BLIND SPEECH SEGMENTATION USING SPECTROGRAM IMAGE-BASED FEATURES AND MEL CEPSTRAL COEFFICIENTS

Adriana Stan[†]*, Cassia Valentini-Botinhao*[‡]*, Bogdan Orza*[†]*, Mircea Giurgiu*[†]

[†] Communications Department, Technical University of Cluj-Napoca, Romania
[‡] The Centre for Speech Technology Research, University of Edinburgh, UK
{adriana.stan,bogdan.orza,mircea.giurgiu}@com.utcluj.ro, cvbotinh@inf.ed.ac.uk

ABSTRACT

This paper introduces a novel method for blind speech segmentation at a phone level based on image processing. We consider the spectrogram of the waveform of an utterance as an image and hypothesize that its striping defects, i.e. discontinuities, appear due to phone boundaries. Using a simple image destriping algorithm these discontinuities are found. To discover phone transitions which are not as salient in the image, we compute spectral changes derived from the time evolution of Mel cepstral parametrisation of speech. These so-called image-based and acoustic features are then combined to form a mixed probability function, whose values indicate the likelihood of a phone boundary being located at the corresponding time frame. The method is completely unsupervised and achieves an accuracy of 75.59% at a -3.26% over-segmentation rate, yielding an F-measure of 0.76 and an 0.80 R-value on the TIMIT dataset.

Index Terms— blind segmentation, unsupervised segmentation, phoneme segmentation, destriping, image processing

1. INTRODUCTION

Recent advances in speech technology, such as highly accurate speech recognition and high quality synthesis, drive the ambition for systems that can operate in any spoken language.[1] For this goal to be achieved, the required manual collection and labelling of data is not feasible.[2] An alternative is to devise unsupervised or lightly supervised methods to build and extract the necessary data and features from recordings of any language. Phone-level segmentation of an utterance is one such data and for many speech processing applications it is in fact an essential requirement, without which the underlying models could not be trained. Although determining the phonetic transcription of a text is fairly simple in most languages, dividing an utterance into its phonetic constituents is not as trivial. Pronunciation variations, both inter- and intra-speaker, as well as individual physiological constraints add to its complexity [1].

Methods to obtain automatic phone-level segmentation are generally classified into two major categories: *constrained* and *unconstrained*. For the *constrained* methods, the number and sequence of phones in the utterance is available, and the task is to determine the exact location of their boundaries in time. The most common procedure for performing constrained phone-level segmentation is to use an acoustic model, usually a hidden Markov model, or dynamic time warping to perform the so-called forced-alignment [2–5]. Recent work in this area is based on deep belief networks [6] that are trained to estimate the posterior probabilities of phone categories and to then locate the boundaries of frames where phone class assignment is uncertain to an extent. These methods can achieve performances similar to the inter-labeller agreement percentages, i.e. 93% [3]. However, correct phonetic transcription of an utterance is difficult to obtain as it requires a lot of manual and expert work. In the second category, the *unconstrained*, phonetic content is not known apriori. There are two types of methods in this category: methods that use pre-trained models, similar to automatic speech recognition, to identify and locate the phonetic boundaries [7–10]; and methods that determine salient acoustic changes from spectral or temporal features, and use them to estimate the number and location of the phone boundaries, but without providing phone identity, commonly referred to as *blind segmentation* methods [11–15].

In this paper we propose a blind segmentation method motivated by the observation of how human labellers perform the phonetic annotation of speech. In most cases, labellers base their boundary assignments on both the *acoustic* signal, by listening to the sample, as well as on the *image* evidence, through spectrogram inspection. This observation led us to explore ways in which we could extract alternative features from the spectrogram by interpreting it as a static image. The majority of the phone transitions consist

The research leading to these results has received funding from the Romanian Ministry of Education, under the grant agreement PN-II-PT-PCCA-2013-4 Nº 6/2014 (SWARA), and from the EPSRC Programme Grant EP/IO31022/1, Natural Speech Technology.

