TELEINFORMATICS REVIEW
No. 3-4, 2019

Experimental research on the impact of similarity function selection on the quality of keyword spotting in speech signal

Łukasz LASZKO

Institute of Teleinformatics and Cybersecurity, Faculty of Cybernetics, MUT
ul. gen. Sylwestra Kaliskiego 2, 00-908 Warsaw, Poland
lukasz.laszko@wat.edu.pl

ABSTRACT: The paper describes an evaluation of the application of selected similarity functions in the task of keyword spotting. Experiments were carried out in the Polish language. The research results can be used to improve already existing keyword spotting methods, or to develop new ones.

KEYWORDS: keyword spotting, signal similarity, quality of detection, dynamic time warping, textual query

1. Introduction

The task of keyword spotting (KWS) consists of query-by-example[1] in the registered spontaneous speech signal. The purpose of the task is achieved by indicating the points in the speech signal where the given word occurs. These indications should usually minimise the probability of false peace and false alarm [22].

The task of KWS is part of the field known as *information retrieval* [50][2]. In this field, it is defined as follows:

a) a speech signal, which is by definition generated by different speakers,

b) the searched word that is set in text form,

[1] The following terms are also used: *keyword or key-word spotting*, *key-phrase detection* [74] or *spoken term detection* [59].

[2] Specifically in the field of sound KWS is sometimes considered part of *Audio IR* [15], *Multimedia IR* [63], [56]. Yet another view is presented in [29].

c) the reference signal which is obtained by converting text-to-speech by using recordings of natural speakers or by speech synthesisers,

d) pattern search in the speech signal which is based on comparing the tested signal with the reference signal,

e) the comparison that applies to signals, not text (string of phonetic symbols).

One of the essential problems to solve is determining the similarity between the models of two signals: utterance and reference signal (the so-called query) [17]. An analysis of publications from the last twenty years has allowed the author to observe that usually this similarity is established in the metric space of speech signal features R^N. The features applied are acoustic coefficients such as mel-frequency cepstral coefficients (MFCC). The assessment of similarity between signal models is based on the distance between them in R^N, with the shorter distance meaning greater similarity. The most commonly applied metric in KWS tasks is the cosine metric [28], [77], [68].

The choice of metric is usually arbitrary and not discussed in publications by researchers. As noted in [17], this may be caused by the properties of the metric itself. However, significant differences in interpretation occur for Euclidean and cosine metrics, for example. This has had an impact on the direction of research described herein.

The purpose of the author's research was to determine the impact of similarity function selection on the quality of keyword spotting in speech signal. This article describes the results of comparative research obtained by the author for using the keyword spotting method introduced in paper [42]. The research was conducted for the Polish language analogously to the research reported in [44], using the same corpus of Polish speech [35].

2. Similarity of words in a speech signal

2.1. Similarity assessment methods

The following approaches can be distinguished for setting the similarity of two speech signals [64], [27][3]:

[33] Own study on [64] pp. 190-193, [27] pp. 22-37. Other classification of approaches is show in [1], for example.

Categorical (ontological) similarity – making an assessment based on a classification according to known conceptual categories (e.g. voiced sound).

Similarity of attributes – where analysed words have identical or similar features (properties), and the numerical values of the features show slight differences (i.e. are similar), such as formant frequencies.

Similarity of relations – where there are identical or similar relations, such as proportions, between the analysed words.

Similarity of causal (semantic) relations – where the analysed words have the same (similar) contexts, e.g. given words define the same subject in a sentence.

In the case of keyword spotting tasks, similarity is usually set according to speech signal attributes (i.e. similarity of attributes). Such attributes (speech signal features) are most often acoustic coefficients, such as: MFCC [55], human-factor cepstral coefficients (HFCC) [74], relative spectral-perceptual linear prediction (RASTA-PLP) [71], [32], [19] and others, referred to in paper [53], for example. The issue of selecting the similarity function may depend on the adopted features that represent the compared signals.

2.2. Similarity assessment

The solution of KWS task can be approached in two ways: using speech recognition methods [72] or speech processing methods [59].

Through the use of speech recognition methods, proper keyword spotting is done in the sphere of text (string of phonetic symbols) obtained by analysing words from the recording. Determining the similarity of words then comes down to calculating the distance between the strings of symbols, based on the Levenshtein distance, for example, as in paper [79]. In this case, the word with the lowest Levenshtein distance from the textual query is indicated.

Other measures are used instead of the Levenshtein distance, such as:

- Damerau-Levenshtein distance[4],
- Jaro-Winkler distance [70],
- Hamming distance [75] and
- LCS (longest common subsequence) [42].

When speech processing methods are used, keyword spotting is done in the sphere of signal. The speech signal for the given textual query is obtained through *text-to-speech* synthesis. The resulting signal sample vector is converted into a feature vector. Further, depending on the signal model, there are the following approaches to assessing word similarity:

1) If the signal representation is a single vector (e.g. MFCC), the similarity assessment is based on:
    a) distances between vectors, typically a cosine distance, although other distances are also used, such as:
        - Euclidean distance [34], [25],
        - cosine-Euclidean distance [22],
        - log-cosine distance [18],
        - Manhattan distance [20],
        - sigma distance[18],

    b) correlation coefficient (with zero meaning no similarity); typically this is the Pearson correlation, although Kendall or Spearman correlations are also used[4] [33], [48], [39].
2) If the signal model is a group (cluster) of vectors (such as a set of frame group features), inferring about the similarity of two signals requires defining similarity between clusters. The similarity assessment is based on the distance between clusters, while the *distance* understood in this way does not usually meet the metrics axiom[5]. The following approaches are used in this case:
    a) setting the distance based on cluster elements (e.g. between central elements of clusters), for which Euclidean distance or other distances based on Minkowski distance [67] are applied,
    b) setting the distribution of elements in the cluster based on the distance, including a probabilistic model, for which the Kullback-Leibler distance is often used [26], [30], even though others are also used, such as:
        - Bhattacharyya distance[1], [16], [5], [3], [31],
        - Mahalanobis distance [3], [38],
        - Hellinger distance [45], [23], [31], [58] and
        - divergences: *f-divergence*, Jensen–Shannon divergence, etc. [57], [62].

This article describes research in relation to the latter approach, i.e. speech processing methods are used to solve the KWS task (cf. [42]), and the research task is to choose the similarity function.

[4] Also known as rank correlation. Ranks are the numbers of subsequent observations in the ordered statistical sample.

[5] Cf. e.g. [78] p. 39.

2.3. Similarity function assessment

In Table 1 there is a list of similarity functions used in the described research. The similarity function is one of the important components of the methods applied in KWS tasks and has a direct impact on the quality of keyword spotting. It is therefore appropriate to apply the similarity function which results in the highest quality results when used in a particular method.

Tab. 1. List of similarity functions tested

No.	**Basis for defining the similarity function[6]:**
1	Bhattacharyya distance (K_{bha})
2	Chebyshev distance (K_{che})
3	correlation-based distance (K_{cor})
4	cosine distance (K_{cos})
5	Euclidean distance (K_{euc})
6	Hellinger distance (K_{hel})
7	symmetrical Kullback–Leibler distance(K_{skl})
8	Manhattan distance (K_{man})
9	Mahalanobis distance (K_{mah})
10	Minkowski distance (K_{min})
11	standardized Euclidean distance (K_{seu})
12	Spearman distance (K_{spr})

2.3.1. Indicators of keyword spotting quality in KWS tasks

The quality of spotting can be measured using basic indicators directly related to the number of results achieved [61]. These include:

- TP (true positive) – number of correct indications (hits),
- TN (true negative) – number of correct rejections,
- FP (false positive) – number of incorrect indications (Type I errors – ‘false alarms’),
- FN (false negative) – number of incorrect rejections (Type II errors, misses – ‘false peace’),

[6] The similarity function symbol is put in brackets.

