Segmentation of a Speech Spectrogram using Mathematical Morphology

Raphael Steinberg, Douglas O'Shaughnessy - Fellow

INRS-Telecommunications
800 de la Gauchetiere Ouest,
Montreal, H5A 1K6, Canada

ABSTRACT

It has been shown that speech spectrograms can be read by trained experts. In this work, we regard the speech spectrogram image as a written text in some unknown language and perform segmentation in order to capture the energy associated with each formant. We propose an algorithm based on Mathematical Morphology operators and mainly on the watershed transform. The result is robust segmentation for wideband speech spectrograms that can be later used for automatic speech recognition. We show results of experimental runs for different phoneme classes.

Index Terms— Speech recognition, Morphological operations, Image segmentation, Optical character recognition

1. INTRODUCTION

The sound spectrogram was invented in the 1940s to help break enemy codes and to help detect submarines. Speech spectrograms soon followed and have been used in the study of speech [1]. Previous attempts to recognize speech using the speech spectrogram as input have produced mixed results. Spectrogram reading involves recognition of phonemes, sub-phonemes and multiple-phonemes and requires from the reader hours of training and good knowledge and understanding of speech.

Human experts can read speech spectrograms with a high level of accuracy. Spectrogram reading requires a combination of different sources of knowledge such as articulatory movement, phonotactics, linguistics and acoustic phonetics [2]. Prof. Victor Zue of MIT has spent over 2500 hours learning spectrogram reading and has reached impressive recognition rates. Zue and Cole [3] have given encouraging results for automatic speech recognition based on speech spectrograms. Different experiments demonstrate recognition rates in the range of 85%. Such encouraging recognition rates motivate the development of an automatic tool to perform the reading task.

In an effort to mimic the human experts' behavior we choose a large time interval on the order of 1 second in order to capture several phonemes that may be related through co-articulation. Speech signals can be modeled as non-stationary signals. Movements of the vocal tract can be well represented using a wideband spectrogram. The wideband spectrogram is generated using a relatively short time window that gives good time resolution but less specified frequency resolution.

Previous attempts to extract information from speech spectrograms have been made. We note here the work of [4] that used morphological skeletons to extract information. While in general it is possible to extract information through a skeleton-based approach, we believe it is necessary to identify and segment the speech spectrogram into *Binary Large Objects* (BLOBs). The uncertainty principle, as demonstrated by the Heisenberg–Gabor inequality, stipulates that the extent to which a particular frequency can be localized is inversely proportional to the length of the time interval chosen. Attempting to track down frequency changes with time using a single pixel skeleton path is futile when the time interval is too short to allow single pixel localization.

In [5], an expert system based on spectrogram reading knowledge was devised with an objective to segment speech into different phonemes. It deals with voiced/unvoiced fricatives, voiced/unvoiced stops, nasals and liquids. A rule-based expert system reports recognition rates of about 90% for the aforementioned phoneme classes. These results motivate us to focus on the classes that are more difficult to recognize by a rule-based segmentation, namely the vowels and glides.

2. TIME FREQUENCY REPRESENTATION

Time-Frequency Representation (TFR) differs from a spectrogram representation by calculating either the spectral power or spectral energy of the signal instead of its spectral absolute value logarithm. Nadine Martin examined various algorithms for TFR segmentation [6, 7]. In [6] two algorithms for TFR segmentation were suggested. The first is based on morphological filters and the watershed transform and the second is based on tracking using a Kalman filter. Another interesting segmentation scheme based on statistical features of a spectrogram is presented in [7]. Since the signal characteristic and origin are assumed unknown throughout the algorithm the segmentation is blind toward the analyzed signal. Tuning is not required. Assuming in both cases [6, 7] a deterministic signal corrupted by additive Gaussian noise, a probability model is developed to allow for local segmentation of objects.

Wideband speech spectrograms are indeed noisy images. Vertical lines that striate the spectrogram show that it may be inappropriate to model the noise as a log-normal distribution as would be the case if we apply the algorithms developed in [6, 7] for TFR to the spectrogram. The vertical striating lines are caused by the opening and closure of the vocal cords. These lines appear in a spaced distance that can be used as a rough approximation to the fundamental frequency f_0, also known as the pitch. Another caveat for using a blind method as proposed in [6, 7] is the difficulty in adjusting it to recognize specific types of information present in the speech spectrogram. It is possible to calculate the fundamental frequency and filter-out the vertical lines by either tracking them on the image spectrogram or by filtering the original signal. For example, the pitch can be calculated from the cepstral coefficients [8] that can be easily extracted from the image

spectrogram. However the morphological tools and filtering that are used in the proposed algorithm perform this task indirectly. Using median smoothing and a local threshold helps in reducing the fundamental frequency vertical striating lines. Due to the windowing effect, energy from different formants is smeared in wide frequency bands. In the following we demonstrate the importance of selecting a suitable window.

