iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VIII **Month of publication:** August 2017

DOI: http://doi.org/10.22214/ijraset.2017.8209

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VIII, August 2017- Available at www.ijraset.com

Spoken Arabic News Classification Based on Speech Features

Naser S. A. Abusulaiman[1], Mohammed A. Alhanjouri[2]
[1, 2] *Islamic University of Gaza (IUG), Gaza, Palestine*

Abstract*: One of the most important consequences of what is known as the "Internet era" is the widespread of varied electronic data. This deployment urgently requires an automated system to classify these data to facilitate search and access to the topic in question. This system is commonly used in written texts. Because of the huge increase of spoken files nowadays, there is an acute need for building an automatic system to classify spoken files based on topics. This system has been discussed in the previous researches applied to spoken English texts, but it rarely takes into consideration spoken Arabic texts because Arabic language is challenging and its dataset is rare and not suitable for topic classification. To deal with this challenge, a new dataset is established depending on converting the common written text (ALJ-NEWS) which is widely used in researches in classifying written texts. Then, keywords extraction method is implemented in order to extract the keywords representing each class depending on using DTW. Finally, topic identification, based on (MFCC, PLP-RASTA) as speech features and (DTW, HMM) as identifiers, is created using a technique that is different from the traditional way, using ASR to extract the transcriptions. Regarding the evaluation of the system, F1-measure, precision and recall are used as evaluation metrics. The proposed system shows positive results in the topic classification field. The F1-measure for topic identification system using DTW classifier records 90.26% and 91.36% using HMM classifier in the average. In addition, the system achieves 89.65% of keywords identification accuracy.

Keywords: Topic Identification, Natural language processing, Keywords extraction, speech segmentation, Speech classification HMM, DTW.

I. INTRODUCTION

Over the Internet enormous quantities of spoken files are existing. Indexing, retrieving and browsing spoken files are highly required and reliable approaches for these issues are really needed. A common organization mission is categorizing, or classifying various speech into different topics. The term spoken document classification is defined as the problem of labelling a speech segment with the suitable topic from a group of possible predefined topics. It is closely related to spoken document retrieval, where a query is given then a list of spoken files is returned in response to this query. Classification and retrieval can also be combined to improve the user experience [1]. A system for automatic spoken document classification can, for instance, be used in applications such as classification of broadcast news reports into topics like “economy” and “sport”, and classification of conversations into either “criminal” or “innocent” actions. Spoken files classification is an efficacious approach in order to manage and index the speech files. This task focuses on tagging a pre-defined topic (class) to a spoken file based on its contents. Topic classification or identification, which known as a single-label categorization process, supposes that a pre-detected group of topics has been established and each spoken file is categorized in order to be appropriate to only one topic from this group. Classification or identification is one of the important fields in machine learning searches, such as speech and language processing, data filtering, data cleaning, …, etc. This paper has a valuable significance as a result of the demanding classification of the huge number of electronic spoken Arabic data, which are available via the rapid and increasing growth of the Internet. Correspondingly the manual classification of such huge data is time and effort consuming. There are a number of classification algorithms applied to spoken data. Nevertheless, most of them go through two steps: transcriptions creation which outcomes from automatic speech recognition (ASR) model and the application of text classification algorithms to the transcriptions. This traditional algorithm has given good classification accuracy, yet it depends on ASR model accuracy. It is worth mentioning that accuracy of ASR model for Arabic language is not accurate as English language model. Spoken Arabic data classification is a problematic issue because of the complexity of Arabic language structure. Besides, reasonably eliminating ASR step increases the performance of classification. This paper provides a speech processing solution with the purpose of enabling content based access to spoken files and proposes a topic classification approach to spoken Arabic news (SAN). The paper main goal is to present a new methodology in order to design an automatic system which identifies the underlying topics discussed in spoken Arabic news files based on speech features directly.

