Automatic Syllable Repetition Detection in Continuous Speech Based on Linear Prediction Coefficients

Adam Kobus, Wiesława Kuniszyk-Jóźkowiak and Ireneusz Codello

Abstract The goal of this paper is to present a syllable repetition detection method based on linear prediction coefficients obtained by the Levinson–Durbin method. The algorithm wrought by the authors of this paper is based on the linear prediction spectrum. At first the utterance is automatically split into continuous fragments that correspond with syllables. Next, for each of them the formant maps are being obtained. After dimension reduction by the K-means method they are being compared. The algorithm was verified based on 56 continuous utterances of 14 stutterers. They contain fluent parts, as well as syllable repetitions on Polish phonemes. The classifying success reached 90 % of sensitivity with 75–80% precision.

1 Introduction

The occurrence and similarity of stuttering can be observed in many languages. This disorder can be divided into several types of dysfluencies: prolongations, blocks, interjections, repetitions of syllables or even parts of words [15]. This paper deals with the syllable and fragment repetitions in continuous speech. The analysis of the possibility of automatic dysfluency detection in continuous speech is a very important topic. Stuttering people have to deal with not only the interpersonal communication problems, but also with the more and more spreading voice actuation. Additionally, this analysis increases the knowledge of the nature of stuttering and helps to prepare even better stuttering therapies.

A. Kobus (✉) · I. Codello
Institute of Computer Science, Marie Curie-Skłodowska University,
Pl. M. Curie-Skłodowskiej 1, 20-031 Lublin, Poland
e-mail: adam.kobus@poczta.umcs.lublin.pl; kobus.adam@gmail.com

W. Kuniszyk-Jóźkowiak
Faculty of Physical Education and Sport in Biała Podlaska, Józef Piłsudski University
of Physical Education in Warsaw, ul. Akademicka 2, 21-500 Biała Podlaska, Poland

R. Burduk et al. (eds.), *Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015*, Advances in Intelligent Systems and Computing 403, DOI 10.1007/978-3-319-26227-7_28

1.1 Related Work

In the research on the automatic dysfluency detection, the initial listening selection of speech fragments into fluent and dysfluent parts is made [1, 5, 7, 8, 16, 19–21, 24–29]. This research aims at detecting dysfluent fragments continuous utterances. First dysfluency detection trials without the initial dysfluent words extraction were undertaken in the works of Howell [9–11], afterwards in the research of Suszyński [23], Wiśniewski [30, 31], Codello [2–4] and Kobus [13, 15]. Those works have as their input the four-second-long recordings initially classified as fluent or dysfluent. The comparison of differences between the results of various analyses of the fluent and dysfluent samples allows to notice the significant disparities between these two types of samples. Thanks to this fact, the prolongation and phoneme repetitions' detection gives good results. The syllable repetition detection is a more complex issue. Still, it was addressed in the Suszyński's research [25], based on the worse of Hiroshima [7] and Codello [4]. For this aim Suszyński performed a comparison of spectrograms. He applied the correlation of the one-third octave spectrums in the connection with the procedures for the time–amplitude file structure analysis to detect the repeating fragments. The classification was made by exceeding the obtained correlation coefficient of the two fragments. This analysis allows to detect and localise syllable repetitions with 70 % efficiency. Codello's research [4] was based on the application of CWT in the Bark scale. The results were split into vectors and compared by the correlation method. This analysis reached 80 % of efficiency. For the speech dysfluency analysis the authors programmed the "Dabar" application. Beside the implementation of the prolongation [15] and block [13] detection algorithms the possibility of syllable and word fragment detection was also analysed. Amongst the many functionalities of the application are the following: dimension reduction by the Kohonen networks, evaluation of linear prediction and PARCOR coefficients, the spectrum evaluation on the basis of LPC and the creation of the elliptic model of the vocal tract [14]. Previous works allow to ascertain that the representation of the speech signal by means of linear prediction coefficients serves perfectly for detecting plosive repetition [13] and phoneme prolongation [15]. For this research, the splitting of utterances into syllables and/or speech fragments was implemented, together with the K-means centres obtaining method.

