

An Image-Based System for Spoken-Letter Recognition

Khalid Saeed¹ and Marcin Kozłowski²

¹ Białystok University of Technology, Faculty of Computer Science
Wiejska 45a, 15-351 Białystok, Poland
aidabt@ii.pb.bialystok.pl

<http://aragorn.pb.bialystok.pl/~zspinfo/>

² Department of Informatics, Statistical Office in Białystok

Krakowska 13, 15-959 Białystok, Poland

mkoz@interia.pl

<http://www.stat.gov.pl/urzedz/bialystok/index-en.htm>

Abstract. A new trial on speech recognition from graphical point of view is introduced. Isolated spoken-letters and color-names words are considered. After recording, the speech signal is processed as an image by Power Spectrum Estimation. For feature extraction, classification and hence recognition, the algorithm of minimal eigenvalues of Toeplitz matrices together with other methods of speech processing and recognition are used. A number of examples on applications and comparisons are presented in the work. The efficiency of the method is very high in the case of the six Polish vowels and English color-names, and the results are encouraging to extend the algorithm to cover more word classes.

Keywords: Speech, Image Processing and Recognition, Burg's and Toeplitz Models

1 Introduction

Most of speech classification algorithms are based on neural network theory, among them is that of Burr's [1]. In his paper, Burr used neural networks to classify and recognize written and spoken letters. Another and often used method of recognition is Hidden Markov Models, described for example in [2].

This paper, however, is based on algorithmic approaches and looks at the spoken letter as an image, that is, the speech signal is treated graphically. It is the aim of this paper that the approaches, achieved by the first author [3], which were successfully used in description and classification of written scripts and texts, are applied on speech recognition. Particular attention is paid to the application of Toeplitz forms and the minimal eigenvalues of their determinants for shape description. However, this approach cannot be applied directly because of the speech complicated nature. That is why the authors are using some other methods of speech pre-processing for better feature extract of voice image before entering Toeplitz-based algorithms. The processing methods that gave good results in most cases are LPC – Linear Predictive Coding coefficients [4], Spectral Moments [5] and Zero-Crossing method [6]. All of them have their advantages and also drawbacks. Seeking a simple and also a more stationary way for the graphical speech description, we applied the frequency spectral

estimation method based on linear prediction model introduced in [7] and will thereafter be referred to as Burg's model. This method seems to be very useful for image-spectral pre-processing. The obtained signal spectrum forms the basis to further analysis for spoken-letter classification and recognition.

2 Spoken-Letter Waveform Pre-processing

The input to the system is a recorded speech waveform (Fig. 1). This sound contains silence region before and after the under-analysis signal. The sound also consists of excessive number of samples, which must be reduced without losing the most important feature.

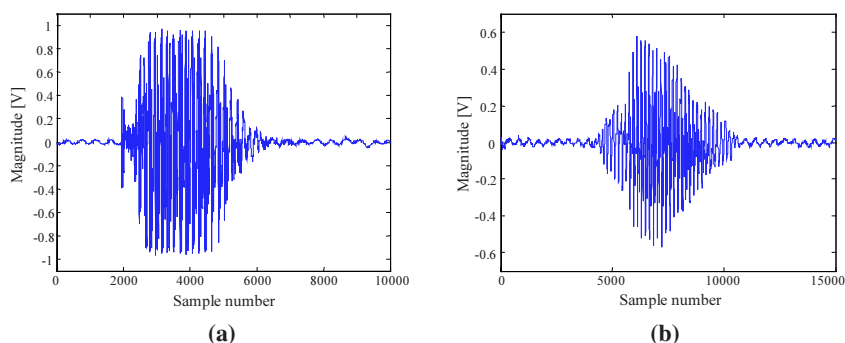


Fig. 1. The waveforms of two recorded speech signals: (a) The Polish vowel *a* - pronounced like 'a' in 'car', (b) The English word 'blue'.

The process of the speech preparation for feature extraction is given in Fig. 2.

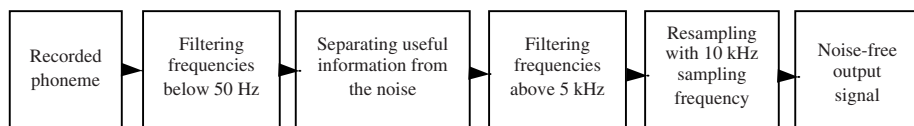


Fig. 2. The flow chart of speech signal preparation for recognition process.

