovered by HMMs. The results were indicated by *(raw)* in Table 2. This simplistic fusion obviously generated many false alarms as detections that were one column apart were also considered as different boundaries, which was definitely unrealistic since no segments can be only one column wide.

In addition, we did not double count true detections in our evaluation. If a boundary had been determined to be less than 20ms from a true boundary, a match for the true boundary had been fulfilled and accounted for in the true detection count. Even if another detection that was within 20ms from the same true boundary was found, the second detection will be calculated as a false alarm. This explains the extremely high false alarm counts for the *(raw)* results. The purpose of the results however is to highlight the improvement in CDR that fusing results from *ISeg* can bring out.

Table 2. Comparison of results obtained from fusing *ISeg* with HMM baselines, against *HMM-rec* and *HMM-fa*.

Method	**CDR(%)**	**OS(%)**	**FA(%)**
ISeg	78.07	7.74	27.54
HMM-rec	84.57	7.73	21.50
HMM-fa	86.60	0.00	13.40
ISeg-HMM-rec (raw)	94.90	104.24	53.53
pruned	89.75	33.98	33.01
strict (f=0.75)+pruned	88.98	27.17	30.03
ISeg-HMM-fa (raw)	96.28	97.04	51.14
pruned	91.85	30.25	29.48
strict (f=0.75)+pruned	91.29	23.06	25.82

As observed from Table 2 above, when we fused the results of

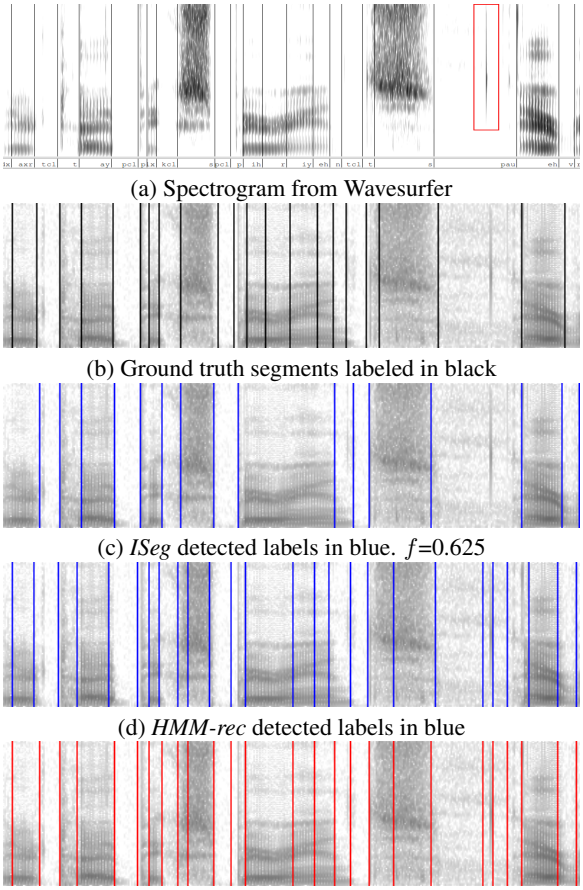
HMM-rec with *ISeg*, the achieved CDR was 94.90%, which was about 10% higher than that obtained by *HMM-rec* alone. Similarly, when the results of *ISeg* was fused with those from *HMM-fa*, close to 10% improvement of CDR was also observed. This confirmed our hypothesis that visual cues on the spectrogram indeed provided an additional feature representation to existing ones for speech processing. Comparing the detections obtained by *ISeg* with that obtained by HMM in Fig. 3, it could be observed that the boundaries obtained by *ISeg* in Fig. 3(c) snap closer to the true phoneme boundaries that are shown in Fig. 3(b). The ones obtained from *HMM-rec* in Fig. 3(d) are very often a few columns off the true boundaries. Visual cues therefore gave a more accurate segmentation of phoneme boundaries which contributed to the improvement of the CDR in the fused results. Note also in Fig. 3(a) that there was a distinct noise spike exhibited as a black vertical line, which we highlighted in a red box. *ISeg* did not generate a false detection of it as another segment due to median filtering of the spectrogram.