2 Methodology

The goal of this research is the analysis of the possibility of automatic syllable and speech parts' repetition detection using linear prediction method. In each utterance, the places where the repetition of syllable or speech fragment occurs were marked.

2.1 Speech Data

For the analysis, the several seconds long utterances were used. All 56 files of the total length of 3 min 47 s are in *WAV* format. Fourteen speakers were recorded—nine men and five women. Men were 11–25 years old (30 recordings), women were 13–24 years old (26 recordings). All recordings contain 262 pairs of syllables, 117 of which were dysfluent. Materials were recorded with Creative Wave Studio using the SoundBlaster card with 22050 Hz frequency and 16 bits per sample.

2.2 Preprocessing

The independent recordings were split into non-overlapping frames with 512 samples. Each of the frames was multiplied by the Hanna window function (1). Based on the previous research, it allows to achieve frequency spectrums with the best characteristic of those obtained by the linear prediction method [13].

$$w(n) = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right), \tag{1}$$

where $N = 512$ is the frame size.

2.3 Feature Extraction

2.3.1 Linear Prediction

The most basic characteristic of the linear prediction method is that it stores the knowledge of the speech signal change in a vector of a few coefficients. Consecutive speech signal samples vary to a small degree [18], thus the approximation of the next sample can be evaluated from the p previous samples using the proper linear prediction coefficients α. This idea is expressed by the following equation:

$$\tilde{s}(m) = \sum_{k=1}^{p} \alpha_k s(m-k) \tag{2}$$

where $\tilde{s}(m)$ is the mth value of the predicted voice sample, $s(m)$ is the mth value of the input speech sample, p is a prediction order and α_k are the obtained linear prediction coefficients. Basing on the linear prediction coefficients, the continuous frequency spectrum with an arbitrary number of stripes may be evaluated. The evaluation is expressed by Eq. (3) derived from the definition of the LP (Eq. 2):

$$H(f_i) = \frac{G}{1 - \sum_{k=1}^{p} \alpha_k \left(\cos\left(k\frac{f_i}{f}2\pi\right) - i\sin\left(k\frac{f_i}{f}2\pi\right)\right)} \tag{3}$$

where f is the chosen number of the spectrum stripes, f_i is the number of the chosen spectrum stripe, where $0 \leq f_i \leq f$, and G is the gain factor.

2.3.2 Analysis

In this algorithm the linear prediction coefficients were evaluated using the Levinson–Durbin method [18] with G gains from the obtained frames. The number of coefficients for each of them was 15, which is sufficient for good signal characteristics [22]. Next, the frequency spectrum was evaluated from Eq. (3) for each of the frames. In the order to obtain a precise spectrum, the number of stripes was set to 300. The evaluated coefficient vectors were used as the input vectors in the method of splitting speech into fragments Sect. 2.4. They were also the basis of the values parameterising the speech for further analysis Sect. 2.6.

2.4 Segmentation

The algorithm of splitting the speech into fragments was implemented for the purpose of this research. The algorithm is based on the frequency spectrum obtained from the linear prediction coefficients. From the obtained spectrums, the average value of the powers of all the m spectrum stripes values was evaluated for each frame. On the basis of the obtained values, the threshold was defined. It was evaluated as the average of the two primary values from the vector increased by the 3 db.

2.5 Pairing

Each fragment of the examined utterance was compared with the succeeding fragment. If the succeeding fragment was longer than the preceding one, it was cut into the length of the first fragment. This is due to the fact that the succeeding fragment may also contain the further, fluent part of the utterance and the repeated fragment is in its initial phase.

2.6 Formants Extraction

For each of the frames the formants are defined [6, 17, 22]. They are determined by those maxima F from the local maximum points of the frequency spectrum which fulfil the bandwidth condition [12, 22]:

$$B_j = -(f_j/\pi)\ln(r_0) < B \tag{4}$$

where B_j is the 3 db bandwidth, j is the ordinal number of the maximum from the chosen frame, f_j is the frequency of the jth spectrum maximum and r_0 are the roots of the polynomial $A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k}$, where $z = e^{i2\pi f_j/f_s}$, f_s is the sampling frequency. B is the maximum bandwidth permissible for the formant. For the purpose of this analysis $B = 1000\,\text{Hz}$. The points have three coordinates: frequency, amplitude and time. For each frame several formants were obtained. As a result of that, a fragment is represented by a set of P points. Two of those dimensions, the frequency and the amplitude, are the basis for the K-means algorithm. The main goal of this method is to reduce m characteristic points from the P set to k characteristic centres for the whole fragment.