[1] https://www.kth.se/en/forskning/artiklar/kth-hjalper-wikipedia-borja-prata-1.631897

[2] The are around 6500 languages spoken worldwide, and for each of them a group of experts should be available.

of abrupt changes or shifts in the spectrum, which translate to discontinuities in the vertical axis of the spectrogram image, or stripes. There are several reasonably accurate image destriping algorithms available [16–19], that could provide a good basis for the image-based automatic phonetic segmentation. As some acoustic transitions are not as prominent in the spectogram image, we combine the image-based information with information extracted from the trajectories of Mel cepstral coefficients [20]. The combined feature sets are mapped to a probability function that indicates how likely it is that a phone boundary is located at a certain time index. The evaluation shows that despite its very simple processing steps, the method outperforms all the previously published blind speech segmentation methods evaluated on the same dataset.

The paper is organised as follows: Section 2 describes the image destriping method used in this work. Section 3 describes how we extract image-based and acoustic features and the manner in which we combine them. Experimental results and discussions are presented in Section 4, followed by conclusions in Section 5.

2. IMAGE DESTRIPING

The automatic image-processing method which can locate and correct the discontinuities, or stripes which appear in the spectrogram, are the so-called image *destriping* algorithms. Their main purpose is to reduce scanning or sequential image composition defects [16–19].

For our work, we selected a simple but effective destriping algorithm which is also available in the open source GIMP Image Processing Software.[3] The algorithm starts by identifying the stripes as local deviations from an average value computed over a fixed length window centred on the current vertical image slice.[4] The deviation is computed independently for each colour component, i.e. red, green and blue (RGB), and represents a single 3x1 RGB vector for each vertical slice. From these deviation values it is possible to create the so-called *negative pattern* of the image by concatenating the deviation obtained for each image slice. In Figure 1(b) we show the negative pattern associated with the image from Figure 1(a). For visualisation purposes, we stacked the pattern horizontally multiple times. This pattern is then applied to the original image, i.e. the RGB values of each vertical slice of the original image are multiplied by the negative pattern associated with that slice. The result is an image in which the discontinuities are smoothed, as it can be observed in Figure 1(c).

3. PROPOSED PHONE SEGMENTATION METHOD

In the same way as the human labellers use multiple stimuli to segment an utterance, we extract both image-based and acoustic features. We expect the acoustic features to contribute most in the case of slow phone transitions, such as with diphthongs, where spectral changes are not as abrupt and therefore not as easily identifiable by the image-based features.

[3]GIMP 2.8 available online: `https://www.gimp.org/`

[4]The width of the vertical slice is usually set to 1 pixel.

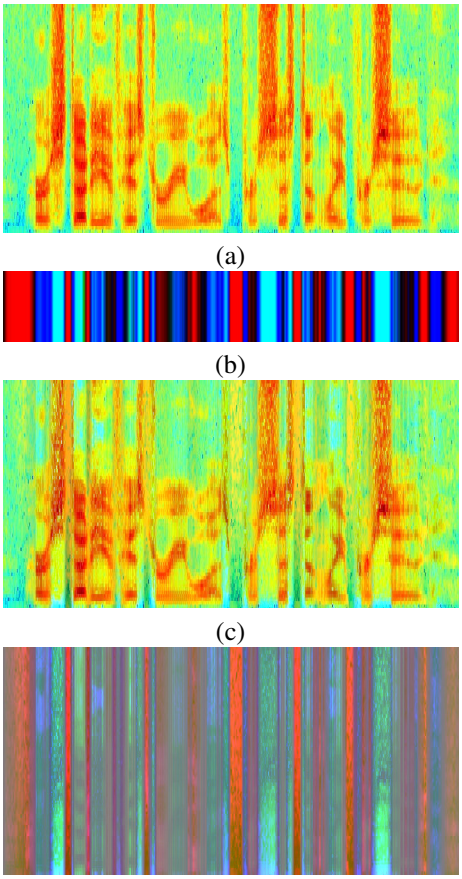

Fig. 1. The spectrogram (a), negative pattern (b), destriped spectrogram (c) and visiogram (d) of a realization of the utterance *"Her study of history was persistently pursued."*

3.1. Image-based features

Through a simple visual inspection of a spectrogram, even non-experts can identify most phoneme boundaries as image discontinuities, see Fig. 1(a). We are however not interested in smoothing the image but rather identifying the areas where the algorithm takes maximum effect, i.e. the discontinuities. To identify these regions we add a weighted version of the negative pattern to the spectrogram image. The weight applied to the pattern is global and bigger than one, so that the pattern is weighted more in the summation. The result of this process can be observed in Figure 1(d), and will be referred to as the *visiogram*. Although at first glance, it might seem that the visiogram is very similar to the negative pattern, on closer inspection, the finer details obtained in this last image are important in ranking the potential phone boundaries.