A digital filter is applied to cut frequencies below 50 Hz. Almost all of the recorded speech signals contain this 50-Hz component. After filtering this frequency, the signal is free of such noise. The next step is to extract the valuable information from the rest of the signal. This is very important part of the signal pre-processing. The noise-free signal forms the input data to the algorithm of description and classification.

2.1 Algorithm

Now we can proceed with our algorithm whose input is the amplified signal with frequencies above 3 kHz [8]. This is applied because consonants generally consist of two parts. The first one, with the lower magnitude, consists of the most important

features, while the second one sounds like a vowel (the letter *b*, for example, is pronounced ‘*be*’ in Polish, and is analyzed as a consonant plus a vowel - *b+e*). Therefore, when amplifying higher frequencies we separate the first most important part of consonant signal. After this process we still have the signal sampled with 22 kHz frequency so that it consists of frequencies from 0 to 11 kHz, because sampling frequency, according to Nyquist theory, is at least twice higher than the highest frequency in the bandwidth.

Resampling. As we are trying to decrease the time of processing, and hence to speed the algorithm up, it is of great benefit then to reduce the number of samples and also number of computing operations. To realize that, in addition to the low-pass filtering of frequencies below 5 kHz, we resample the signal into one of less number of samples, without losing its essential-for-recognition characteristics.

In our case, after reducing the maximal frequency to 5 kHz, the sampling frequency must be at least 10 kHz in order to meet the requirements of Nyquist theory of sampling. As a result of this resampling process, the number of samples is reduced to less than half its original value (from 22000 to 10000 per second), and the further steps of the analysis will be run on minimal data. This causes the system speed to increase twice. Comparison of the signal with and without resampling is given in Fig. 3.

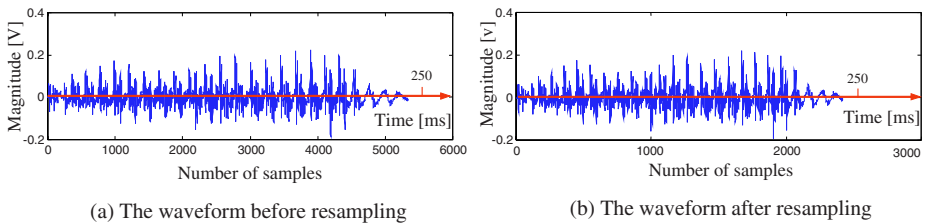


Fig. 3. Waveform resampling.

After resampling, the new signal is now prepared to be applied as an input to the next block, namely, Speech Signal Processing. This step is the most important one in the whole process of classification where the features of the acoustic images are extracted.

2.2 Speech Signal Processing

Among many methods of Speech Signal Processing, the authors have chosen the method based on spectrum analysis. This method allows speech image-feature extract from spectral analysis in a simple way. However, the processing is a difficult task because of the very complicated model of speech creation and perception. The advantage of spectral analysis processing is the possibility of analyzing the particular frequencies contained in the speech signal, which are articulation dependent and hence in some manner to allow identifying such components like phonemes. Following this, the authors are working on finding a method for smoothing irregular spectral shape resulting from applying FFT (Fast Fourier Transform). Experiments showed that

power spectrum estimation of Burg's model is the best for this purpose. It is based on the Linear Predictive Coding approach (Fig. 4). The theory of linear predictive coding is given in [9], where the sample $u(n)$ can be approximated as the linear combination of the P previous samples, with $n > 0$.

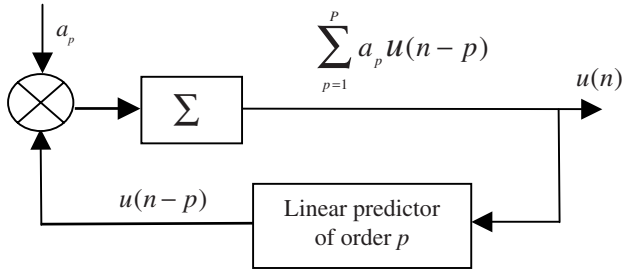


Fig. 4. The main idea of the prediction.