In other words, *ISeg* was able to overcome transient impulse noise spikes in the speech with suitable filtering. However, looking at the output of *HMM-rec* in Fig. 3(d), there is a false detection in that same segment. *ISeg*, of course, has its drawbacks. At a further observation of Fig. 3, we noticed the series of miss detections in the middle portion of Fig. 3(c). Comparing to the ground truth in Fig. 3(a) and Fig. 3(b), that portion corresponds to a sequence of vowels and approximants which exhibit themselves very similarly in spectrograms as discussed in [27]. Therefore, it makes it very challenging to find vowel-vowel or approximant-vowel boundaries as no sharp transitions on the image spectrogram could be observed. Perhaps *ISeg* would be more effective in segmenting syllabic languages with clear consonant-vowel-consonant structures.

Despite the improvement in CDR brought about by fusion, the simplistic merging of the two set of boundaries generated too many false alarms to deem the performance improvement useful. Therefore, we removed some of those unrealistic false alarms by running a sliding window with a size of 5 columns through all the detected boundaries. If the window contained more than one boundary point, we selected the one with the highest intensity change at the boundary. If the intensity change at this selected boundary was less than a threshold τ, we rejected this selected boundary as it did not correspond to a location with high intensity changes, and should thus be a false alarm. τ is set to 3 here. This eradicates segment boundaries that are less than 5 columns from each other and the results of such a false alarm pruning is illustrated in Fig. 3(e).

The performance of such a pruning strategy on the evaluation metrics are shown in rows indicated by *pruned* in Table 2. This post-processing helped to remove some of the false alarms, but still a significant amount remained. This suggests that a better enhancement step should be performed on the spectrogram so that the extracted features are more invariant to spectral changes and thus these spurious detections could be reduced. A better fusion method is also required to exploit the advantages of both *ISeg* and HMM without generating too many false alarms.

If a slight loss in CDR can be tolerated, a stricter version of *ISeg* can also be used. In rows indicated by *strict (f=0.75), pruned* in Table 2, the peak selection threshold in *ISeg* was set to f=0.75, higher than the optimal f=0.625 as discussed previously. The results of this configuration was fused with HMM segmentation and false alarm pruning was also performed. A slight degradation in CDR was observed, but both OS and FA were lowered with this setup. The CDR now was still above that achieved by *ISeg* or HMM alone.



(a) Spectrogram from Wavesurfer

(b) Ground truth segments labeled in black

(c) *ISeg* detected labels in blue. f=0.625

(d) *HMM-rec* detected labels in blue

(e) Fused results of *ISeg* and *HMM-rec* with false alarm pruning

Fig. 3. Comparison of segment boundaries obtained from *ISeg* and *HMM-rec*

4. CONCLUSION AND FUTURE WORK

In summary, we have presented an image processing approach to speech phoneme segmentation. Results are comparable to other unsupervised phoneme segmentation methods which proved that image-based feature can be a possible alternative to phoneme segmentation. We also fuse the results of the proposed segmentation algorithm with the HMM results and observed an improvement to the segmentation accuracy, which proves that image features does contain additional information that can be harnessed to aid the performance of speech processing tasks.

In the future, we wish to explore better image features that are more invariant to spectral variations in order to reduce the false alarms. Different kind of spectrogram representations can also be studied. Fusing the results of the proposed algorithm with other language-independent detectors is also a possible future direction. One such detector would be the manner and place attribute detector. As speech research focus increasingly on minority languages, it is believed that language-independent speech processing techniques would become more important. Since the image processing framework described here is free of any language resources, we hoped that it can contribute to low-resource language processing.

5. REFERENCES

[1] D. H. Klatt and K. N. Stevens, *On the Automatic Recognition of Continuous Speech: Implications from a Spectrogram-Reading Experiment*, IEEE Transactions on Audio and Electroacoustics, Vol. AU-21, No. 3, June 1973.

[2] F. Ingermann and P. Mermelstein, Haskin Laboratories, *Speech Recognition Through Spectrogram Matching*, Journal of the Acoustical Society of America, Vol 57, No 1, January 1975.

[3] P. W. Nye, F. S. Cooper, and P. Mermelstein, *Interactive Experiments with a Digital Pattern Playback*, Haskin Laboratories: Status Report on Speech Research SR-49, 1977.

[4] V. W. Zue and R. A. Cole, *Experiments on Spectrogram Reading*, IEEE Conference Proceedings, ICASSP, Washington D. C., 1979, pp. 116-119.