Hypothesised segmentation locations are signalled through

colour variations in the visiogram. Higher variations correspond to more abrupt spectral changes. To detect and rank these variations we use the standard CIEDE2000 perceptual colour distance [21].[5] The distance is computed between every two consecutive vertical slices, and then normalised to one. The result is therefore a positive number between zero and one, which can be interpreted as a probability function associated with the probability that a phone boundary is located at a certain vertical slice or time index. Figure 2(a) shows an example of this probability function for a short utterance and the manually annotated phone boundaries. It can be observed that this function has higher values in the vicinity of the time frames where phone boundaries occur.

It can be argued that the negative pattern and the visiogram calculation could also be obtained from the values of the spectrogram function directly, rather than the RBG values associated with its image representation. The reasoning for choosing to use RGB values instead is motivated by the idea that the quantized representation, the image, is the additional information used by human labellers when performing the task.

3.2. Acoustic Features

Based on the positive results obtained in previous phone-level segmentation studies [22], we selected the Mel cepstral coefficients (MCEP) as acoustic features. Similar to the visual features' processing steps, we aim to detect spectral variations from the variations of MCEP values over time. These variations are computed by calculating a distance measure between MCEPs of consecutive time frames. The distance measure that provided best results on a small validation set was the Manhattan distance, defined as follows:

$$d(X,Y) = \sum_{i=1}^{n} |x_i - y_i| \tag{1}$$

where $X = (x_1, x_2, ...x_n)$ and $Y = (y_1, y_2, ..., y_n)$ are the two consecutive Mel cepstral coefficients vectors of order n.

The acoustic distance is computed for every pair of consecutive analysis frames, and then normalised to 1. Again, we can associate this value to the probability that a phone boundary is located at a certain time index. Figure 2(b) shows the acoustically-derived phone boundary probability values for a short utterance. We can observe that the peaks occuring close to phone boundaries are not as pronounced as in Figure 2(a) but the presence of local peaks could still be used as an indication of a boundary.

The calculation of MCEPs does not impose a substantial additional computational load to the method since the spectral representation of the utterance is also required to construct the visiogram. However, in our current implementation we use separate processes to compute the spectrum, but would like to unify them in the future.

[5] Due to the complex formulation of the distance, we do not present its formal definition, and refer the reader to `http://www.brucelindbloom.com/index.html?Eqn_DeltaE_CIE2000.html`

3.3. Combining image-based and acoustic features

We hypothesize that phone boundaries should be assigned to the regions where both visual and acoustic probabilities are high, or where one of them is significantly higher than the neighbouring values. Hence, a mixed probability value is computed as the product of the two probabilities at each time index, and summed over a window centred around the current time frame as follows:

$$P_b(t) = \frac{1}{2N+1} \sum_{i=-N}^{N} P_v(t+i) P_a(t+i), \tag{2}$$

where P_b is the mixed probability at time t, P_v and P_a are the visual and acoustic probabilities, respectively, and $2N+1$ is the window length. By examining the mixed probability function plotted in Figure 2(c), we can notice that its peaks are better indicators of the phone boundaries. Yet in order to minimise over-segmentation, a minimum peak height and distance are used in the peak detection process.

4. SEGMENTATION RESULTS

This section presents the evaluation data and results obtained with the proposed method. We also compare our results with other published methods on the same dataset, and discuss the limitations and future developments.