The n^{th} sample estimator $\tilde{u}(n)$ is defined to be:

$$\tilde{u}(n) = -\sum_{p=1}^P a_p u(n-p) \quad (1)$$

where a_p – prediction coefficients, $p = 1, 2, \dots, P$, and P – prediction order. The difference between $u(n)$ and $\tilde{u}(n)$ is called the prediction error $e(n)$. Hence,

$$e(n) = u(n) - \tilde{u}(n) = u(n) + \sum_{p=1}^P a_p u(n-p) \quad (2)$$

Burg's model together with the whole software implementation is included in [10]. This step is the starting point for further analysis. The analysis is based on minimizing forward and backward prediction errors according to Levinson – Durbin recursion [11,12]. For most of the phonemes, the spectral estimation furnishes a smooth envelope, while local extremes can be seen clearly (Fig. 5).

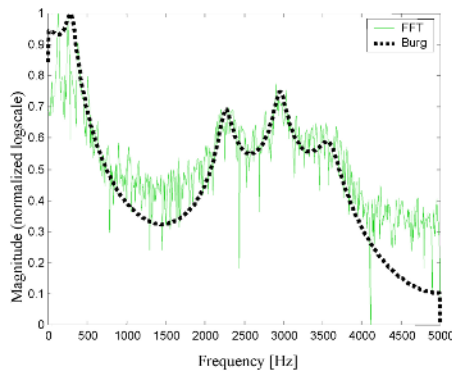


Fig. 5. Spectral analysis for standard FFT and Burg's method.

The obtained power spectrum acoustic images are then analyzed using the known methods of classification and recognition. In this work we use both classical point-to-point algorithms as well as the one based on minimal eigenvalues of Toeplitz matrices [3]. This algorithm has proved to be efficient in image feature extraction of scripts, texts [13,14] and a number of other various image-recognition processes [3,15]. Therefore, Toeplitz algorithm is applied to recognize spoken waveforms looked at as acoustic images. Of course, there are some factors which directly affect the accuracy of the pre-processing algorithms. For example, when applying the method of estimation, one needs to specify the prediction order and the FFT length. The FFT length must give the smoothest shape of the spectrum (the more samples we have, the smoother shape we get), and it cannot be a case when too many samples are considered. This, as very well known, would definitely lower the efficiency of the algorithm. Prediction order is also an important parameter. When it is too low, the envelope doesn't match with FFT shape, and when it's too high, it causes the speed of the algorithm to fall. So it's very important, although very difficult, to choose the best prediction order.

Again, when considering Toeplitz matrices and their minimal eigenvalues, it is essential to know how to select the characteristic points from the spectral estimation curve in order to apply as the feature data for the input to the classifying system. Figure 6 shows this procedure. In Fig. 6a out of 250 points only 25 are selected, that is every tenth point is considered. However, Fig. 6b shows larger number of samples as they are taken every fifth sample. This decreases the number of samples to only 50.

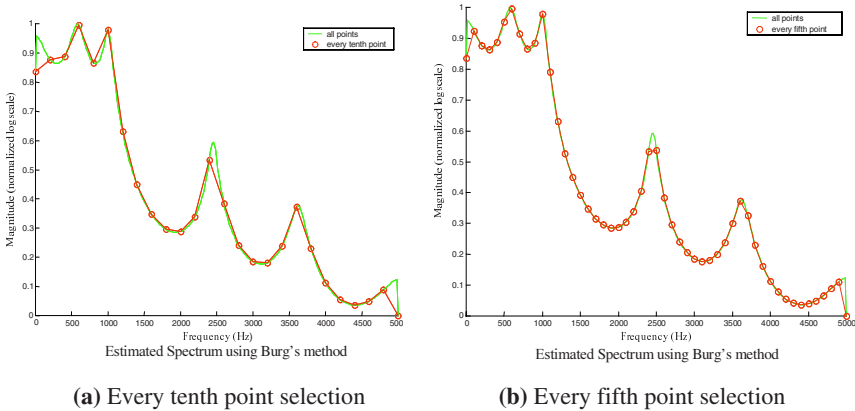


Fig. 6. Characteristic and feature point selection

As can be seen from this figure, the choice of every-tenth-point selection (Fig. 6a) leads to a distorted curve, while that of Fig. 6b does not cause any mentioned changes or distortion. This means that the last choice (Fig. 6b) is sufficient to assign and deliver the necessary minimized number of characteristic points as the input data to image classifying system based on Toeplitz approach of description. This would certainly decrease the computation size taking place in evaluating the necessary Toeplitz form determinants by at least 5 times.

