DOI: 10.15439/2015F341
ACSIS, Vol. 5

Word detection in recorded speech using textual queries

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

Abstract—**The paper presents unsupervised method for word detection in recorded spoken language signal. The method is based on examining signal similarity of two analyzed media description: registered voice and a word (textual query) synthesized by using Text-to-Speech tools. The descriptions of media were given by a sequence of Mel-Frequency Cepstral Coefficients or Human-Factor Cepstral Coefficients. Dynamic Time Warping algorithm has been applied to provide time alignment of the given media description. The detection involved classification method based on cost function, calculated upon signal similarity and alignment path. Potential false matches were eliminated in the algorithm by comparing costs of the path subsequences to a threshold value. The results of the work could provide incentives to build affordable commercial or non-commercial solutions for specific and multilingual applications.**

Index Terms—**Speech processing, speech analysis, pattern matching, keyword search, audio information retrieval**

I. Introduction

CURRENTLY one can observe increasing use of methods and techniques of digital processing of sound information for simple daily tasks. Many of these methods and techniques have been implemented in various types of mobile devices, having as a matter of principle, relatively small memory resources and being not efficient enough to fulfill their requirements to full extent. Two common tasks in this field are connected with speech recognition and speech synthesis. Due to the limited resources two models, especially in recognition, are being observed: low quality local processing for specific usage, high quality remote (Internet service-oriented) processing[1] for wider usage. Regarding the described trend, in this paper the combination of both approaches were being adopted to word detection in recorded speech task.

Word detection relates to problem of searching for a given word in a speech medium (container or stream). In this paper only the solid container, such as: WAV, MP3 is concerned. The detection is usually given by the two [1] coupled values: position in medium (i.e. time code) and quality ratio. This problem is also recalled in contemporary literature as “keyword spotting[2]” (KWS) [2], [3] or “spoken term detection” [4] and usually refers to continues, unconstrained vocabulary speech.

Classical solutions to this problem address two-step-supervised approach where models such as hidden Markov model (HMM) or support-vector machine (SVM) are trained like in a typical automatic speech recognition (ASR) system, using Large Vocabulary Continuous Speech Recognition (LVCSR) methods [5] at the first step [6]. At this step the speech signal is divided into segments of equal-size, from which speech features are extracted. At the second step, appropriate algorithm is employed to determine the type of signal present in each segment.

Different approaches like this described in [1] base on the fact that for some applications it is not possible to have model trained, either due to lack of relevant training data or due to time-specific limitations. Moreover as maintains [7] by exploiting the structure of repeating patterns within the speech signal, unsupervised recognition task is made possible directly from an un-transcribed audio stream.

Under the concept of the unsupervised matching process lay suitable speech features and a classification strategy. Speech signal cepstrum-based features like Mel-Frequency Cepstral Coefficients (MFCC) are those used extensively in nowadays ASR [8]. Interesting study on MFCC and its two siblings can be found in [9]. However that work hasn’t considered Human-Factor Cepstral Coefficients (HFCC), which had been introduced in [10] one year later. In [1] HFCCs are described as the closest to human perception system and therefore seen as more robust in this task.

Concerning classification strategy mainly two are considered: HMM with Viterbi algorithm [2] and Dynamic Time Warping (DTW) [1]. Advantages and disadvantages of

This work was supported by Cybernetics Faculty (Military University of Technology, Warsaw, Poland) under the grant no. RMN/953/2014

[1] See: www.nuancehealthcaredeveloper.com, as an example of speech recognition/synthesis service and www.shazam.com, as an example of music excerpts recognition service. For legal notice see footnote 3.

[2] Spotting task is strictly connected with pattern matching but usually without explicit requirement for further (a posteriori) verification. The word “keyword” usually occurs within a clause (not alone) or/and describes its context.

978-83-60810-66-8/$25.00©2015, IEEE

these approaches are discussed in [11]. Further in this paper the approach based on Dynamic Time Warping (DTW) for pattern matching has been chosen, as it does not require any modeling or training as compared to the HMM, but still enables one to mitigate temporal differences.