4.1. Data

Evaluating phone segmentation methods is quite a hard task in itself, as there are not that many manually labelled speech corpora available. Most of the studies published so far have used the TIMIT corpus [23] for training, and we therefore selected it for the evaluation, as well. Our method does not require any training data, but in order to compare our results directly to the other methods, we use the test subset of the corpus for our evaluation. The subset contains 1344 utterances[6] from 168 speakers (approx. 1.5 hours of data), sampled at 16kHz with a 16 bit resolution. The complete 61 phone set was maintained, including pauses and separate symbols for the closure and release intervals of the stops. However, the beginning and end silence segments were trimmed to 50ms, so that the high number of spectral discontinuities which commonly occur in these areas do not bias our results.[7]

To avoid overtuning on the TIMIT test set data, we also evaluate our method using a 300 utterance subset (approx. 28 minutes) of the Italian read speech corpus, CLIPS-

[6] The *sa dialect calibration utterances were excluded.

[7] Excluding the silence from the phonetic segmentation evaluation is common, as a voice activity detection method could be employed as a preprocessing step.

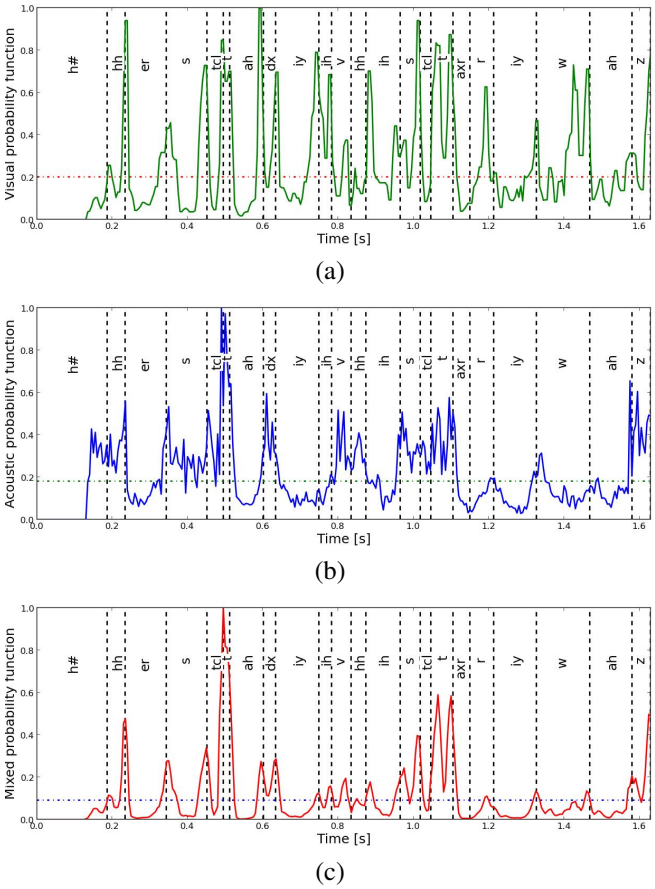

Fig. 2. (a) Image-based, (b) acoustic, and (c) mixed probability function plots for the utterance *"Her study of history was [...]"*. The vertical dotted lines represent the reference phone boundaries. The horizontal lines represent the minimum peak height threshold.

Letto [24].[8] The subset contains 15 speakers, and it is sampled at 16kHz with 16 bit resolution. The data is manually labelled at phone-level using the 43 symbols of the Italian SAMPA phoneset. The same silence trimming as for the TIMIT corpus was applied.

4.2. Phone Segmentation configuration

The complete segmentation process is as follows: each utterance is first pre-emphasised using a first order high-pass filter. The spectrograms are then computed using a 128 length fast Fourier transform and 5ms window shift using Matplotlib's *specgram* function[9], and plotted as a colour mesh. The image resolution was set such that each pixel had a width which equated to 5ms on the time axis. The destriping algorithm used a window of length 100. To alleviate accidental high-frequency variations in the resulting image, a 5 pixel wide Gaussian smoothing process was applied to the visiogram.

The probability function associated with the image-based features was estimated at every 5ms time index, and smoothed using a third order median filter. The probability function associated with the acoustic features was derived from a 12 MCEP plus energy feature set extracted from the STRAIGHT spectrum [25] using the SPTK tool.[10] Initial tests showed that applying a median filter on this probability function does not improve the performance of the method, and we therefore did not use it in the evaluation. The combination of these two probability functions was computed over a 15ms window ($N = 1$ in Eq. 2). The peak detection algorithm had a minimum peak height set to 80% of the mixed probability's average value across the utterance, and a minimum of 20ms distance between peaks. These numbers were adjusted on a small validation set taken from the TIMIT training dataset.