These indicators are often set into an error/confusion table/matrix [69][7]. Precision of indications and other indicators that allow for referencing the results obtained (e.g. to compare two methods) are also important in KWS tasks. These include derived indicators. The following indicators were selected for the research:

- **precision**, marked PPV,
- **accuracy**, marked ACC,
- **recall**, **true positive rate**, marked TPR,
- **specificity**, **true negative rate**, marked TNR,
- **F-measure**, **F_1Score**, marked F_1S) [9], [65] and
- **Youden's J statistic**, marked YJS [73].

Based on the PPV it can be assessed whether a given method (using a given similarity function) gives repeatable results, characterized by a small spread. The ACC value makes it possible to assess whether a given method always gives results close to true (real) results. The TPR indicates the ability of the method to correctly detect (indicate a result) where the value sought actually exists. On the other hand, the TNR specifies the ability of the method to correctly reject results (the so-called selectivity). F_1Score is used to assess the method *reliability*, i.e. a feature demonstrating the authenticity of the results obtained (both indications and rejections). However, the YJS is used to assess the method effectiveness[8] and to select the best method parameters in the ROC analysis (cf. Chapter 5.1).

2.3.2. Vector assessment scalarization

It is assumed in the paper that the vector assessment of the similarity function will be made using six derivative indicators listed above. It should be noted that the indicators described above have the same range of values. It is a number range [0,1], where an indicator value of *one* characterises a good method (which is the most precise, most accurate, etc.).

A scalar assessment was made by adding the best results of each quality indicator to arrange the vector assessments in order and at the same time select the best function. The above assumptions result from the author's observation that these results strictly depend on the experiment conditions. In particular, in

[7] Based on: https://en.wikipedia.org/wiki/Confusion_matrix (visited: 19.08.2019).

[8] The method effectiveness, shown by the YJS, indicates its sensitivity when false results exist in the set of results obtained by the given method.

the case of high variability of the tested material, there is insufficient justification for statistical quality assessment, e.g. the number of slots significantly depends on the detected word. Therefore, the ‘competition’ method was adopted. It consists in assessing the tested function through the best result obtained (in the whole research series).

3. Research experiment

The research consisted in using the method presented in paper [42]. This method is aimed at the use of patterns derived from the TTS synthesizer; such patterns were the main focus of interest. Research was conducted for the Polish language, the CLARIN-PL Mobile Corpus (EMU) [35], to the extent and as per the procedure described in paper [44]. Table 2 shows the values of the method parameters unchanged in relation to [44] and changed values adopted for the similarity functions not tested in paper [44].

For comparative purposes, additional tests were carried out using patterns from real speech recordings. They are marked in the results as *real*.

4. Results

4.1. Basic quality indicators

The results of 120 tests are presented as charts and tables. The main results are the number indicators obtained directly from the experiment: TP, TN, FP, FN. They were the basis for determining the derived indicators described above.

Table 3 presents sample test results when the similarity function was based on the Bhattacharyya distance. The values in the table, in the following lines, present the results for the query extracted from the real speech recording (*real*) and the synthesized textual query (*TTS*). The number of analysis slots, designated as *Slots*, is the number of all units the method extracted in the analysed speech signal. The number depends on the query length, hence its difference in test for the same session. The slot is not an analysis window, but the length of the pattern sought (cf. Table 2).

Other test results (for other similarity functions) are presented in a cumulative manner in figures 1 and 2.

Tab. 2. Parameters of the KWS method used in the described tests

	Parameter name	Parameter values											
Unchanged values	Number of FFTs	8192											
	Analysis window size	1024											
	Overlap percentage	33%											
	Number of HFCCs	15											
	Signal frequency range	[300, 3400]											
	Query length rate	1.5											
	Query match rate	0.5											
	Path threshold value	0.6											
Changed values	Similarity measurement method[9]	*bha*	*che*	*cor*	*cos*	*euc*	*hel*	*skl*	*man*	*mah*	*min*	*seu*	*spr*
	Normalisation method[10]	-	HE	HE	HE	HE	-	HE	HE	HE	HE	HE	HE
	Sequence threshold value (real/TTS)[11]	89/78	80/70	77/76	65/54	73/65	82/85	85/60	75/55	97/97	78/68	68/67	75/72
	Other[12]	NAN=1	NAN=1	NAN=0	NAN=0	NAN=1	NAN=1	NAN=0	NAN=1	ABS, NAN=1	NAN=1	NAN=1	NAN=1

Both charts show cumulative values for all selected sessions used in speech corpus research. The charts give the opportunity to compare the results for different similarity functions. They also show that despite the lack of proper method calibration, in each case, the method results are useful, i.e. true results (TP and TN) are always in total in the majority (i.e. more than 50% of all results). Undesirable false results (FP and FN) are partly the result of the said lack of calibration, although they also show the imperfection of the method, which depends on the dependence on the data itself (i.e. recordings), as mentioned in [42]. More information on the results can be found in the derivative indicator values presented in the next section.

[9] Designations as in Tab. 01.

[10] HE - normalisation by means of histogram equalization.

[11] In papers: [42], [43] and [44] the value is defined as the recognition quality threshold. It is used after marking detected sequences as suspicious, i.e. after applying the *path threshold*, which is clearly shown in paper [42].

[12] NAN - interpretation of non-numeric values, ABS - absolute value.

Tab. 3. Test results for ten selected recording sessions using the speech corpus. The similarity function is based on the Bhattacharyya distance.

		1	2	3	4	5	6	7	8	9	10
Real	**Slots**	80	56	88	56	53	40	50	55	48	77
	TP	22	10	25	12	16	12	10	26	12	26
	FP	17	14	32	29	3	13	30	13	29	37
	TN	36	28	28	15	29	14	10	14	7	14
	FN	5	4	3	0	5	1	0	2	0	0
TTS	**Slots**	43	26	39	26	38	26	36	36	29	53
	TP	21	12	22	10	21	14	10	24	10	27
	FP	6	4	7	10	6	7	14	9	18	23
	TN	13	9	7	5	9	3	2	3	1	3
	FN	3	1	3	1	2	2	0	0	0	0

4.2. Quality indicators obtained

Table 4 shows an example of indicator values for the results obtained in the tests for the Hellinger distance-based similarity function. The row for the sensitivity rate (TPR) is marked in the table. It shows the ability of the method to detect (indicate a result) where the value sought actually exists. Values close to one demonstrate the high sensitivity of the classifier. In the presented case, there were sessions for which virtually all searched words were found with a small percentage of false rejections (TN).