2.7 *Clusterisation*

The main rule of the K-means algorithm is to gradually adjust the midpoints (centres) so that they would divide the space in the best possible way into groups of points concentrated around each other.

2.7.1 The Normalisation of Space of Characteristic Points

The normalisation of space of characteristic points consists in scaling each of the dimensions with respect to the average value and the variance of the value of a set of points, that is, on the basis of the Mahalanobis distance:

$$\overline{p_j} = \frac{p_j - \mu}{\sigma} \tag{5}$$

where $\overline{p_j}$ is a normalised point in the R^N space, p_j is a point in the R^N space, μ is a point of average values for p_j points and δ is a vector of variances for these points.

2.7.2 Input Data

The P set of the normalised points p_j is given, where $0 < j < m, m$ is the number of the P set elements, as well as the vector $\mathbf{k}$ of the normalised initial centres c_i, where $0 < i < k$ and k is the length of the $\mathbf{k}$ vector. Both the set of points and the vector of the centres should be normalised. In order to measure the distance between two points in N-dimensional space the space normalisation is needed for the comparison of distances. The vector $\mathbf{k}$ of the centres should be also initialised by the values of the coordinates.

2.7.3 Centres Initialisation

The random initialisation of the centres is often applied; however, thanks to the knowledge of the structure of the set of points, the initialisation based on this knowledge may be applied. The uniform distribution method was used for the purpose of this research. The basic data for the initialisation are points with two dimensions: amplitude and frequency. The input points are sorted in the order of frequency. Next, k_p points distributed uniformly in frequency are chosen. If there are only few of such points, these points are being duplicated up to the k number so that they would not affect the result of measurement of the distances between the vectors.

2.8 Distance Measuring

For the analysis of the distances between the vectors $\mathbf{k}$ and consequently between the points, a few distance measures were applied.

2.8.1 Metrics

For measuring the distance between the points two metrics were applied.

Euclidean Metric

This metric was applied in the measurement of the distance between points in the K-means algorithm, as well as in the classification for measurement of the distance between the tested pair and the threshold between the fluent and dysfluent pairs (**MEDIAN**, **AVERAGE**).

Mahalanobis Metric

This metric was applied in the classification for examining the distance between the tested pair and the group of fluent and dysfluent pairs (**MAHALANOBIS**).

2.8.2 Distance Between Vectors

After the K-means analysis, each fragment is represented by a vector of centres. They are combined into pairs and the distance between them in two dimensions is being evaluated. The evaluation consists in obtaining the sum of minimal Euclidean distances between the centres of the succeeding and preceding fragment.

2.9 Classification

The obtained distances were used for the classification of 262 pairs of fragments into dysfluent (fragment repetitions) and fluent ones. The previously classified pairs were randomly divided into three parts. Two of them were used as a training part (172 pairs) and one as a testing part (90 pairs). Three methods were used for classification:

AVERAGE the minimal distance from the threshold in the Euclidean metric—the threshold was obtained experimentally as $n + (f - n)/5$, where n is an arithmetic average of the distances between contiguous recurring fragments and f is an arithmetic average of the distances between contiguous nonrecurring fragments (fluent speech).

MEDIAN analogous to the above mentioned, but the average distance was substituted with a median.

MAHALANOBIS the minimal distance from the group in a Mahalanobis metric.

For the evaluation of the average and the median, the training part was used.

3 Results and Discussion

The results refer to the testing part of extracted pairs. The result figures contain a number of parameters related to the quality assessment of the classification.

1. Sensitivity—$100 * TP/(TP + FN)$—where TP—true positive—is a number of correctly classified dysfluencies, FN—false negative—a number of not recognised dysfluencies
2. Precision—$100 * TP/(TP + FP)$—where TP is a number of non-fluent pairs with recognised non-fluency, FP—false positive—a number of fluent pairs classified as non-fluent
3. K—a number of the centres in a representation vector
4. Categorisation method Sect. 2.9: AVERAGE, MEDIAN, MAHALANOBIS (Figs. 1, 2 and 3).