II. LITERATURE REVIEW

Until the early 1990s, topic identification (TID) researches on spoken data had not been started seriously due to the lack of suitable spoken data. In Arabic language, the existing dataset for speech-based TID is significantly smaller. Reference [2] presents one of the most primitive researches in speech-based TID is done using only a small group of 510 speech monologues, with the length of 30-second, distributed over six dissimilar scenarios. Speech-based TID applications borrow most of the commonly used techniques in text processing area and adapted them in order to be applicable to the speech-based classification systems. Reference [3] introduces the general TID techniques used in written text applications. More additional descriptions and techniques can be found in a book chapter by [4]. One of the most common written text classification approach used in TID approach is the naive Bayes approach and it has been borrowed in speech-based systems too ([5], [6], [7] and [1]). Nevertheless, many of the topic classification methods are applied to written Arabic news and most of the studies and papers rarely discuss TID in spoken Arabic files. Spoken file processing faces many challenges and opportunities when compared to written text processing [8]. One of the rarely studies on TID based on spoken Arabic language that uses ASR capabilities is done by [9]. They propose an approach that aims to form an automated task-oriented Arabic dialogue system which is capable of determining the topic of spoken question asked by telecom provider customers. The system is built on an Arabic adapted CMU sphinx ASR. The best performance of suggested overall system is 76.4% accuracy with haphazard forest classifier given by Weka toolkit tested on 750 questions recorded by 30 speakers with the Palestinian dialect. Nonetheless, under resource-limited conditions, the manually transcribed speech needed for developing standard ASR systems can be severely limited or unavailable [10]. Reference [11] uses minimum classification error (MCE) training in order to enhance traditional approaches to TID. A fundamental element of novel MCE training methods is their capability to professionally implement jack-knifing or leave-one-out training to achieve enhanced models which generalize better to invisible data. Sizeable enhancements observed in TID accuracy using the new MCE training techniques. Reference [12] investigates alternative unsupervised solutions to obtain tokenization's of speech in terms of vocabulary of automatically exposed word-like or phoneme-like units, without relying on the supervised training of ASR systems. They prove that a convolutional neural network based framework for learning spoken text representations affords competitive performance compared to a standard bag-of-words representation. Another application of unsupervised acoustic unit discovery for TID of spoken texts is presented by [13]. The acoustic unit discovery method is built on a nonparametric Bayesian phone-loop model that segments a speech utterance into phone-like groups. The exposed phone-like (acoustic units) are additional fed into the conventional TID framework. Using multilingual bottleneck features for the acoustic unit discovery outperforms other systems that are built on cross-lingual phoneme recognizer.

III. PROPOSED METHODOLOGY

Obviously, the clearest way with the intention of performing TID process on speech collection is to handle these data by ASR system. The offered transcript resulted from the ASR system is passed directly to process them by contemporary text-based TID systems. In general, there is a need for a system that is able to process spoken files with limited (or zero) resources without a reliance on ASR model. Speech processing directly without a dependency on ASR results is proposed as an improbable alternative [14]. In the figure below, the main components of the presented topic identification methodology, without the necessity of ASR component, are described. The first one includes creating, categorizing and storing the dataset in SAN database. The second one is the processing of SAN. The processed SAN is used by the keywords extraction component, which is used to extract the most dominant keywords for each category. Then, the keywords extraction component puts them in an organized manner in which the selected KW database can be suitable. The last component is topic identification which uses KW database with the purpose of identifying the SAN clip with unknown topic label.

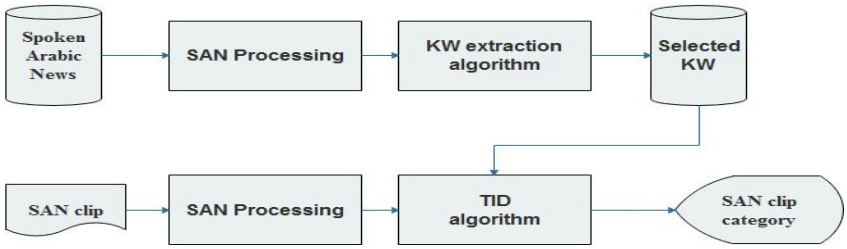


Fig. 1 General boxes for topic identification methodology

The first step in SAN identification system is SAN processing step, which includes two main processes: speech pre-processing and speech segmentation process. The pre-processing steps are performed on a speech signal with the purpose of emphasizing the effective frequency. Original speech signal is normalized to be at some programmed amplitude range, scale all values in one SAN clip to its maximum. Then, the normalized speech signal is pre-emphasized by amplifying the higher frequency components of the speech signal.