4.3. Evaluation Metrics

In [26] Rasanen et al. proposed a metric that is considered to be more appropriate for the evaluation of blind segmentation algorithms than the conventional F-measure statistics. In blind segmentation over-segmentation is inevitable and this increase in the number of hypothesised boundaries leads to an increase in recall. The measure introduced by Rasanen establishes an ideal operating point of a segmentation algorithm to be at 100% recall and 0% over-segmentation. The metric uses the following three quantities: the total number of boundaries in the reference segmentation N_{ref}, the number of boundaries correctly detected N_{hit} and the number of boundaries hypothesised by the proposed method N_{alg}. Using these quantities, intermediate values for hit rate HR and over-segmentation rate OS are computed as follows:

$$HR = \frac{N_{hit}}{N_{ref}} \times 100; \quad OS = \left(\frac{N_{alg}}{N_{ref}} - 1\right) \times 100 \qquad (3)$$

The final metric, called the *R-value* is defined as:

$$R = 1 - \frac{|r1| + |r2|}{200} \qquad (4)$$

where:

$$r1 = \sqrt{(100 - HR)^2 + (OS)^2}; \;\; r2 = \frac{-OS + HR - 100}{\sqrt{2}} \qquad (5)$$

To compute N_{hit} it is common to consider a hit accuracy margin. This means that the estimated boundary can lie within an interval equal to twice the accuracy margin, centred around the reference boundary location. The most common value for this accuracy margin is 20ms.

Because not all previously published results report the R-value, we also use the F-measure to present the performance of our speech segmentation method.

[8] Available online: `http://www.clips.unina.it/it/`

[9] `http://matplotlib.org/api/mlab_api.html#matplotlib.mlab.specgram`

[10] Speech Signal Processing Toolkit (SPTK) 3.9 `http://sp-tk.sourceforge.net/`

Table 1. *Segmentation results for the **image-based**, **acoustic** and **mixed** probability functions, expressed in terms of hit rate (HR), over-segmentation rate (OS), F-measure and R-value, at various accuracy margins.*

Features	Margin	HR [%]	OS [%]	F-meas.	R-Value
Acoustic	20ms	86.03	53.88	0.68	0.48
Image-based	20ms	76.37	12.83	0.72	0.74
Mixed	5ms	47.94	-3.26	0.50	0.57
Mixed	10ms	63.80	-3.26	0.65	0.70
Mixed	20ms	75.59	-3.26	0.76	0.80
Mixed	50ms	83.06	-3.26	0.84	0.86

Table 2. *Comparison with other blind speech segmentation methods which report results on the TIMIT corpus in terms of F-measure and R-value.*

Method	F-meas.	R-value
Dusan et al. (2006) [7]	0.71	0.73
Esposito and Aversano (2005) [14]	0.75	0.74
Khanagha et al (2014) [15]	0.73	0.77
Rasanen et al. (2009) [13]	0.76	0.78
Estevan et al. (2007) [12]	0.76	0.80
Proposed method	0.76	0.80

4.4. Results

As a first evaluation we calculated the performance metrics when using the two probability functions, acoustic and image-based separately as the input of the peak detection module that identifies phone boundaries. Table 1 shows these results at an accuracy margin of 20ms. It can be noticed that neither function achieves a satisfactory accuracy on its own. However, the image-derived probability function is significantly more accurate than the acoustic one. It should be noted that no individual optimisation was performed for either of the two functions, and therefore the results could potentially be improved. Nonetheless, when combining the two functions into the mixed probability function, the results increase notably. Table 1 shows these results at several accuracy margins. At the 20 ms accuracy margin, results on the TIMIT test set achieve a 0.76 F-measure, and an 0.80 R-value. As these results could be a result of overtuning the configuration on the TIMIT data, we used the same setup on the CLIPS-Letto subset. The results are similar: 0.77 F-measure and 0.79 R-value, and perhaps could be improved by adapting the configuration parameters on a small validation set from this data.