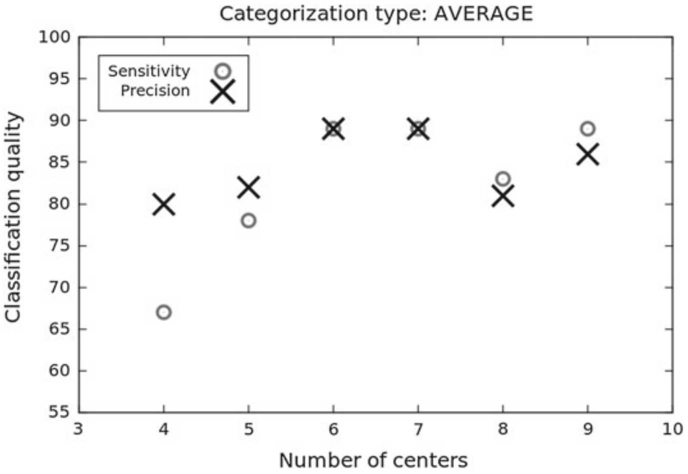

Fig. 1 Classification quality percentage graph for the best precision and sensitivity in relation to the number of centres for the AVERAGE type of categorisation

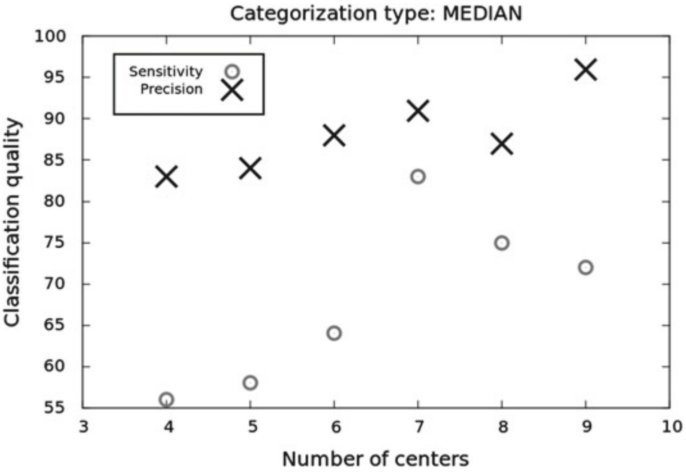

Fig. 2 Classification quality percentage graph for the best precision and sensitivity in relation to the number of centres for the MEDIAN type of categorisation

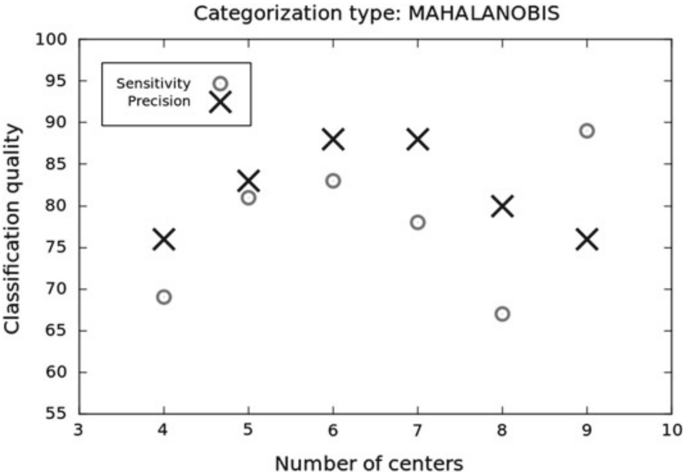

Fig. 3 Classification quality percentage graph for the best precision and sensitivity in relation to the number of centres for the MAHALANOBIS type of categorisation