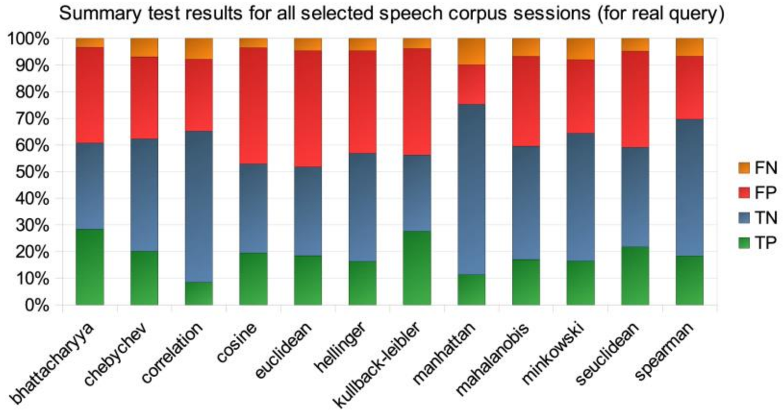


Fig. 1. Results for the real query - percentage value

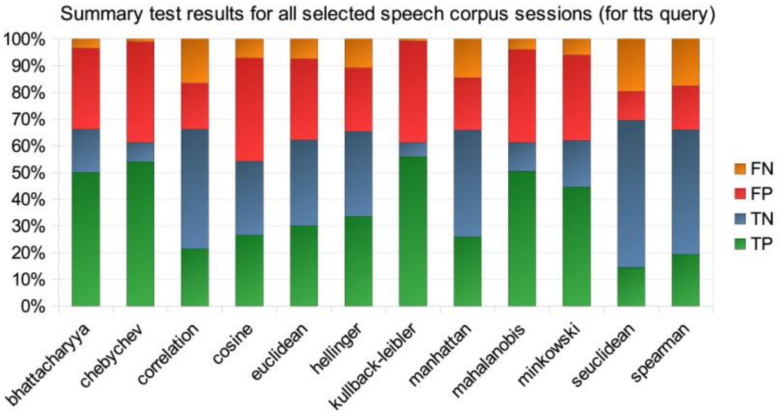


Fig. 2. Results for the TTS query - percentage value

Mean values: $\overline{TPR_{real}} = 0{,}74$, $\overline{TPR_{TTS}} = 0{,}75$, i.e. for the so-called *average case*, show that this similarity function can be successfully used in a situation where the researcher is primarily interested in maximizing the number of detections (true indications, TP), completely ignoring false positive (FP) values.

Tab. 4. Quality indicators for the method using the Hellinger distance-based similarity function. The results of 10 test sessions are presented.

		1	2	3	4	5	6	7	8	9	10
Real	**PPV**	0.55	0.10	0.43	0.14	0.38	0.33	0.08	0.54	0.13	0.45
	ACC	0.71	0.60	0.66	0.41	0.62	0.66	0.38	0.66	0.43	0.59
	TPR	**0.89**	**0.57**	**0.80**	**0.73**	**0.69**	**0.77**	**0.57**	**0.85**	**0.75**	**0.77**
	TNR	0.62	0.60	0.61	0.36	0.60	0.64	0.36	0.55	0.39	0.49
	F1S	0.68	0.17	0.56	0.24	0.49	0.47	0.14	0.66	0.22	0.57
	YJS	0.51	0.17	0.41	0.09	0.29	0.41	-0.07	0.39	0.14	0.26
TTS	**PPV**	0.67	0.88	0.94	0.47	0.83	0.64	0.36	0.70	0.32	0.51
	ACC	0.70	0.81	0.82	0.59	0.82	0.62	0.42	0.72	0.38	0.57
	TPR	**0.56**	**0.64**	**0.74**	**0.70**	**0.67**	**0.64**	**0.89**	**0.84**	**0.89**	**0.96**
	TNR	0.80	0.93	0.94	0.53	0.91	0.58	0.18	0.59	0.15	0.24
	F1S	0.61	0.74	0.83	0.56	0.74	0.64	0.52	0.76	0.47	0.67
	YJS	0.36	0.57	0.68	0.23	0.58	0.23	0.07	0.43	0.04	0.20

For the other functions, the calculated values of indicators are presented graphically. The first summary shows PPVs and ACCs (Fig. 3). Four similarity functions were selected, for which the mean indicators were the highest. They should be analysed simultaneously, as then they can indicate the possible direction of the detection method calibration. Based on these results, it can be stated that the KWS method applied is accurate, as the ACC obtained quite high values, and at the same time they are characterised by a low spread (which can be seen in charts c and d). At the same time, the method is not very precise, i.e. for some of the analysed recordings it does not detect the fragments it should detect (low PPV), and detects it for others (PPV close to one) - charts a and b).

Figure 3 b) shows that the PPV set using the Bhattacharyya distance-based similarity function is not characterised by such a big difference in value in subsequent tests (for other data) than *better* function based on Spearman's correlation at some points. This demonstrates that the first similarity function is less dependent on the specific data used in the test, and therefore the robustness of the whole spotting method is higher.

The level of *reliability* to the applied detection method can be concluded based on the second summary (Fig. 4). The *TTS* synthesised query was used in the tests. In this case, a *reliable* method is understood as one that does not maximise the number of false results, but detects and rejects what it should, according to the facts.

The third summary (Fig. 5) shows the calculated Youden's J statistics for the average and maximum cases. The results obtained are presented in an orderly manner relative to the mean value. The best similarity functions, as per the indicator, are those based on Spearman, Bhattacharyya and Manhattan distances.

4.3. Qualitative assessment of similarity function

The similarity function ranking shown below in Table 5 is a summary of the tests described in the article. It was based on qualitative assessment for all test samples, as per the method described in item 2.3.2. The final result presented in the table was obtained through the previously described scalarization. The test results for *real* query are also included for comparison.

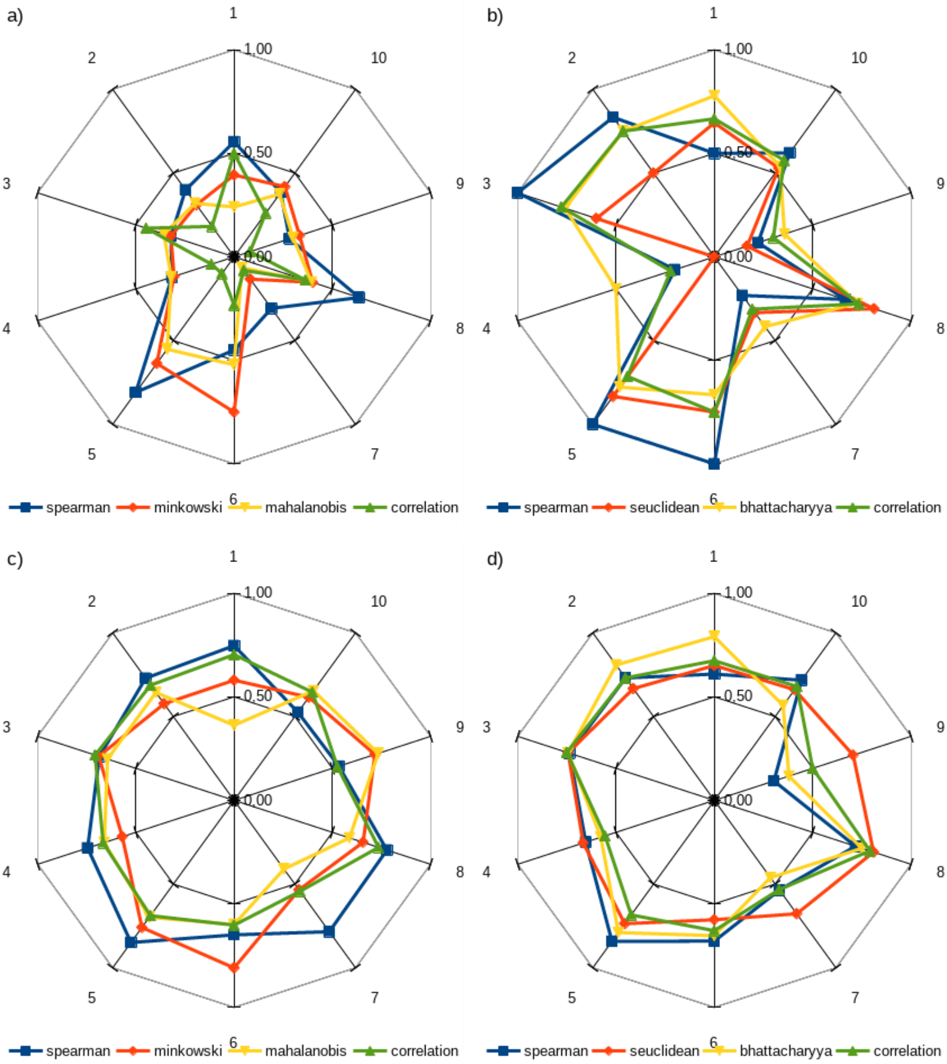


Fig. 3. Summary of PPV and ACC indicators for selected similarity functions. a) PPV for *real* query, b) PPV for *TTS* query, c) ACC for real query, d) ACC for *TTS* query; the results were obtained in subsequent test sessions (1 to 10)