II. PROBLEM STATEMENT

A. *Scenario and speech features*

In the paper the following scenario is considered. To an operator (a user) a container with recorded speech is provided. The operator has to search the speech content for the existence of specific words. The words are either not known in advance or are changing frequently. Moreover the recorded voice origin and its language is out of operator's knowledge. The sound quality is low and its characteristics (especially the environment) are changing during analyzed period like in live telephone conversation. Still the detecting process is time-sensitive as the existence of specific words will result in more detailed examination and perhaps in appropriate human activity (like calling police or fire brigade).

The presented scenario provides considerably limited usage of classical LVCSR methods because of too little knowledge of speech signal. The approach speculated here is directed to unsupervised methods, resulting in coarse detection, connected either with user interaction (i.e. hearing) or involving precise detecting methods.

Proposed approach applied to the scenario has been taken from [1] but the innovative reference queries strategy has been proposed in this paper. The approach assumes at this point the choice of appropriate speech signal features. In the research two types of feature vectors have been used:

- Mel-Frequency Cepstral Coefficients (MFCC),
- Human-Factor Cepstral Coefficients (HFCC).

MFCCs has been computed according to the following algorithm:

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

2) each segment has been processed by short-time Fourier transform (STFT) with length of 51 ms and step size of 10 ms;

3) then the triangular filter bank has been developed with 40 equally spaced mel-scale center frequencies f_i, $i = 1,...,40$ and with uniform bands controlled by the neighbor center frequencies $f_{i\pm1}$ (see Fig. 1);

4) in this step the actual filtering (or spectral smoothing) has been done, by multiplication of each STFT segment (representing magnitude spectrum) with magnitude spectrum of bands for MFCC;

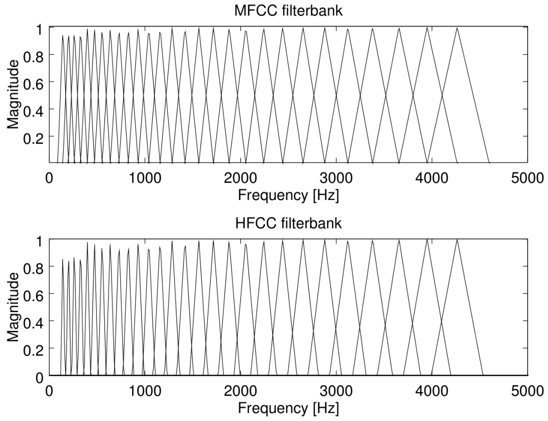

Fig. 1. Exemplary filter banks for MFCC and HFCC. HFCC filters corresponding to MFCC filters have narrower bandwidth, determined by (1).

5) the result has been then decorrelated using Discrete Cosinus Transform (DCT). Finally only 15 the most decorrelated vectors (MFCC coefficients) have been kept.

The concept of Human-Factor Cepstral Coefficients introduced in [10] employs congenial algorithm to the abovementioned. The essence of HFCC lays in filter design stage. In MFCC filter bands are determined by the spacing of the center frequencies which are equally distributed in mel frequency scale. In HFCC design filter center frequencies are still equally spaced in mel frequency scale, but unlike in MFCC filter bandwidth is treated as a parameter. This parameter determines filter bands' endpoints (cut-off frequencies) using the following measure, called Equivalent Rectangular Bandwidth (ERB) [10]:

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

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

As a result of using the approach, HFCCs are perceived as a better approximation of human auditory model [1], [10] than MFCCs and by incorporating appropriate scaling to ERB, by some factor, HFCC-based speech recognition can be under some circumstances treated as robust speech features to chosen applications [12].

B. *Textual query*

For the recalled scenario two reasonable approaches are emerging:

1) Profile (or design and implement new) ASR system: write down a set of words likely to be searched for, gather relevant training set, then prepare, train and verify model of each word, according to chosen pattern learning method. Set the system to recognize only selected words from the trained set. In case of positive recognition register the time of related speech segment.