As it is common to report using the 20ms accuracy margin we present in Table 2 our results using this margin against results obtained by other methods on the TIMIT data set and on the same task, i.e. blind segmentation. We can see that the spectrogram destriping method outperforms all other methods. Although the results are similar to the ones described in [12], the authors reported a hit rate of 76% with 0% over-segmentation, resulting in a 0.795 R-value, while our method achieves an 0.802 R-value. This difference, however, could have been caused by slight variations in the evaluation procedure, and it is not statistically significant.

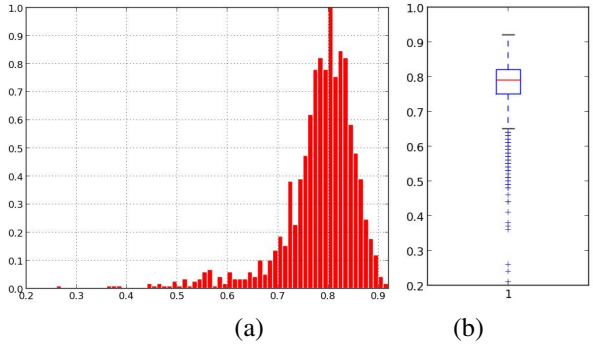

Fig. 3. Normalised histogram (a) and boxplot (b) of the TIMIT testset utterance-level R-values .

It is important to note that we used a 20ms minimum distance between peaks for our peak detection method. This means that phone boundaries located less than 20ms from each other are omitted. If we exclude these boundaries from the test set (approx. 4% of the total number of boundaries in the test corpus), our hit rate is equal to 78.72% at an over-segmentation rate of -3.26%. This results in a R-value of 0.82. However, if the minimum distance is reduced to 5ms, the over-segmentation rate increases dramatically. For this reason we believe that it is better to maintain the 20ms value.

The F- and the R-values are considered good metrics for the evaluation of the blind segmentation methods but they do not inform how consistent a method is across the evaluation data. To provide this information we present in Figure 3 the histogram and boxplot of the R-values computed for each individual utterance. It is worth observing that these are concentrated around the 0.8 value with very few outliers, 90% of the utterances are above 0.7, and 48% are above 0.8. This is a good indication that that the proposed method performs well, independent of the speaker or phonetic content, even though it is completely unsupervised and unconstrained. Upon closer inspection, it seems that the low R-values are determined by a high degree of over-segmentation, which in turn can be caused by spectral artefacts, or clipping in the speech data.

5. CONCLUSIONS

This paper introduced a novel method for blind phonetic segmentation of speech that combines two sets of features. One set is composed of the Mel cepstral coefficients, while the other one is derived from the spectrogram image via the use of an image destriping algorithm. The method is completely unsupervised and simple to compute as the additional image-based features are also derived from the short term

Fourier transform of the speech signal. Results obtained using this method were found to be better than results obtained by any other unsupervised segmentation evaluated on the TIMIT test dataset, with an accuracy of 75.59% at -3.26% over-segmentation rate, resulting in a 0.802 R-value. Similar accuracies were obtained on a separate dataset of Italian read speech corpus, 0.793 R-value, even though the method's parameters were chosen using a subset of the validation material from TIMIT. The results also show that the proposed method is consistent across the entire evaluation dataset with very few outliers in terms of the sentence-level R-values distribution. As future work we would like to investigate if the visiogram is also suited for noisy conditions, phone alignment, as well as perhaps phone recognition. It would also be interesting to study the voiced-voiced transitions into more detail, and to determine a more suitable distance, or representation for this type of boundary.

6. REFERENCES

[1] Casey O'Callaghan, "Auditory Perception," in *The Stanford Encyclopedia of Philosophy*. 2014.

[2] F. Brugnara and D. Falavigna and M. Omologo, "Automatic segmentation and labeling of speech based on Hidden Markov Models.," *Speech Communication*, vol. 12, no. 4, pp. 357–370, 1993.

[3] John-Paul Hosom, "Speaker-independent phoneme alignment using transition-dependent states," *Speech Communication*, vol. 51, no. 4, pp. 352–368, Apr. 2009.