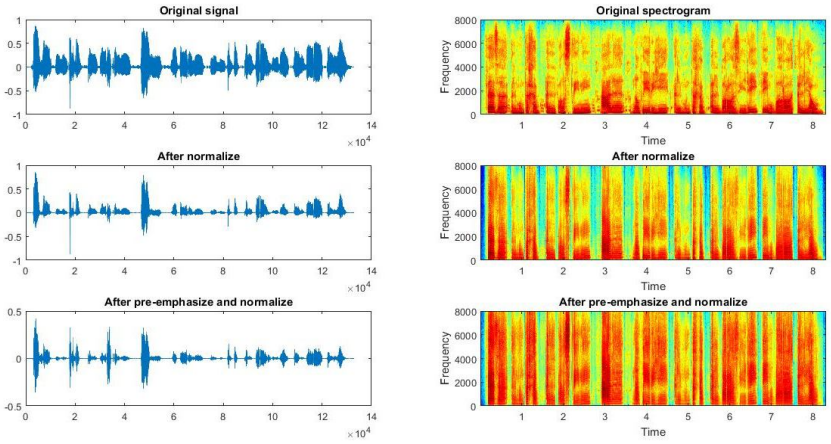


Fig. 2 Applying pre-emphasize and normalization on speech signal.

After pre-processing steps, the processed SAN clip is divided into regions, each of which corresponds to only one word. This process is known as speech segmentation at the word level. It can be achieved by a change point detection method or a word divider, such as the space. The co-articulation, a phenomenon which may happen between adjacent words as within a single word, is the main challenge in speech segmentation across languages especially in Arabic language. In this paper, a novel algorithm to segment SAN clip into its lexical items based on distinctive boundary normalized features is proposed. The feature vectors are extracted from the processed SAN clip using Relative Spectral Transform - Perceptual Linear Prediction (RASTA-PLP) method [15]. With some calculations which are performed on each feature vector, the frames (frame length is 10ms) with the lowest value are identified and the space between each frame are obtained by the lowest value. The main idea of the proposed segmentation methodology is the signal change detection, which determines when significant continuous changes occur in the SAN signal and without forgetting the suitable length of the word. The detected changes define the boundaries of the resulting segments. The proposed SAN clip segmentation system has six major steps, as shown in Figure 4.3.

A. Speech Pre-Processing
B. Speech Features Extraction (RASTA-PLP)
C. Histogram Computation (Normalization of Features 0 or 1 depends on some factors)
D. Define boundaries based on some criteria (Frame length and number of continuous frames with 1's values (neighbors)
E. Extract the Mel-frequency Cepstral Coefficients (MFCC) feature for each identified word
F. Post-Processing

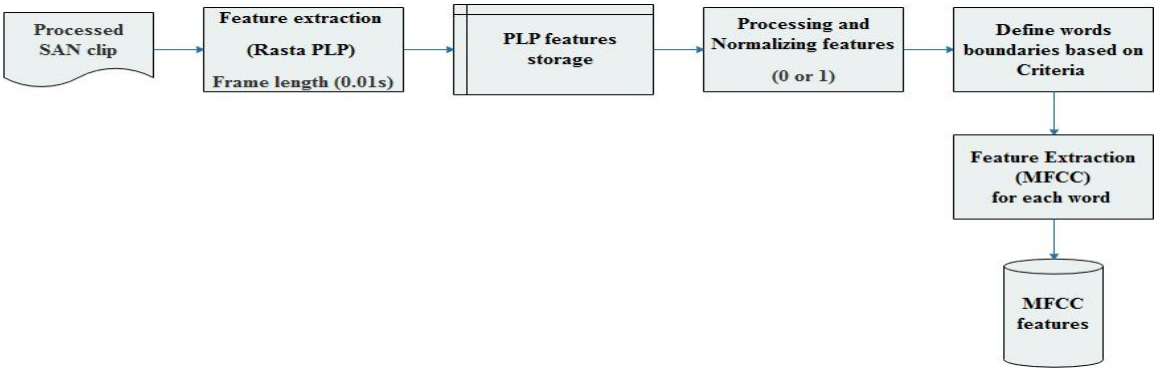


Fig. 3 Segmentation processes.

