Communication Papers of the Federated Conference on Computer Science and Information Systems pp. 123–127
DOI: 10.15439/2018F178
ISSN 2300-5963 ACSIS, Vol. 17

Developing keyword spotting method for the Polish language

Łukasz Laszko
Cybernetics Faculty,
Military University of
Technology,
ul. Gen. W. Urbanowicza 2,
00-908 Warsaw, Poland
Email:
lukasz.laszko@wat.edu.pl

***Abstract*—The paper presents the application of unsupervised method to word detection in recorded speech for the spoken Polish language. The method utilizes similarity measure between analyzed speech and a pattern synthesized from pure text. Dynamic time warping algorithm is applied for time alignment and the resulting alignment path defines an input to the classifier. The classification process involves calculation of cost function and extraction of the projected sequence of Human-Factor Cepstral Coefficients, both of which are compared with the threshold values. The results obtained after application of the method to the CLARIN-PL Mobile Corpus are encouraging to develop this method for the Polish language.**

I. Introduction

THE hankering for good and robust method of automatic speech recognition or word spotting for the Polish language has been observed for years in Polish scientific, as well as business field. Recently much effort has been put to appropriate language modeling [1], considering the nature of the Polish spoken language. These studies reveal the complexity of the modeling process for general purposes and indicate particularly difficult attributes of the language to model, such as its inflection, non-positionality and frequent occurrence of short words. Deep insight into speech signal shows also that the existence of high-frequency, low-energy consonants like fricatives and plosives, restricts the adopting of widely used methods and tools good for the English language [3]. Although to simple daily tasks one could employ with success the grammar-based ASR's [2], which firstly require primitive ontological relations to be built for a class of sentences in the given field. Either HMM or various classes of neural networks, built on specific acoustic features, are the most common models used in this area.

Considering the evolution of speech recognition and speech processing tools available for the Polish language the trend to exploit open-source technologies is observed. One example is SARMATA [4], the aboriginal Polish ASR system, which has recently (version 2.0) being under departing from its own engine to Kaldi toolkit. The system in its pre-2.0 versions was able to be used in industry, recognizing up to 1000 learned words. The new version is very likely to be much more versatile, because of using available in the toolkit large number of possible to use speech models and techniques of speech processing, as well as massive GPU processing implemented in the toolkit.

Contemporarily, the most-growing Polish set of tools, as it seems to be, is provided by CLARIN-PL. This is actually much more than the set of software, but it is seen as a speech platform for processing, visualizing and depositing language data [5]. This platform provides cloud-based research infrastructure (type B) with corpora, tools (via web services[1]) and metadata. It also enables users to make available their own products like tools or corpora[2].

Nor the reader can miss the *de facto standard* of Google SpeechRecognizer, which engine is integrated to most Android mobile devices used today, and has a support for 119 languages including the Polish language. Moreover its API is freely available to developers of Android applications.

In this field the author propose the adoption of keyword spotting method (abbr. KWS) introduced in [6] for the spoken English language, to the Polish language. The method is designed to search for specific words only and does not analyze the structure of speech at higher levels than the acoustic features, i.e. the language or the grammar. There is also no supervised model training step, apart from that one do need to assess a few threshold values. Although applied for the English language, the method gives relatively high detection rate, about 80%.

II. Problem statement

A. Method background

The precise description of the method as well as alternatives could be found in [6]. In the nutshell, this method is searching through a speech medium (database) fragment by

This work was supported by Cybernetics Faculty of the Military University of Technology, under the grant no. RMN/813/2016

[1] For tools availability, see: https://clarin-pl.eu/en/services/

[2] E.g. *Acoustic Data Building Toolset*, about 29 hours (17 GB) corpus of annotated Polish speech, together with software.

fragment and comparing the description of each fragment with the same class of description of a search pattern. The description is understood as a sequence of acoustic features. The comparison is done in the similarity space, which contains implicit information about correlation between the two descriptions. The strength of similarity is measured by applying Dynamic Time Warping (abbr. DTW) algorithm and extracting the best projection path of the search pattern description to the description of the analyzed fragment of speech, which is called in the method the alignment path. DTW fulfills therefore two important tasks: (1) time alignment between search pattern and analyzed fragment, (2) cost counting, which provides values to the classification process (see Fig 1).