[5] J. Johannsen, J. MacAllister, T. Michalek and S. Rose, *A Speech Spectrogram Expert*, ICASSP, Boston, 1983.

[6] V. W. Zue and L. F. Lamel, *An Expert Spectrogram Reader: A Knowledge-based Approach to Speech Recognition*, ICASSP, Tokyo, 1986.

[7] J. H. Connolly, E. A. Edmonds, J. J. Guzy, S. R. Johnson and A. Woodcock, *Automatic Speech Recognnition based on Spectrogram Reading*, International Journal of Man-Machine Studies, Vol. 24, pp 611-621, 1986.

[8] K. Hatazaki, Y. Komori, T. Kawabata and K. Shikano, *Phoneme Segmentation using Spectrogram Reading Knowledge*, ICASSP, Glasgow, 1989.

[9] L. F. Lamel *A Knowledge-based System for Stop Consonant Identification based on Speech Spectrogram Reading*, Computer Speech and Language, 1993.

[10] H. C. Leung and V. W. Zue, *Visual Characterization of Speech Spectrograms* ICASSP, Tokyo, 1986.

[11] J. F. Hemdal and R. M. Lougheed, *Morphological Approaches to the Automatic Extraction of Phonetic Features*, IEEE Transactions on Signal Processing, Vol. 39, No. 2, February 1991.

[12] R. Steinberg and D. O'Shaughnessy, *Segmentation of a Speech Spectrogram using Mathematical Morphology*, ICASSP, 2008.

[13] B. Pinkowski, *Multiscale Fourier Descriptors for Classifying Semivowels in Spectrograms*, Pattern Recognition, Vol. 26, No. 10, pp 1593-1602, 1993.

[14] B. Pinkowski, *Principal Component Analysis of Speech Spectrogram Images*, Pattern Recognition, Vol. 30, No. 5, pp 777-787, 1997.

[15] K. Schutte and J. Glass, *Speech Recognition with Localized Time-frequency Pattern Detectors*, Automatic Speech Recognition and Understanding Workshop, 2007.

[16] L. Rabiner and B. H. Juang, *An Introduction to Hidden Markov Models*, ASSP Magazine, IEEE, Vol. 3, Issue 1, pp 4-16, 1986.

[17] D. Torre, L. A. H. Gomez and L. V. Grande *Automatic Phonetic Segmentation*, IEEE Transactions on Speech and Audio Processing, Vol. 11, Issue 6, pp 617-625, 2003.

[18] J. Adell, A. Bonafonte, J. A. Gomez and M. J. Castro, *Comparative Study of Automatic Phone Segmentation Methods for TTS*, ICASSP, 2005

[19] S. Dusan and L. Rabiner, *On the Relation between Maximum Spectral Transition Positions and Phone Boundaries*, Interspeech, 2006.

[20] Y. P. Estevan, V. Wan and O. Scharenborg, *Finding Maximum Margin Segments in Speech*, ICASSP, 2007.

[21] D.R. Van Niekerk and E. Barnard, *Acoustic Cues Identifying Phonetic Transitions for Speech Segmentation*, Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), 2008

[22] Y. Qiao, N. Shinomura and N. Minematsu, *Unsupervised Optional Phoneme Segmentation: Objectives, Algorithm and Comparisons*, ICASSP, 2008.

[23] N. J. Shah, B. B. Vachhani, H. B. Sailor and H. A. Patil, *Effectiveness of PLP-based Phonetic Segmentation for Speech Synthesis*, ICASSP, 2014.

[24] J. B. Allen, *Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform*, ICASSP 1977.

[25] R. E. Woods and R. C. Gonzales, *Real-time Digital Image Enhancement*, Proceedings of IEEE, Vol 69, Issue 5, pp 634-654, 1981.

[26] W. K. Pratt, *Median Filtering*, Semiannual Report, Image Processing Institute, University of Southern California, pp 116-123, 1975.

[27] R. Hagiwara, (2009, Sep, 19), *How to read a spectrogram*, [Online], Available: http://home.cc.umanitoba.ca/ robh/howto.html.

[28] P. G. Garn-Nunn and J. M. Lynn, *Calvert's Descriptive Phonetics, Third Edition*, NY:Thieme Medical Publishers, Inc., 2004, pp 127-128.