4 Conclusion

The results of the research proved the effectiveness of the syllable and fragment repetition detection using the proposed method. Regarding the number K of the centres, the best results were reached for $5 \leq K \leq 7$. The results for $8 \leq K \leq 9$ were good and for $K = 4$, a little bit worse. It may result from the number of formants which represent the centres. If K is too small, some of the formants may not be duly represented, e.g. in the case of two or three phonemes. If K is too big, the representation is good but the excess of points has an influence on the distortion of the results for the lower number of the phonemes. Fortunately, the differences were not significant. The second aspect which was examined was categorisation. It was observed that regardless of its type, over 80 % of sensitivity was achieved. The best results were obtained using the AVERAGE categorisation. The MAHALANOBIS and MEDIAN categorisations were a little bit worse but similar to each other. The results of the analysis lead to the conclusion that detecting speech fragment repetitions, using the K-means method as a dimension reducer, the distance analysis as a classifier and the linear prediction coefficients as the input values, is possible and reaches 89 % precision with 89 % sensitivity. The applied method of automatic speech segmentation into syllables and fragments allows to determine the places where the syllable, phoneme or fragment repetition occurs with an 89 % efficiency. The algorithm described in this paper, equipped with a database of recordings with predefined dysfluencies' occurrences, may also be applied in a real-time detection of repetitions.

References

1. Chia Ai, O., Hariharan, M., Yaacob, S., Chee, L.S.: Classification of speech dysfluencies with MFCC and LPCC features. Expert Syst. Appl. **39**, 2157–2165 (2012)
2. Codello, I., Kuniszyk-Jóźkowiak, W., Smołka, E., Kobus, A.: Disordered sound repetition recognition in continuous speech using CWT and Kohonen network. J. Med. Inform. Technol. **17**, 123–130 (2011)
3. Codello, I., Kuniszyk-Jóźkowiak, W., Smołka, E., Kobus, A.: Automatic prolongation recognition in disordered speech using CWT and Kohonen network. J. Med. Inform. Technol. **20**, 137–144 (2012)
4. Codello, I., Kuniszyk-Jóźkowiak, W., Smołka, E., Kobus A.: Automatic disordered syllables repetition recognition in continuous speech using CWT and correlation. In: Proceedings of the 8th International Conference on Computer RecognitionSystems CORES 2013, Advances in Intelligent Systems and Computing, vol. 226, pp 865–874 (2013)
5. Czyżewski, A., Kaczmarek, A., Kostek, B.: Intelligent processing of stuttered speech. J. Intell. Inf. Syst. **21**(2), 143–171 (2003)
6. Halberstam, B., Raphael, L.J.: Vowel normalization: the role of fundamental frequency and upper formants. J. Phon. **32**, 423–434 (2004)
7. Hiroshima, S.M.: A spectrographic analysis of speech disfluencies: characteristics of sound/syllable repetitions in stutterers and nonstutterers, In: Proceedings 24th IALP Congress Amsterdam, pp. 712–714, Amsterdam (1999)
8. Geetha, Y.V., Pratibha, K., Ashok, R., Ravindra, S.K.: Classification of childhood disfluencies using neural networks. J. Fluen. Disord. **25**(2), 99–117 (2000)
9. Howell, P., Sackin, S.: Automatic recognition of repetitions and prolongations in stuttered speech. In: Proceedings of the First World Congress on Fluency Disorders, pp. 1–4 (1995)
10. Howell, P., Sackin, S., Glenn, K.: Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: I. Psychometric procedures appropriate for selection of training material for lexical dysfluency classifiers. J. Speech, Lang. Hear. Res. **40**(5), 1073–1084 (1997)
11. Howell, P., Sackin, S., Glenn, K.: Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: II. ANN recognition of

repetitions and prolongations with supplied word segment markers. J. Speech, Lang. Hear. Res. **40**(5), 1085–1096 (1997)