3. WINDOW SELECTION

A common tradeoff in window selection is the main lobe width versus side lobe roll-off rate. A Hann window is used often due to its good roll-off properties: 60 dB/decade. The Hamming window has a lower roll-off of 20 dB/decade but a lower main lobe width since its maximum side lobe level is -43 dB as apposed to -32 dB for the Hann window [9]. Choosing a narrow main lobe reduces the uncertainty in frequency and allows us to better distinguish between formants that have a small frequency difference. The lower roll-off introduces dependencies on previous and future speech samples resulting in a noisier image. However, the watershed transform can produce better results since it can better capture low energy regions in particular on rising and falling formants as seen by comparing fig. 1(c) and fig. 1(d). Therefore we choose to work with a hamming window for the specified short time interval of 6.5 ms.

4. SPEECH SPECTRORAMS

A speech signal sampled at 16 kHz is transformed using a 1024 point *Fast Fourier Transform* (FFT). The FFT is windowed using a 6.5 ms Hamming window with 50% overlap. The logarithm of the absolute value of the FFT is then taken. Finally, histogram equalization is performed on rectangular tiles. The rectangular tiles are then tapered using the bilinear transform. A Gamma transform with an exponent value of 0.8 is applied to enhance image brightness. The Gamma correction value can differ depending on the hardware (monitor/printer) used.

The tiled histogram equalization operates on rectangular regions and generates more homogeneous energy values for the different formants. This method of equalization differs from the pre-emphasis filter, since it is performed on a rectangular tile and not on particular vertical lines/time instances. The result is an image spectrogram that clearly shows the first four speech formants, f_1 to f_4.

5. MATHEMATICAL MORPHOLOGY

Mathematical Morphology (MM) enables mathematical characterization of geometrical shapes based on lattice theory and topology. It was invented by Jean Serra and Georges Matheron in 1964 [10]. The basic building blocks of MM are Dilation (Minkowski addition) and its dual, Erosion [11]. Using these two dual operators, more complex operators can be constructed such as opening, closing, skeleton, *Skeleton by Influence Zones* (SKIZ), thinning, thickening, Hit-Or-Miss, watershed transform and more [12]. MM was originally used for binary images and later extended to gray scale images.

Our goal is to extract from the Speech Spectrogram the first four formants of each phoneme. The speech spectrogram gives us the energy concentration blurred due to the windowing effect. After binarization each formant will appear as a numbered BLOB.

6. WATERSHED TRANSFORM

The *Watershed Transform* (WT) is a morphological-based image processing segmentation algorithm. First proposed by Digabel and Lantuéjoul [13] and later extended to gray scale images by Beucher and Lantuéjoul [14] the watershed transform has been studied from theoretical, practical and algorithmic points of view. Currently, thanks to the work of Soille and Vincent [15] an efficient fast algorithm for computing watersheds exists and enables practical implementation of segmentation tasks.

A watershed can be classified as a region-based segmentation approach; it takes its reasoning from a natural phenomena occurring in geography: consider the image to represent a 3D topographical surface. Multiple local minima terrain points are flooded with water. At points where water from different basins meets, dams are built. The water is confined within catchment basins and the dams which are called watershed lines or simply watersheds, are the separators between different segmented pieces. When the water reaches the highest point in the landscape, the process is stopped resulting in a labeled and segmented image. Before running the watershed transform a gradient of the image is calculated. Normally, a morphological gradient also known as Beucher gradient is used. The Beucher gradient is defined as:

$$g(f)=(f\oplus B)-(f\ominus B), \quad (1)$$

where $\oplus$ is a morphological dilation and $\ominus$ is the morphological erosion, both using the same structuring element. Since the SKIZ consists of all points which are equidistant (in a geodesic sense) to at least two nearest connected components, we have in the continuous case an identity between the watersheds and the SKIZ [16]. A good example of different uses of the WT can be found in [17].

7. LOCAL VS. GLOBAL THRESHOLD

We would like to obtain a binary image from the grayscale image spectrogram. In order to obtain a binary image we need to perform some sort of quantization. The naïve approach to quantization is selecting a global threshold level for the entire image spectrogram. Since lower formants tend to have higher energy concentration than higher formants and since the spectrogram image contains much detailed information regarding different formants, simply using a global threshold will not yield good results. A local threshold is used to isolate each BLOB from its surrounding. A global threshold is used to clean the image from noise. Combining the locally threshold image with the globally threshold image using a logical OR will yield the desired result.

8. ALGORITHM DESCRIPTION

1. Median filter is used once on a 3 by 20 rectangular and 4 times using a 20 pixel horizontal line.
2. Run a 2D Gaussian window (Gabor filter).
3. Smooth using a 2D Wiener filter. The local mean and variance are estimated in a 16 by 16 square around each filtered pixel.
4. Apply local threshold on (3).
5. Apply global threshold on (3).
6. Combine the results of (4) and (5) using a logical OR.
7. Dilate with a disk as a structuring element in order to disconnect thin lines and eliminate small areas in the image.

8. Use morphological connectivity to disregard small sections that contain less than 40 pixels or that have a maximum width that is less than 20 pixels.
9. Perform an 8-connectivity watershed algorithm.

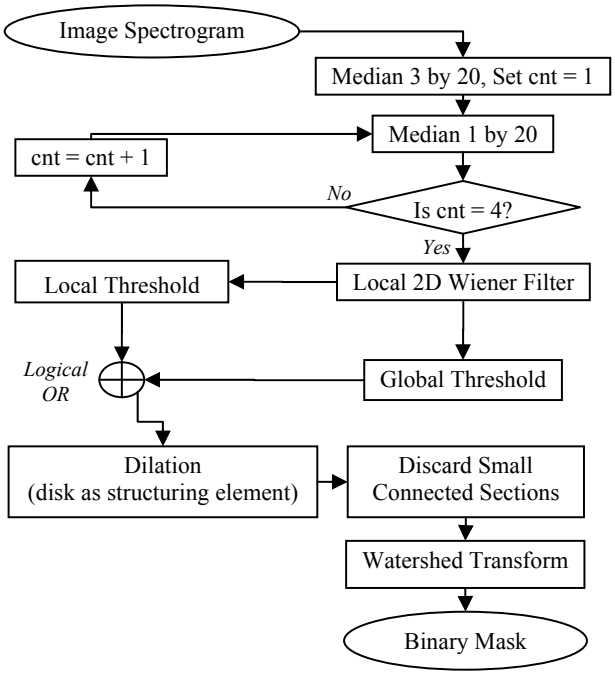


Figure 1: Algorithm flow chart

We obtain a segmented and labeled image. Two notable algorithmic improvements to [18] are the combination of a local and global threshold that allows the utilization of a threshold and the use of morphological image processing to tackle the segmentation problem. Therefore we are not limited to segmenting vowels as in [18] and we obtain robust results even when formants are near one another.

9. RESULTS

The algorithm was tested on different speech samples from the TIMIT database. Results were robust, the segmentation performed well on different speakers and different sentences. The TIMIT database contains female and male speakers from 7 different dialect regions in the United States. The speakers repeated sentences specially designed at SRI, MIT and TI to exemplify different speech characteristics such as accent, co-articulation and different combinations of phonemes. Orthographic transcription and time-aligned phonetic transcription are included for every sentence.

Our first example uses the meaningful sentence "However, the litter remained, augmented by **several dozen** lunchroom suppers"; bold face fonts indicate 1 second of speech that in this example is displayed in fig. 1(a). We obtain good segmentation for the first and second formants for all voiced phonemes. For the third and fourth formant, the segmentation misses part of the phoneme /r/ but the general direction is preserved. In this example, all four formants are well aligned and ready to be recognized by an appropriate system.

Our second example presents a more challenging scheme. We examine a different section of the same sentence: "However, **the litter remained**, augmented by several dozen lunchroom suppers." As seen in fig. 1(b), the algorithm has difficulty in segmenting the second and third formant of /r/. Since these formants are very close together it is hard to distinguish between them and to segment them as different objects. In addition, high energy levels for f_3 make it more difficult to separate it from f_2. Another difficulty arises in the identification of the nasal /m/. The low spectral density makes it hard to segment the phoneme correctly. The low spectral density is caused by a spectral zero that reduces the second formant. One other problem is small segments that do not represent a formant but still appear in the image (false positives). This problem can be solved by changing the constant in step #8 of the algorithm. However, changing the constant to accept only stronger energies would result in losing some real formants. In general, the algorithm manages to perform well when the formant energies are strong.

As a last example, we choose: "Don't ask me to carr**y an oily rag** like that." As seen in fig. 1(c), we obtain several cases in which formants are segmented into more than one BLOB. Even though over-segmentation was tackled in the watershed algorithm we still have remainders in the form of small binary objects that can cause problems in the recognition stage. On the other hand as was also noticed in the previous examples, BLOBs associated with f_1 sometimes relate to more than one phoneme. This phenomenon occurs in some cases for the higher formants as well.

In order to check the algorithm behavior in a more systematic fashion we test the results on multiple runs. The criteria for which we judge the performance is the fuzzy variable 'Grade' that takes the values {'Perfect', 'Good', 'Average', 'Below Average', 'Poor'} for the segmentation results. We assign numbers to each descriptor where 'Perfect' takes the highest value of 5, 'Poor' takes the lowest value of 1 and it is believed that 'Average' which takes the value of 3 contains enough information for automatic recognition. We select 10 phonemes and run 20 different tests for each phoneme, in total 200 different speech segments. The results including the mean and variance of the visual measurements are presented in Table 1.

After examining the algorithm we see that in general the algorithm obtains good segmentation results for the formant energy levels throughout different phonemes. The algorithm obtains better segmentation results when the phoneme duration is longer. Since more information is available and since our segmentation algorithm is searching for large objects we tend to miss small concentrations of energy. In general, the vowels are well-recognized. The nasal /m/ and the glide /l/ have lower segmentation results due to the difficulty of tracking diagonal lines in the spectrogram. It is possible to extend the algorithm to detect diagonal lines either by adding tracking procedure such as a Kalman filter or by a diagonal line emphasizing median filter. The semivowel /w/ is better segmented on short duration phonemes since there is a higher energy concentration that enables better segmenting of f_3 and f_4.

10. CONCLUSION

A robust algorithm for speech spectrogram segmentation was presented. By using morphological image processing techniques, we are able to obtain reliable segmentation of formants in most cases. The algorithm performs well for all voiced phonemes and has better segmentation results than the algorithms described in the

introduction; however, difficulties occur when formant frequencies are close together or when there is a low-energy formant that is rapidly going up or down in frequency. Some suggestions such as changing the threshold level were made to improve or tune the algorithm. These results can be used as input to an automatic speech recognition system or in other general uses of speech spectrograms. It is in the authors' belief that a spectrogram-based speech recognition system can complement an existing recognition system by incorporating human expert knowledge into the recognition task.

11. FUTURE WORK

We intend to create a feature vector from the BLOBs that can later be used by an Expert System. It seems that a system based on a combination of Fuzzy Logic with an adaptive neural network can give good results while avoiding the problems encountered in [4]. We intend to construct membership functions that would contain the rules needed to read a spectrogram and to train the system to perform speech recognition automatically.

12. REFERENCES

[1] Steinberg, J.C. and French, N.R. "The Portrayal of Visible Speech". J. Acoustical Society of America, vol. 17, pp. 4-18. 1946.

[2] Lamel, L. F. and Zue, V.W. "An Expert Spectrogram Reader: A Knowledge-Based Approach to Speech Recognition". ICASSP, vol. 11, pp. 558-561, 1986.

[3] Zue, V.W. and Cole, R.A. "Experiments on Spectrogram Reading". ICASSP, vol. 4, pp. 116-119, 1979.

[4] Hemdal, J.F. and Lougheed, R.M. "Morphological Approaches to the Automatic Extraction of Phonetic Features". Digital Object Identifier, vol. 39, Issue 2, pp. 490-497, Feb 1991.

[5] Shikano, K. et. al. "Phoneme Segmentation Using Spectrogram Reading Knowledge" pp:393 – 396, vol.1 Digital Object Identifier ICASSP. May 1989.

[6] Leprettre, B. and Martin ,N. "Extraction of Pertinent Subsets from Time-Frequency Representations for Detection and Recognition Purposes," Signal Process., vol. 82, no. 2, pp. 229–238, Feb. 2002.

[7] Hory, C., Martin, N. and Chehikian, A. "Spectrogram Segmentation by Means of Statistical Features for Non-Stationary Signal Interpretation" IEEE Transactions On Signal Processing, vol. 50, No. 12, Dec 2002.

