Proceedings of the Federated Conference on Computer Science and Information Systems pp. 797–801
DOI: 10.15439/2016F518
ACSIS, Vol. 8. ISSN 2300-5963

Using formant frequencies to word detection in recorded speech

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 considers increasing the precision of detection of words in unsupervised keyword spotting method. 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 descriptions. 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 applying two-staged verification, using the Longest Common Subsequence algorithm and analyzing formant frequencies of eleven English monophthons. The use of formant frequencies at the stage of verification increased overall detection precision by about 10% as compared to original algorithm.**

***Index Terms*—keyword spotting, formant frequency analysis, pattern matching, audio information retrieval**

I. Introduction

INCREASING use of digital sound processing methods to simple daily tasks is currently very popular due to widespread availability of mobile devices having implemented this type of methods. Regarding this trend [1] considers an approach that could be used to detect words in recorded speech of unknown language without training, by using publicly available, free of charge online translation services with Text-To-Speech support e.g.: Google Translate, Bing Translator, Yandex Translate[1].

The problem of word detection consists in searching for given words in a speech medium, which is either solid container or a stream. The detection is usually given by the two coupled values [2]: time code of the beginning of the word and a quality ratio. This problem in contemporary literature is usually called "keyword spotting[2]" (shorten as KWS) [3], [4].

Classical solutions to this problem address 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]. In consequence the speech signal is divided into segments of equal-size, from which speech features are extracted. Next, an appropriate algorithm is employed to determine the type of signal in each segment. As a result, recognized words, together with the corresponding indexes are stored in a database. Then, a text query is performed within the indexed data [6].

Based 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, different unsupervised approaches to the problem have been developed [2], [7].

Under the concept of unsupervised matching lay suitable speech features and a classification strategy. The approach presented in [1] employs cepstrum-based features: Mel-Frequency Cepstral Coefficients (MFCC) and Human-Factor Cepstral Coefficients (HFCC). As for classification strategy that approach points Dynamic Time Warping (DTW) algorithm.

However results of applying the method shows relatively high overall rate of false positives: 13,51% for MFCC and 14,86% for HFCC.

In this paper, the author propose an approach to minimize this insufficiency, by adding additional verification stage, based on the analysis of formant frequencies. The motivation for this came from [12]. Using formant frequencies analysis, as shown below, has a positive influence on the results, but limits the versatility of the KWS method to specific language only. Different improvement techniques used for KWS could involve combining of multiple features, like described by Mitra et al. in [5].

II. Problem statement

A. Method background

This paper considers the same use case as in [1]. In this approach the KWS method supports human operator in searching for specific words in a given speech medium. For

This work was supported by Cybernetics Faculty (Military University of Technology, Warsaw, Poland) under the grant no. RMN/765/2015

[1] Names of the products have been presented in this paper only in relation to the contemporary, publicly available technology, not for marketing purposes.

[2] Also „spoken term detection" (shorten STD).

978-83-60810-90-3/$25.00©2016, IEEE

this scenario, the sound queries are synthesized directly from text. The method gives coarse detection, prior to involving precise detecting methods or just hearing by the operator.

B. Speech features model

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) and Human-Factor Cepstral Coefficients (HFCC). MFCCs have been 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 (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 , $i = 1,\dots,40$ and with uniform bands controlled by the neighbor center frequencies $f_{i\pm1}$;

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

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

The same model has been applied to Human-Factor Cepstral Coefficients, with a change in point 3). In HFCC, center frequencies are still equally spaced in mel frequency scale, but unlike MFCC filter bandwidth is treated as a parameter, which determines filter bands' cut-off frequencies, using measure called Equivalent Rectangular Bandwidth (ERB) [8]:

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

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

C. Textual query

In the presented method Text-to-Speech (TTS) system is exploited to generate synthetic voice from a text query. Next, the query (pattern) is transformed to chosen speech feature space. Then a chunk of speech signal from a given source is read. This chunk is transform to the same speech feature space. Then a classification strategy is applied. In case of the pattern matched, time code of the corresponding sound segment is registered.

Using textual query and TTS make it easy to extend the approach to reflect language variations assumed in the scenario, to search for the same word translated to several languages [1].

D. Similarity and time alignment

DTW is used in the method to compare two feature vectors of different length (analyzed voice and the reference pattern) and to find an optimal alignment path P of both. P is usually calculated upon the local distance matrix (similarity matrix) from the minimal indexes (usually lower left corner) to maximal indexes (usually 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. For details of applying DTW to exemplary speech features vectors, see [1].

Building similarity matrix $D_{A,R}$ where A stands for analyzed voice feature vector and R stands for reference pattern feature vector, is the first step considered in speech classification. Feature vector consists of either MFCC or HFCC coefficients computed for segments $s_1 \dots s_N$. 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:

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

Next, the two-staged cost path algorithm is executed. The first stage stands for the calculation of an accumulator $C_{A,R}$ (where C is of size D). The resulting structure contains at each of its element $c(a,r)$ the value of accumulated lowest transition cost to this element from its neighbors, including the cost of lowest transition to the neighbors from theirs consequent neighbors until the starting element $c(1,1)$. The computation retains directional constraints, according to the recursion:

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

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

The second stage stands for an optimal aligning of analyzed voice and the reference pattern. In this stage the path P is created. Its creation is based on the 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.

E. Classification and verification

After applying DTW, to proper classification an additional matching procedure is proposed in the method, see Fig. 1.

This procedure assigns weight values v based on referring points of matrix $D_{A,R}$ and a path threshold T_P to the path P , satisfying inequality (4).

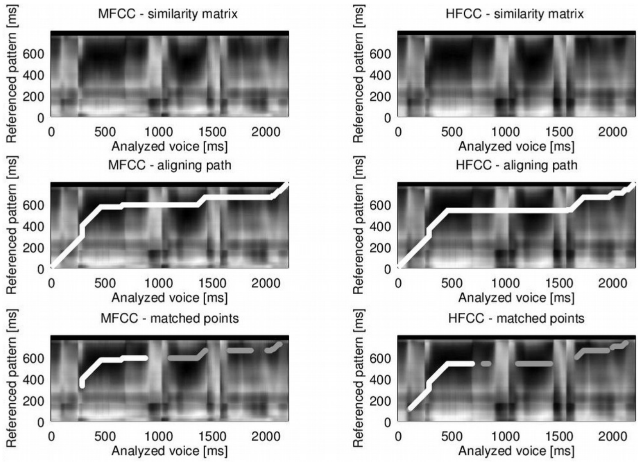

Fig. 1. Pattern matching procedure. Upper images present computed similarity between the pattern (the word "school" - synthesized woman's voice) and analyzed voice (recorded man's voice "school is closed today"). Images in the middle present global alignment path. Bottom images present resultant match: white strip is for the best match, gray strips are for remaining matches.

$$T_P \leq v := 1 - d \leq 1 \qquad (4)$$

where T_P controls the number of points suspected to indicate detected words.

Assuming, that several possible word could occurre, which is indicated by subsequences: $p_{k_1}^{(l)} \cdots p_{k_{N_P}}^{(l)}$, $l = 1,2\ldots$, the verification step is executed.

Originally this step consisted of applying Longest Common Subsequence (LCS) with maximization criterion of cumulative weights of subsequences, i.e. the longest subsequence with the highest sum of weights wins.

The number of possible word reoccurrence was controlled by sequence threshold value, which restricted the minimal cumulative cost for a subsequence.

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 (e.g. white strips in the bottom images of Fig. 1).

F. Format frequencies analysis

Conclusions from experiments described in [1] indicate that relatively high level of false positive results could be minimized after applying less speaker-dependent speech features. By following that suggestion and motivated by [12], in this paper the author propose to extend verification step with formant frequencies analysis. Formant frequencies are defined either as an acoustic resonance of the human vocal tract or more technically as local maxima of the envelope of the signal spectrum. As stated in [12] they are important in determining phonetic content of speech sounds, but they are not quite good speech features. This is because of, on one hand: their little speaker dependency (assuming the same speakers gender) and existence of cataloged form for specific language (usually including frequency range for a specific phonetic unit). On the other hand: their strict connection with high signal energy of phonetic units, like vowels, and unreliable measure of purely defined signals (silence, weak fricatives, etc).

Nevertheless in this paper it was hypothesized that knowledge of at least a part of speech segment will positively influence the quality of detection.

III. KWS SUPPORTED BY FORMANT FREQUENCIES ANALYSIS

A. Formants estimation

In the described research formant frequencies have been estimated only for English vowels, as for all other phones such frequencies either do not exist or are difficult to be identified. The following phonetic convention of the vowels has been adopted (see table 1) in the presented research.

TABLE 1. PHONETIC CONVENTION OF ENGLISH VOWELS SELECTED FOR THE RESEARCH[3]

Classification	*IPA[4] notation*	*Own[5] notation*	*Example*	*Choice*
short vowels	/ʌ/	a	cup	Yes
	/æ/	ae	cat	Yes
	/e/	e	bed	Yes
	/ə/	e_	about	No
	/ɪ/	y	hit	Yes
	/i/	i	happy	No
	/ɒ/	o	hot	Yes
	/ʊ/	u	good	Yes
long vowels	/ɑː/	aa	arm	Yes
	/ɜː/	ee	bird	Yes
	/iː/	ii	see	Yes
	/ɔː/	oo	call	Yes
	/uː/	uu	food	Yes

In English language, there are 5 vowels: <a, e, i, o, u>, but their pronunciation depends on a variety of factors, resulting in several distinguishable monophthongs[6]. Usually for phonetic analysis from 8 to 13 monophthongs are chosen [9]. According to [10] for further analysis 11 monophthongs have been chosen, the choice is marked in table 1, in the far right column[7].

TABLE 2. AVERAGE VALUES OF F_1, F_2 AND F_3 IN HZ [10].

	Male			Female		
	F_1	F_2	F_3	F_1	F_2	F_3
/iː/	280	2249	2765	303	2654	3203
/ɪ/	367	1757	2556	384	2174	2962
/e/	494	1650	2547	719	2063	2997
/æ/	690	1550	2463	1018	1799	2869
/ʌ/	644	1259	2551	914	1459	2831
/ɑː/	646	1155	2490	910	1316	2841
/ɒ/	558	1047	2481	751	1215	2790
/ɔː/	415	828	2619	389	888	2790
/ʊ/	379	1173	2445	410	1340	2697
/uː/	316	1191	2408	328	1437	2674
/ɜː/	478	1436	2488	606	1695	2839

[3] Own work based on [9] as well as other materials available at official IPA website: https://www.internationalphoneticassociation.org/

[4] International Phonetic Alphabet

[5] Own notation was used because of programming and results presentation reasons.

[6] Single and the smallest phonetic unit; pure vowel sound.

[7] The choice was caused by the most recent research found in this area, which published a comprehensive list of results.

B. Algorithm

In general overview the KWS algorithm discussed in this paper is presented in Fig. 2. Description of presented processing blocks can be found in [1], but the verification stage is described in details below. The input to the verification stage is given from the LCS algorithm, resulting in a few best matches (e.g. in the bottom images of Fig. 1 there is one best match colored white, the other stripes, colored gray, present remaining matches).

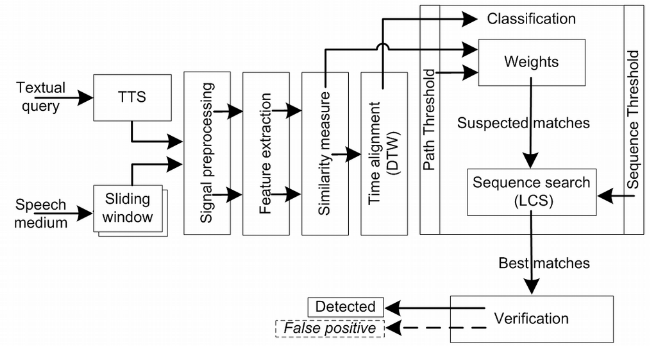

Fig. 2. Unsupervised key detection algorithm.

Let's assume there is only one best match and the verification stage is to decide if this match is a true detection or a false positive detection. Then the resultant match is subjected to verification, which is based on the analysis of formant frequencies: the reference pattern vector of formant frequencies and the analyzed voice vector of formant frequencies. It is worth noting that these two signals are after the whole procedure time-aligned, thus could be extracted and analyzed separately. Then for these excerpts the algorithm described in [11] is carried out to estimate formant frequencies.

As for the reference pattern excerpt the resultant formant sequences (exactly two formant frequencies F_1, F_2 have been chosen) are averaged and compared with the values described in table 2 to estimate appropriate monophthong.

The estimation of monophthong is performed by comparison of averaged estimated frequency F_1 from the excerpt, to all F_1 values from the table 2. The difference between the values (its absolute value), measured in the sense of Euclidean, is treated as the quality of the detection (monophthong cost). The smaller the difference the better the detection. The same estimation is done for F_2. As a result of this procedure monophthongs are detected (if any) for the reference pattern. From the collection of all detected monophthongs only the one with the smallest cost is chosen. The example of detection of /u:/ monophthong for the reference pattern is presented in the Fig. 4. on the vertical axis[8].

This estimation is carried out in an analogous manner to the analyzed voice excerpt. Finally, if the estimated monophthong for reference pattern matches analyzed voice monophthong, then according to the scenario assumed in the research, the excerpt with analyzed voice is played back to the human operator.

C. Experiments

A series of preliminary experiments have been conducted with regard to the presented algorithm. The target was to determine the influence of the proposed verification stage based on formant frequencies analysis to the quality of detection.

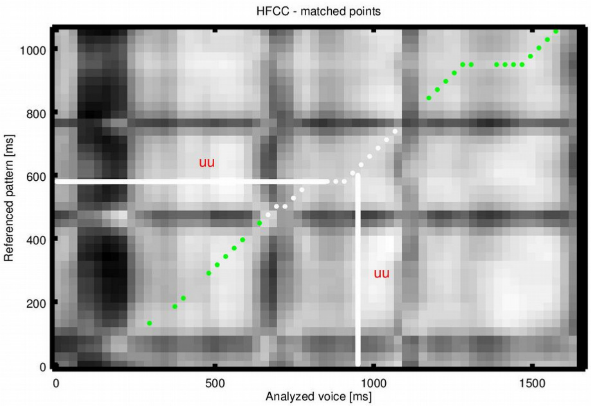

Fig. 3. An example of detecting monophthong /u:/ in the word "school". The dots represent optimal alignment path (as a result of DTW). White dots represent best matched sequence (as a result of LCS). While the white lines with the description "uu" represent the (averaged) position of the detected monophthong in the matched sequence.

The experiments have been conducted on the same research material as the original method [1], but only male voices have been chosen. Therefore the material consisted of five short (from 1 – 5 seconds) sentences in English language: spoken by one man (natural speech) and synthesized by six TTS systems with fourteen men voices. This material has been stored on a hard drive in the WAV containers.

The queries have been produced online by one chosen TTS system, different from these used in prepared research material. The TTS was accessed through the World Wide Web via HTTP protocol.

During examination all used sounds were resampled to 8000 Hz. The model (Fig. 2) was configured according to the same guidelines as in the original examinations.

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. 2).

D. Results

A series of preliminary results were obtained. They were compared to the results of the original method [1]. This allowed to determine the influence of the new verification approach on the quality of word detection. These results are shown in Table 3.

As it was hypothesized the percentage of false positives decreased. The decrease is about 40%. Surprisingly this had also positive effect to the percentage of detected words and had no negative effect on misses ("No detection").

The results showed that the unsupervised detection of word in a given set is possible with relatively high detection rate. Moreover, new verification method has given satisfactory results, approaching the method to industry

[8] It should be noted that the time position indicated by white lines has only illustrative value and not necessarily reflect the detection of a monophthong. This is due to successive averages made in the algorithm.

standards. Comparing the actual results with the original results of the method (MFCC: 82.43%, HFCC: 85.14%), it can be concluded, that MFCC recorded an increase of 8.32 percentage points, and HFCC recorded an increase of 8.66 percentage points. This gives an overall increase in precision of about 10% for both MFCC and HFCC features.

However taking into account small research material involved in the examination, these results do not allow for general statements.

TABLE 3. **Overall results by speech features**

	Detected words	**No detection**	**False positive (new method)**	**False positive (original method)**	**Overall increase in the elimination of false positives**
MFCC	90,75%	4,05%	5,19%	13,51%	38,41%
HFCC	93,8%	0,00%	6,2%	14,86%	41,72%

IV. CONCLUSIONS AND LESSON LEARNED

Results of the work shows that the inclusion in the KWS formant frequencies analysis, increases its quality. This partly proves the hypothesis that knowledge of a part of speech segment has a positive influence on the quality of detection. The main advantage of this approach is the elimination of false positives. However the presented approach limits the application of the method only to one language, due to the requirement of having a catalogue of formant frequencies. The requirement of possessing 3 formant frequency models (for male, female and children) for each language is also a strong one.

During the examination the author also encountered the problem of covering (overlapping) formant frequencies that lie in close proximity to each other (in the spectrum), also well known from the literature (see [12]). For example, for F_1 and F_2 in the spectrum only one peak is perceptible. In the described research it was noticed that generally F_2 overlaps F_1, therefore F_2 becomes F_1 and F_3 becomes F_2 in consequence.

One solution to this problem was to properly assign these frequencies as F_2 and F_3, leaving F_1 undetected (or arbitrary setting its value to 0).

Leaving the problem unresolved significantly deteriorates estimation of vowels, to the extent, that the result becomes random.

The problem in the presented approach (especially while applying DTW and LCS) that cannot be fully circumvent is the determination of the threshold values. According to literature search, the most popular technique for solving this problem in KWS, is to parallel true positive rate with false positive rate for several chosen threshold values, to create Receiver Operating Characteristic (ROC) and to find optimal threshold value by using graphical method.

REFERENCES

[1] 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

[2] 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.

[3] 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.

[4] 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

[5] 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

[6] J. Tejedor, D. T. Toledano, et al., “Access Spoken term detection ALBAYZIN 2014 evaluation: overview, systems, results, and discussion”, EURASIP Journal on Audio, Speech, and Music Processing (2015) 2015:21, DOI 10.1186/s13636-015-0063-8

[7] A. S. Park and James R. Glass, (Cited in [2]) “Unsupervised pattern discovery in speech,” IEEE Trans. on Audio, Speech and Language Processing, vol. 16, no. 1, pp. 186–197, 2008.

[8] 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.

[9] G. Hunter, H. Kebede, Formant frequencies of British English vowels produced by native speakers of Farsi. Societe Francaise d’Acoustique, Acoustics 2012, Apr 2012, Nantes, France

[10] D. Deterding (1997). The Formants of Monophthong Vowels in Standard Southern British English Pronunciation, Journal of the International Phonetic Association, 27, pp. 47-55

[11] R. Snell, F. Milinazzo, Formant location from LPC analysis data, IEEE Transactions on Speech and Audio Processing. Vol. 1, Number 2, 1993, pp. 129-134.

[12] J. Holmes, W. Holmes, P. Garner, Using formant frequencies in speech recognition, Eurospeech, Vol. 97, pp. 2083-2087, http://www.idiap.ch/~pgarner/pubs/holmes1997.pdf