The next step is to determine the Toeplitz matrices and calculate their corresponding determinants.

3 Classification

For the purpose of classification, the required comparison between the reference pattern Ψ and the resulting one from the input data Φ , is led by the Absolute Deviation classifier [3], defined by the summation of the absolute-values of the differences between Ψ_i and Φ_i elements for $i = 1, 2, \dots, P$, that is,

$$D = \sum_{i=1}^P |\Psi_i - \Phi_i| \quad (3)$$

The input data Φ_i to this equation is taken from the feature vector of the minimal eigenvalues extracted by the following criterion. To explain how the algorithm of minimal eigenvalues works, we first introduce the way of calculating Toeplitz-matrix determinants.

According to the method given in [3,15], the under-test object-features are described by the following rational function:

$$H(s) = \frac{P(s)}{Q(s)} \quad (4)$$

where $P(s)$ and $Q(s)$ are n -degree polynomials in the complex variable s whose coefficients are the coordinates of the feature points of Fig. 6. They are treated as pairs of complex numbers $s_i = x_i + jy_i$ with $i = 1, 2, 3, \dots, n$, n being the number of points considered. Therefore,

$$H(s) = \frac{x_0 + x_1s + x_2s^2 + \dots + x_ns^n}{y_0 + y_1s + y_2s^2 + \dots + y_ns^n} \quad (5)$$

Apply the bilinear transformation $s = \frac{1-z}{1+z}$ to have H in another simpler form $H(z)$.

To create Toeplitz matrices and their determinants, evaluate Taylor series for $H(z)$:

$$T(z) = c_0 + c_1z + c_2z^2 + \dots + c_nz^n + \dots \quad (6)$$

$$\text{where, } c_i = \frac{1}{x_0^{i+1}} \begin{vmatrix} y_i & x_1 & x_2 & \dots & x_i \\ y_{i-1} & x_0 & x_1 & \dots & x_{i-1} \\ y_{i-2} & 0 & x_0 & \dots & x_{i-2} \\ y_{i-3} & 0 & 0 & \dots & \dots \\ \dots & \dots & \dots & \dots & x_1 \\ y_0 & 0 & 0 & \dots & x_0 \end{vmatrix} \quad \text{for } i = 1, 2, \dots, n \quad (7)$$

These coefficients form the elements of Toeplitz forms:

$$D_i = \begin{vmatrix} c_0 & c_1 & c_2 & \dots & c_i \\ c_1 & c_0 & c_1 & \dots & c_{i-1} \\ c_2 & c_1 & c_0 & \dots & c_{i-2} \\ \dots & \dots & \dots & \dots & \dots \\ c_i & c_{i-1} & c_{i-2} & \dots & c_0 \end{vmatrix}, \quad i = 0, 1, 2, \dots, n \quad (8)$$

Determine the minimal eigenvalues of the forms in *Eq. (8)*:

$\lambda_{\min} \{D_i\} = \lambda_{\min_i} = \lambda_i$ for $i = 1, 2, \dots, n$. Hence, the following feature vector is found:

$$\Phi_i = (\lambda_0, \lambda_1, \lambda_2, \dots, \lambda_n) \tag{9}$$

Eq.(9), acts as the input data to the classification algorithms when applying the known methods of similarity and comparison for the sake of recognition. The characteristic behavior of this equation lies in the fact, that it forms a monotonically non-increasing series whose limit has something common with the minimal value of the rational function in *Eq.(5)* at $s = jy$. The behavior of this function and its derivatives is beyond the topics of this work. Again, the feature vector of *Eq.(9)* presents a very useful tool in describing an image within a class of similar objects. Simply, each has its own series of minimal eigenvalues Ψ_i descending to a definite limit of specific value differing from the series and their limits of other ones. The obtained results from the theory of both Burg's and Toeplitz according to the authors' approach, are given in the following section.