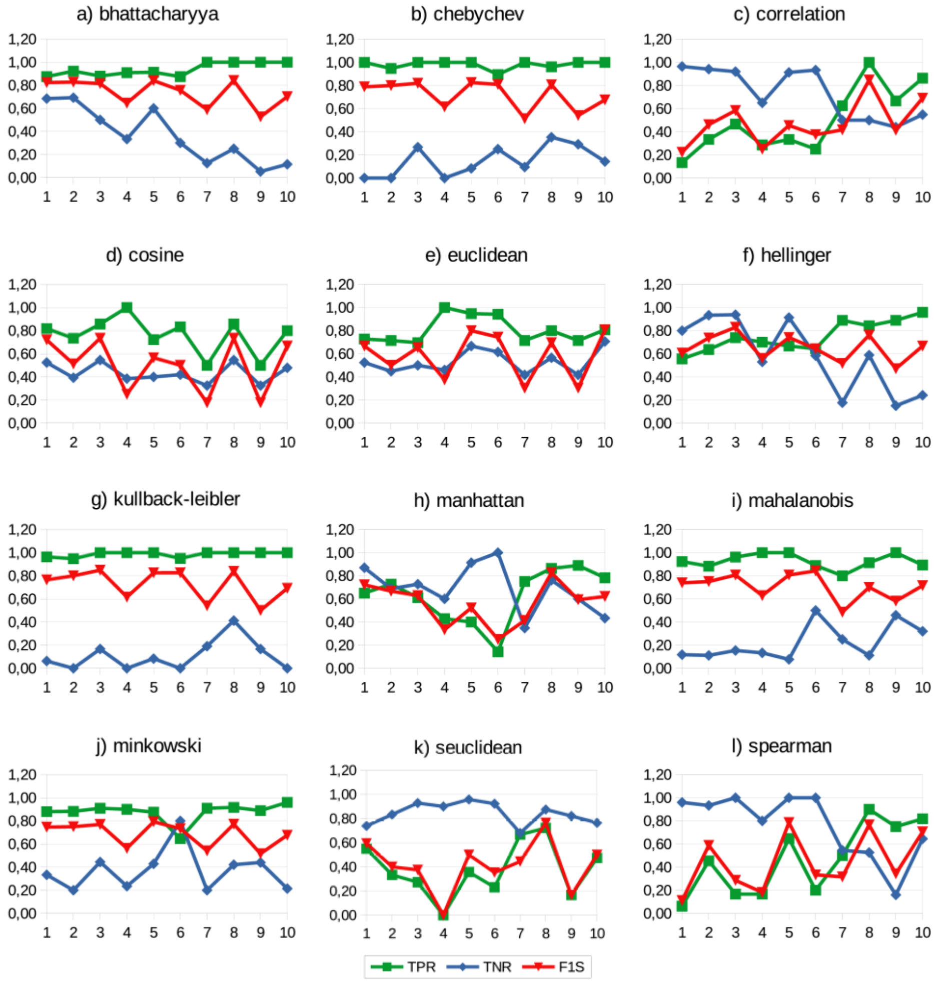


Fig. 4. Summary of indicators demonstrating the *reliability of* the detection method. The *TTS* synthesised query was used in the tests.

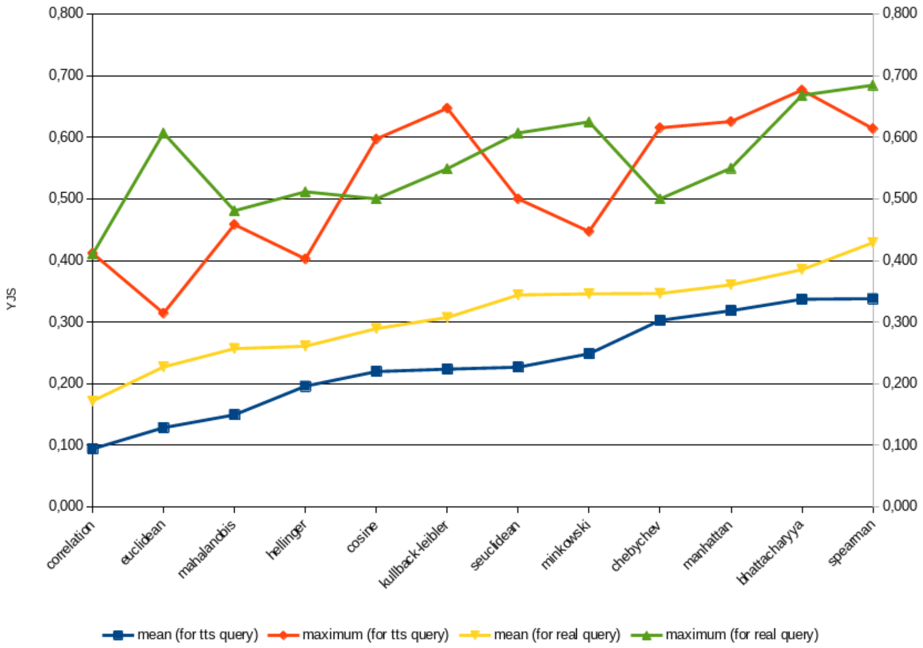


Fig. 5. List of Youden's J statistics (YJS). Blue squares and yellow triangles show average cases. Red diamonds and green triangles show best cases

5. Additional tests

5.1. ROC curve analysis

The numerical indicators used to select the signal similarity function describe only a certain momentary state of test. To learn how the keyword spotting method behaves in a wider range, a Receiver Operating Characteristic curve analysis was conducted [14], [60]. The analysis was carried out only for the selected (best) Spearman similarity function. The ROC curve is made as a set of indicating TPR and FPR values, obtained for several tests and repeated at different threshold values (see Fig. 6). Where:

$$FPR = 1 - TNR \tag{1}$$

is a fallout, false positive rate.

Tab. 5. Similarity function ranking

	TTS		real	
No.	**Similarity function**	**Indicator rating**	**Similarity function**	**Indicator rating**
1	K_{spr} (Spearman)	5.175	K_{bha} (Bhattacharyya)	5.066
2	K_{hel} (Hellinger)	5.167	K_{spr} (Spearman)	5.028
3	K_{man} (Manhattan)	5.154	K_{min} (Minkowski)	4.869
4	K_{cor} (correlation)	4.880	K_{man} (Manhattan)	4.782
5	K_{bha} (Bhattacharyya)	4.735	K_{seu} (standardized Euclidean)	4.670
6	K_{euc} (Euclidean)	4.726	K_{euc} (Euclidean)	4.537
7	K_{seu} (standardized Euclidean)	4.685	K_{che}(Chebyshev)	4.439
8	K_{min} (Minkowski)	4.556	K_{mah} (Mahalanobis)	4.046
9	K_{mah} (Mahalanobis)	4.370	K_{skl} (symmetrical Kullback-Leibler)	4.031
10	K_{skl} (symmetrical Kullback-Leibler)	4.176	K_{hel} (Hellinger)	3.971
11	K_{cos} (cosine)	4.023	K_{cos} (cosine)	3.957
12	K_{che}(Chebyshev)	3.957	K_{cor} (correlation)	3.808

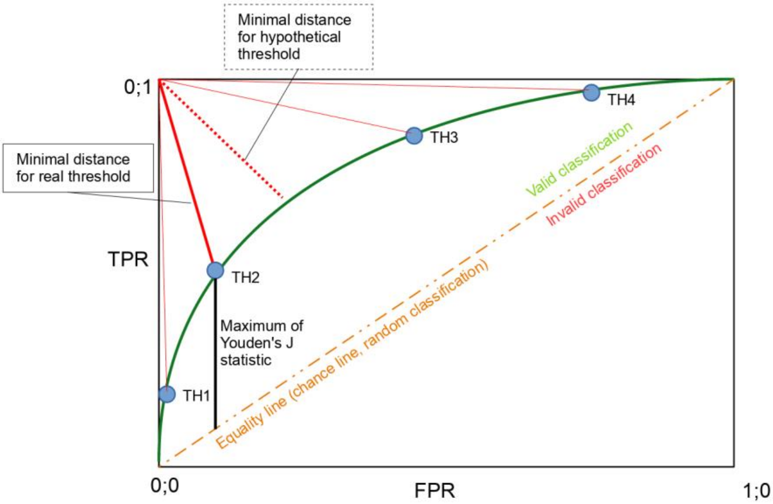


Fig. 6. Schematic representation of the ROC curve and the method of determining the threshold value. Thresholds TH 1 to 4 are applied in place of actual TPR and FPR values resulting from the measurement. TH can have any value range depending on the method. Youden's J statistic is marked similarly and its value is higher for the TH2 test than for the tests with other THs. The chart also shows hypothetically best TH value, which can be determined graphically, for example by comparing two adjacent threshold values (in this case TH2 and TH3)

The tests were conducted for the *TTS* query case. A total of 250 tests were conducted for the method presented in paper [42]. Method parameters adopted are the same as in Tab. 2; only the threshold value of the sequence is changed in the range of 50 to 98 with step 2 (i.e. for 25 values of this threshold). The tests were carried out for all selected recording sessions using the analysed speech corpus. TPR and FPR values were based on the results obtained and included on charts. The charts below show:

- Fig. 7: detailed analysis of the ROC curve for the selected session, including the method of selecting the threshold value that maximises TPR and minimises FPR,
- Fig. 8: curve analyses conducted for the remaining sessions with indicated best threshold value.

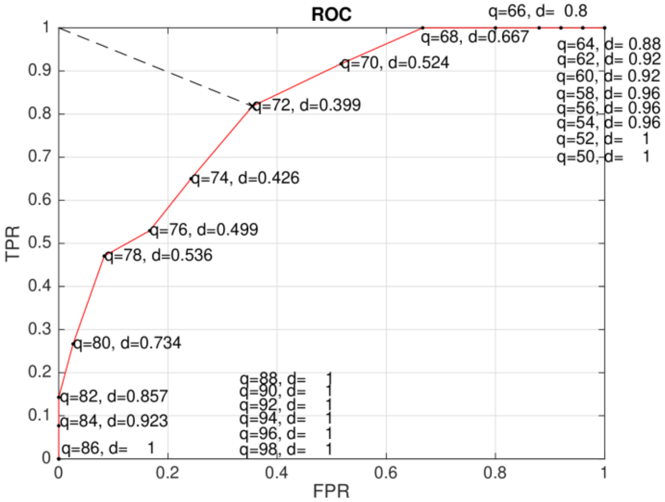


Fig. 7. ROC curve analysis for the selected session. The measurement points for the threshold value (q) are included on the chart with the determined distance value (d)

5.2. Matthews correlation coefficient

Subsequent tests were aimed to assess the impact of the selected similarity function on the random prediction of the detection method. The random prediction of the method means that it produces equally true and false results (cf. Fig 6). This is a very undesirable feature of the method, which is associated with its imperfection or lack of calibration. The Matthews correlation coefficient was used to achieve this goal [49]. This indicator takes into account the values of all four basic indicators (cf. formula 2), and its values are interpreted as follows [8]:

- '1' perfect prediction (zero false detections and rejections),
- ‘-1’ total disagreement (zero true values),
- '0' random prediction.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}} \tag{2}$$

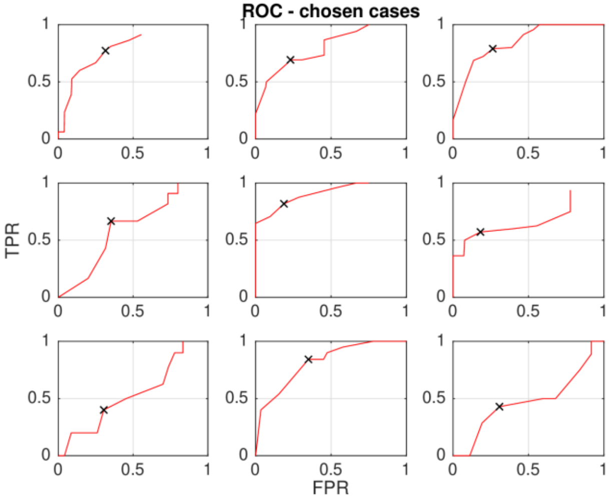


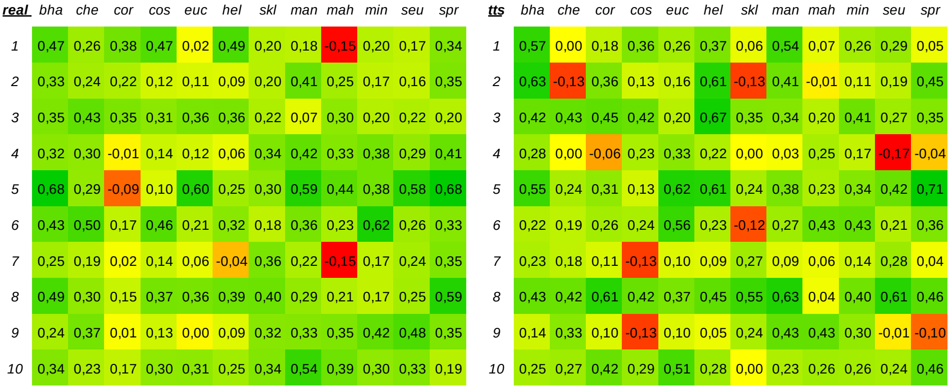
Fig. 8. Selected cases of ROC curve analysis for various sessions

Calculated values of the MCC are shown in Fig. 9. The values for other similarity functions are also included for comparison. It should be noted that the presented matrices are not correlation matrices. The MCC applies to mutual relation between true (TP, TN) and false (FP, FN) values of the method.

The test results confirmed the lack of random prediction for the detection method that uses similarity function K_{bha} and partly for the method that uses function K_{spr}.

6. Experiment conclusions

In the task of word spotting in speech signal, the choice of the signal similarity function is not obvious. The main aspect is the dependence of the similarity function on data, i.e. recordings of speech signal and its representation. This relationship translates into the quality of detection, as observed by comparing differences in results for *real* and *TTS* queries. The selection of the similarity function may come down to indicating the function which will be the most robust to data change. In the tests conducted, such a similarity function was based on the Spearman distance (K_{spr}).



real	*bha*	*che*	*cor*	*cos*	*euc*	*hel*	*skl*	*man*	*mah*	*min*	*seu*	*spr*
1	0,47	0,26	0,38	0,47	0,02	0,49	0,20	0,18	-0,15	0,20	0,17	0,34
2	0,33	0,24	0,22	0,12	0,11	0,09	0,20	0,41	0,25	0,17	0,16	0,35
3	0,35	0,43	0,35	0,31	0,36	0,36	0,22	0,07	0,30	0,20	0,22	0,20
4	0,32	0,30	-0,01	0,14	0,12	0,06	0,34	0,42	0,33	0,38	0,29	0,41
5	0,68	0,29	-0,09	0,10	0,60	0,25	0,30	0,59	0,44	0,38	0,58	0,68
6	0,43	0,50	0,17	0,46	0,21	0,32	0,18	0,36	0,23	0,62	0,26	0,33
7	0,25	0,19	0,02	0,14	0,06	-0,04	0,36	0,22	-0,15	0,17	0,24	0,35
8	0,49	0,30	0,15	0,37	0,36	0,39	0,40	0,29	0,21	0,17	0,25	0,59
9	0,24	0,37	0,01	0,13	0,00	0,09	0,32	0,33	0,35	0,42	0,48	0,35
10	0,34	0,23	0,17	0,30	0,31	0,25	0,34	0,54	0,39	0,30	0,33	0,19

tts	*bha*	*che*	*cor*	*cos*	*euc*	*hel*	*skl*	*man*	*mah*	*min*	*seu*	*spr*
1	0,57	0,00	0,18	0,36	0,26	0,37	0,06	0,54	0,07	0,26	0,29	0,05
2	0,63	-0,13	0,36	0,13	0,16	0,61	-0,13	0,41	-0,01	0,11	0,19	0,45
3	0,42	0,43	0,45	0,42	0,20	0,67	0,35	0,34	0,20	0,41	0,27	0,35
4	0,28	0,00	-0,06	0,23	0,33	0,22	0,00	0,03	0,25	0,17	-0,17	-0,04
5	0,55	0,24	0,31	0,13	0,62	0,61	0,24	0,38	0,23	0,34	0,42	0,71
6	0,22	0,19	0,26	0,24	0,56	0,23	-0,12	0,27	0,43	0,43	0,21	0,36
7	0,23	0,18	0,11	-0,13	0,10	0,09	0,27	0,09	0,06	0,14	0,28	0,04
8	0,43	0,42	0,61	0,42	0,37	0,45	0,55	0,63	0,04	0,40	0,61	0,46
9	0,14	0,33	0,10	-0,13	0,10	0,05	0,24	0,43	0,43	0,30	-0,01	-0,10
10	0,25	0,27	0,42	0,29	0,51	0,28	0,00	0,23	0,26	0,26	0,24	0,46

Fig. 9. Summary of Matthews correlation coefficient (MCC) values. The left side of the figure shows the *real* query, the right side shows the *TTS* query

The method of choosing the best similarity function proposed in the paper is based on six quality indicators. Therefore, the selected similarity function is not assessed unilaterally.

The analysis of the ROC curve conducted as part of the additional tests showed that the detection quality can be significantly impacted by the selection of the appropriate threshold value (marked q in Fig. 7). It should be noted that completely bad results (i.e. more false detections and rejections than true results), using similarity function K_{spr}.

It is worth noting that the differences in the values of quality indicators obtained for different similarity functions are small. Choosing a similarity function based only on a single quality indicator value can be deceptive. Therefore, when choosing the similarity function, it is justified to carry out at least several tests for different data. The analysis of quality indicators for such tests gives more complete knowledge and it can be then expected that the chosen similarity function will give correct results for different data.

References

[1] AMGOUD L., DAVID V., DODER D., *Similarity Measures Between Arguments Revisited.* In: Kern-Isberner G., Ognjanović Z. (eds) Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU 2019, Lecture Notes in Computer Science, Vol. 11726, pp. 98-107, DOI 10.1007/978-3-030-29765-7_1

[2] BHATTACHARYYA A., *On a measure of divergence between two statistical populations defined by their probability distributions.* Bulletin of the Calcutta Mathematical Society. Vol. 35, 1943, pp. 99-109.

[3] BASENER W., FLYNN M., *Microscene evaluation using the Bhattacharyya distance.* Proc. of SPIE 10780, Honolulu, 2018, DOI 10.1117/12.2327004

[4] BOYTSOV L., *Indexing methods for approximate dictionary searching: Comparative analysis.* Journal of Experimental Algorithmics, Vol. 16, Article 1.1, May 2011, pp. 1-91, DOI 10.1145/1963190.1963191

[5] CHANG H.Y., *An SVM Kernel With GMM-Supervector Based on the Bhattacharyya Distance for Speaker Recognition.* IEEE Signal Processing Letters, 2009, Vol. 16, Issue 1, pp. 49-52, DOI 10.1109/LSP.2008.2006711

[6] CHEN B., WANG H.-M., CHIEN L.-F. LEE L.-S., *A*-Admissible Key-Phrase Spotting With Sub-Syllable Level Utterance Verification.* The 5th International Conference on Spoken Language Processing, Incorporating The 7th Australian International Speech Science and Technology Conference, Sydney, Australia, 1998, pp. 783-786.

[7] CHEN Y.-I., WU CH.-H., YAN G.-L., *Utterance Verification Using Prosodic Information for Mandarin Telephone Speech.* 1999 IEEE International Conference on Acoustics, Speech and Signal Processing. Keyword Spotting Proceedings, ICASSP '99, Vol. 2, Phoenix, AZ, USA, pp. 697-700, DOI 10.1109/ICASSP.1999.759762

[8] CHICCO D., *Ten quick tips for machine learning in computational biology*. BioData Mining, Vol. 10, No. 35, 2017, pp. 1-17, DOI 10.1186/s13040-017-0155-3

[9] CHINCHOR N., *MUC-4 Evaluation Metrics*. In Proceedings of the Fourth Message Understanding Conference, 1992, pp. 22-29, http://www.aclweb.org/anthology-new/M/M92/M92-1002.pdf

[10] DEB K., *Introduction to Evolutionary Multiobjective Optimization.* In: Branke J., Deb K., Miettinen K., Słowiński R. (eds) Multiobjective Optimization. Lecture Notes in Computer Science, Vol. 5252, 2008, Springer, Berlin, Heidelberg, pp. 59-96, DOI 10.1007/978-3-540-88908-3_3

[11] DUIN R.P. W., PĘKALSKA E., *The Dissimilarity Representation for Structural Pattern Recognition.* Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 2011, pp. 1-24, DOI 10.1007/978-3-642-25085-9_1

[12] DUIN R.P.W., PĘKALSKA E., *Non-euclidean dissimilarities: Causes and informativeness.* In proc. Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), 2010, LNCS, Vol. 6218, Springer, Heidelberg, pp. 324-333, DOI 10.1007/978-3-642-14980-1_31

[13] DUBUISSON M.P., JAIN A.K., *A Modified Hausdorff distance for object matching.* In ICPR94, Jerusalem, Israel, 1994, pp. 566-568.

[14] FAWCETT T., *An Introduction to ROC Analysis.* Pattern Recognition Letters, Vol. 27, No. 8, 2006, pp. 861-874, DOI 10.1016/j.patrec.2005.10.010

[15] FOOTE J., *An Overview of Audio Information Retrieval.* ACM Multimedia Systems, Vol. 7, 1998, pp. 2-10, DOI 10.1.1.39.6339

[16] FUKUNAGA K., *Introduction to Statistical Pattern Recognition.* 2nd Edition, Elsevier Inc, 1990, DOI 10.1016/C2009-0-27872-X

[17] GÜNDOĞDU B., *Keyword Search for Low Resource Languages.* PhD Thesis, Bogazici Universit, 2017.

[18] GÜNDOĞDU B., SARAÇLAR M., *Distance metric learning for posteriorgram based keyword search.* 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, 2017, pp. 5660-5664, DOI 10.1109/ICASSP.2017.7953240

[19] GUPTA K., GUPTA D., *An analysis on LPC, RASTA and MFCC techniques in Automatic Speech Recognition.* 2016 6th International Conference - Cloud System and Big Data Engineering System (Confluence), Noida, 2016, pp. 493-497, DOI 10.1109/CONFLUENCE.2016.7508170

[20] GUPTA P., PUROHIT G.N., RATHORE M., *Number Plate Extraction using Template Matching Technique.* International Journal of Computer Applications, Vol. 88, No. 3, 2014, pp. 40-44, DOI 10.5120/15336-3670

[21] HAASDONK B., BAHLMANN C., *Learning with distance substitution kernels.* In Pattern Recognition – Proc. of the 26th DAGM Symposium, 2004, pp. 220-227, DOI 10.1007/978-3-540-28649-3_27

[22] HAFEN R.P., HENRY M.J., *Speech information retrieval: a review.* Multimedia Systems, Vol. 18, No. 6, 2012, pp. 499-518.

[23] HELLINGER E., (in German) *Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen.* Journal für die reine und angewandte Mathematik, Vol. 136, 1909, pp. 210–271, DOI 10.1515/crll.1909.136.210

[24] HENRIKSON J., *Completeness and total boundedness of the Hausdorff metric.* MIT Undergraduate Journal of Mathematics, 1999, pp. 69-80.

[25] HIGGINS A., WOHLFORD R., *Keyword recognition using template concatenation.* IEEE International Conference on Acoustics, Speech and Signal Processing,

ICASSP '85, Tampa, FL, USA, 1985, pp. 1233-1236, DOI 10.1109/ICASSP.1985.1168253

[26] HOBSON A., CHENG B-K., *A comparison of the Shannon and Kullback information measures.* Journal of Statistical Physics, Vol. 7, No. 4, 1973, pp. 301–310, DOI: 10.1007/BF01014906

[27] HOLYOAK K.J., THAGARD P., *Mental Leaps: Analogy in Creative Thought.* A Bradford Book series, MIT Press, 1996.

[28] JANSEN A., DURME VAN B., *Efficient Spoken Term Discovery Using Randomized Algorithms*. 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, Waikoloa, HI, 2011, pp. 401-406, DOI 10.1109/ASRU.2011.6163965

[29] JANSEN B., RIEH S.Y., *The Seventeen Theoretical Constructs of Information Searching and Information Retrieval.* In Journal of the American Society for Information Science and Technology, Vol. 61, No. 8., 2010, pp. 1517-1534, DOI 10.1002/asi.21358

[30] JENSEN J.H., ELLIS D.P. W., CHRISTENSEN M.G., JENSEN S.H., *Evaluation of Distance Measures Between Gaussian Mixture Models of MFCCs.* Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, Vienna, 2007, pp. 107-108.

[31] KAILATH T., *The Divergence and Bhattacharyya Distance Measures in Signal Selection.* IEEE Transactions on Communication Technology, 1967, Vol. 15, No. 1, pp. 52-60, DOI 10.1109/TCOM.1967.1089532

[32] KAMIŃSKA D., SAPIŃSKI T., ANBARJAFARI G., *Efficiency of chosen speech descriptors in relation to emotion recognition.* EURASIP Journal on Audio, Speech, and Music Processing (2017), Vol. 3, pp. 1-9, DOI 10.1186/s13636-017-0100-x

[33] KASSAMBARA A., *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning (Multivariate Analysis).* Vol. 1, CreateSpace Independent Publishing Platform, 2017.

[34] KESHET J., GRANGIER D., BENGIO S.A., *Discriminative keyword spotting.* Speech Communication, 2009, Vol. 51, No. 4, pp. 317-329, DOI 10.1016/j.specom.2008.10.002

[35] KORŽINEK D., MARASEK K., BROCKI Ł., WOŁK K., *Polish Read Speech Corpus for Speech Tools and Services.* Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016, CLARIN Common Language Resources and Technology Infrastructure, No. 136, Linköping University Electronic Press, Linköpings universitet, 2017, pp. 54-62.

[36] KULLBACK S.; LEIBLER R.A. *On information and sufficiency*. Annals of Mathematical Statistics, Vol. 22, No. 1, 195, pp. 79-86, DOI 10.1214/aoms/1177729694

[37] KULLBACK S., *Information theory and statistics.* Dover Books on Mathematics, New Edition, 1997.

[38] KWIATKOWSKI W., (in Polish) *Klasyfikacja metodą grupowania cech z uwzględnieniem ich wzajemnej korelacji.* Biuletyn Instytutu Automatyki i Robotyki, Nr 14, 2000, pp. 139-146.

[39] KWIATKOWSKI W., (in Polish) *Metody automatycznego rozpoznawania wzorców.* Instytut Automatyki i Robotyki, WAT, Wydanie I, Warszawa, 2001.

[40] KWIATKOWSKI W., (in Polish) *Wykrywanie anomalii bazujące na wskazanych przykładach*. Przegląd Teleinformatyczny, Nr 1-2, 2018, pp. 3-21.

[41] KWIATKOWSKI W., (in Polish) *Wstęp do cyfrowego przetwarzania sygnałów. BEL Studio,* WAT, Warszawa, 2003.

[42] LASZKO Ł., *Word detection in recorded speech using textual queries*. Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 5, pp. 849-853, DOI 10.15439/2015F341

[43] LASZKO Ł., *Using formant frequencies to word detection in recorded speech*. Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 8, pp. 797-801, DOI 10.15439/2016F518

[44] LASZKO Ł., *Developing keyword spotting method for the Polish language*. Communication Papers of the 2018 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 17, pp. 123-127, DOI 10.15439/2018F178

[45] LEBRET R., COLLOBERT R., *Word Embeddings through Hellinger PCA*. 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL, 2014, pp. 482-490, DOI 10.3115/v1/E14-1051

[46] LI H., HAN J., ZHENG T., ZHENG G., *Mandarin keyword spotting using syllable based confidence features and SVM.* 2nd International Conference on Intelligent Control and Information Processing, Harbin, 2011, pp. 256-259, DOI 10.1109/ICICIP.2011.6008243

[47] LI W., BILLARD A., BOURLARD H., *Keyword Detection for Spontaneous Speech.* 2nd International Congress on Image and Signal Processing, Tianjin, 2009, pp. 1-5, DOI 10.1109/CISP.2009.5303824

[48] LIU D., CHO S., SUN D., QIU Z., *A Spearman correlation coefficient ranking for matching-score fusion on speaker recognition.* TENCON 2010 - 2010 IEEE Region 10 Conference, Fukuoka, 2010, pp. 736-741, DOI 10.1109/TENCON.2010.5686608

[49] MATTHEWS B.W., *Comparison of the predicted and observed secondary structure of T4 phage lysozyme*. Biochimica et Biophysica Acta (BBA) – Protein Structure, Vol. 405, No. 2, 1975, pp. 442-451, DOI 10.1016/0005-2795(75)90109-9

[50] MANNING CH.D., RAGHAVAN P., SCHÜTZE H., *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[51] MIETTINEN K., *Introduction to Multiobjective Optimization: Noninteractive Approaches*. In: Branke J., Deb K., Miettinen K., Słowiński R. (eds) Multiobjective Optimization. Lecture Notes in Computer Science, Vol 5252, 2008, Springer, Berlin, Heidelberg, pp. 1-26, DOI 10.1007/978-3-540-88908-3_1

[52] MIETTINEN K., RUIZ F., WIERZBICKI A.P., *Introduction to Multiobjective Optimization: Interactive Approaches*. In: Branke J., Deb K., Miettinen K., Słowiński R. (eds) Multiobjective Optimization. Lecture Notes in Computer Science, Vol 5252, 2008, Springer, Berlin, Heidelberg, pp. 27-57, DOI 10.1007/978-3-540-88908-3_2

[53] MITRA V., HAUT VAN J., FRANCO H., VERGYRI D., *Feature Fusion for High-Accuracy Keyword Spotting*. Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference Lei Y., et al. on, 2014, pp. 7143-7147.

[54] MOHAMED S.S., ABDALLA A., JOHN R.I., *New Entropy-Based Similarity Measure between Interval-Valued Intuitionstic Fuzzy Sets*. Axioms, Vol. 8, No. 2, 2019, Article-Number 73, DOI 10.3390/axioms8020073

[55] MUSCARIELLO A., GRAVIER G., BIMBOT F., *Audio keyword extraction by unsupervised word discovery*. In Proceedings of the Interspeech, 2009, pp. 2843–2847.

[56] MÜLLER M., *Information Retrieval for Music and Motion*. Springer Berlin Heidelberg New York, 2007.

[57] NIELSEN F., *A generalization of the Jensen divergence: The chord gap divergence*. arXiv preprint, 2017, pp. 1-13, https://arxiv.org/abs/1709.10498

[58] PARDO L., *Statistical Inference Based on Divergence Measures*. Statistics: A Series of Textbooks and Monographs, 1st Edition, Chapman and Hall/CRC, 2006.

[59] PARK A.S., GLAS J.R. *Unsupervised pattern discovery in speech*. IEEE Trans. on Audio, Speech and Language Processing, 2008, Vol. 16, No. 1, pp. 186-197.

[60] PONTIUS R.G., KANGPING S., *The total operating characteristic to measure diagnostic ability for multiple thresholds*. International Journal of Geographical Information Science, Vol. 28, No. 3, 2014, pp. 570-583, DOI 10.1080/13658816.2013.862623

[61] POWERS D.M.W., *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation*. Journal of Machine Learning Technologies, Vol. 2, No. 1, 2007, pp. 37-63.

[62] QIAO Y., MINEMATSU N., *A Study on Invariance of f-Divergence and Its Application to Speech Recognition.* IEEE Transactions on Signal Processing, 2010, Vol. 58, No. 7, pp. 3884-3890, DOI 10.1109/TSP.2010.2047340

[63] RAIELI R., *Introducing Multimedia Information Retrieval to libraries.* Italian Journal of Library, Archives, and Information Science, Vol. 7, No. 3, 2016, pp. 9-42, DOI 10.4403/jlis.it-11530

[64] SAMMUT C., WEBB G.I. (eds.), *Encyclopedia of Machine Learning and Data Mining.* 2nd Edition, Springer, 2017.

[65] SASAKI Y., *The truth of the F-measure.* 2007, 5 pages, Web resource available at https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf

[66] SCHÖLKOPF B., *The Kernel Trick for Distances.* Advances in neural information processing systems, Vol. 13, 2000, pp. 301-307.

[67] SINGH A., YADAV A., RANA A., *K-means with Three different Distance Metrics.* International Journal of Computer Applications, Vol. 67, No.10, 2013, pp. 13-17, DOI 10.5120/11430-6785

[68] SINGHAL A., *Modern Information Retrieval: A Brief Overview.* Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Vol. 24, No. 4, 2001, pp. 35-43.

[69] STEHMAN S.V., *Selecting and interpreting measures of thematic classification accuracy*. Remote Sensing of Environment, Vol. 62, No. 1, 1997, pp. 77-89, DOI 10.1016/S0034-4257(97)00083-7

[70] TABIBIAN S., AKBARI A., NASERSHARIF B., *Improved dynamic match phone lattice search for Persian spoken term detection system in online and offline applications.* International Journal of Speech Technology, March 2019, Vol. 22, Issue 1, pp 205-217, DOI 10.1007/s10772-019-09594-w

[71] TÜSKEA Z., NOLDEN D., SCHLÜTERA R., NEY H., *Multilingual MRASTA features for low-resource keyword search and speech recognition systems*. 2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2014, pp. 7349-7353.

[72] WILPON J. G., RABINER L.R., LEE C., GOLDMAN E.R., *Automatic recognition of keywords in unconstrained speech using hidden Markov.* IEEE Transactions on Acoustics, Speech and Signal Processing, 1990, Vol. 38, No. 11, pp. 1870-1878, DOI 10.1109/29.103088

[73] YOUDEN W. J., *Index for rating diagnostic tests.* Cancer, Vol. 3, 1950, pp. 32–35, DOI 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

[74] ZEDDELMANN VON D., KURTH F., MÜLLER M., *Perceptual audio features for unsupervised key-phrase detection.* Proc. ICASSP2010, 2010, pp. 257-260, DOI 10.1109/ICASSP.2010.5495974

[75] ZHANG Y., *Unsupervised Speech Processing with Applications to Query-by-Example Spoken Term Detection.* PhD thesis, Massachusetts Institute of Technology, 2013.

[76] ZHANG Y., GLASS J.R., *Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams.* 2009 IEEE Workshop on Automatic Speech Recognition & Understanding, Merano, 2009, pp. 398-403, DOI 10.1109/ASRU.2009.5372931

[77] ZHU X., PENN G., RUDZICZ F., *Summarizing multiple spoken documents: finding evidence from untranscribed audio.* Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Vol. 2, 2009, pp. 549-557.

[78] ZIELIŃSKI T.P., (in Polish) *Cyfrowe przetwarzanie sygnałów od teorii do zastosowań.* Wydawnictwa Komunikacji i Łączności, Warszawa, 2005.

[79] ZIÓŁKO B., GAŁKA J., SKURZOK D., JADCZYK T., *Modified Weighted Levenshtein Distance in Automatic Speech Recognition.* Krajowa Konferencja Zastosowań Matematyki w Biologii i Medycynie, Krynica, 2010, s. 116-120.

Eksperymentalne badanie wpływu wyboru funkcji podobieństwa na jakość wykrywania słów w sygnale mowy

STRESZCZENIE: W pracy przedstawiono ocenę zastosowania wybranych funkcji podobieństwa w zadaniu wykrywania słów kluczowych. Przeprowadzono eksperymenty dla języka polskiego. Wyniki badań można wykorzystać do ulepszenia już istniejących metod wykrywania słów kluczowych lub do opracowania nowych.

SŁOWA KLUCZOWE: wykrywanie słów kluczowych, podobieństwo sygnałów, wskaźniki jakości wykrycia, odkształcanie skali czasu, kwerenda tekstowa

Received by the editorial staff on: 27.11.2019