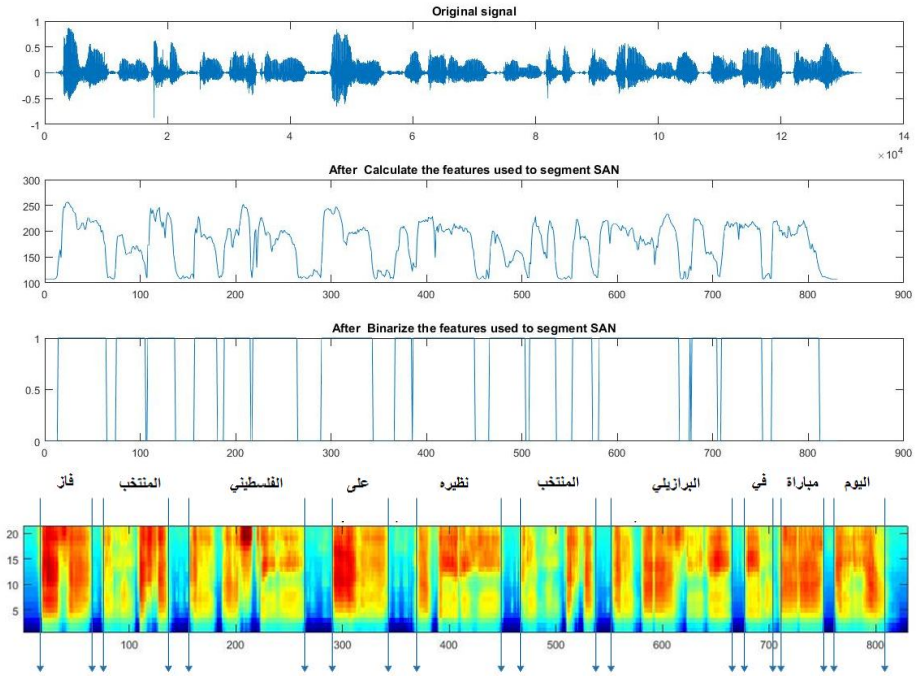


Fig. 4 Segmentation of SAN clip example.

The next main box in TID system is keywords extraction module. It consists of three stages, practically similar to existing systems for keywords extraction. Generating candidate keywords is the first stage which includes filtered segmented words (each word in the filtered words storage considered as candidate keywords). The second stage involves extracting features for each candidate keywords. These features capture the frequency of a word. The frequency of a word is computed by matching candidate word with other words in order to find the distance between them. By using threshold distance, the considered matched words are counted. The final stage is a keywords selector (word ranking). The selector examines the rank of word *p* for each candidate keyword. The candidate is selected as a keyword if *p* is high enough. Figure 5 shows the main components in the keywords extraction method.

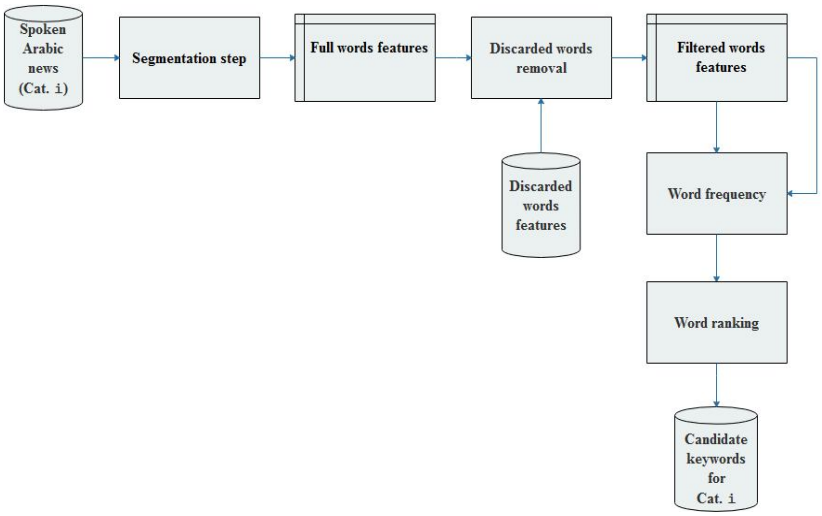


Fig. 5 Keywords extraction processes.

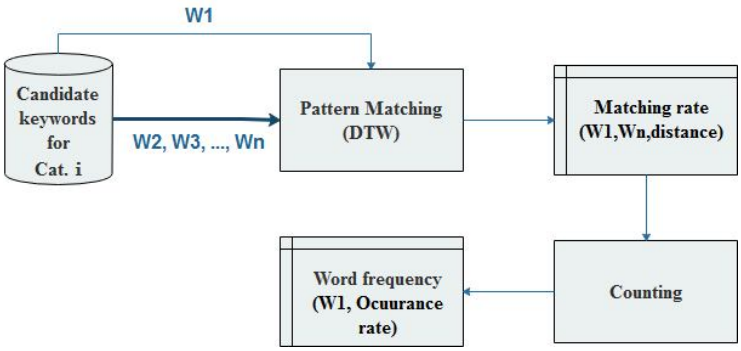


Fig. 6 Word frequency processes.