4 Recognition – Experiments and Result

A number of experiments and calculations have been performed for varieties of parameters and different possibilities of data introducing to *Eq. (4)* through the choice of the feature points and their data-defining variables $s_i = x_i + jy_i$, $i = 1, 2, 3, \dots n$.

Experiment 1. Here, a direct application to the results obtained from Burg's estimation spectrum is used for classification of characteristic points extracted from the under-test spoken-letter image. The recognition rate is high, and the number of samples is high, too. Table 1 shows the results of this experiment.

Table 1. Recognition results of Experiment 1.

| Burg's model prediction order: $P=12$; FFT length: 500; number of input samples: 1000 | | | | | | | | | | | | | | | | | | | | | | Overall Recognition | Recognition Rate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|------------------------|---------------------|
| Pattern set | a | b | c | d | e | f | g | h | i | k | l | m | n | o | p | r | s | t | u | w | y | z | |
| Number of recognized samples out of 5 iterations | 5 | 3 | 4 | 5 | 5 | 5 | 5 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 3 | 5 | 5 | 5 | 4 | 5 | 4 | | 102/110 |
| | | | | | | | | | | | | | | | | | | | | | | | 93% |

Experiment 2. This time and before applying the classification methods, apply the minimal eigenvalues algorithm of Toeplitz considering only the characteristic points extracted from the transfer function of *Eq.(4)*. The first results showed a high efficiency level (88%) for the 14 pronounced-in-Polish letters a, c, d, e, l, m, n, o, p, r, s, t, u, z, but a lower rate for the others, which reduced the overall recognition rate to 65% (see Table 2).

The recent work has shown higher recognition rate especially when the input data to *Eq.(5)* was modified leading to other more distinguishable series of minimal eigenvalues. The results were encouraging to extend the algorithm and test word groups by the Toeplitz-based algorithm. The first tested groups were the color names in two languages Polish and English. Our research group has been working on other spoken-word classes like days of the week, months of the year, and so forth.

Table 2. Recognition results of Experiment 2.

| Burg's model prediction order: $P=12$, FFT length: 50; number of input samples: 1000 | | | | | | | | | | | | | | | | | | | | | | | Overall Recognition | Recognition Rate | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---------------------|------------------|-----|
| Pattern set | a | b | c | d | e | f | g | h | i | k | l | m | n | o | p | r | s | t | u | v | w | y | z | | |
| Number of recognized samples out of 5 iterations | 4 | 2 | 4 | 5 | 5 | 1 | 0 | 0 | 5 | 2 | 2 | 5 | 5 | 5 | 3 | 5 | 3 | 5 | 0 | 2 | 3 | | | 71/110 | 65% |

Experiment 3. This experiment shows the results of recognition seven color-names in Polish and another seven-color group in English. The recognition rates are very high; they are **96%** for Polish group and **94%** for the English group (see Table 3 for the English group).

Table 3. Recognition results of English seven-color names group of Experiment 3.

| Burg's model prediction order: $P=12$; FFT length: 50; number of input samples: 1000 | | | | | | | | Overall Recognition | Recognition Rate |
|---|-------|--------|------|-----|--------|------|-------|---------------------|------------------|
| Pattern set | Green | Yellow | Pink | Red | Orange | Blue | White | | |
| Number of recognized samples out of 10 iterations | 10 | 6 | 10 | 10 | 10 | 10 | 10 | 66/70 | 94% |

5 Conclusions

The image-based approach in speech classification and processing has proved to be as good as other known methods in their recognition efficiency. In the case of using 22 letter-pattern classes, a recognition accuracy of above 93% was achieved. Experiments have shown that spectrum estimation parameters are very important for the recognition quality. Moreover, results improvement lies in choosing the suitable sampling window, for example Blackman's window [10,16] is adequate to our needs. Other methods are based on various modifications of the minimal eigenvalues algorithm based on Toeplitz matrices [3,13]. This algorithm has given very good results in case of scripts, texts and two-dimensional object images [3,13,15]. In the case of speech signals and their spectral images, this kind of image is more complicated. Speech images, no matter what form they are described in, they always represent a complex of a lot of various elements and they need special and specific ways of processing.