In the classification stage the calculation of cost function and sequence search according to the extracted aligning path are done and compared with the threshold values. This provides the decision on the class of analyzed fragment of speech as match or no-match. Then upon the results of assessing values of cost function and found sequences, the quality of the matches is calculated, which provides the numerical control of matches.

Next stages depend on the application and could involve, but are not limited to the verification by listening or thorough search of the area pointed by best matches.

It is worth noting that one input to the model from Fig 1 , comes from the Text-to-Speech generation. This is the valuable attribute of the presented method, which according to [6], makes the method versatile.

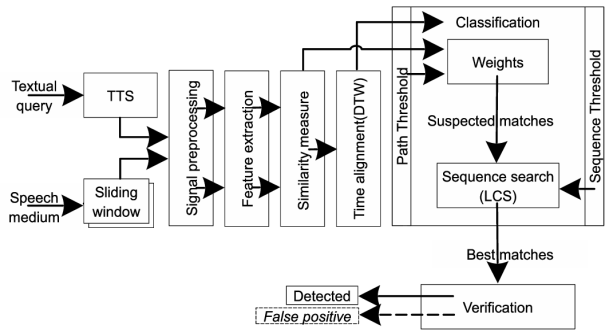

Fig 1. Overview of the unsupervised detection method

B. Mathematical model of speech description

In these research Human-Factor Cepstral Coefficients (abbr. HFCC) have been chosen as the description of speech signal. HFCCs are computed according to the following algorithm:

1) given signal S has been windowed by Hamming window resulting in N segments, $s_1 \dots s_N$;

2) each segment has been processed by short-time Fourier transform (abbr. STFT) with length of 64 ms and a fixed step size of 5 - 10 ms;

3) then the triangular filter bank has been developed with 40 equally spaced mel-scale center frequencies f_i, $i=1,\dots,40$ with bands controlled by the measure called Equivalent Rectangular Bandwidth (abbr. ERB):

$$ERB(f)=6.23f^2+93.39f+28.52 \text{ Hz}; \quad (1)$$

where f states for filter center frequency, expressed in kHz.

4) next, the filtering has been done, by multiplication of each STFT segment with magnitude spectrum of bands for HFCC;

5) finally, the result has been decorrelated using Discrete Cosinus Transform (abbr. DCT), keeping only 15 the most decorrelated coefficients.

C. Measuring similarity of two signals

Let the matrix $D_{A,R}$, where A stands for analyzed voice feature vector and R stands for reference pattern feature vector, hold the information on similarity between A and R . Then the individual element $d(a,r)$ of the matrix, where a,r stand for specific element of vector A and vector R respectively, is given by inner product:

$$d(a,r)=\frac{\langle A_a, R_r\rangle}{\|A_a\|\,\|R_r\|} \quad (2)$$

D. Applying DTW

Let the $C_{A,R}$ be the accumulator matrix of size D . Then the accumulation in each element $c(a,r)$ holds the value of lowest transition cost to this element from its neighbors, including the cost of the lowest transition to the neighbors from theirs consequent neighbors until the starting element $c(1,1)$. The computation is given by the recursion:

$$c(a+1,r+1)=d(a+1,r+1)+\min\begin{cases}c(a-1,r)\\ c(a,r)\\ c(a,r-1)\end{cases} \quad (3)$$

where: $a,r\geq 1$ and $c(1,1)=d(1,1)$.

The stage of applying DTW gives the calculation of optimal aligning P of analyzed voice description and the reference pattern description. P is created based on the accumulator elements traceback, starting from its last element $c(N_A,N_R)$ and ending in $c(1,1)$ recursively, by searching across all allowable predecessors to each element. Because C has been built of costs of the lowest transitions, the actual calculation of the path is based on choosing the next element from the closest elements with minimal cost value.