After extracting candidate keywords (the top ones) for each category, the candidate keywords for categories go through mutually exclusive process (MEP). MEP selects the keywords for each category in order to ensure that keywords in one category cannot be found in any other categories. To conclude. the general steps for keywords extraction methodology are:

Segment all SAN clips for each category.

Clean the segmented words (candidate keywords) by removing punctuations and stop words.

Compute the frequency of the words.

Rank the word (candidate keywords).

Mutual exclusive process is applied to all categories

The last step of the SAN identification system is the TID process which is based on pattern comparison and scoring techniques. Two main modules form the pattern matching system which are feature extraction and feature matching. The purpose of feature extraction module is to extract a feature vector using MFCC extractor . In the feature matching module, the extracted MFCC feature vector from unknown spoken word sample is compared to acoustic model (keywords). The model successes when it has a max score (the model with low distance successes). The output of this system is considered as a matched word. In the acoustic model, Dynamic Time Warping (DTW) or Hidden Markov Model (HMM) is used with the purpose of scoring the model by distance or Logliklehood metric.

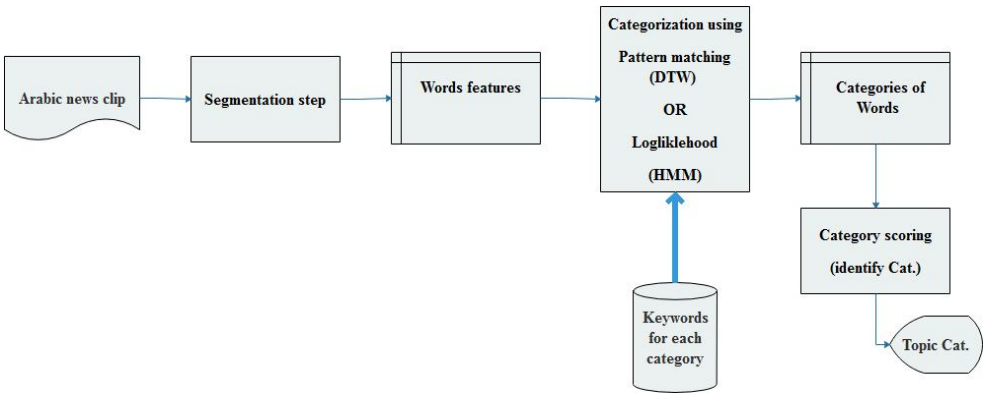


Fig. 7 Topic identification processes.

The steps of pattern matching stage are summarized as:

Read the input file (SAN clip **C**) then preprocessing and segmentation steps are implemented in order to obtain (**C** = w_1, w_2,…, w_n) where w_n is the separated word.

After applying the segmentation step, MFCC features for each segmented word are generated.

For each word w, one of the pattern match algorithms is applied in order to decide to which category w belongs. If w_1 is the MFCC features of the first segmented word, and $\mathbf{kw_x}$ is the MFCC features of keywords in category x. For example, $\mathbf{kw_s}$ are the MFCC

features of keywords in the sport category (**kw_s** = kw_1, kw_2, …, kw_n). kw_1 is a container of several different images of kw_1 spoken by different persons.

By using DTW, the distance between w_1 and each kw in the keywords database is calculated to elect the least distance, if the least distance exceeds the set threshold, the word is omitted from calculations.

By using HMM, the logliklehood is calculated to elect the most logliklehood, if the most logliklehood exceeds the set threshold, the word is omitted from calculations For instance, w_1 matches some images of kw_1 in **kw_s** (sport category). In order to elect the most suitable pattern, the kw with least distance to w_1 is set to be as a representative of its category. This is done for all kw in **kw_s** and other categories. The best pattern, which gives the least distance, is selected as the candidate of the category.

After selecting the most suitable representatives, which belongs to one of the categories and represents the segmented word w_1. The selected category is stored in the categories of words box.

Repeat the steps above from one to four for all segmented words.

Rank the frequency (count) of the categories in the categories of words storage; the maximum category count is selected as the topic of this SAN clip.