Researches, focused on developing methods based on the minimal eigenvalues algorithm are still being working on. The accuracy of these methods is mainly affected by the minimization of the number of characteristic points. Another important factor that may lower the recognition accuracy, is the fact, that power spectrum is too similar for different phonemes. This is why more efforts are employed to apply Toeplitz approach directly on recorded voice just after filtering and noise removing so that the recognition is processed in the real time. The successful application of the Toeplitz-based algorithm to color-names classes in two languages, Polish (96% recognition rate) and English (94%), has proved that the presented work is promising. Experiments are being done on Arabic classes, as well. Some recent results [17] seem to be helpful in the pre-processing steps, which will certainly improve the level of accuracy and give much better results. Moreover, the results achieved in [18] are being used for

better feature extract and improved voice-signal image processing. The near future publications will show more interesting results, hopefully of higher recognition rate and wider applications.

Acknowledgement

This work was supported by the Rector of Białystok University of Technology (grant number W/II/3/01).

References

1. D. J. Burr, "Experiments on Neural Net Recognition of Spoken and Written Text," IEEE Transactions on Acoustic, Speech, and Signal Processing, vol. 36, July 1988.
2. J. L. MacDonald, W. Zucchini, W. Zucchi, "Hidden Markov and Other Models for Discrete-Valued Time Series," CRC Press, Jan. 1, 1997.
3. K. Saeed, "Computer Graphics Analysis: A Criterion for Image Feature Extraction and Recognition," Vol. 10, Issue 2, 2001, pp. 185-194, MGCV - International Journal on Machine Graphics and Vision, Institute of Computer Science, Polish Academy of Sciences, Warsaw.
4. R. W. Schafer, L. R. Rabiner, "System for Automatic Formant Analysis of Voiced Speech," J. Acoust. Soc. Amer. Vol.47, Feb. 1970.
5. L. Grad, "Obrazowa reprezentacja sygnału mowy," Biuletyn IAI R WAT, nr 11, Warsaw 2000.
6. Cz. Basztura, "Modele analizy i procedury w komputerowym rozpoznawaniu głosów," Prace naukowe ITiA Politechniki Wrocławskiej, nr 30, Wrocław 1989.
7. L. S. Marple, "Digital Spectral Analysis," Englewood Cliffs, NJ: Prentice Hall, 1987.
8. K. Saeed, M. Kozłowski, A. Kaczanowski, "Metoda do rozpoznawania obrazów akustycznych izolowanych liter mowy," Zeszyty Politechniki Białostockiej (in Polish), 1-1/2002, pp.181-207, Białystok 2002.
9. R. Tadeusiewicz, "Sygnał mowy," WKiŁ (in Polish), Warsaw 1988.
10. V. K. Ingle, J. G. Proakis, "Digital Signal Processing Using MATLAB," Brooks Cole, July 1999.
11. N. Levinson, "The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction," Journal Math. Phys. Vol. 25, 1947.
12. J. Durbin, "Efficient Estimation of Parameters in Moving Average Models," Biometrics, Vol. 46, part 1, 2, 1969.
13. Khalid Saeed, "Experimental Algorithm for Testing The Realization of Transfer Functions," Proceedings of the Fourteenth IASTED International Conference, Austria 1995.
14. R. Niedzielski, "Kryterium do rozpoznawania znaków maszynowych alfabetu łukowego," MSc Thesis, Ins. Informatyki PB, Białystok 1999.
15. K. Saeed, A. Dardzinska, "Language Processing: Word Recognition without Segmentation," *JASIST - Journal of the American Society for Information Science and Technology*, Volume 52, Issue 14, 2001, pp. 1275-1279, John Wiley and Sons.
16. R. G. Lyons, "Wprowadzenie do cyfrowego przetwarzania sygnałów," WKiŁ (in Polish), Warsaw 1999.
17. Sadaoki Furui, "Digital Speech Processing, Synthesis, and Recognition," Marcel Dekker, Inc. 2001.
18. K. Saeed, M. Rybnik, M. Tabędzki, "More Results and Applications about the Algorithm of Thinning Images to One-Pixel-width," 9th CAIP Int. Conference on Computer Analysis of Images and Patterns, Sept. 5-7, 2001, pp. 601-609, Springer-Verlag, Warsaw 2001.