E. Classification and quality measure

P and D hold then the full information on the similarity strength between analyzed fragment and referenced pattern. Upon this v is computed based on referring costs

of matrix $D_{A,R}$, where A, R are taken from P. Then v is equated to path threshold T_P, producing suspected matches M. To this result the Longest Common Subsequence (abbr. LCS) algorithm is applied to reject the least valuable sequences according to sequence threshold T_S. For all accepted results the quality measure is computed according to (4).

$$Q_M = \frac{v(M)}{LCS(M, T_S)} 100 \quad (4)$$

F. Variables

The method has many variables which values decide on the usability of the method. There is a need for: calculation of the width of analysis window; HFCC computation parameters which are used in point B.1, deciding on the feature space dimension discussed in point 5; P-specific calculations in DTW algorithm (direction variation, analyzed area in D, etc); threshold values: T_P and T_S which decide on the resultant matches, as well as the minimal quality value satisfactory for specific applications.

III. Experiments

A. Research material

The experiments have been conducted on CLARIN-PL Mobile Corpus (EMU) [8][3]. This is a Polish speech corpus of read speech recorded over a phone. It contains 554 sessions of many speakers reading a few dozen different sentences. Each recorded speech is annotated. Total corpus length is about 13 hours (12 GB uncompressed). Sound quality is at medium level (16 kHz, 32 bits/sample, mono) stored in WAV containers.

The queries have been generated using Google Text-to-Speech engine, available via Google Translate, based on a textual input.

B. Procedure

According to Fig 1, the experiments started by preparing queries and signal preprocessing. Then in accordance with point II.B, HFCC features were computed for the query as well as for the analyzed fragment. Fragments lengths were in these experiments 1.5 times longer in time than the queries. The last analyzed fragment was complemented with zeros.

Sliding windows were produces with fixed step size without overlapping of neighboring windows. The overlapping was included at the stage of computing HFCCs.

Then the D matrix was computed, before applying it DTW. Based on the results of DTW the classification of the

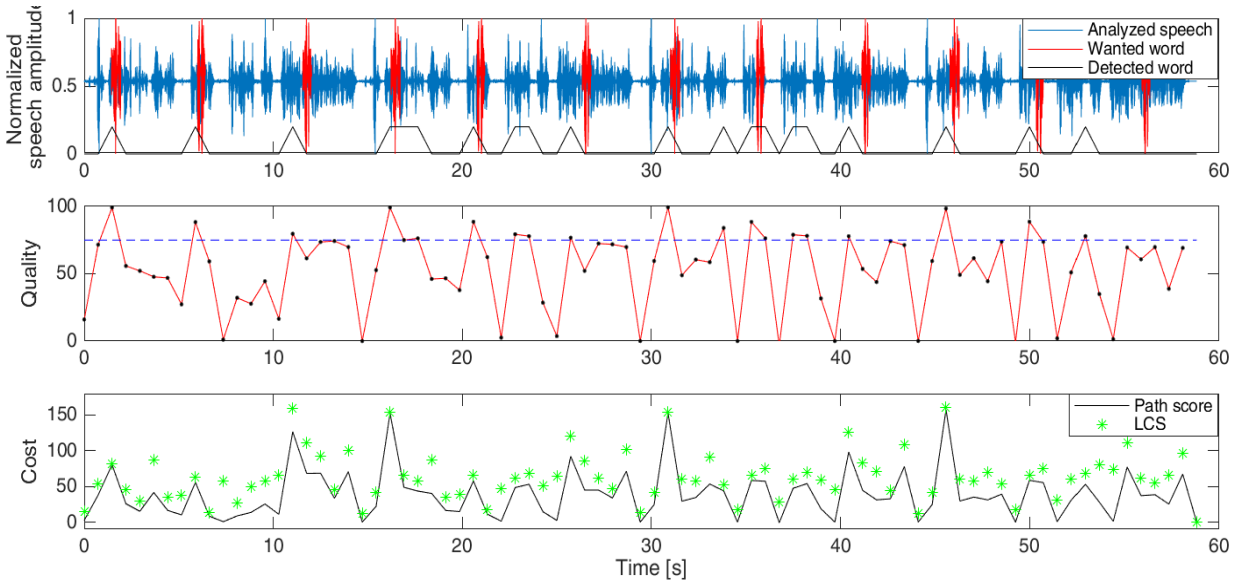

Fig 2.Results of word spotting on sentence 1 of session 1 of the CLARIN-PL Mobile Corpus. The chosen word 'senator' occurs 12 times in the sentence (in different inflections), which is marked in the upper chart. The upper chart has also markers placed around the bottom axis, to indicate the best matches obtained by the discussed method. Medium chart presents quality of detection, with the satisfying cutoff level of 75%. Bottom chart presents costs computed during classification stage.

[3] For corpus imprint see: https://clarin-pl.eu/dspace/handle/11321/237

fragment was done thus obtaining the list of potential matches of sequences allegedly detected in the fragment.

Finally for accepted matches, the detection quality was assessed based on (4).

The procedure was repeated two times for the same sentence with the change of reference pattern. For method verification, the `synthesized query` was exchange for `the excerpt of the same material` with the same content.

C. Results and discussion

Overall results have been presented in Table 1. These results concern whole research material, which includes chosen sessions from 554 available sessions in the speech corpus, without distinction to the gender of speaker. Unfortunately only one TTS system has been used in the research (female voice of medium quality), which probably caused understating of the percentage of detected words.

High word detection rate has been observed. Concerning real speech pattern results, more values have been obtained over the presented mean value (negative skewness). Although false detection rate also maintained at rather high level, these results do not seem to correlate (correlation coefficient, CC equals: -0,2).

Referring to TTS results, positive impression is given, not only by high detection rate, but also by the maximal value for detection. This means that synthetic speech has perfectly been aligned to the real (unknown) speech in some experiments. Unfortunately this seems to correlate with false detection rate (CC equals: 0,6).

Fig 2 presents the exact results of an exemplary analysis. The analyzed speech has been manually replicated four times for the sake of observing method correctness for the relating fragments of speech. As presented in the upper chart, the best matches indicated by markers placed around the bottom axis, are in the area close to the place where the wanted word is spoken. These markers show eleven areas out of twelve occurrences of the word in the analyzed speech. Four markers point faulty and four other markers are redundant, because of pointing the area being already pointed.

The presented markers come from the quality assessment of the corresponding matches, which is presented entirely in the medium chart.

The bottom chart of this figure shows the costs computed during classification stage. Path score plot presents the v vectors of the corresponding path P, while the green stars present chosen subsequences extracted from M.

Table 1. Overall results by speech source

	Detected words	No detection	False detection
Real speech	**82,92%**	17,08%	**56%**
min	*37,5%*	*0%*	*26,7%*
mode	*91,7%*	*0%*	-[4]
max	*100%*	*62,5%*	*75,7%*
Skewness	*-1,4*	*1,4*	*-0,6*
TTS	**74,17%**	25,83%	**41,76%**
min	*50%*	*0%*	*0%*
mode	*50%*	*50%*	*50%*
max	*100%*	*50%*	*66,7%*
Skewness	*0*	*0*	*-1,1*
Standard Error of the Mean: ~6%			

Chosen fragments of the analyzed speech during the searching for the word ‘senator’ have been presented in Fig 3. Times in the titles of each charts indicate real time range related to the speech presented in the upper chart of Fig 2. Presented steps 3, 9 and 16 show the best matches for the word ‘senator’ found in analyzed speech. Time steps of the best matches are not presented for the sake of readability. Although this outline shows that length of matches are different (i.e. the red stripes vary in length).

Steps 2, 8, 10 and 15 show larger sections of the fragments with silence in speech. Normally DTW algorithm includes this in the alignment path, causing matches that not necessarily carrier important information.

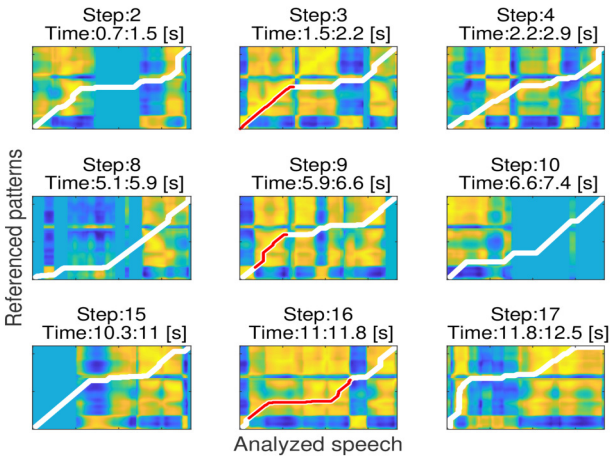

Fig 3. Operation of the discussed method presented on the selected fragments of analyzed speech. White stripes show optimal alignment paths between referenced pattern and analyzed fragment. Red stripes show the best matches selected after classification stage.

[4] Only unique values were observed.

During performing the experiments using TTS query, some method variables have been recalculated, although during experiments with different sessions of the corpus, all variables haven't been changed.

IV. Conclusion

Results of the work presented in this paper are satisfactory, but the overall performance, as comparing to the original application of the method to the English language [6], is lower (especially for TTS-generated queries), which shall be further investigated. Possible improvement of the performance the author sees in employing formant frequencies analysis in the verification step of the method, as it is described in [7].

Additional study on TTS generation for the Polish language and its influence to the detecting properties of the method shall also be further investigated.

The method has many variables which are depended on the analyzed data. The optimization of the variables values has to be done according to applications.

References

[1] J. Sas, A. Żołnierek, "Pipelined language model construction for Polish speech recognition" in International Journal of Applied Mathematics and Computer Science, vol. 23, no. 3, 2013, pp. 649-668, DOI: 10.2478/amcs-2013-0049

[2] D. Koržinek, Ł. Brocki, "Grammar Based Automatic Speech Recognition System for the Polish Language" in R. Jabłoński, M. Turkowski, R. Szewczyk (eds), Recent Advances in Mechatronics, Springer, Berlin, Heidelberg, 2007, ISBN 978-3-540-73956-2, pp. 87-91, DOI: 10.1007/978-3-540-73956-2_18

[3] M. Ziółko, J. Gałka, B. Ziółko, T. Jadczyk, et al., "Automatic Speech Recognition System Dedicated for Polish" in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2011, pp. 3315-3316.

[4] B. Ziółko, T. Jadczyk, D. Skurzok, P. Zelasko, et al, SARMATA 2.0 Automatic Polish Language Speech Recognition System, Conference: Interspeech 2015, Dresden, Germany, 2015.

[5] M. Pol, T. Walkowiak, M. Piasecki, "Towards CLARIN-PL LTC Digital Research Platform for: Depositing, Processing, Analyzing and Visualizing Language Data" in I. Kabashkin, I. Yatskiv, O. Prentkovskis (eds), Reliability and Statistics in Transportation and Communication, RelStat 2017, Lecture Notes in Networks and Systems, vol 36, Springer, Cham, 2018, pp. 485-494, DOI: 10.1007/978-3-319-74454-4_47.

[6] Ł. Laszko, "Word detection in recorded speech using textual queries", Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, 2015, pp. 849-853, DOI: 10.15439/2015F341.

[7] Ł. Laszko, "Using formant frequencies to word detection in recorded speech", Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, 2016, pp. 797-801, DOI: 10.15439/2016F518.

[8] D. Koržinek, K. Marasek, Ł. Brocki, K. Wołk, "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, number 136, Linköping University Electronic Press, Linköpings universitet, 2017, pp. 54–62.