IV. RESULTS AND DISCUSSIONS

A new Arabic speech dataset is established based on a widely used dataset in the text classification field. Two datasets are used as the base of this study. The first one is from Aljazeera News Arabic dataset which is a collection of 1500 Arabic news files obtained from Aljazeera online news agency. These files are distributed among five categories (300 files for each category, 240 files for keywords extraction step and 60 files for TID step) which are politics, art, science, economic and sport. The second dataset is from a collection of Arabic weather news files obtained from online news websites to construct the weather category. All these written text files are converted to spoken files using a special recorder by multiple speakers (30 speakers). The conversion is made by speakers of various genders, ages and intonations.

TABLE I DATASET DESCRIPTION

	Keywords extraction step		TID step	
Category	Almost Speaking Time (Minutes)	Words	Almost Speaking Time (Minutes)	Words
Weather	640	23,965	240	10626
Sport	1488	66,485	344	15154
Science	1638	86,449	390	16334
Politics	1431	72,050	387	16298
Economy	990	54,312	320	13900
Art	1470	67,188	403	17193

To approve the accuracy of the used machine learning algorithm, F1-Measure metric is used in the field of information retrieval. The value of F1-Measure is governed by two factors which are precision and recall. The figure below shows the evaluation metrics (precision, recall and F1-Measure) results of using DTW and HMM methods as identifieres in SAN classification system. The results of applying speech based identification algorithm to Alj-News dataset reveal that they have recorded a good performance. The results are motivating as the overall F1-measure is 90.26% and 91.36% for DTW and HMM in sequence. The results show that the high performance are scored using HMM classifier as shown in the figures below.

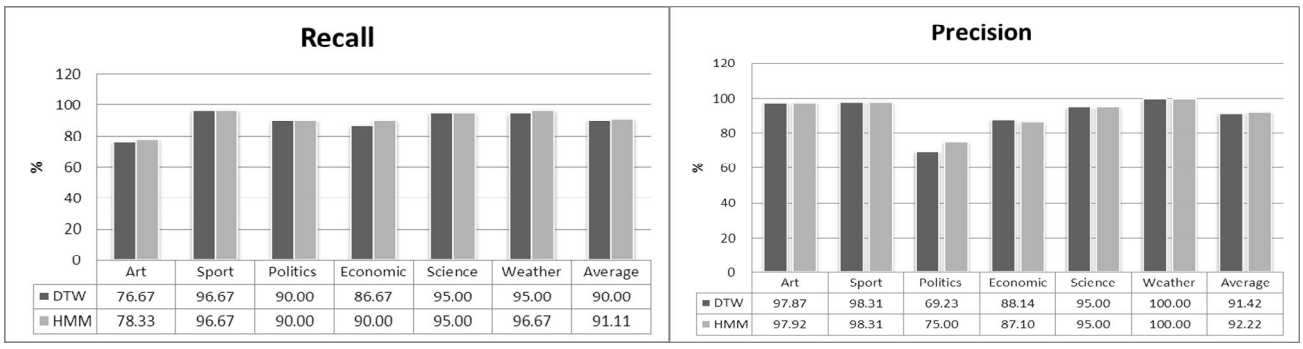


Fig. 8 Precision, Recall using DTW and HMM method

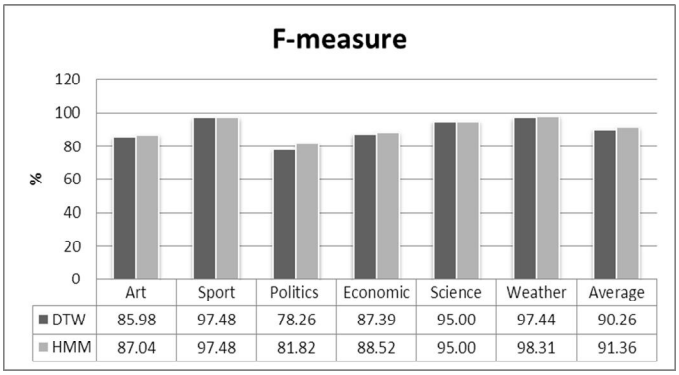


Fig. 9 F-measure using DTW and HMM method

The top performance is in 'Weather' and 'Sport' classes with (100% and 98.31%) Precision , (96.67% and 96.67%) Recall, (98.31% and 97.48%) F-measure while the lowest F-measure performance recorded is 81.82 % in 'Politics' class. This is due to the lack of confusion of news included in weather or sport classes in contradiction of news included in the classes of politics, art, economy and science. The news in these classes might be ambiguous or intersected. As a result, the classification F-measure achieved using speech based techniques are :

90.26% of classification accuracy is achieved by using DTW matching technique which uses a distance as a matching measurement between patterns.