[4] Andreas Stolcke, Neville Ryant, Vikramjit Mitra, Wen Wang, and Mark Liberman, "Highly Accurate Phonetic Segmentation Using Boundary Correction Models and System Fusion," in *Proc. ICASSP*. May 2014, pp. 5552–5556, IEEE SPS.

[5] Montri Karnjanadecha and Stephen A. Zahorian, "Toward an Optimum Feature Set and HMM Model Parameters for Automatic Phonetic Alignment of Spontaneous Speech," in *Proc. of Interspeech*. 2012, pp. 2290–2293, ISCA.

[6] Ozlem Kalinli, "Combination of auditory attention features with phone posteriors for better automatic phoneme segmentation," in *Proc of Interspeech*. 2013, pp. 2302–2305, ISCA.

[7] Sorin Dusan and Lawrence R. Rabiner, "On the relation between maximum spectral transition positions and phone boundaries.," in *Proc. of Interspeech*. 2006, ISCA.

[8] Yu Qiao, Naoya Shimomura, and Nobuaki Minematsu, "Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons," in *Proc. of ICASSP*, 2008, pp. 3989–3992.

[9] RaviShankar Prasad and B. Yegnanarayana, "Acoustic segmentation of speech using zero time liftering (ZTL)," in *Proc. of Interspeech*. 2013, pp. 2292–2296, ISCA.

[10] Vijayaditya Peddinti and Kishore Prahallad, "Exploiting Phone-Class Specific Landmarks for Refinement of Segment Boundaries in TTS Databases," in *Proc. of Interspeech*. 2011, pp. 429–432, ISCA.

[11] Odette Scharenborg, Vincent Wan, and Mirjam Ernestus, "Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries," *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 1084–1095, 2010.

[12] Y.G. Estevan, V. Wan, and O. Scharenborg, "Finding maximum margin segments in speech," in *Proc. of ICASSP*, April 2007, vol. 4, pp. IV–937–IV–940.

[13] Okko Räsänen, Toomas Altosaar, and Unto Laine, "Blind segmentation of speech using non-linear filtering methods," *INTECH Open Access Publisher*, 2011.

[14] Anna Esposito and Guido Aversano, *Text Independent Methods for Speech Segmentation*, pp. 261–290, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[15] Vahid Khanagha, Khalid Daoudi, Oriol Pont, and Hussein Yahia, "Phonetic segmentation of speech signal using local singularity analysis," *Digital Signal Processing*, Dec. 2014.

[16] Yi Chang, Luxin Yan, Houzhang Fang, and Chunan Luo, "Anisotropic spectral-spatial total variation model for multispectral remote sensing image destriping," *IEEE Transactions on Image Processing*, vol. 24, no. 6, pp. 1852–1866, 2015.

[17] J. van Gumster and R. Shimonski, *GIMP Bible*, Wiley Publishing, 2010.

[18] Shu-wen W. Chen and Jean-Luc Pellequer, "DeStripe: frequency-based algorithm for removing stripe noises from AFM images," *BMC Structural Biology*, vol. 11, no. 1, pp. 1–10, 2011.

[19] Liu Qihua and Feng Jing, "A destriping method combining strong filter with weak filter based on image divided and adapative strip noise detection," *Physics Procedia*, vol. 25, pp. 2103 – 2108, 2012.

[20] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, San Francisco, USA, March 1992, vol. 1, pp. 137–140.

[21] Gaurav Sharma, Wencheng Wu, and Edul N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color Research and Application*, vol. 30, no. 1, pp. 21–30, 2005.

[22] Iosif Mporas, Todor Ganchev, and Nikos Fakotakis, "Phonetic segmentation using multiple speech features," *International Journal of Speech Technology*, vol. 11, no. 2, pp. 73–85, 2008.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," 1993.

[24] Renata Savy and Francesco Cutugno, "Diatopic, diamesic and diaphasic variations in spoken Italian," in *Proceedings of the 5th Corpus Linguistics Conference: CL2009*, 2009.

[25] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveign, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187 – 207, 1999.

[26] Okko J. Räsänen, Unto K. Laine, and Toomas Altosaar, "An improved speech segmentation quality measure: the R-value," in *Proc. of Interspeech*, Sept. 2009, pp. 1851–1854.