[8] Schafer, R.W. and Rabiner, L.R. "System for Automatic Formant Analysis of Voiced Speech," J. Acoust. Soc. Amer., vol.47, pp. 634-648, Feb 1970.

[9] Porat, B. "A Course in Digitial Signal Processing" ISBN 0-471-14961-6. John Wiley and Sons. Oct 1996.

[10] Matheron, G. and Serra, J. "The Birth of Mathematical Morphology". Jun 1998.

[11] Serra, J. "Lecture Notes on Morphological Operators" http://cmm.ensmp.fr/~serra/cours/T-34.pdf

[12] Dougherty, E. R., "Mathematical Morphology in Image Processing". CRC Press. ISBN: 0824787242. Sep 1992.

[13] Beucher, S. and Lantuéjoul, C. "Use of watersheds in contour detection" in Proc. Int. Workshop Image Processing, Real-Time Edge and Motion Detection/Estimation, Rennes, France, Sep. 1979.

[14] Digabel, H. and Lantuéjoul, C. "Iterative Algorithms" in Proc. 2nd European Symp. Quantative Analysis of Microstructures in Material Science, Biology and Medicine, Caen, France, Oct. 1977 (1978) J.-L. Chermant, Ed. Stuttgart, Riederer, Verlag, pp. 85-99.

[15] Vincent, L. and Soille, P. "Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations". IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 13, No. 6, June 1991.

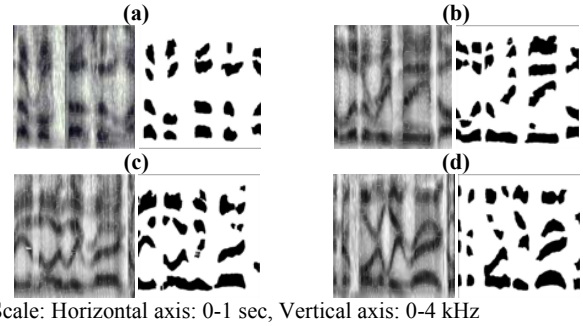
[16] Roerdink, J. and Meijster, A. "The Watershed Transform: Definitions, Algorithms and Parallelization Strategies". IOS Press Fundamenta Informaticae 41, pp. 187-228, 2001.

[17] Beucher, S. The Watershed Transform Page http://cmm.ensmp.fr/~beucher/wtshed.html

[18] Leung, H.C. and Zue V.W. Visual "Characterization of Speech Spectrograms". ICASSP, vol. 11 pp. 2751-2754, 1986.

Test #	Phoneme									
	aa	ae	Eh	ux	ow	oy	r	l	m	w
1	5	5	5	5	5	5	5	3	5	1
2	5	5	4	4	4	4	3	4	5	2
3	2	3	3	2	2	3	3	1	4	5
4	4	5	2	5	4	5	5	5	3	3
5	5	4	5	5	5	3	5	5	2	4
6	2	5	5	5	3	5	5	2	5	5
7	5	4	4	5	5	4	3	5	1	5
8	4	5	5	5	5	3	3	1	2	2
9	5	4	2	2	4	4	4	1	1	3
10	3	3	5	5	5	5	5	1	4	3
11	5	4	1	5	5	4	4	2	1	3
12	5	5	3	4	3	4	5	1	1	4
13	5	4	5	5	4	3	5	5	3	5
14	5	4	4	3	2	3	3	3	1	2
15	2	2	5	3	3	3	4	5	2	2
16	5	4	5	4	5	2	5	5	2	5
17	4	3	5	5	5	3	5	2	1	2
18	2	5	5	5	3	5	4	5	2	2
19	4	5	5	5	5	5	4	5	3	5
20	5	5	5	5	4	4	4	3	5	5
Mean	4.1	4.2	4.15	4.35	4.05	3.85	4.2	3.2	2.65	3.4
Variance	**1.46**	**0.8**	**1.61**	**1.08**	**1.10**	**0.87**	**0.69**	**2.91**	**2.34**	**1.94**

Table 1: Results of a visual inspection. The grades describe the accuracy of the segmentation algorithm for each phoneme.



Scale: Horizontal axis: 0-1 sec, Vertical axis: 0-4 kHz

Figure 2: Image Spectrograms before (left) and after segmentation.(a) "several dozen". (b) "the litter remained". (c) "an oily rag" using a Hamming Window. (d) "an oily rag" using a Hann Window.