91.36% of classification accuracy is achieved by using HMM matching technique which uses a distance as a matching measurement between patterns.

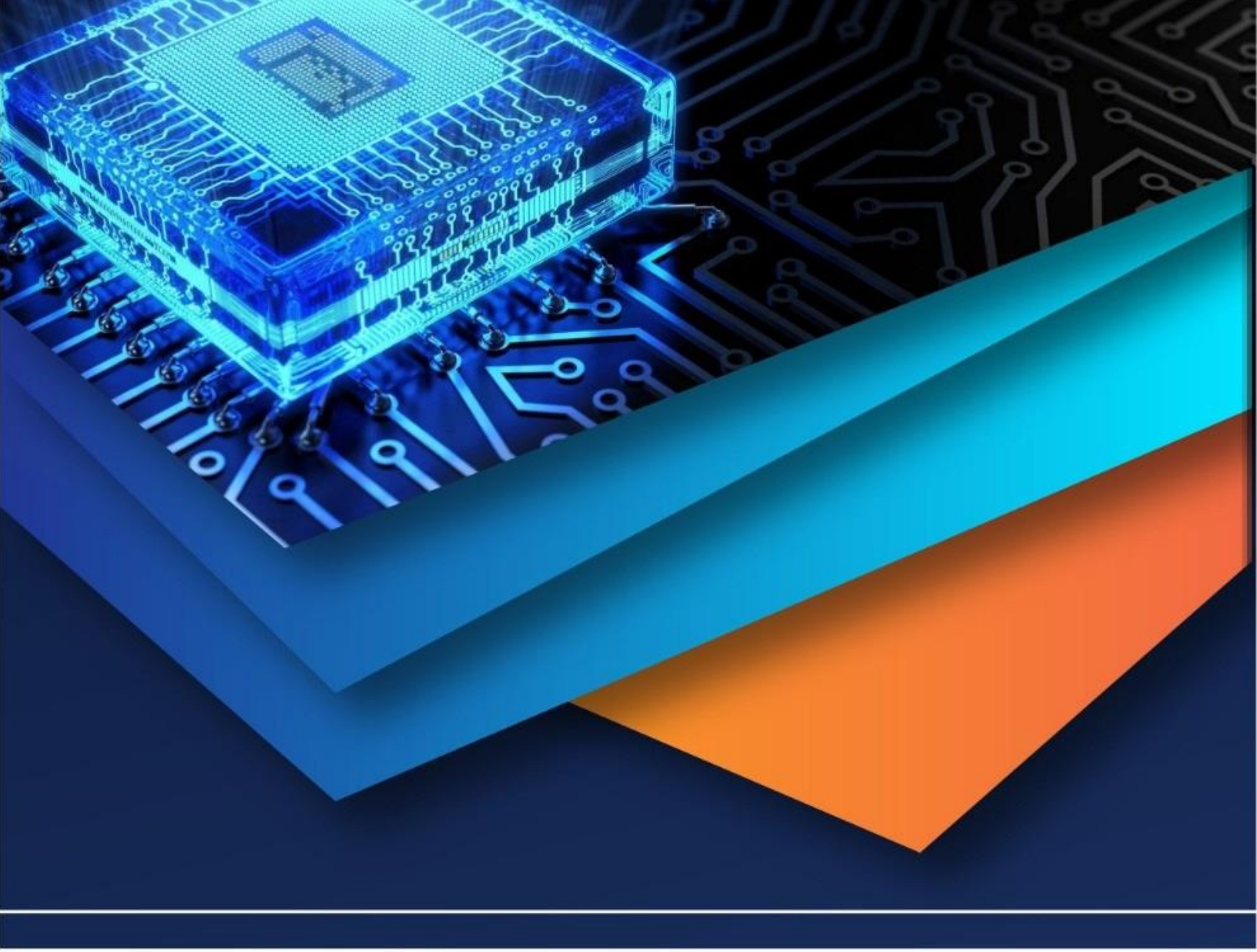
V. CONCLUSIONS

This paper proposes a new methodology dealing with classification of spoken Arabic files directly using speech features, while the traditional methods rely on converting the spoken files into transcriptions. The proposed approach works well in this area after extracting the experimental results. Also, it can be used in the field of Arabic speech search engines and the classification of huge numbers of spoken Arabic texts into various classes. The keywords extraction system is proposed in order to extract the keywords for each class depending on the DTW matching technique . After the keywords selection is done, an automatic topic identification system is suggested. The first process in this system is the segmentation of a connected spoken file and the extraction of features of each separated word. Then, the SAN system uses one of the two techniques (DTW or HMM) in order to identify to which class this word is related. The top match scoring between the separated spoken word and keywords in all classes based on the distance value using DTW algorithm is the first method in order to classify this word. The second one is to assign the separated spoken word to a model (which characterizes one of the keywords that represents one of the classes) based on logliklehood value using HMM algorithm. It is worth mentioning that choosing the best matching or assigning depends on specific selected threshold value. In the last stage, the keywords frequency sorting method is used with the purpose of identifying the class of the SAN clip. To evaluate TID approach, an Arabic news dataset is established via various speakers who read text clearly in order to create a usable spoken dataset which can be useful in the topic classification area. F1-Measure is used as an evaluation metric to estimate the accuracy of the classification process which comes from two factors which are precision and recall factors. The values of F1-Measure for topic identification system with the DTW classifier are with an average 90.26% and 91.36% with HMM classifier. This system is suggested to eliminate the step of converting the spoken texts into transcriptions and benefit only from the speech features. Also, it is built to be applicable to Arabic language in the time that Arabic data are rare. Many systems can be built depending on the methodologies discussed in this system such as creating spoken files search engines.

REFERENCES