2) Exploit Text-to-Speech (TTS) system to generate synthetic voice from the text (query). Transform the produced speech query to chosen speech feature's space (the query becomes a pattern). Read a chunk of speech signal from the given source, transform it to the same speech feature space and apply appropriate classification strategy.

In case of detection (pattern matched) of the given word register the time of related speech segment.

These approaches shall be amplified (or duplicated) to reflect language variations assumed in the scenario. Implementation of the selected approach could then take advantage of parallel processing techniques and search for the same word, translated to several languages, in the given speech signal.

The second approach makes impression of being much more innovative and promising than the first one. While using TTS there is rather no limitation in producing new query, in ASR-based approach this task will be probably the most demanding, and resource consuming. Supporting argument is the existence of publicly available, free of charge online translators with speech synthesis features and accessible APIs, e.g.[3]: Google Translate, Bing Translator, Yandex Translate, etc. Based on this argument several queries can be created in less than a second.

C. Similarity and alignment path

Dynamic Time Warping is a known and still used with success speech classifier [3]. It is popular for readability of implementation and analysis as well as for relatively high recognition accuracy similar to HMM.

In the overall look on DTW used in speech recognition it is to compare two feature vectors of different length (analyzed voice and the reference pattern) and to find an optimal alignment path P of both by stretching them with respect to time. P is usually calculated upon the local distance matrix (similarity matrix) from lower left corner to upper right corner of the matrix. Optimal means here the lowest cost path P for passing from one point of matrix to another within given constraints. Application of DTW to exemplary speech features vectors is presented in Fig. 2. The similarity refers to speech signal of the same phrase, spoken by the same speaker. Double reduced utterance tempo is observed in analyzed voice part.

Building similarity matrix $D_{A,R}$ where A stands for analyzed voice feature vector and R stands for referenced pattern feature vector, is the first step considered in speech classification. Feature vector consists for either MFCC or HFCC coefficients computed for segments $s_1 \ldots s_N$. Individual element $d(a,r)$ of similarity matrix, where a,r stands for specific element of vector A and vector R respectively, is given by inner product [6]:

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

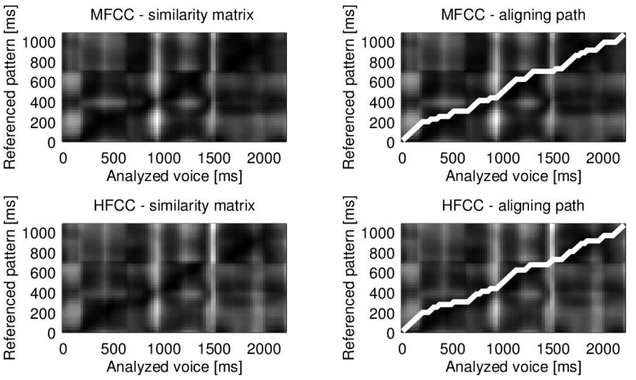

Fig. 2. Exemplary similarity matrices for MFCC and HFCC as well as corresponding aligning paths, received after applying DTW. The difference in image contrast between MFCC and HFCC indicates a higher MFCC features fuzziness.

In the following step DTW algorithm is exploited to perform cost path computation. The algorithm is two-staged[4]. At the first stage the calculation of an accumulator $C_{A,R}$ is performed (where C is of size D). The accumulator is a structure that contains at each of its point $c(a,r)$ the value of accumulated lowest transition cost to this point from its neighbors, including the cost of lowest transition to the neighbors from theirs consequent neighbors until the starting point $c(1,1)$, retaining directional constraints, according to the recursion:

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

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

At the second stage the optimal aligning path P is created. Its creation is based on accumulator traceback, starting from its last point $c(N_A, N_R)$ and ending in point $c(1,1)$ recursively by searching across all allowable predecessors to each point. Because each point holds the value of the lowest transition cost to itself, the actual calculation of the path is based on choosing the next point upon the minimal value.

Presented algorithm is a type of conventional Dynamic Time Warping and it is regarded as slow and memory consuming [11], especially for aligning large sequences, therefore a practical usage could engage its specific modifications, like a multiscale approach [14]. Nevertheless in this for DTW has been used in the research described in the paper.

D. Classification and verification

DTW presented in preceding paragraph refers to global alignment of one feature vector to another (in time domain), such as these presented in Fig. 2 (on the right). In general this is not enough to classify the part of analyzed voice as detected word, with regard to referenced pattern. Moreover

[3] Author of the paper would like to strongly accent having nothing in common with the companies whose products were listed. The author is far from rating, comparing and criticizing these products. Names of the products have been presented in this paper only in relation to the contemporary, publicly available technology, not for marketing purposes.

[4] The DTW algorithm used in the reported research, is based on the code originated with the project described in [13].

as commonly $N_A >> N_R$ holds, additional matching procedure shall be applied. The matching procedure is presented in figure Fig. 3.

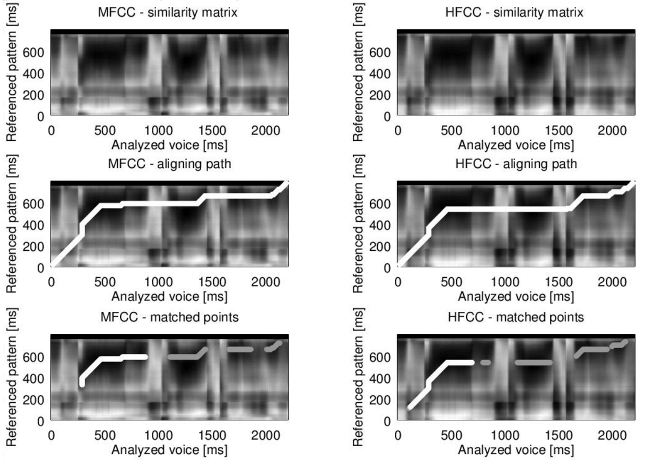

Fig. 3. Pattern matching procedure. Upper images present computed similarity between the pattern (the word "school" - synthesized woman's voice) and analyzed voice (recorded man's voice "school is closed today"). Images in the middle present global alignment path. Bottom images present resultant match: white strip is for the best match, gray strips are for remaining matches.

After computation of aligning path P, the points $p_1 ... p_{N_P}$ lying on the path are being assigned weight values v based on referring points of matrix $D_{A,R}$ and a path threshold T_P, satisfying inequality (4).

$$T_P \le v := 1 - d \le 1 \tag{4}$$

T_P controls the number of points suspected to indicate detected words. In Fig. 3 they are presented in the bottom images (as gray strips). As there could be several alleged word occurrences detected, indicated by subsequences: $p_{k_1}^{(l)} ... p_{k_{N_P}}^{(l)}$, $l = 1,2...$, the verification step is executed. This step consists of computing Longest Common Subsequence (LCS) with maximization criterion of cumulative weights of subsequences, i.e. the longest subsequence with the highest weights sum wins. It is worth noting that the minimal cumulative cost for a subsequence should be restricted by the second threshold value (called here sequence threshold), controlling the number of possible word reoccurrence.

As a result of matching procedure, assuming only one occurrence of the searched word, the best match is projected to the analyzed voice time domain (see white strips in the bottom images of Fig. 3) and, according to the scenario assumed in research, presented to the operator as a speech signal.

III. UNSUPERVISED KEY DETECTION

A. Algorithm

On the basis of literature search, especially [1], [2], [3], [11], [12] and performed experiments, the following algorithm has been proposed to unsupervised key detection.

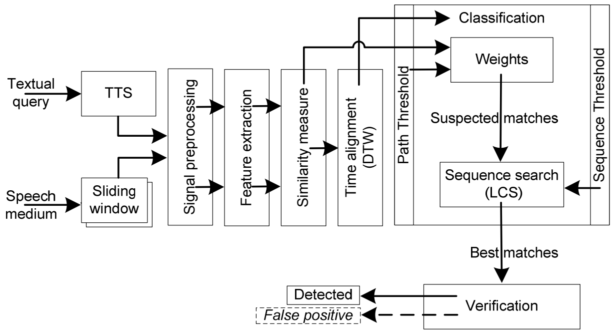

Fig. 4. Unsupervised key detection algorithm.

Description of respective processing blocks, presented in Fig. 4 is the content of the previous paragraphs, except of the sliding window and signal preprocessing. Sliding window block represents the abstract of feeding the model (or the program etc.) with respectively short signal, which the model is able to process. The size of the window is to be determined before the start of examination. It is assumed to be the function of size of the query.

Signal preprocessing is understood as applying standard digital signal processing techniques, at least but not limited to: silence cancellation, resampling and pre-emphasis filtering.

B. Experiments

A series of preliminary experiments have been conducted on the basis of methods and techniques of speech signal processing addressed in the paper, with regard to the presented algorithm. The target was to detect a word in a given speech medium by examining signal similarity of two analyzed media description: registered voice and referenced pattern, where the descriptions were given by a sequence of MFCC or HFCC coefficients.

Dedicated research material has been prepared, which consists of five short (from 1 – 5 seconds) sentences in English language: spoken by one man (natural speech) and synthesized by six (free of charge) TTS systems with fourteen men voices, and nine women voices. This material has been stored on a hard drive in the WAV containers.

The queries have been produced online by three TTS systems, different from these used in preparing research material. One TTS was the part of local operating system while two others were available through the World Wide Web via HTTP protocol. Nevertheless query generation was taking less than one second.

The research material as well as the queries were of the following (different[5]) properties: PCM (lossless) codec, one channel, 16 bits per sample, sampling frequency: 8000 – 22050 Hz, bit rate: 256 – 352 kbps. During examination all used sounds were resampled to 8000 Hz. The model (Fig. 4) was configured according to the guidelines from paragraph II.A. For HFCC ERB scaling factor of value 3 was used.

[5] During material acquisition (except of the natural voice recording) there was not possible to influence sound properties.

Experiments have been conducted according to the following strategy: selected word to find (textual query) has been sent to the TTS system to obtain speech signal, the signal then has been read by program and compared with the entire research material according to the algorithm (Fig. 4). This has covered both situations: the existence and inexistence of the word in the examination set.

C. Results

Overall results have been presented in Table 1. "No detection" phrase used in the results means the percent of false negatives (lack of detection, when the word actually existed in the analyzed signal). Higher detection rate for HFCC, as well as lack of false negatives for these features is noticeable. Nevertheless MFCC present lower false detections.

TABLE 1. OVERALL RESULTS BY SPEECH FEATURES

	Detected words	No detection	False positive
MFCC	82,43%	4,05%	13,51%
HFCC	85,14%	0,00%	14,86%

The results showed that the unsupervised detection of word in a given set is possible with relatively high detection rate. It is worth noting the lack of "no detection" results when using HFCC.

In Table 2 the discrimination on the basis of speech source has been presented. The results give the overview that either male or female synthesized voice can be with success use to detect words in speech.

TABLE 2. MEAN RESULTS BY SPEECH SOURCE

	Detected words	No detection	False positive
Real speech	95%	0%	5%
TTS (male)	80%	15%	15%
TTS (female)	70%	12,5%	17,5%

HFCC based detection gives better overall results in comparison with the MFCC method, but the results of the research don't allow to make general statement for this.

IV. CONCLUSION

Results of the work presented in this paper are satisfactory, but still far from industrial standard. It is difficult to make direct comparisons with other related works as this work has been conducted partly on synthesized material. The following work shall be considered: (1) applying multiscale approach (like presented in [13]) for bypassing fully different signal parts, and analyzing these initially consonant parts on higher resolution, (2) parallel processing: for searching the multilingual queries, as well as increasing detection reliability based on information fusion.

Moreover additional study on less speaker-dependent features should be done, to limit the number of false positives.

The main problem in the approach (especially while applying DTW and LCS) that cannot be fully circumvent is the determination of the threshold values. The mitigation of the problem could employ computation of the values based on query or signal properties.

In general view of this work, its results could provide incentives to build affordable multilingual commercial or non-commercial solutions for specific applications.

ACKNOWLEDGMENT

Author would like to thank the team (of the year 2010) of KOM department, Fraunhofer FKIE, Wachtberg, Germany, especially Dirk, Frank and Paul (NUWC), who gave me the inspiration of the topic, as well as the leader of the department Dr. Markus Antweiler, who enabled my scientific internship at FKIE.

REFERENCES

[1] D. von Zeddelmann, F. Kurth, and M. Müller, "Perceptual audio features for unsupervised key-phrase detection," Proc. ICASSP2010, 2010, pp. 257-260, DOI:10.1109/ICASSP.2010.5495974.

[2] S. Tabibian, A. Akbar, B. Nasersharif, "A fast search technique for discriminative keyword spotting," Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on, pp.140-144, 2-3 May 2012, DOI:10.1109/AISP.2012.6313733.

[3] M. Sigmund, "Search for Keywords and Vocal Elements in Audio Recordings", Elektronika ir elektrotechnika, ISSN 1392-1215, vol. 19, no. 9, pp. 71-74, 2013

[4] V. Mitra, J. van Hout, et. al., "Feature Fusion for High-Accuracy Keyword Spotting", Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 7143-7147 2014

[5] A. Manos and V. Zue, "A segment-based word spotter using phonetic Filler models," in proceeding of ICASSP, pp. 899–902, 1997.

[6] W. Kwiatkowski, "Methods of automatic pattern recognition" (in polish: „Metody automatycznego rozpoznawania wzorców"), Instytut Automatyki i Robotyki, WAT, ISBN 83-912747-7-2, Wydanie I, Warszawa 2001, pp. 185-191; 118-121 .

[7] A. S. Park and James R. Glass, (Cited in [1]) "Unsupervised pattern discovery in speech," IEEE Trans. on Audio, Speech and Language Processing, vol. 16, no. 1, pp. 186–197, 2008.

[8] D. Eringis, G. Tamulevičius. "Modified Filterbank Analysis Features for Speech Recognition", Baltic J. Modern Computing, Vol. 3 (2015), No. 1, 29-42.

[9] B. J. Shannon, K. K. Paliwal, "A Comparative Study of Filter Bank Spacing for Speech Recognition", Microelectronic Engineering Research Conference 2003.

[10] M. D. Skowronski and J. G. Harris, "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition," The Journal of the Acoustical Society of America (JASA), vol. 116, no. 3, pp. 1774–1780, 2004.

[11] M. S. Barakat, C. H. Ritz , D. A. Stirling, "Keyword spotting based on the analysis of template matching distances", Signal Processing and Communication Systems (ICSPCS), 2011 5th International Conference on, pp.1-6, 12-14 Dec. 2011, DOI:10.1109/ ICSPCS.2011.6140822.

[12] R. Wielgat, T. P. Zieliński, T. Potempa, A. Lisowska-Lis, D. Król, "HFCC based recognition of bird species", In: Signal Processing: Algorithms, Architectures, Arrangements, and Applications, ISBN-13 978-83-913251-8-6, pp. 129–134, Poznań 2007.

[13] R. Turetsky and D. Ellis, "Ground-Truth Transcriptions of Real Music from Force-Aligned MIDI Syntheses", 4th International Symposium on Music Information Retrieval ISMIR-03, pp. 135-141, Baltimore, October 2003.

[14] A. Zinke and D. Mayer, "Iterative Multi Scale Dynamic Time Warping", In: Computer graphics technical reports, CG-2006/1, ISSN 1610-8892, Universität Bonn, 2006.