12. Kim, C., Seo, K., Sung, W.: A robust formant extraction algorithm combining spectral peak picking and root polishing. EURASIP J. Appl. Signal Process. **2006**, 1–16 (2006)
13. Kobus, A., Kuniszyk-Jóźkowiak, W., Smołka, E., Codello, I.: Speech nonfluency detection and classification based on linear prediction coefficients and neural networks. J. Med. Inform. Technol. **15**, 135–144 (2010)
14. Kobus, A., Kuniszyk-Jóźkowiak, W., Smołka, E., Suszyński, W., Codello, I.: A new elliptical model of the vocal tract. J. Med. Inform. Technol. **17**, 131–140 (2011)
15. Kobus, A., Kuniszyk-Jóźkowiak W., Smołka, E., Codello, I., Suszyński W.: The prolongation-type speech non-fluency detection based on the linear prediction coefficients and the neural networks. In: Proceedings of the 8th InternationalConference on Computer Recognition Systems CORES 2013, Advances inIntelligent Systems and Computing, vol. 226, pp. 885–894 (2013)
16. Kuniszyk-Jóźkowiak, W., Suszyński, W., Smołka, E., Dzieńkowski, M.: Automatic recognition and measurement of durations of fricative prolongations in the speech of persons who stutter. Speech Lang. Technol. **8** (2004). (in polish)
17. Millhouse, T., Clermont, F., Davis, P.: Exploring the importance of formant bandwidths in the production of the singer's formant. In: Proceedings of the 9th Australian International Conference on Speech Science & Technology, Melbourne, pp. 373–378 (2002)
18. Rabiner, L.R., Schafer, R.W.: Theory and Applications of Digital Speech Processing Chap. 9. Pearson Higher Education Inc (2011)
19. Ravikumar, K., Reddy, B., Rajagopal, R., Nagaraj, H.: Automatic detection of syllable repetition in read speech for objective assessment of stuttered disfluencies. In: Proceedings of world academy science, engineering and technology, pp. 270–273 (2008)
20. Ravikumar, K.M., Rajagopal, R., Nagaraj, H.C.: An approach for objective assessment of stuttered speech using MFCC features. ICGST Int. J. Digit. Signal Process., DSP **9**(1), 19–24 (2009)
21. Ravikumar, K.M., Ganesan, S.: Comparison of multidimensional MFCC feature vectors for objective assessment of stuttered disfluencies. Int. J. Adv. Netw. Appl. **02**(05), 854–860 (2011)
22. Snell, R.C., Milinazzo, F.: Formant location from LPC analysis data. IEEE Trans. Speech Audio Process. **1**(2), 129–134 (1993)
23. Suszyński, W.: Automatic detection of speech non-fluencies. In: 50th Opened Acoustic Seminar, pp. 386–390 (2003). (in polish)
24. Suszyński, W., Kuniszyk-Jóźkowiak, W., Smołka, E., Dzieńkowski, M.: Automatic recognition of nasals prolongations in the speech of persons who stutter. Structures—Waves—Human Health, pp. 175–184 (2003)
25. Suszyński, W.: Computer analysis and speech dyspluency recognition. Doctoral dissertation, Politechnika Śląska, Gliwice (2005). (in polish)
26. Szczurowska, I., Kuniszyk-Jóźkowiak, W., Smołka, E.: The application of Kohonen and multilayer perceptron networks in the speech nonfluency analysis. Arch. Acoust. **31**(4), 205–210 (2006)
27. Szczurowska, I., Kuniszyk-Jóźkowiak, W., Smołka, E.: Speech nonfluency detection using Kohonen networks. Neural Comput. Appl. **18**, 677–687 (2009)
28. Świetlicka, I., Kuniszyk-Jóźkowiak, W., Smołka, E.: Artificial neural networks in the disabled speech analysis. Computer Recognition Systems 3, vol. 57/2009, Springer, Heidelberg, pp. 347–354 (2009)
29. Tian-Swee, T., Helbin, L., Ariff, A.K., Chee-Ming, T., Salleh, S.H.: Application of malay speech technology in malay speech therapy assistance tools. In: International Conference on Intelligent and Advanced Systems, pp. 330–334 (2007)
30. Wiśniewski, M., Kuniszyk-Jóźkowiak, W., Smołka, E., Suszyński, W.: Automatic detection of disorders in a continuous speech with the hidden markov models approach computer recognition systems 2, Vol. 45/2008, pp. 445–453. Springer, Berlin (2007)
31. Wiśniewski, M., Kuniszyk-Jóźkowiak, W.: Automatic detection and classification of phoneme repetitions using HTK toolkit. J. Med. Inform. Technol. **17**, 141–148 (2011)