[1] Sandsmark, H. "Spoken document classification of broadcast news", M. Eng. thesis, Norwegian University of Science and Technology, Norway, 2012.
[2] Rose, R., Chang, E. and Lippman, R., "Techniques for information retrieval from voice messages", Proceedings of the International Conference on Acoustics, Speech, and Signal Processing . Toronto, Ont., Canada, 1991.
[3] Sebastiani, F. "Machine learning in automated text categorization". ACM Computing Surveys, 34(1), pp. 1–47, doi:10.1145/505282.505283, 2002

[4] Manning, C. and Schütze, H. *Foundations of statistical natural language processing*, MIT Press Cambridge, MA, USA chapter Text Categorization, pp. 575–608, 1999
[5] Hazen, T., Richardson, F., and Margolis, A. "Topic identification from audio recordings using word and phone recognition lattices", Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding . Kyoto, Japan., 2007.
[6] Lo., Y-Y. & Gauvain, JL. "Tracking topics in broadcast news data", Proceedings of the ISCA Workshop on Multilingual Spoken Document Retrieval . Hong Kong, 2003.
[7] McDonough, J., Ng, K., Jeanrenaud, P., Gish, H. and Rohlicek, J.R. "Approaches to topic identification on the switchboard corpus. In Acoustics, Speech, and Signal Processing, ICASSP-94., IEEE International Conference on, volume i, pp. I/385 –I/388 vol.1, doi: 10.1109/ICASSP.1994.389275, 1994
[8] Rosenberg, A. "Challenges and opportunities in spoken document processing: Examples from keywords search and the use of prosody", The Journal of the Acoustical Society of America 140, 3010 (2016); doi: http://dx.doi.org/10.1121/1.4969335, 2016.
[9] Qaroush, A., Hanani, A., Jaber, B., Karmi, M. & Qamhiyeh, B. "Automatic spoken customer query identification for Arabic language". ICIME 2016, doi:10.1145/3012258.3012261, pp. 41-46, 2016.
[10] Liu, C., Yang, J. , Sun, M., Kesiraju, S., Rott, A., Ondel, L., Ghahremani, P., Dehak, N., Burget, L. & Khudanpur, S. "An empirical evaluation of zero resource acoustic unit discovery", in Proc. ICASSP, Feb., 2017.
[11] Hazen, T. J., "MCE training techniques for topic identification of spoken audio documents", IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 8, pp. 2451–2460, Nov., 2011.
[12] Liu, C., Trmal, J., Wiesner, M., Harman, C. & Khudanpur, S. "Topic identification for speech without ASR", Computation and Language, Publication: eprint arXiv:1703.07476, Mar., 2017.
[13] Kesiraju, S., Pappagari, R., Ondel, L. & Burget, L. "Topic identification of spoken documents using unsupervised acoustic unit discovery", in Proc. ICASSP, 2017
[14] Dredze, M., Jansen, A. , Coppersmith, G. & Church, K. "NLP on spoken documents without ASR," in Proc. of EMNLP, 2010.
[15] Matlab audio processing examples, (2012) [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY