The Islamic University–Gaza
Research and Postgraduate Affairs
Faculty of Engineering
Master of Computer Engineering

الجامعـــــــة الإســـلاميـة – غـزة
شئون البحث العلمي والدراسات العليا
كليـــة الهندســة
ماجستير هندسة الحاسوب

Automatic Topic Classification System of Spoken Arabic News

النظام الآلي للتصنيف الموضوعي للأخبار المنطوقة باللغة العربية

Naser S. A. Abusulaiman

Supervised by

Dr. Mohammed Alhanjouri

Associate Prof. of Artificial Intelligence and Digital Signal Processing

**A thesis submitted in partial fulfilment
of the requirements for the degree of
Master of Computer Engineering**

September/2017

إقـــــرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

Automatic Topic Classification System of Spoken Arabic News
النظام الآلي للتصنيف الموضوعي للأخبار المنطوقة باللغة العربية

أقر بأن ما اشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وأن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل الاخرين لنيل درجة أو لقب علمي أو بحثي لدى أي مؤسسة تعليمية أو بحثية أخرى.

Declaration

I understand the nature of plagiarism, and I am aware of the University's policy on this.

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted by others elsewhere for any other degree or qualification.

Student's name:	ناصر صادق أبو سليمان	اسم الطالب:
Signature:	ناصر صادق أبو سليمان	التوقيع:
Date:	06/09/2017	التاريخ:

بسم الله الرحمن الرحيم

الجامعة الإسلامية غزة
The Islamic University of Gaza

عمادة البحث العلمي والدراسات العليا هاتف داخلي 1150

الرقم: ج س غ/35/ Ref:

التاريخ: 2017/09/06 Date:

نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة عمادة البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحث/ **ناصر صادق عبدالله ابوسليمان** لنيل درجة الماجستير في **كلية الهندسة/ قسم هندسة الحاسوب** وموضوعها:

النظام الآلي للتصنيف الموضوعي للأخبار المنطوقة باللغة العربية

Automatic topic classification system of spoken Arabic news

وبعد المناقشة التي تمت اليوم الأربعاء 15 ذو الحجة 1438هـ، الموافق2017/9/06م الساعة الواحدة ظهراً، اجتمعت لجنة الحكم على الأطروحة والمكونة من:

د. محمد أحمد الحنجوري	مشرفاً ورئيساً	.....
د. وسام محمود عاشور	مناقشاً داخلياً	.....
د. إيهاب صلاح الدين زقوت	مناقشاً خارجياً	.....

وبعد المداولة أوصت اللجنة بمنح الباحث درجة الماجستير في **كلية الهندسة / قسم هندسة الحاسوب.**

واللجنة إذ تمنحه هذه الدرجة فإنها توصيه بتقوى الله ولزوم طاعته وأن يسخر علمه في خدمة دينه ووطنه.

والله ولي التوفيق،،،

عميد البحث العلمي والدراسات العليا

أ.د. مازن اسماعيل هنية

الملخص

إن أهم العواقب الرئيسية لما يعرف ب "عصر الانترنت" هو الانتشار الواسع للبيانات المتنوعة نوعاً وموضوعاً. هذا الانتشار بحاجة ملحة إلى نظام آلي لتصنيف هذه البيانات لتسهيل عملية البحث. مثل هذا النظام معمولٌ به بشكل كبير في البيانات النصية المكتوبة ولكن مع ازدياد حجم البيانات النصية المنطوقة (الصوتية) بشكل كبير تظهر الحاجة إلى نظام آلي لتصنيف البيانات النصية المنطوقة بشكل مباشر دون الحاجة لتحويلها لنصوص مكتوبة ومِنْ ثمَّ تطبيق الخوارزميات المتبعة في النصوص المكتوبة. النظام المباشر تم مناقشته بشكل بسيط في الأبحاث السابقة على النصوص المنطوقة باللغة الإنجليزية وبشكل يكاد ينعدم على نظيراتها باللغة العربية نظراً لصعوبة التعامل مع اللغة العربية إضافة لعدم توفر مجموعة بيانات نصية منطوقة باللغة العربية تصلح لعملية التصنيف الموضوعي. هذا البحث يرتكز بشكل رئيسي على إنشاء نظام متكامل ابتداء من استخراج الكلمات المفتاحية آلياً لكل صنف على حدى. ولاحقاً يتم الاستفادة من هذه الكلمات في نظام تصنيف النص المنطوق بشكل مباشر وذلك اعتماداً على الخصائص الصوتية للكلمة وليس بالطريقة المعتادة عبر تحويله إلى نص مكتوب. تم تحويل النصوص المكتوبة باللغة العربية المستخدمة كثيرا في التصنيف الموضوعي للنصوص المكتوبة (ALJ-NEWS) إلى نصوص منطوقة عبر متحدثين متنوعين للاستفادة منها في هذا البحث. في عملية استخراج الكلمات المفتاحية تم الاعتماد على DTW كطريقة لقياس تكرار الكلمة المنطوقة داخل الصنف عبر مقارنة الخصائص المستخرجة (MFCC) لكل كلمة. في هذا البحث تم الاعتماد على (HMM and DTW) كطرق لتصنيف الكلمة المنطوقة اعتمادا على الخصائص المستخرجة (MFCC and PLP-RASTA) من الكلمة المنطوقة. تم اقتراح آلية جديدة لعمل تقطيع للملف الصوتي إلى كلمات منفصلة في هذا البحث. وبالنظر إلى تقييم الأداء التصنيفي للنظام. تم استخدام معايير عدة: (F1-measure, Accuracy, Precision and Recall). النظام المقترح أعطى نتائج جيدة في عملية التصنيف حيث سجل نظام التصنيف ما متوسطه 90.26% باستخدام DTW و 91.36% باستخدام HMM على مقياس F1-measure بالإضافة إلى أن دقة تحديد الكلمات المفتاحية كانت 89.65%.

الكلمات المفتاحية :التصنيف الموضوعي للنصوص المنطوقة، معالجة اللغة الطبيعية، استخراج الكلمات المفتاحية، تقطيع الصوت.

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. To deal with this challenge, a new dataset is established depending on converting the common written text (ALJAZEERA-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 dynamic time warping. Finally, topic identification, based on (Mel-frequency Cepstral Coefficients and Relative Spectral Transform - Perceptual Linear Prediction) as speech features and (Dynamic Time Warping and Hidden Markov Models) as classifiers, is created using a technique that is different from the traditional way, using an automatic speech recognition to extract the transcriptions. Segmentation method is proposed to deal with the segmentation of spoken files into words. Regarding the evaluation of the system, accuracy, F1-measure, precision and recall are used as evaluation metrics. The proposed system shows positive results in the topic classification field. The F1-measure metric for topic identification system using dynamic time warping classifier records 90.26% and 91.36% using hidden Markov models classifier in the average. In addition, the system achieves 89.65% of keywords identification accuracy.

Keywords: Speech classification, Topic classification, Speech segmentation Keywords extraction, NLP, Spoken Arabic News, Speech features, HMM, DTW, MFCC.

Dedication

To My Mother, Father, Sisters, Brothers and Friends
Each of whom has a special place in my heart.
My wife is incredible. She has provided support in every possible way both in emotional terms as well as in practical ways.
With special gratitude to Leen, the best daughter I can imagine. You have been the gift from the beginning

Acknowledgment

The accomplishment of this thesis could not have been possible firstly without the Allah, the author of knowledge and wisdom. Also, I am grateful for the participation and assistance of so many people whose names may not all be mentioned. Their contributions are highly appreciated and thankfully acknowledged. However, I would like to express my deep gratitude and appreciation mainly

To my research supervisor, Associate Prof. Mohammed A. Alhanjouri for his boundless support, kind and understanding spirit during thesis presentation.

To the discussion committee, Dr. Wesam M Ashour and Dr. Ihab Zakout, for their guides and commentaries.

To all relatives, friends and others who in one way or another shared their support, either morally, financially and physically, thank you.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AGu	Unigram Adaptor Grammar
AR	Autoregressive
ASR	Automatic Speech Recognition
AUD	Acoustic Unit Discovery
CA	Classical Arabic
DA	Dialectal Arabic
DARPA	Defense Advanced Research Projects Agency
DCT	Discrete Cosine Transform
DiBS	Diphone-Based Segmentation
DP	Dynamic Programming
DSP	Digital Signal Processing
DTW	Dynamic Time Warping
EM	Expectation-Maximization
FFT	Fast Fourier Transform
FN	False Negative
FP	False Positive
HMM	Hidden Markov Models
IDF	Inverse Document Frequency
IDFT	Inverse Discrete Fourier Transformer
IFT	Inverse Fourier Transform
KW	Keyword
LPC	Linear Predictive Coding
LVCSR	Large Vocabulary Continuous Speech Recognition
MCE	Minimum Classification Error
MFCC	Mel-Frequency Cepstral Coefficients
MSA	Modern Standard Arabic
NLP	Natural Language Processing
OOV	Out-Of-Vocabulary
PLP	Perceptual Linear Prediction
RASTA-PLP	Relative Spectral Transform - Perceptual Linear Prediction
SAN	Spoken Arabic News
S-DTW	Segmental Dynamic Time Warping
SOUs	Self-organizing Units
STD	Spoken Term Detection
SUs	Sound Units
SVM	Support Vector Machines
TDT	Topic Detection and Tracking
TF	Term Frequency
TF-IDF	Term Frequency- Inverse Document Frequency
TID	Topic Identification
TN	True Negative
TP	True Positive
TPs	Transitional Probabilities
WER	Word Error Rate

Chapter 1
Introduction

Chapter 1
Introduction

1.1 Topic Area

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 these files into different topics. The term spoken document classification is defined as the problem of labeling 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 experiences (Sandsmark, 2012). 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 to manage and index the speech records. This task focuses on tagging a pre-defined topic (class) to a spoken file based on its contents. *Topic classification 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. This pattern is called a *single-label categorization*. The area of topic classification includes other related problems in addition to classic detection and classification responsibilities. Over the passage of time, new topics may be risen. For instance, new events in spoken newsletters occur frequently demanding the creation of new topic classes for labeling coming newsletters related to these events. Essential works have been stated on topic classification from texts field, but a little amount of works on spoken files are reported, particularly using information from speech signals. This thesis provides a speech processing solution to enable content based access to spoken files and proposes a topic classification approach for Spoken Arabic News (*SAN*).

1.2 Topic Identification (TID)

The process of topic identification technique requires initially to define the topic and topic tags which should be built and assigned. The definition of topic identification or classification is the process of identifying or classifying the targeted topic that is relevant to a recorded speech. A binary tag (relevant or irrelevant) for each topic works as a description of the speech segment. It is not easy to define the weight of relevance of a specific topic to a specific speech segment. Topic relevancy standard definitions are not existent and most of their measurement tools are subjective. Therefore, the labels or tags of a group of data are inconsistence and vary depending on taggers or labelers. Although there are many works on measuring the relevancy which might be steadily tagged or labeled, they probably require a vital amount of effort to gather and to classify speech datasets. Two common styles of TID that are commonly studied are: topic classification and topic detection.

A pre-defined group of topics or subjects has been determined in *Topic classification style* and each speech segment is identified as having its place in one and only one topic or subject from this group. *Single-label categorization* is another name of topic classification . When the main mission assigns the speech data into distinctive holders or directing them to certain systems or audiences, the task that is frequently used is a single-label categorization

Nevertheless, sometimes a speech segment can be related to any number of topics or subjects. Hence, in what is called *topic detection style* an independent decision is needed to discover the occurrence or absence of each related topic or subject. The alternative name of this style is *multi-label categorization*. For easier filtering, organizing, examining, and retrieving of a required speech segment, the *multi-label categorization* can be used. For instance, spoken files might be labeled or tagged with more than one topic labels or tags and this would let users quickly find and view certain spoken files about topics or subjects of their concern (Hazen, 2011). In this thesis the system is built based on topic classification concept which refers to TID and they are used interchangeably throughout this thesis .

1.3 Thesis Significance

Spoken data classification or spoken data categorization is a fundamental problem in computer science. TID is used in order to assign spoken data to one class or category. Classification or identification is one of the important fields in machine learning searches, such as speech and language processing, data filtering, data cleaning, …, etc. During information retrieval, classification is used in many subtasks of the search pipeline: Preprocessing, content filtering, ordering, ranking, etc.

Today, spoken files classification is acute need due to the big number of spoken files that are needed to deal with every day. This thesis has a valuable significance, for the result of demanding classification of the huge number of electronic spoken Arabic files, which are available via the rapid and increasing growth of the Internet. Correspondingly, the manual classification of such huge files is time and effort consuming.

There is a number of classification algorithms applied to spoken files. 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.

Most of the existing systems for TID of spoken files use, as the pre-processing step, word or phoneme based ASR. The implementation of techniques developed by the text retrieval community is used at the second stage. Training ASR systems for English on large amount of data and software is possible. However, not every language is rich in resources for building ASR systems; therefore, developing techniques that are useful for other languages, with low resources, are needed.

1.4 Research Questions

The main research problem is identifying or classifying the topics which are discussed in SAN clips. There are a number of researches that aimed at classifying spoken files using text methods applied to transcriptions resulting from ASR model.

- How to classify SAN clips based on speech features without using text methods or depending on ASR model.
- How to use related features of speech data in order to enhance the results.
- How to select the main features with the intention of reducing dimensionality.
- How to handle uncertainty in classification decisions.
- What is the impact of pattern matching methodology precision using hidden Markov models (HMM) or dynamic time warping (DTW) on topic classification precision.

1.5 Thesis Goal

To design an automatic system which identifies the underlying topics that are discussed in spoken Arabic news files with high accuracy and pace depnding on speech features..

1.6 Thesis Contributions

1. A new database for Arabic news broadcasts with its transcription is created.
2. An automatic keywords extraction methodology is proposed and implemented.
3. Topic labels (Keywords) database is built using an automatic keywords extraction methodology.
4. A complete system for automatic spoken Arabic news classification based on speech features without using text algorithms is created and implemented.
5. Segmentation algorithm is proposed and enhanced to deal with Arabic speech.
6. Mixed features of speech, PLP and MFCC, are proposed in order to enhance the results.

1.7 Limitations

- The TID and keywords extraction speed needs a further improvement.
- The segmentation algorithm in this thesis deals with specific conditions, such as clear and understandable spoken words and a reasonable speed while speaking.
- The created topic labels depend on standard Arabic speech.

1.8 Thesis Organization

The rest of this thesis is divided into four chapters prearranged as follows:

Chapter 2 introduces the related works; it shows some researches that have been done in classifying spoken Arabic news. Because of the lack of works on spoken Arabic news, other languages are focused on. The related works are divided into two categories. The first one shows some researches working on TID using ASR model in their classification algorithms. The second one shows some researches using some features of speech in order to enhance TID process.

Chapter 3 overviews background theory in which three main parts are discussed. The first part shows segmentation algorithms. The second part shows keywords extraction algorithms. The last part of this chapter explains the classification concept and the main parts of it including pre-processing module, feature extraction module and HMM and DTW classifiers used in topic classification or identification systems.

Chapter 4 presents the proposed work used in the thesis including a description of preprocessing steps, Feature processing and extraction, keywords extraction, segmentation and classification processes.

Chapter 5 discusses the tools, dataset and evaluation metrics. It also presents and analysis the experimental results of the proposed work.

Chapter 6 concludes the research in which recommendations , remarks, and some notes are stated.

1.9 Related Publications

Abusulaiman, N. and Alhanjouri, M. (2017) Spoken Arabic News Classification Based on Speech Features, Volume 5, Issue VIII, *International Journal for Research in Applied Science and Engineering Technology (IJRASET)* Page No: 1486-1491 , ISSN: 2321-9653, www.ijraset.com

Chapter 2
Related Works

Chapter 2
Related Works

A lot of topic classification or identification systems for several written text applications are built. As a consequence of text classification achievements, speech-based TID applications borrow most of the commonly used techniques in the text processing area and adapted them in order to be applicable to the speech-based classification systems. (Sebastiani, 2002) introduces general TID techniques used in written text applications. More additional descriptions and techniques can be found in a book chapter by (Manning & Schütze, 1999). 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 (Hazen et al., 2007; Lo & Gauvain, 2003; McDonough et al., 1994; Rose et al., 1991). Nevertheless, many of topic classification methods are applied to written Arabic news and most of the studies and papers rarely discuss TID in spoken Arabic news. Spoken files processing faces many challenges and opportunities when compared to written files processing (Rosenberg, 2016).

2.1 Traditional TID Systems

Until the early 1990s, 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. (Rose et al., 1991) 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. (e.g., map-reading, photographic, doll, …, etc.). Despite the fact that larger dataset tends to be available during the 1990s, three different types of data classes which are broadcast news sections, customer service calls and human-human dialogues have begun to be mainly used in the researches lately. An instance of dataset that is used in speech-based TID and provides comparable outcomes to text-based TID is organized news broadcasts which presented by (Fiscus et al., 1999). These prepared written news reports are read via professional speakers in which they record clips using high-quality recording tools and in a unique acoustic

circumstances. Subsequently, (Pallett et al., 1999) argues that working on this dataset tends to give very low error rates when using modern powerful ASR systems. Many of human-human dialogues are recorded through telephone lines to establish the Switchboard and Fisher dataset (Cieri et al., 2003). They are produced with a topic specific label. Thus, a number of different researches use these dataset for TID researches (Carlson, 1996; Gish et al., 2009; Hazen et al., 2007; McDonough et al., 1994). The dataset is pre-tagged with the topic label. Speech-based TID tasks are originated from the project of Topic Detection and Tracking (TDT) from Defence Advanced Research Projects Agency (DARPA) which started in 1998 and it is still working for many years into the succeeding period (Wayne, 2000).

A spoken file classifier is suggested by (McDonough et al., 1994) in which they use a multinomial naive Bayes classifier linked with expected word counts feature extraction. They utilize a two feature selection to choose the classification vocabulary. Their outputs on the switchboard corpus show that using expected word counts is better than 1-best, both for feature (word) selection and for classification. They recommend against feature selection from 1-best data by debating that selecting a word because it differentiates well based on 1-best might not work well since it could be the issue that it cannot be surely detected or it has a large insertion probability by the speech recognizer. The expected word counts are not counted immediately from a recognition lattice, but rather they are counted by using a HMM word spotter, which is a special-purpose speech recognizer that aims to only recognize words selected by the feature selector. They use a forward-backward algorithm in order to compute the posterior probabilities for the task of classification vocabulary words at all-time steps. To compute the expected word counts, instead of summing all the posteriors for each word, they only encompass a posterior in the sum when it reaches a local maximum on the time axis. Spoken document identification is closely associated with spoken document retrieval where a collection of documents is returned depends on a specified request.

With the growing of spoken files in different languages and dialects, there is an essential requirement of systems performing automatic and unsupervised classification and searching on spoken files streams in a file retrieval phase. The simplest way of constructing spoken files classification system is concatenating an

ASR model with a text-based classification algorithms. The common approach has been transferred from manual rule-based methods to more recent automatically trained probabilistic methods (Sebastiani, 2002). A lattice-based method to spoken document retrieval is suggested in (Chelba, 2007) where expected word counts are estimated from the lattice and subsequently used to generate a ranking of the documents. They report a considerable relative improvement by using expected counts from lattices instead of 1-best word counts.

Nevertheless, under resource-limited conditions, the manually transcribed speech needed for developing standard ASR systems can be severely limited or unavailable (Liu et al., 2017). Unfortunately, the accuracy of current speech recognizers degrade remarkably when faced with unconstrained conversational speech or varying acoustical environments (Garofolo et al., 2008). Because the relative advantage of considering various hypotheses is expected to boost with decreasing recognition reliability (Mamou et al., 2006) (Chelba et al., 2007), there has been much work going into the field of spoken files classification.

When utilizing an unsuitable ASR system, TID suffers from possibly some major problems. One of these problems is many of the important related words, which are the labels for a topic, may be missed in the dictionary used by the ASR system. The ASR system in this situation cannot assume these missing words. This problem stated as out-of-vocabulary (OOV) word problem. Many approaches can be used in order to address the OOV problem in several speech applications, such as spoken file retrieval application. A widespread approach is working on the processing of speech phonetic parts (Chelba et al. 2008). In these circumstances, the key process extracts the primary sequences of phonetic parts (or phones) as words representative. Hence, for the TID problem, the “car” , for example, is a word for a certain TID assignment. The system needs to find out that the sequence of phonetic units [c a: r] includes related information once detected in a spoken file.

In some situations, TID capability is desired in a fresh language where there is partial or zero transcribed data to train a suggested phonetic system. In this situation, (Hazen et al., 2007) suggest that carrying out TID using a phonetic system which recognizes phonetics from another language is probable. (Gish et al., 2009) present an in-language phonetic system trained without the transcriptions in a fully

unsupervised style. A phonetic system can produce posterior lattices, in a similar manner to word recognition, for speech segments.

(Cerisara, 2009) tries to learn both words and topics from spoken files only by using transcriptions resulting from a phonetic ASR system. In order to find possible shared words from texts, fuzzy phonetic sequence matching process is used. Then, the system describes texts by the distances resulting from best matching of these discovered words and segments within texts. Resulting distance vectors are used in order to cluster the topics under the supposition that texts with small minimum distances to the same assumed word are interrelated. The unsupervised learning research field might continue to be more significant where the level of complexity in the algorithms might increase. The relation between TID systems and on-line learning techniques is expected to be taken more into account. Thus, it should concern the study of topics in data and over time how the representative indications of topics are changed and developed. (Katakis et al., 2005) discuss some studies of this topic on written text data and similar researches should be followed on spoken files.

For spoken files (Hazen et al., 2011) present both supervised and unsupervised topic modelling using phonetic information. If word-based recognition is absent or infeasible, phonetic information can be engaged to indirectly learn and capture information delivered by relevant words. In the case of a transcribed data is absent or insufficient, a supervised training of the same-language phonetic recognition system can be inhibited. In these situations, cross-language prototypes or self-organizing units (SOUs) learned in a fully unsupervised approach can be utilized. They present novel enhancements using phonetic information. They show cconsidering outcomes using new techniques for TID used in parallel with novel SOU learning enhancements. An initial inspection of the use of this topic modelling for unsupervised detection of topics and relevant words from phonetic information is also shown in their studies.

The problem of TID using phonetic speech recognition outputs have been investigated by several former researches. However, this approach is inadequate for picking up the quasi-regular structure of speech, which causes substantial recognition

failure in noisy surroundings (Wright et al., 1996; Kuhn et al., 1997; Nöth et al., 1997; Theunissen et al., 2001; Belfield et al., 2003).

As mentioned above, related works on TID or classification process has looked at various ASR capabilities which are word-based and phonetic-based recognition and ASR lattice output. It is worth mentioning that most of them are applied to English language.

It is known that most of speech transcriptions include errors, but speech conveys information beyond the words that are said. A modified TF-IDF feature weighting calculation is suggested by (Wintrode & Kulp, 2009) in order to reduce the impact of ASR errors on the accuracy of TID. Their approach provides important validity under various recognition error conditions. It is worth mentioning that they examine the impact of ASR errors in informal telephone speech, so 1-best and lattice outputs are generated using a single recognizer tuned to be carried out at various speeds. By using SVM classifier they implement TID on the output of ASR. They work on classifiers to be crucially more valid to errors.

(Hazen, Nov, 2011) 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 are generalized to be better for invisible data. Sizeable enhancements are observed in TID accuracy using the new MCE training techniques.

Innovative approach to an unsupervised training of speech recognizers is suggested by (Siu et al., 2014). Their approach regulates sound units (SUs) which are enhanced for the acoustic field, so the speech recognizers can be used for implementations in speech fields where there are zero transcriptions. HMM-based speech recognizers is recommended by (Siu et al., 2014) when the transcribed data are not existing. This is done by expressing the HMM training as an enhancement over parameter and transcription sequence space. At that moment spoken files are transcribed into these SOUs and they test the value of SOUs on the task of TID on the Switchboard and Fisher dataset.

In order to deal with transcription errors, a spoken file should be viewed as a structure of viable hypotheses not as an absolute transcription. In addition, the

manner in which words are spoken and their prosody can be mined for information about the speaker and his or her intention. (Rosenberg, 2016) uses a keywords search as a case study that requires operating under errorful and adverse circumstances.

In zero resource situations, earlier works on the analysis of spoken files are based on identifying recurrent patterns of speech (spoken words) where DTW based algorithms are used (Flamary et al, 2011; Harwath et al., 2013). Nonetheless, these methods are not scalable to large quantities of data. An alternative method is to use phone recognizers from other languages. This idea is discovered for the task of TID in (Hazen et al., 2011). Under limited resource conditions, i.e., with limited vocabulary for training an ASR, TID of spoken files is explored in (Wintrode, J. & Khudanpur, S., 2014).

Based on the automatic patterns discovery in the speech (Flamary et al, 2011) suggest an automatic speech summarization technique. Firstly, the system extracts recurrent acoustic patterns from the spoken file. Then, they are clustured and graded in relation to the frequency of the patterns in the clip. A "Spoken WordCloud" is formed due to the likeness with text-based word-clouds. Using a small dataset of connected spoken words, their approach reaches a cluster purity up to 90% and an inverse purity of 71%.

(Harwath et al., 2013) show zero-resource system, an automatic examination of a group of speech-based data in a fully unsupervised manner without the advantage of any transcriptions of the speech data. The system automatically finds keywords and phrases included in a spoken files dataset. (Harwath et al., 2013) suggest a method that utilizes a segmental dynamic time warping (S-DTW) procedure for the sake of acoustic pattern recognition in connection with a probabilistic technique which gives the related topic. Using information and Expectation-Maximization (EM) algorithm, acoustic outlines of the central topical themes of the spoken files group are created.

By using statistical approach, (Wintrode, J. & Khudanpur, S., 2014) detect topic keywords in which they study the overall word error rate (WER) and TID performance relation. WER measures keyword-specific discovery performance. Their study is established on the Fisher Spanish dataset and they focus on various

datasets ranged from large vocabulary continuous speech recognition (LVCSR) to limited-vocabulary keywords. They show that WER and low-precision term detection are impediments to TID.

Multilingual speech data are the case study of (Caranica et al., 2016) research. They focus on retrieving information from spoken files such as broadcast newsletters or even telephone dialogues. Vocabulary-independent search is the ultimate goal of a spoken file retrieval system. They present methodology in order to catch spoken "queries" over large collections of speech data. There are two cases. The first one is when the language is recognized, so the task is comparatively clear. It can be solved by using a LVCSR tool to produce accurate word transcripts, index them and retrieve a query from the index. The second case is when the language is unidentified where the recognizer's words are not included in the query, so the system cannot retrieve the relevant spoken files. Consequently, the texts retrieved are not related to the request. They use the input features taken from multi-language resources in order to investigate that if it helps the process of unsupervised spoken word discovery without considering the language existence. Furthermore, they join both language recognition and LVCSR based search, with unsupervised Spoken Term Detection (STD). They try to use multiple open-source tools with the aim of recommending a language independent spoken files retrieval system.

2.2 TID Systems without ASR

(Liu et al., Feb, 2017) automatically detect a categorical acoustic unit inventory from speech and produce a matching acoustic unit tokenization using Acoustic Unit Discovery (AUD). A major route for unsupervised acoustic model training in zero resource surroundings is provided by AUD where expert-provided linguistic knowledge and transcribed speech are unobtainable. They propose that AUD assessments are feasible using diverse alternative language resources when only a subset of these assessment resources can be existing in typical zero resource implementations. Another paper published later by (Liu et al., Mar, 2017) investigate alternative unsupervised solutions in order to obtain tokenization's of speech in terms of a word 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 AUD for TID of spoken files is presented by (Kesiraju et al., 2017). The AUD method is built on a non-parametric Bayesian phone-loop model that segments a speech utterance into phone-like groups. The exposed phone-like (acoustic) units are additionally fed into the conventional TID framework. Using multilingual bottleneck features for the AUD outperforms other systems that are built on cross-lingual phoneme recognizer.

2.3 Arabic TID Systems

A number of Modern TID systems for speech exploit ASR model to produce speech transcripts, and to perform a supervised classification on such ASR results. One of the rare studies on TID based on spoken Arabic language that uses ASR capabilities is done by (Qaroush et al., 2016). 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 modified CMU sphinx ASR and they apply Arabic ASR to the formal Arabic speech. The recognized text is used in order to define the question category using supervised machine learning methods so that it can take a preferred action, such as routing customer call to the suitable destination. 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 (Qaroush et al., 2016).

Chapter 3
Background Theory

Chapter 3
Background Theory

In Chapter 3, some background theories are overviewed, three main parts are discussed in this chapter. In the first part, speech segmentation is explained. In the second part, keywords extraction methods are shown. In the last part, the classification concept and the main phases in the TID systems are discussed.

3.1 Speech Segmentation

Segmentation is a processing phase that is essential for some of speech analysis applications. The aim is to split a connected speech signal into segments of homogeneous content that has meaning as sentences, words or characters. The borders between sentences, syllables, words, characters or phonemes in spoken natural languages can be discovered using speech segmentation process. The process of segmentation is applied to the mental processes which are used by humans, and to artificial processes of NLP. This process can be done at various levels in the connected speech. Segmentation at the lexical level is defined as the way of finding word boundaries in spoken language when the underlying vocabulary is still ambiguous (Gold & Scassellati, 2006).

Classification of speech segmentation methodologies can be done in a number of ways, but one of the very known groupings is the division of the segmentation methodologies to two main groups which are aided and blind segmentation methodologies (Räsänen, 2007). An essential difference which can be stated between aided and blind algorithms is the amount of utilizing the formerly obtained data or external knowledge to process the expected speech using the segmentation algorithm. A number of systems can use statistical data to learn and adapt to signals that they are considered as the input of the systems, or they can be taught previously with changeable methods. Finding statistically related information from the speech is the major focus of learning so that it can be used in order to improve the quality of the segmentation. In ASR, The most used statistical methods

are HMMs, where leading features of the signal are utilized for pattern matching or recognition process (Knill & Young, 1997).

Looking inside the *aided segmentation algorithms*, it can be inferred that they utilize some types of outside linguistic knowledge of the connected speech with the aim of segmenting the connected speech into some consistent segments of the preferred form. Generally, the meaning of that is how to use a written or phonetic transcription as a parallel input with the speech data, or how to train the algorithm previously by such data.

Whenever there is no previous information about linguistic features, such as the full phonetic series, of the speech signal which are needed to be segmented the procedures used in this case are fall under the *blind segmentation algorithms*. Some of implementations somewhere blind segmentation are needed to be applied are listed in (Sharma & Mammone, 1996) study. They show that speech data segmentation and labeling, speech recognition systems, speaker verification systems and language identification systems are highly suggested to be processed under the blind segmentation algorithms. Nevertheless, using some former information about features of the connected speech data by the system, which commonly refers to some types of machine learning that cannot be excluded, is possible (see, e.g., Bishop, 2007). Blind segmentation depends completely on the existing acoustical features in the speech signal because of the shortage of outside or top-down information. This is called bottom-up processing which is regularly based on the front-end parameterization of the connected speech. Also, the extracted features such as LP-coefficients, MFCC, or FFT spectrum are used (SaiJayram et al., 2002). The behavior of selected parameters can lead to indications for discovering the segment border. To show changes in the speech signal and often with relatively low computational overheads, the reduced extracted features, which are suitable to be a basis for detecting probable lexical or phone borders, can be used. On one hand, these reduced extracted features describe each speech segment in a demonstrative way in which other similar speech segments can be grouped and categorized together by matching their reduced extracted features. On the other hand, the divergent speech segments will be excluded.

Around the idea of executing the steps of classification and segmentation, several algorithms are evolved such as (Eriksson, 1989; Siegler et al.,1997; Cettolo et al., 2005; Aguilo et al., 2009; Bhandari et al., 2014). Traditionally, three groups form the segmentation approaches, that are mainly based on three different concepts which are energy-based, model-based and metric-based. The first algorithm is based on energy and it only depends on the running power in the time-domain. In contrast, the model-based in addition to the metric-based methods rely on statistical models. The advantage of the energy-based model is that it can be simply applied in which silence times are expected to be the segment boundaries. Using a predefined threshold and energy rate, the silences in the speech signal can be identified. Nevertheless, there are some challenges for many implementations when using this method (Bhandari et al., 2014). The techniques that are classified in this category can be divided into two groups (Giannakopoulos & Pikrakis, 2014):

- Signal change detection where the results simply includes the endpoints of the detected segments. Information related to the individual labels is not returned by the algorithm.
- Segmentation with clustering where the identified segments are clustered and the resulting clusters are used to assign labels.

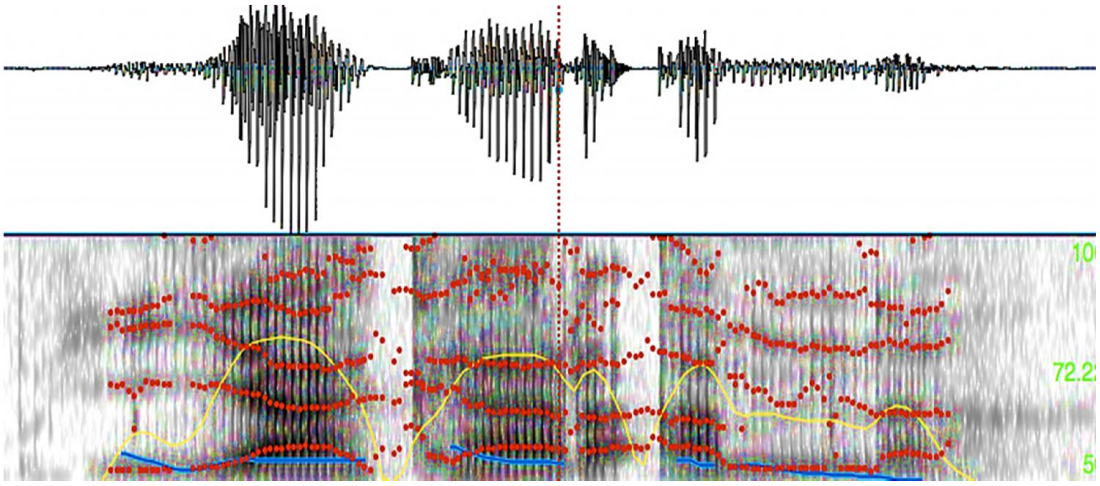


Figure (3.1): Word segmentation (a sequence of discrete words).

The core idea of signal change detection, as shown in Figure 3.1, is to define when substantial changes arise in the content of the connected speech signal. The detected changes determine the boundaries of the resulting segments. Depending on the nature of the signals being studied, the detection of endpoints may use a wide range of features and detection methods.

In lexical segmentation level, professional listeners hear connected speech as a sequence of discrete words, but there are no clearly pauses in the waveform. 8-month-olds can segment the connected speech into word units using only statistical computation as several studies have been discussed (Aslin et al. 1997, 1998; Mattys et. al, 1999). The child has no lexicon to back up, so the knowledge which the child is exposed to must be the same as what is used to learn how to segment properly.

A variety of computational approaches have been used to model the detection of word forms from continuous speech by infants. (Larsen et al., 2017) present the word segmentation algorithms as follows:

Diphone-Based Segmentation (DiBS) algorithm, based on phonotactics properties, which keeps the frequency of two phones occurring together in memory then it decides to place a boundary between them by computing Bayes' Theorem, which defines the probability of an event, relies on previous knowledge of conditions that might be related to the event. The model has several assumptions to do that. The first one is when the learner knows the phonetics categories. The second assumption is that the learner able to detect utterance boundaries. The third assumption assumes phonological independence across word boundaries. The forth assumption tracks context-free distribution of diphones and the last one is when the learner knows the relative frequency of word forms that have been already learned.

Transitional Probabilities (TPs) over syllables which suggest a boundary between two syllables if its co-occurrence probability is locally lowest (relative threshold) or is lower than an absolute threshold that are generally computed by taking the averaged value of syllables pairs.

Unigram Adaptor Grammar (AGu) which is an optimal learner model such as when a learner has an infinite memory and a batch process and looks at the whole corpus before segmenting.

3.2 Keywords Extraction

Automatic identification of terms is the key task of keywords extraction in which they best describe the subject of a topic. Key segments, key terms, key phrases or just keywords are the lexicon which is used for describing the terms that represent the most relevant information contained in the topic.

As in written text based, (Oelze, 2009) divides existing keywords extraction methods into four groups: simple statistics, linguistics, machine learning and mixed approaches.

Simple Statistics Approaches

The name of these approaches indicates that they are simple and they need limited requirements without needing the training data. The focus in these approaches is on speech non-linguistic features, such as inverse document frequency, term frequency and the position of keywords. In order to identify the keywords in the spoken files, some statistics knowledge can be used. N-Gram statistical information is commonly used in the direction of indexing the document automatically. Word frequency, word co-occurrences, …,etc. are presented by (Matsuo & Ishizuka, 2004) as statistics methods that can be used. The fact that the pure statistical methods generally ensure that the production of good outcomes and their ease of use are the advantages of these methods.

Linguistics Approaches

The linguistic features of the words, sentences and the whole files are used in these approaches. The methods which benefit from linguistic features which are syntactic structure, semantic qualities and part-of-speech are liable to add value. A number of researchers compare the advantages of using the lexical resources to a statistical method and relative frequency ratio (Plas et al., 2004). Different methods of integrating linguistic features into keywords extraction are studied by (Hulth, 2003). Three commonly used features evaluate the keywords which are term frequency (TF), Inverse Document Frequency (IDF) and the position of the word. Linguistic features lead to notable enhancement of the automatic keywords extraction. Actually, linguistic methods can be mixed or combined with some common statistical approaches in order to create new enhanced approaches.

Machine Learning Approaches

In general, keywords extraction is considered as a supervised learning method. The machine learning methodology procedure firstly provides a set of training spoken files to the system, each of which has a range of selected keywords by human. In order to find keywords from new spoken files, the gained knowledge is applied. Using spoken language processing methods, (Suzuki et al., 1998) select the keywords from spoken files. As a guide for relevance, the researchers use an encyclopedia. They separate the process into a term-weighting part and a keywords extraction part. Feature vectors are produced from different encyclopedia fields and the same process is done on the data of newspaper articles. The produced vectors from both articles and encyclopedia are compared using a likeness calculation in order to classify the article vectors into different fields. Extraction of keywords is done by analysing a segment where the most relevant field is selected for it by using the produced feature vectors in the second part. The technique of phoneme recognition is engaged to ensure the analysis. The keywords are allocated to the spoken files segment when the system selects the best fitting field.

Mixed Approaches

The methods, which are mentioned above, are mainly combined or some heuristic knowledge can be used in the task of keywords extraction, such as the length, position, html tags around of the words, layout feature of the words, …,etc. The resulting approaches about keywords extraction are called mixed approaches. The indication of the relevant works tells how the automatic keywords extraction methods are less costly and faster than if the extraction is done by human intervention. Moreover, the automatic keywords extraction reaches the precision of the human. Nevertheless, the proposed methods for automatic keywords extraction need some requirements like training examples or some specific information of domain in interest. Mixed methods present some enhancements in dealing with keywords extraction because it benefits from advantages of the involved approaches.

3.3 Topic Identification Concept

TID concept is based on the assumption that a binary tag can be placed on each spoken file for a single topic in order to show whether the topic is relevant or irrelevant. Apparently, the most understandable and commonly used method in performing speech-based TID is an ASR model which is used to handle the speech data (spoken files) and to extract the assuming transcript. The resulting transcription directly passes into the typical TID system that identifies the topic based on widely used text methodologies. Combinations of various packages, such as ASR, machine learning, NLP, information retrieval and text classification can be used to handle the TID for spoken files. Several of these packages, mainly ASR, are regularly based on large linguistic resources. Most of the spoken files identifiers are built based on combining an ASR with a TID algorithm. Such a combination consists of four different units which are the speech recognizer, a feature extractor, a feature transformation and the topic identifier (Sandsmark, 2012). Other spoken files identifiers are built to omit ASR unit in order to improve the outcomes and to handle the ASR production challenges. Figure 3.2 shows both traditional and modern identification of spoken files methodology. The modern spoken files identification methodology discards the ASR unit and only bases on the speech features. The core units are used in this thesis are: feature extraction and topic classification units.

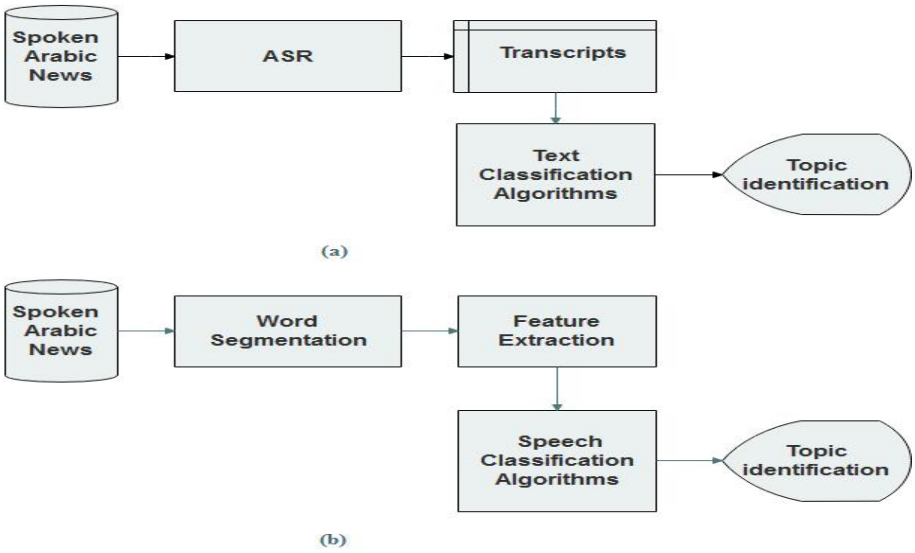


Figure (3.2): (a) ASR-based methodology (b) Without ASR methodology.

3.3.1 Feature extraction module

As it is known, all of the acoustic information are carried in the time domain waveform of a speech signal. It should be taken into consideration that data differs from information and the waveform can only be measured as a raw data. depending on the phonological opinion, very little benefits can be gained relying on the waveform itself. In some pattern recognition systems, one of the first decisions is the selection of the suitable features needed to be used. In order to make the identification algorithm's task easier, the basic signal that needs to be identified must be precisely represented (Prithvi & Kishore-Kumar, 2016). Generally, feature extraction is a critical processing phase in machine learning and pattern recognition tasks. The main objective is extracting a set of feature vectors from the considered dataset. These vectors must be useful in regard to the preferred characteristics of the fundamental data. The feature extraction is able to be seen as a data rate reduction technique for the reason that the small number of features should be used relatively in order to build the analysis algorithms (Giannakopoulos & Pikrakis, 2014).

Feature extraction of speech methodologies are required for the transformation of the original spoken files signals into a series of speech feature vectors measurements that covers only information required for the classification of a predefined topic. As each speech has different distinctive characteristics included in spoken words, these distinctive characteristics can be obtained from an extensive feature extraction methods and are able to be engaged to speech recognition tasks. Nonetheless, certain criteria of the extracted feature ought to be met while dealing with the spoken files signals. One of the criteria examples is that extracted speech features have to be computed effortlessly. Other examples of the criteria are the stability of the extracted features with time and the adaption of the extracted speech features with environment and noise challenges. By using spectral analysis methods, such as Mel-frequency cepstral coefficients (MFCC) and linear predictive coding (LPC) wavelet transforms, the feature vector of speech signals are usually extracted. The main feature extraction techniques used with speech signals are discussed more specifically below:

3.3.1.1 Linear Predictive Coding (LPC)

In the feature extraction for spoken files signals examination phase, one of the most common techniques used is the linear prediction method which obtains the essential factors of speech signal. It considers the human tract construction to make available an accurate evaluation of the speech signal factors when some lexical are spoken. It is categorized as a time domain methodology. LPC examines the spoken files signals via approximating the formants in addition to removing their extraordinary effects from the speech signal and estimates the loudness. Each of them is expressed as a linear arrangement of the preceding samples in LPC system. The primary goal of LPC method is the approximation of a known speech sample as a linear combination of former speech signal segments. A distinctive parameter set coefficients can be found as a result of decreasing the total of squared differences to its minimum rate between the definite speech segments and computed values. The factor values predictor forms the basic block for LPC of speech signal. LPC is a good technique in which it is used to warp or compress the speech signals in digital signal processing (DSP). For calculating the elementary factors of speech, LPC arises to be the most important technique because it provides an accurate estimation of the speech factors in addition to that it is an effective computational model. The following figure shows the processes of LPC feature extraction system presented by (Vimala & Radha, 2012).

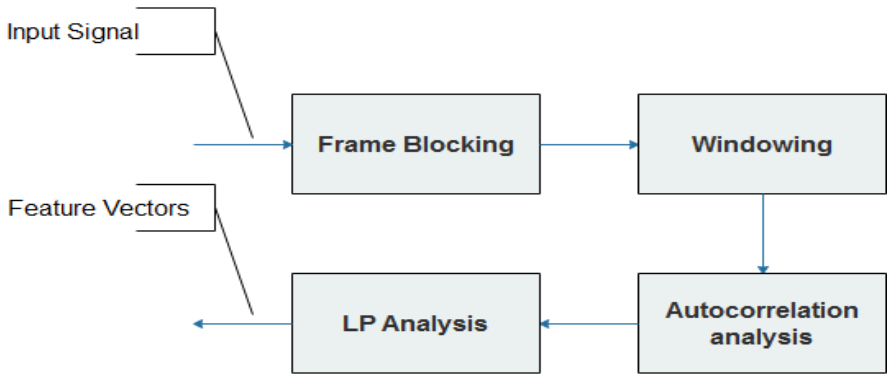


Figure (3.3): Stages of LPC.

Frame blocking process task examines the signal over several segments which should be short, these short segments are called frames. By applying a short-duration

(commonly 20-30 msec) overlapping window (typically Hamming) to the spoken file signal, this signal is segmented into many frames. The measurements of a forward LP (which is called LP coefficients) are determined by using LPC via reducing the prediction error rate in the least squares sense. In order to find the filter coefficients, LPC follows the autocorrelation method of autoregressive (AR) modeling. Although the data series is really an AR procedure of the precise order, the produced filter does not model the process precisely. Because the autocorrelation method indirectly windows the data, it assumes that speech signal samples beyond the length of *s* are zeros.

In the process of the creation of LP in human speech, the vocal tract shape controls the nature of the produced sounds. With the aim of studying the properties quantitatively, a digital all-pole filter models the vocal tract. In z-domain, The next equation describes the filter model transfer function.

$$V(z) = \left(\frac{\mathrm{G}}{1 - \sum_{k=1}^{p} a_k z^{-k}}\right) \tag{3.1}$$

Where *V(z)* is the transfer function of the vocal tract. $\{a_k\}$ is a set of autoregression coefficients called LPC and G is the filter gain. *P*, The upper limit of summation, is the all-pole filter order. LPC set defines the features of the transfer function of the vocal tract. LPC uses autocorrelation method constructing a matrix of autocorrelation of the windowed speech frame and simultaneous equations in order to evaluate the LPC set and the filter gain.

3.3.1.2 Perceptual Linear Prediction (PLP)

The Perceptual linear prediction technique is technologically advanced by the Hermansky. The main objective of PLP is removing the undesirable information of the speech signal, so the speech recognition rate is enhanced. PLP is similar to LPC except for the alteration of spectral features of LPC in order to match the human auditory system characteristics (Kishori et al., 2015) .

The PLP coefficients are created from the LPC coefficients by performing perceptual processing before performing the Autoregressive modeling. In the next figure, the perceptual front-end phases are obtained. Figure 3.4 shows the most processes involved in PLP.

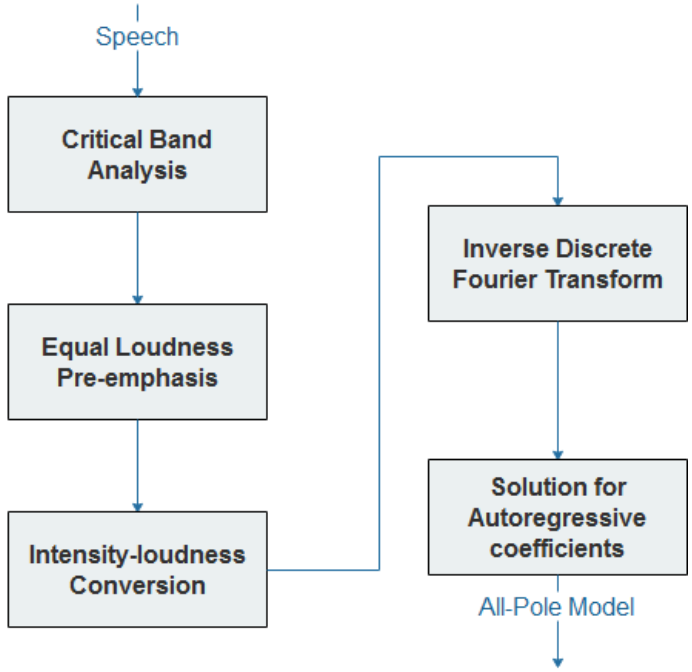


Figure (3.4): Calculating Perceptual Frontend Processing (Hermansky, 1990).

The main steps of PLP are summarized as follows:

- By using the Hamming window, subdivide the speech signal into frames. After that, Fast Fourier Transform (FFT) is performed on the windowed frame in order to get the speech power spectrum.
- In the PLP analysis, a non-linear frequency scale which is called the Bark scale is the basic of the filter bank. The filter weights and sum of the product of each FFT value on the speech spectrum shape the output of a filter.
- Within a number of diverse frequencies, the sensitivity of human hearing system is approximated by the equal-loudness curve. the corresponding equal-loudness weight weighs each filter output in the filter bank.
- By using Intensity-loudness, compression definite signal intensity is approximated. It refers to the non-linearly relationship between speech signal intensity and the perceived loudness.

- LP analysis is performed, a digital all-pole filter can be used with the purpose of modeling the vocal tract. There are several methods for evaluating the LPC feature sets and the gain of the filter. The autocorrelation technique is one of them in which it consists of constructing a matrix of both simultaneous equations and the autocorrelation of the windowed frames.
- Cepstral analysis is performed, this denotes to the procedure of discovering the speech series cepstrum. The cepstrum is equally considered the Inverse Fourier Transform (IFT) of the logarithm of a signal's spectrum.
- The spectral peaks which is located in the auditory-like spectrum is approximated proficiently using autoregressive modeling. With an exponential window, the cepstral measurements of the PLP model are recursively calculated and weighted. Consequently, all coefficients have a range for each input which is similar to a neural network range. For each frame, eight cepstral coefficients, containing log power, of a PLP model are produced.

3.3.1.3 RASTA-PLP

RASTA-PLP stands for Relative Spectral Transform - Perceptual Linear Prediction. As presented previously, Hermansky proposes PLP as a method of warping spectra in order to minimize the differences occurring between various speakers with the concern of maintaining the significant speech information (Prithvi & Kishore-Kumar, 2016). In each frequency sub-band, A band-pass filter is applied to the energy through RASTA in order to smooth across short-term noise differences and to discard any constant offset caused from static spectral pigmentation in the speech signal channel e.g. from a telephone line channel (Hermansky & Morgan, 1994). A special band-pass filter is added to each frequency sub-band in classical PLP procedure with the purpose of smoothing out short-term noise differences and in order to remove any constant offset in the speech channel. The next figure shows the most processes that comprise the RASTA-PLP method. They also contain, as in PLP, the computation of the critical-band power spectrum, the alteration of spectral amplitude over some compressing static nonlinear transformation, the use of the band pass filter, the filtering of the time path for each altered spectral element, the

transformation of filtered speech through enlarging static nonlinear transformations, the imitation of the power law of hearing and finally the calculation of an all-pole model of the spectrum (Hermansky, 1990). An all-pole model is used by PLP model in order to smooth the changed power spectrum; then, the cepstral coefficients are calculated.

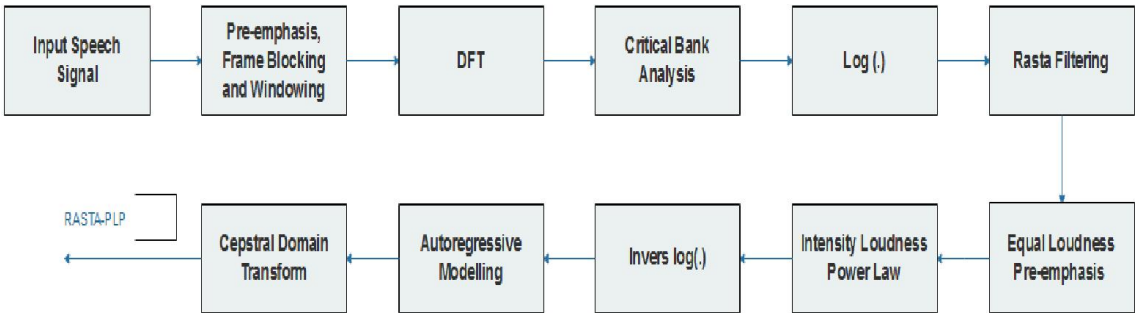


Figure (3.5): RASTA-PLP block diagram (Hermansky, 1990).

RASTA-PLP steps for each analysis frame are summarized by (Hermansky & Morgan, 1994) as follows:

1. Critical-band power spectrum is computed (as in PLP).
2. Spectral amplitude is transformed via compressing static nonlinear transformation (RASTA).
3. The time path of each transformed spectral component is filtered (RASTA).
4. The filtered speech signal representation is transformed via enlarging static nonlinear transformation (RASTA).
5. In order to simulate hearing, Multiply signal by the equal loudness curve and exponentiate by 0.33 (PLP).
6. An all-pole model is computed or Autoregressive (AR) of the outcome spectrum, this is done by go along with the conventional PLP technique (RASTA-PLP).

3.3.1.4 Mel Frequency Cepstral Coefficients (MFCC)

Another standard method can be considered for feature extraction is the use of MFCC. The main purpose of this method is to moderate the frequency information of the speech signal into a less number of factors that matches the effective factors of

the ear in separating critical bands, i.e., it tries to code the information as the human cochlea does. Furthermore, modeling loudness perception as in the human auditory system is done by the logarithmic operation in MFCC. After that, the Mel-frequency scale matches to a linear scale below 1 kHz and algorithmic scale above the 1 kHz (Sivaram & Hermansky, 2011). Regardless of the fact that MFCC is a very basic model of auditory treating system, MFCC is easy and comparatively fast to compute. In ASR systems, about of 20 MFCC coefficients usage is public. While for coding speech, 10-12 coefficients are usually considered to be adequate (see, e.g., Hagen et al., 2003). The most remarkable downside of using MFCC, which makes researchers keep searching for more reliable methods headed for describing the speech signal, is its sensitivity to noise as a result of its reliance on the spectral form. With the intention of overcoming this problem, the information in the periodicity of speech signals should be used although the speech signal includes aperiodic content (Ishizuka & Nakataani, 2006).

The main concentration of using MFCC methodology is on the short-term analysis. The feature vector is calculated from each frame distinctly. The MFCC coefficients are computed by taking speech segment as an input and at that time hamming window is put into operation in order to decrease the incoherence of a signal. At that point, FFT is applied to generate the Mel filter bank. After that, it combines the log total power around the center frequency in a critical band according to Mel frequency warping. After warping process is done, the numbers of coefficients are formerly acquired. Finally, The Inverse Discrete Fourier Transformer (IDFT) is employed for the computation of cepstral coefficients. IDFT performs transformation from the log of the domain coefficients in the direction of the frequency domain where the length of the FFT is (N).

MFCC can be computed by using the formula (Vimala & Radha, 2012).

$$\text{Mel}\,(g) = 2595 * \log10(1 + \frac{g}{700}) \tag{3.2}$$

The next Figure 3.6 shows the included steps in MFCC (Vimala & Radha, 2012).

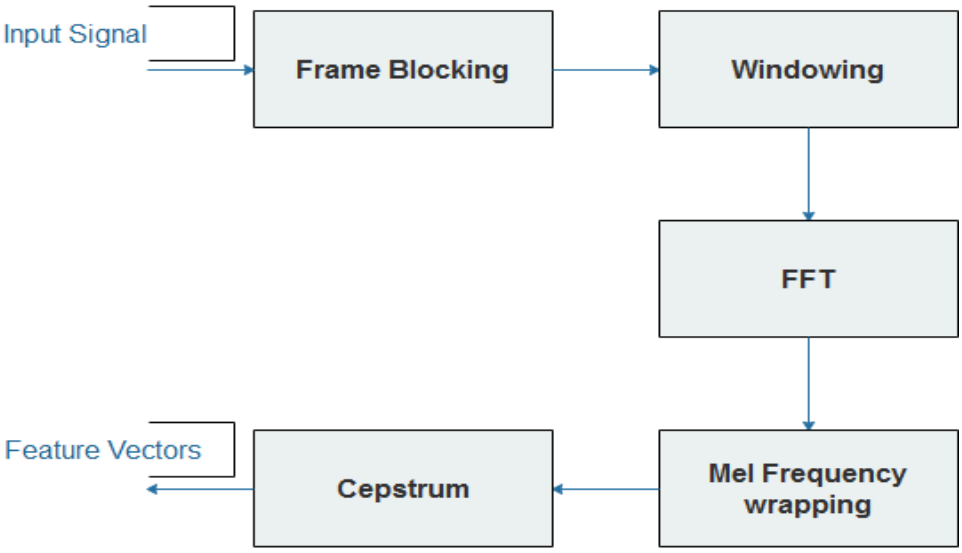


Figure (3.6): Stages of MFCC (Vimala & Radha, 2012).

To sum up, the MFCC extractor has seven processes, most of which are similar to LPCC. First, the speech sequence passes through a first order digital filter for pre-emphasizing step. Second, subdivision of the speech sequence into frames is done using Hamming windows. Third, speech signal spectrum is obtained by operating FFT to each speech frame. Fourth, the human auditory system is mimicked using the Mel frequency filter bank which is a series of triangular band-pass filters in which the Mel-scale is a non-linear frequency scale. Fifth, The filter bank is produced depending on the Mel-scale which is a measure of tones arbitrated by human. Sixth, MFCC algorithm decorrelates and reduces autocorrelation within a signal while preserving other aspects of the signal. Finally, the filter outputs and Discrete Cosine Transform (DCT) is operated on the filter outputs (Cheng et al., 2005).

RASTA-PLP is used for segmentation step, because RASTA-PLP smooth across short-term noise differences and remove any constant offset caused from static spectral pigmentation in the speech signal channel. In order to easily define the word boundaries, A band-pass filter is applied to the energy in each frequency sub-band through RASTA. MFCC is used for TID step, due to the less number of factors used to match the effective factors of the ear in separating critical bands. MFCC is widely used in speech recognition field because it is easy and comparatively fast to compute.

3.3.2 Topic Classification Module

The final step of TID system, once a feature vector is specified, is to generate identification scores and decisions using a TID identifier for each topic. The identifiers which are applied to the TID problem in this thesis are: DTW and HMM.

3.3.2.1 Dynamic Time Warping

DTW is considered as a time series alignment methodology established initially for speech recognition task. It is an algorithm for evaluating similarity among two time-based series that may have speed variances. Because of speaking rates differences, a non-linear variation happens in speech pattern contrasted with time axis that needs to be removed. Dynamic programming (DP) focuses on a pattern matching method that utilizes a time normalization consequence. By using a non-linear time-warping function, the variations in the time axis are modeled. For example, the timing differences of any two-speech patterns can be discarded by warping the time axis of one of them; hence, the maximum score is achieved with the second one. Additionally, a smaller amount of distinction can be obtained between patterns having its place in different groups when the warping function is permitted to take any probable value. Therefore, restrictions are carried out on the warping function slope in order to improve the distinction between patterns having its place in dissimilar groups (“Dynamic time warping”, 2017).

With specific restrictions, generally, DTW is an algorithm that computes a best match between two given time series. "Warping" the sequences non-linearly in the time domain in order to define the distance measuring their likeness ratio are independent of specific non-linear variations in the time aspects. In the time series classification, DTW algorithm is often used. In addition to a similarity amount between the two series, what is called "warping path" is formed and the two signals may be aligned in time by warping according to this path. To clarify the concept, it is assumed that: **X** (original) is the signal with an original group of points, so **Y** (original) is converted to **X** (warped), **Y** (original). “Warp” the time series does not essentially need both points to have the same X-axis value with the intention of measuring the distance between each point; this is the core idea behind DTW. As a substitute, the points can be selected even though they are away in an attempt to

reduce the overall distance between the sequences. The initial and latest points is restricted by DTW algorithm to be the start and the end of each series (Sakoe & Chiba, 1978). Starting from that point, the points matching are imagined as a path on an **n by m** lattice, **n** and **m** are considered as the number of points in each time sequences as obtained in Figure 3.7 ("Reverse Engineering DynamicHedge's Alpha Curves", 2013)

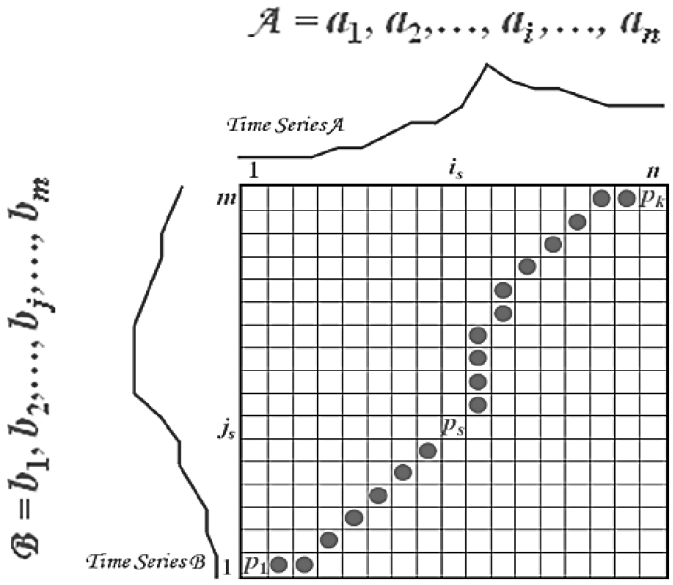


Figure (3.7): Warping path between time series A and B.

A distance value can be placed inside an individual cell. It describes the matching value resulting from comparing the elements of both series. The main purpose of the algorithm is to reduce the total distance value between series. To do that, a sequence needs to discover a path through the lattice. The minimum overall distance value defines the best match or alignment between both series. The process of calculating this total distance contains: finding all possible routes through the lattice and computing the total distance for each one. The total distance is defined as the lowest of the sum of the distances among the individual elements on the created path divided by the sum of the weighting function. In order to control the path length, the weighting function is applied.

For the clarification of DTW algorithm, it is assumed that **A** and B are two sequences (time series) and **P** is a warping function as shown in Figure 3.7. In order to find the best alignment between the both series A and B, one needs to find the path through the lattice

$$P = p_1, \dots, p_s, \dots, p_k$$

$$p_s = (\boldsymbol{i_s}, \boldsymbol{j_s})$$

p_s which minimizes the total distance between A and B.

D (A, B) is the time-normalized distance between both series and calculated by:

$$D(A, B) = \frac{\sum_{s=1}^{k} d(p_s) \cdot w_s}{\sum_{s=1}^{k} w_s}, w_s > 0$$

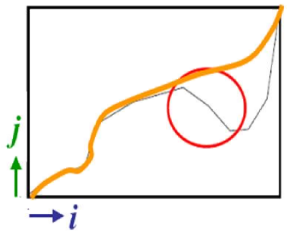
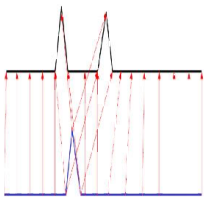
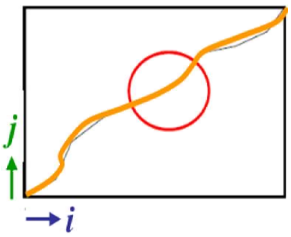
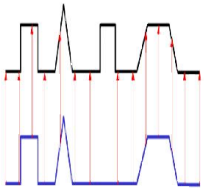
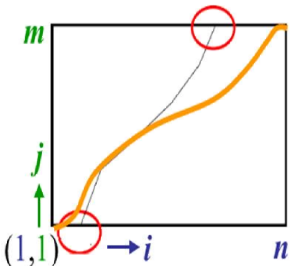
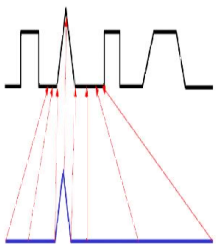
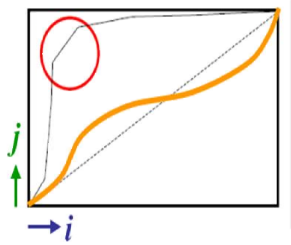
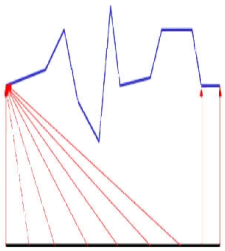
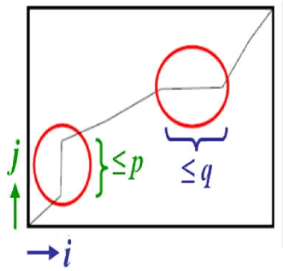
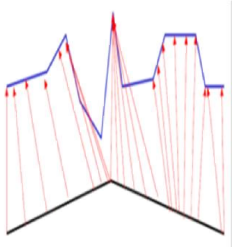
w_s: weighting coefficient, $d(p_s)$: distance between i_s and j_s (3.3)

Best alignment path between A and B: $p_0 = \arg\min_p(D(A, B))$ is independent of the warping function. In order to choose the weighting coefficient, a weighting coefficient function should guarantee that: $C = \sum_{s=1}^{k} w_s$. Thus $D(A, B) = \frac{1}{c} \min_p[\sum_{s=1}^{k} d(p_s) \cdot w_s]$

It is obvious that, for any notable long sequences, the number of probable paths through the grid is very large. The main optimizations or constraints of the DTW algorithm arise from the observations on the nature of acceptable paths through the grid as shown in table 3.1:

- **Monotonic condition**: the path does not turn back on itself. Both i and j indexes never decrease; they either stay the same or rise.
- **Continuity condition**: the path progresses just one-step at a time. So, i and j indexes can only rise in at most one on each step alongside the path .
- **Boundary condition**: at the bottom left, the path is started and at the top right, the path is ended.
- **Warping window condition**: a suitable path is improbable to move far away from the diagonal line. The window width is the distance value which is allowed to move path.
- **Slope constraint condition**: too declined path or too thin path should not be allowed. This is for excluding the short series matching the extreme long series.

Table (3.1): Restrictions on the warping function (Tsiporkova , 2017).

		**Monotonic** **Agreements**: In the alignment, the features are not repeated.
		**Continuity** **Agreements**: Important features must not be omitted.
		**Boundary** **Agreements**: One of the sequences is not considered partially by the alignment.
		**Warping window** **Agreements**: The alignment should not skip dissimilar features and become stuck at similar features.
		**Slope constraint** **Agreements**: The very short parts of the series are prevented from matching very long ones.

The transfers that can be made from any point in the path are limited by the previous mentioned restrictions. Consequently, the paths number that needs to be measured is limited. As a replacement for finding all possible routes over the lattice which satisfy the above restrictions, the DTW algorithm controls the paths by observations of the cost of the best path to each point in the lattice. DTW algorithm tries to find which path has a minimum overall distance path, but this is cannot be possible during the calculation, so DTW algorithm traces back the results when the last point is gotten (Sakoe & Chiba, 1978).
Pseudo-code below describes the core of DTW as (Sakoe & Chiba, 1978) present.

For each line i of the grid from 1 to I

For each row j of the grid from 1 to J

$$\text{compute } D(i,j) = \min_{\substack{i-1 \le k \le i \\ j-1 \le l \le j \\ (k,l) \neq (i,j)}} \{D(k,l) + cost(i,j)\}$$

return $D(I,J)$

3.3.2.2 Hidden Markov Models

Identification by HMM has achieved a significant progress in the past decade. HMM theory has been applied widely in certain fields as speech and speaker recognition. HMM is a statistical Markov model where the system being modeled is supposed to be a Markov process with unseen (hidden) states. An HMM can be presented as the simplest dynamic Bayesian network.

The HMM is a finite set of states, each of which is related to a (generally multidimensional) probability distribution. A set of probabilities controls the transitions among the states which are called transition probabilities. In a specific state, an outcome or observation can be generated along with the associated probability distribution. The state is unseen to an external observer and only the outcome is visible. Therefore, states are “hidden” to the outside.

For clarification, there is the state (x) that changes with time (Markov) and it should be guessed or tracked. Inappropriately, this state cannot directly be seen (hidden). Nevertheless, the state (y) can be observed which is related to it.

The general architecture of an HMM is shown in the figure below. Each oval shape represents a random variable that can adopt any number of values. The random variable **x(t)** is the unseen state at the time **t** with $x(t) \in \{ x_1, x_2, \dots , x_n\}$. The random variable **y(*t*)** is the observation at the time **t** with $y(t) \in \{ y_1, y_2, \dots , y_n \}$. The arrows in the figure represent the conditional dependencies ("Hidden Markov model", 2017).

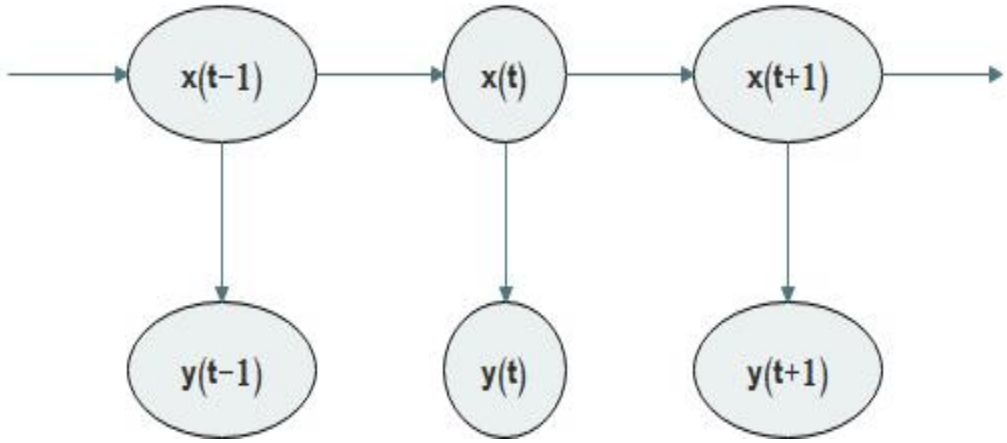


Figure (3.8): Trellis diagram ("Hidden Markov model", 2017).

Each set of states in HMM has a limited number of transitions and emissions. Each transition between states has an assigned probability and each model starts from the starting state and ends in the ending state. This is understandable when looking at Figure 3.9.

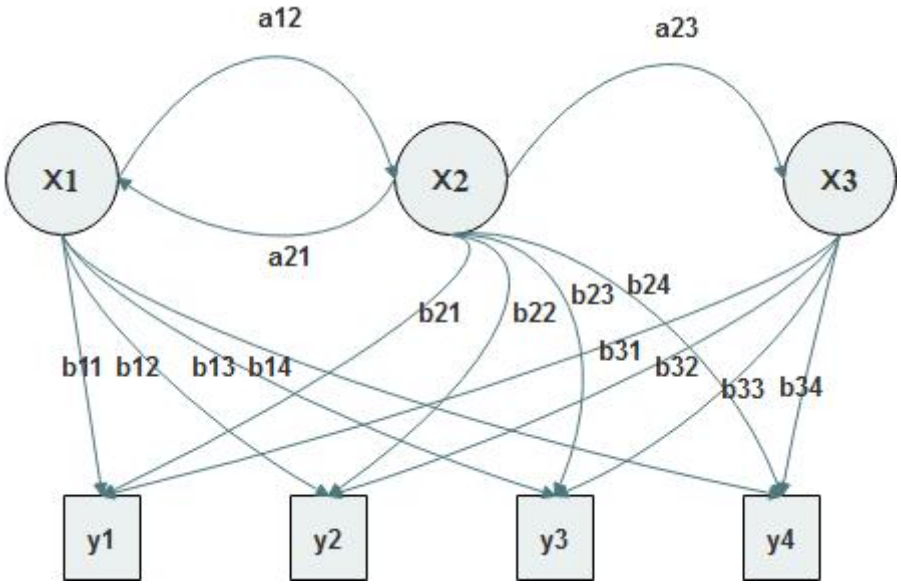


Figure (3.9): Probabilistic parameters of a hidden Markov model (example).

(X) refers to states, (y) refers to possible observations, (a) refers to state transition probabilities and (b) refers to output probabilities

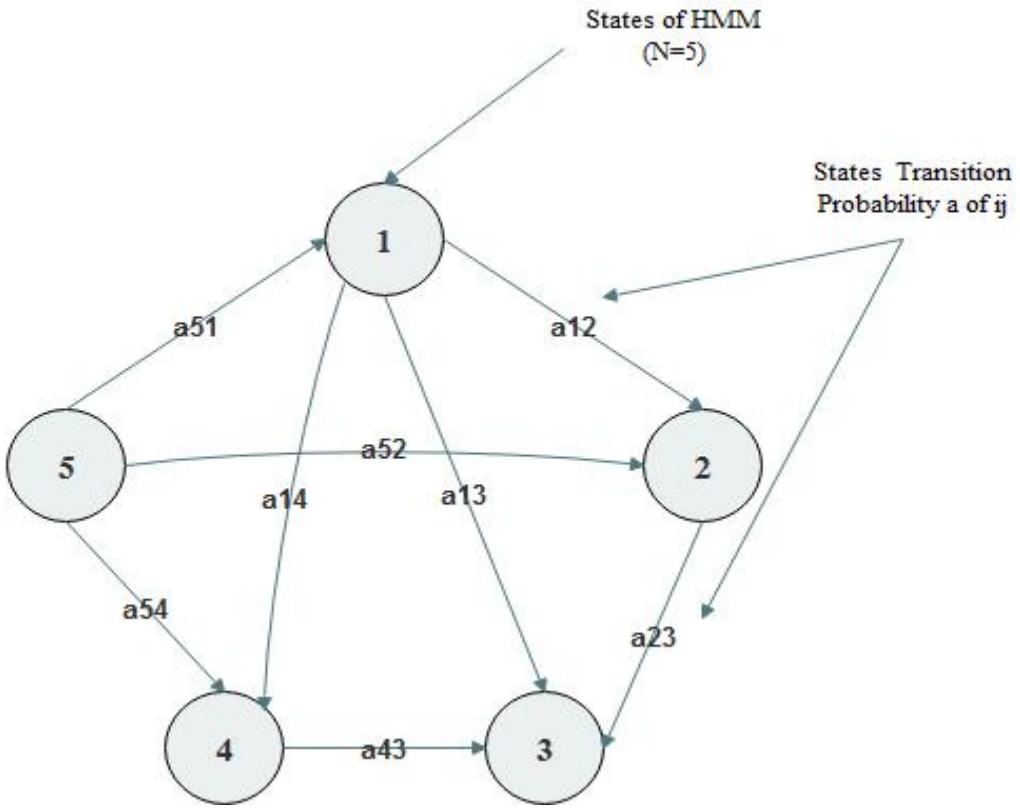


Figure (3.10): Elements of HMM.

As shown in Figure 3.10 HMM is characterized by the following parameters (Saxena, 2015):

1. *N*, the number of hidden states.
2. *V*, the number of distinct observation symbols per state M, $V = V_1, V_2, \cdots, V_M$
3. $\{a_{ij}\}$, the state transition probability distribution.
4. $\{b_{jk}\}$, the observation symbol probability distribution in state j. (the emission probability)
5. $\{\pi_i = P\ [w\ (1) = w_i]$, the initial state distribution.
6. The complete parameter set of the model $\lambda = (\{a_{ij}\}, \{b_{jk}\}, \{\pi_i\})$.

Keep in mind that an HMM is a generative model, that generates or emits sequences. HMM can work as a creator to produce an observation sequences $O = O_1\ O_2\ O_3 \cdots O_T$. depending on the value of *N, M, a, b* and *π*. Therefore, characterizing HMM using suitable value can enhance the HMM-based systems. There are three basic problems related to HMM and they are needed to be solved for real-word applications. These problems are evaluation, decoding, and learning problems.

Decoding

It exposes the hidden part of the problem. In practice, the optimal criterion is usually used in order to find the best possible solution, but there is no precise solution to this problem. Decoding problem answers the question: After being given a sequence of symbols (observations) and a model, what is the most probable sequence of states that are generated the sequence?

GIVEN: HMM $\lambda = (A, B, \pi)$, and observation sequence O

FIND: corresponding state sequence $Q = q_1\ q_2 \cdots q_T$ (q_T is actual state at time t) which is seen as optimal.

Evaluation

Evaluation problem answers the question: if the observation sequence and model are set, how to deal with the calculation of probability value where the observed sequence is produced by the model as a scoring problem. When choosing between many competing models, a model, with maximum probability, gives an enhanced result as shown in Figure 3.11. Thus, the evaluation problem answers the question:

what is the probability of the fact that a specific sequence of symbols is generated by a particular model?

GIVEN: HMM $\lambda = (A, B, \pi)$, and observation sequence O,

FIND: *P (O|λ),* the probability of the observation sequence which gives the best model.

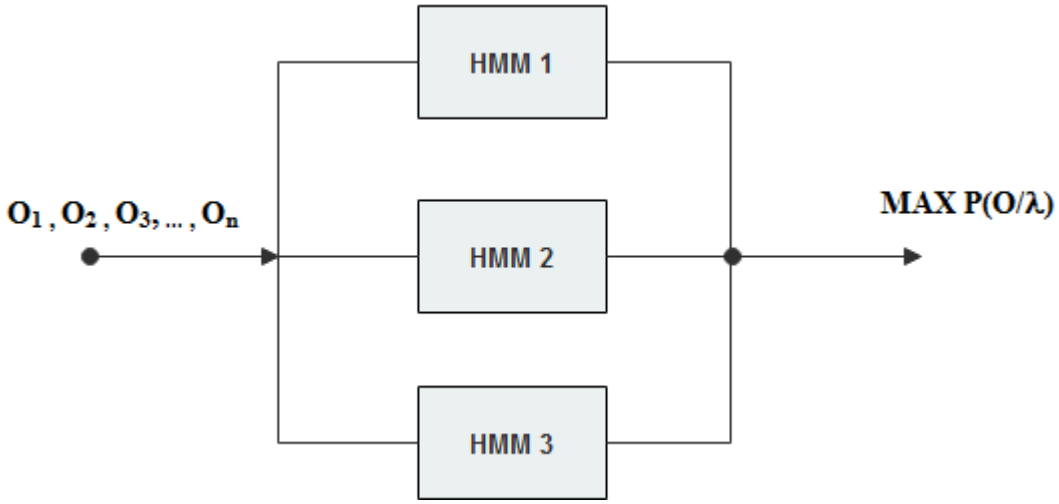


Figure (3.11): Evaluation problem (Different HMM with different parameters).

Learning problems

The observation sequence used here is called "training" sequence as it is used for training HMM. Training is one of the vital elements of HMM. It allows to adjust model parameter to create best model for a given training sequence. Learning problem answers the question: After being given a model structure and a set of sequences, how to discover the model that best fits the data?

GIVEN: HMM λ, with unspecified transition/emission probs

FIND: Model parameter $\lambda = (A, B, \pi)$ in order to maximize $P (O|\lambda)$.

There are several solutions to these problems, such as the Forward–Backward algorithm, the Viterbi algorithm, and the Baum-Welch algorithm (Rabiner, 1989; Meng et al., 2006).

The Forward/Backward algorithm

The Forward/Backward algorithm is a DP algorithm used in HMM in order to powerfully compute the state posteriors over all the hidden state variables. In a posterior decoding, these values, which simply select the state with the maximum

posterior marginal for each position in the sequence, are then used ("Forward-Backward", 2017).

Viterbi Algorithm

The Viterbi algorithm is a DP algorithm. *p(O|λ) is* calculated by summing the overall state sequences. Sometimes it is preferable to approximate *p(O|λ)* which uses all state sequences with *ṕ(O|λ)* which uses the single most likely state sequence. The Viterbi algorithm finds the most likely state sequence ("Hidden Markov Models – Forward & Viterbi Algorithm", 2014).

$$\acute{\rho}(O|\lambda) = \max_{x}[p(O, X|\lambda)]$$
$$where\ X\ is\ the\ most\ likely\ state\ sequence \quad (3.4)$$

Baum-Welch Algorithm

The adjustment of the model parameters (A, B, λ), which are adjusted to maximize the probability of the observation sequence, is the most challenging part of all. Any finite observation sequence are set as training data, but there is no optimal method of estimating the model parameters. Nonetheless, some heuristic procedures try to find a local optimization over global optimization. With the purpose of finding the maximum likelihood, the parameters of HMM are estimated. After being given a set of observed feature vectors, the Baum–Welch procedure utilizes the familiar EM method (Uchat, 2012). In the following steps, the Baum Welch works for each sequence in the training set of sequences as follows:

1. Calculate the forward probabilities using forward algorithm.
2. Calculate the backward probabilities with the backward algorithm.
3. Calculate the contributions of the current sequence to the transitions of the model
4. Calculate the contributions of the current sequence to the emission probabilities of the model.
5. Calculate the new model parameters (start probabilities, transition probabilities, emission probabilities)
6. Calculate the new log likelihood of the model
7. Stop when the change in log likelihood is smaller than a specified threshold or when a maximum number of iterations is reached.

To sum up, the Forward/Backward algorithm scores an observation sequence against the model. The Viterbi algorithm gets the most likely state sequence and the Baum-Welch algorithm catches the parameters of the model from the data.
Details of HMM are not discussed more in this thesis. For more information about HMM, refer to e.g. (Rabiner, 1989).

3.4 Arabic Language Challenges

One of the broadly spoken languages in the world is Arabic Language. It is a supreme language which includes a compound and abundant morphology. It is a vastly adjusted language and can adapt with extreme changes that happen because of the compound morphology that it has (Al-Harbii et al., 2008). 420 million people around the world speak Arabic, making it the sixth most spoken language ("Complete List of Arabic Speaking Countries 2017", 2017). The Arabic language offers researchers and developers of natural language processing (NLP) implementations for Arabic texts and speech with thoughtful challenges. With the evolution of the Arab internet users, the spoken Arabic texts and written Arabic texts have also increased. Arabic language has three forms; Modern Standard Arabic (MSA), Classical Arabic (CA), and Dialectal Arabic (DA). MSA is an adaptive language in which it has an evolving diversity of Arabic with constant borrowings and innovations showing that Arabic reinvents or reshapes itself to meet the changing requirements of its speakers. More than fourteen centuries ago, Arabs speak what is called CA. At the regional level, there are as several Arab dialects as there are different members of the Arab group (Farghaly & Shaalan, 2009).

In addition to forms of the Arabic language challenges, there is another issue which is that Arabic texts include many words which their spelling, in general, tends to be unpredictable in spoken Arabic texts. For example, the word " America" could be spelled 'أمريكا' 'amreeka' , 'أميريكا' 'amereeka', 'أميركا' 'amerka' , 'أمركا' 'amreka' and so on. Another big problem is the lack of a sizable corpus of spoken Arabic texts.

Moreover, Arabic language semantic processing is measured to be more sophisticated than other languages like English. This difficulty comes from the very derivational nature of the Arabic language. One word could have several synonyms

in Arabic language, for example the following word 'وقف ' could have various meanings as قام , ارتاب , منعه عنه ,فهمه و تبينه ,قصر نشاطه على, ... , etc.

The meaning of the spoken word depends on the context, vowelization 'tashkeel' and the pronunciation ways. There are some aspects that slow down the progress in Arabic NLP compared to the accomplishments in English and other languages (Farghaly & Shaalan, 2009).

When looking at the composition of the Arabic language as obtained in table 3.2, it is notable that it consists of 28 alphabet characters. Different styles of Arabic letters are shaped when the letters appear in a word. The style depends on the letter position in the word. Another issue that can shape the style of Arabic letter is the link possibility between the letter to its neighboring letters ("Arabic phonology", 2017)

Table (3.2): Arabic script letter names and sounds ("Arabic phonology", 2017).

7	6	5	4	3	2	1
خاء	حاء	جيم	ثاء	تاء	باء	ألف
khā'	**ḥā'**	**jīm**	**thā'**	**tā'**	**bā'**	**'alif**
14	13	12	11	10	9	8
صاد	شين	سين	زاي	راء	ذال	دال
ṣād	**Shīn**	**sīn**	**zayn / zāy**	**rā'**	**dhāl**	**dāl**
21	20	19	18	17	16	15
قاف	فاء	غين	عين	ظاء	طاء	ضاد
qāf	**fā'**	**ghayn**	**'ayn**	**ẓā'**	**ṭā'**	**ḍād**
28	27	26	25	24	23	22
ياء	واو	هاء	نون	ميم	لام	كاف
yā'	**wāw**	**hā'**	**nūn**	**mīm**	**lām**	**kāf**

The consistent diversity of Arabic which is used in writing and in formal speech is called standard Arabic. 28 consonant phonemes in addition to 6 vowel phonemes form the MSA. Phonemes contrast between "emphatic" consonants and “non-emphatic” consonants. As it is said, only six vowel phonemes are used in MSA. Two diphthongs (shaped by a grouping of short /a/ with the semivowels /j/ and /w/) without allophones are found in CA. Allophony is partly accustomed by consonants (exactly neighboring ones) within the same word. It is notable that a speaker's background defines the pronunciation.

This thesis deals mainly with MSA because MSA is the typical diversity common to educated speakers within Arabic-speaking areas. It is commonly used in writing in formal newspapers and orally in broadcasts and spoken texts.

Chapter 4
Proposed Work

Chapter 4
Proposed Work

In Chapter 4, the methodology and materials which are presented include SAN processing which includes segmentation process, keywords extraction, feature processing and extraction, and TID methods. Also, SAN dataset and evaluation metrics are presented.

4.1 General Steps of the SAN Identification Methodology

Obviously, the clearest way with the intention of performing TID process on speech collection is to handle these data by using ASR system. The offered transcript resulting from the ASR system is passed directly to process them by contemporary text-based TID system as shown in Figure (3.2.a). Inappropriately, this approach works well under the conditions that the speech collection and related transcriptions are similar in style. Another condition is the ability of the ASR system to produce high-quality errorless transcriptions. There is a significant interest in interdisciplinary combinations of many boxes, which are ASR, natural language processing, machine learning, information retrieval and text classification, to handle the speech-based TID problems. Most of these boxes, specifically ASR, rely on considerable linguistic resources. Furthermore, related black boxes in series tend to increase errors, especially when the key terms are OOV. In general, the system needs to be able to process spoken files with limited (or zero) resources. Speech processing without a dependency on ASR results is proposed directly as an improbable alternative (Dredze et al., 2010).

In the Figure 4.1, the main components in the presented TID methodology, without the necessity of ASR component, are described. The first step includes creating, categorizing and storing the dataset in SAN database. Next, SAN goes through processing step with the aim of using the processed SAN by the keywords extraction box. Keywords extraction algorithm is used to extract the most dominant keywords for each category and to store them in an organized manner in which the selected KW database can be suitable for TID process. The KW database is used by TID box with the purpose of identifying the SAN clip which has an unknown topic label.

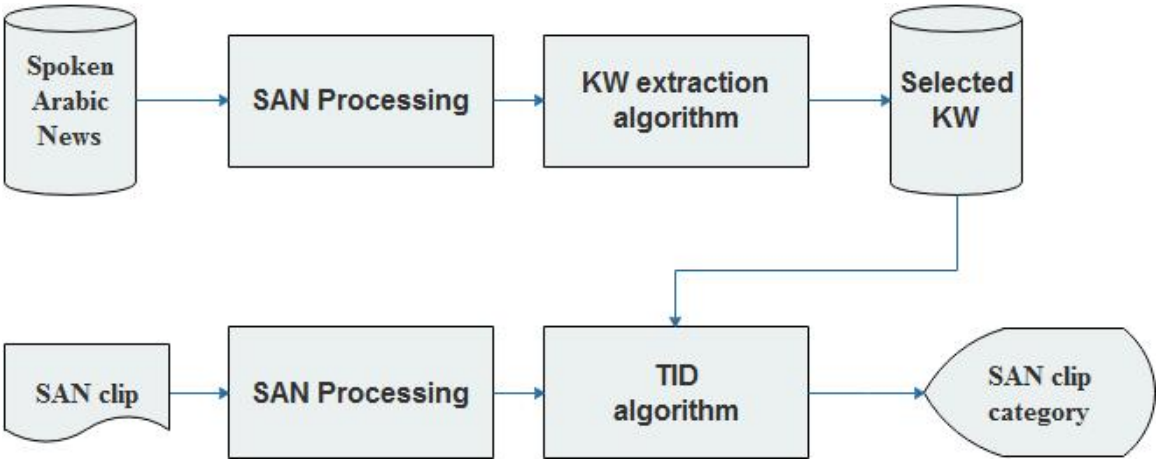


Figure (4.1): General boxes for topic identification methodology.

The next sections describe all boxes in more details:

4.2 SAN Processing

The first step in the SAN identification system is the SAN processing step which includes two main processes: speech pre-processing and speech segmentation process that are described in the subsections below.

4.2.1 Speech pre-processing

The preprocessing steps are performed on the speech signal with the purpose of emphasizing the effective frequency. Preprocessing on speech signals, such as isolating the voiced region from the silence/unvoiced portion of the signal, is usually supported as an essential step in the improvement of a consistent speaker or speech recognition system. This is for the reason that most of the speech or speaker specific attributes are presented in the voiced part of the speech signals (Saha et al., 2005)

4.2.1.1 Normalization:

In this thesis, an original spoken file is normalized to be at some programmed amplitude range (scaling all values in one SAN clip to its maximum)

$[\boldsymbol{a}, \boldsymbol{Fs}] = \boldsymbol{audioread}(\boldsymbol{clip1.wav})$;

After determining the sampled data $\boldsymbol{a}$, the equation 4.1 shows how to create the normalized signal via scaling all values to its maximum

$$normalised\_signal = (a.* \frac{abs(a)}{\max(abs(a))}); \quad (4.1)$$

4.2.1.2 Pre-emphasizing

A common practice for speech recognition is the usage of the pre-emphasis filter. The aim of pre-emphasizing is to amplify the higher frequency components of the speech signal with the intention of imitating the human ear additional sensibility to high frequencies. In general, this is done by using a high pass filter characterized by a specific slope. The speech signal pre-emphasizing has come to be at high frequencies as a standard pre-processing step. Lower frequencies are less significant for signal disambiguation than higher frequencies and in order to get slightly better results, the pre-emphasis becomes widespread. Another advantage of pre-emphasis is that it aids to deal with DC offset that is often present in recordings and it can enhance an energy-based voice activity detection. Pre-emphasis decreases the spoken file spectrum dynamic range in which the factors are estimated in an accurate manner using this method. Signal is passed through a filter with the purpose of emphasizing the higher frequencies and this is done under the pre-emphasis processes. pre-emphasis processes amplify the energy of the signal at higher frequency. Where the amplitudes of lower bands are decreased and the amplitudes of high frequency bands are increased, the pre-emphasis becomes a very simple signal processing method. It can be implemented simply as:

$$y(t) = x(t) + (1 - \alpha)x(t - 1) \quad (4.2)$$

y (t) is the pre-emphasized signal, x (t) is the original signal and typical values for the filter coefficient (α) are 0.95 or 0.97.

The equation 4.2 describes that low frequency signals are sampled at a highly sufficient rate, and are able to yield adjacent samples of similar numerical value. This is because low frequency, basically, means slow variation in time, so the numerical values of a low frequency signal can be subjected to change smoothly or slowly from sample to sample. Via the subtraction, the part of the samples that do not change in relation to its adjacent samples is removed (adjacent is specified by an exponential

window parameterized by α) thus, the part of the signal that changes rapidly is remained, i.e. its high-frequency components. The figure below obtains the SAN clip signal and its spectrogram before and after applying pre-emphasize filter effects.

SAN clip example

فاز المنتخب الفلسطيني على نظيره المنتخب البرازيلي في مباراة اليوم
'Faza almontakhab alphalasteeni ala nazeerehi almontakhab albarazeeli fee mobarati alyawm'

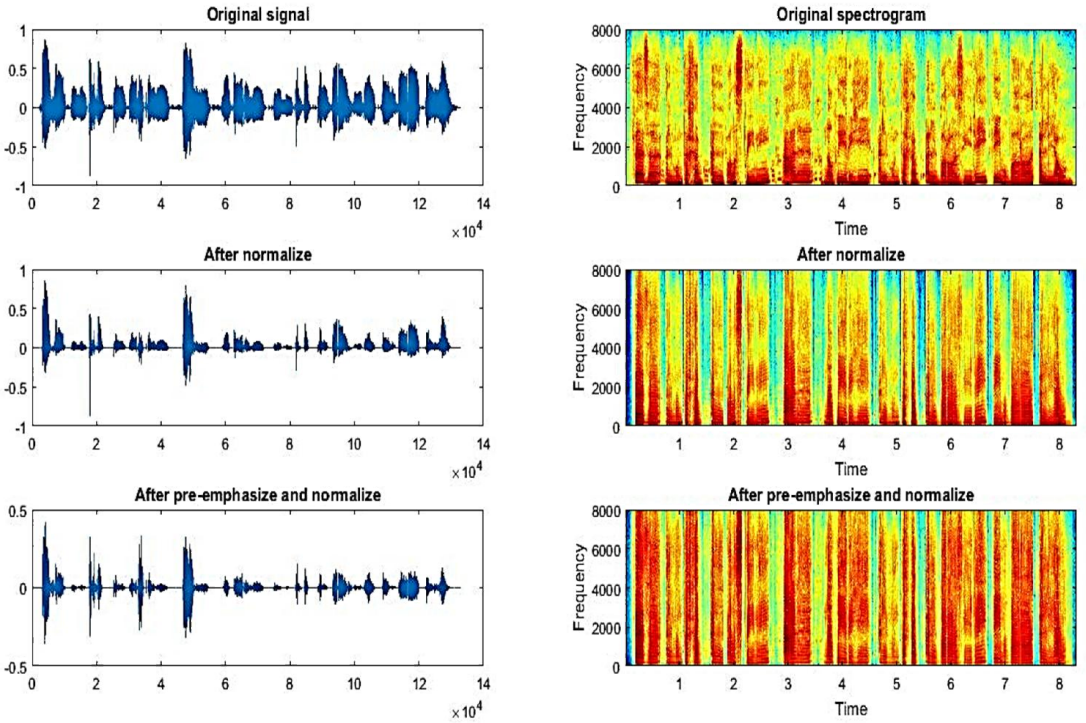


Figure (4.2): Applying pre-emphasize and normalization effects on speech signal.

Figure 4.2 describes the effect of preprocessing steps. As shown in the figure, the energy for low frequency is normalized to be zero while the energy for high frequency is emphasized to be clear. This is done in order to facilitate the segmentation process. To sum up the pre-emphasis filter is valuable because it: balances the frequency spectrum, bypasses numerical problems during the Fourier transform operation and improves the Signal-to-Noise Ratio (SNR).

4.2.2 Speech Segmentation

Speech segmentation process mainly focuses on how to identify the borders between words, syllables, or phonemes in spoken natural languages. In this thesis, the processed SAN clip is divided into regions, each of which corresponds to only one word. This process is known as speech segmentation at 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 the speech segmentation across languages especially in Arabic language. In this thesis, a novel algorithm to segment SAN clip into its words based on distinctive boundary normalized features is proposed. Based on various time and frequency domain features, the boundaries of the homogeneous regions are decided. The main idea of the proposed segmentation methodology is the signal change detection method, which determines when significant continues 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 detection of endpoints may use a wide range of features and detection techniques depending on the nature of the signals being studied. Together, the length of spoken word frame is used to confirm the endpoint of the word.

The general steps of an unsupervised detector of signal changes suppose that the input to this process is just the speech signal (SAN clip) and that the output consists of a list of pairs of endpoints.

1. Extract the feature vectors.
2. For each pair of successive feature vectors, compute a dissimilarity measure (distance function). Accordingly, a sequence of distance values is generated.
3. Detect the local maxima of the previous sequence. The locations of these maxima are the endpoints of the detected segments

The segmentation method proposed in this thesis is a little bit different. It is a bottom-up blind speech segmentation algorithm. The overall principle of the algorithm is to track the amplitude or spectral changes in the signal using short-time energy. The segment boundaries at the locations are detected where amplitude or spectral changes exceed a minimum threshold level, which is selected via trial and

error way (manual chosen). The main features used for segmenting speech signal are the normalized features extracted via PLP methodology.

The automatic speech segmentation system is implemented in Windows environment and MATLAB Tool Kit 2016 is used for developing this application. The proposed SAN clip segmentation system has six major steps as shown in Figure 4.3.

A. Speech Features Extraction (RASTA-PLP).
B. Processing and normalizing features (Binarization of Features 0 or 1 depending on some factors).
C. Define boundaries based on some criteria (Frame length and the number of continuous frames with specified values (neighbors).
D. Extract the MFCC feature for each identified word.

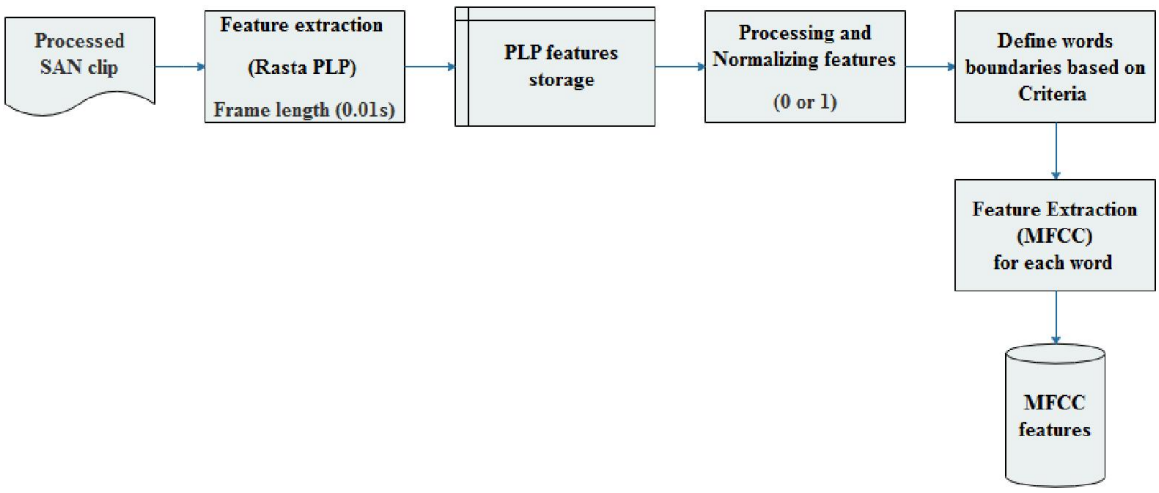


Figure (4.3): Segmentation processes.

The figure above shows the key steps of speech segmentation methodology. First, the feature vectors are extracted using RASTA-PLP ("Matlab Audio Processing Examples", 2012) from the processed SAN clip, produced by speech preprocessing step which is discussed in section 4.2.1. The second step is to perform some calculations on each feature vectors in order to identify the frames (frame length is 10ms) with the lowest value, which describes the space between each frame. This is done after calculating 12^{th} order PLP features without RASTA ("Matlab Audio Processing Examples", 2012), which is very good with the intention of identifying the word boundaries. In order to give more space to low frequencies

RASTA auditory warps the frequency axis and the technique in which RASTA filtering emphasizes the beginnings of static sounds like vowels. In order to locate the frames which are supposed to be spaces, power spectrum of a time series (dB is 10log10), which describes the distribution of power into frequency components composing that signal, is calculated. In relation to Fourier analysis, any physical signal is able to be decomposed into several discrete frequencies, or a spectrum of frequencies, the distribution of the energy of a waveform among its different frequency components, over a continuous range,.

The total of the power of a waveform among its different frequency components shown in Figure 4.4 for each frame is calculated. Then, a binarization process (0 or 1) on each frame total value is performed depending on a threshold which is selected based on the power spectrum values via trial and error technique. To identify the start and the end of the word boundaries, take into account the suitable length of a word and the neighboring frames values.

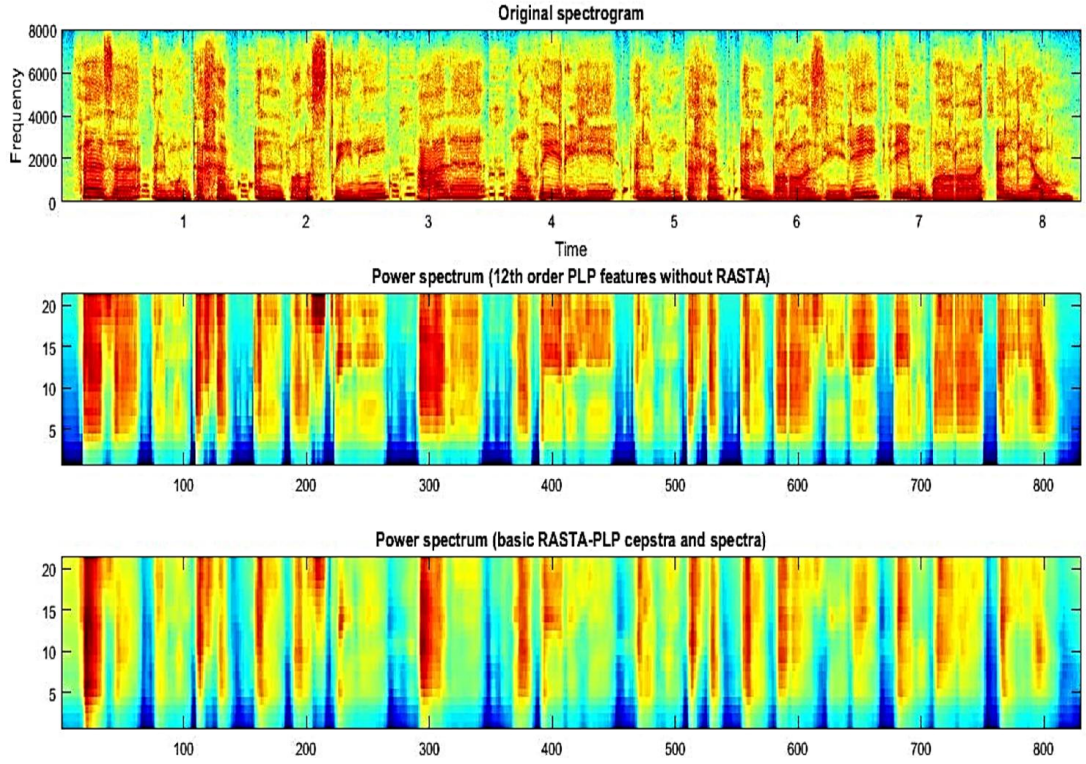


Figure (4.4): With and without RASTA-PLP features.

Figure 4.4 shows the calculated power spectrum that is used in a segmentation step and Figure 4.5 illustrates the binarization step that precedes the

identification of each word boundaries and finds the frame that are supposed to be neglected (space or zero).

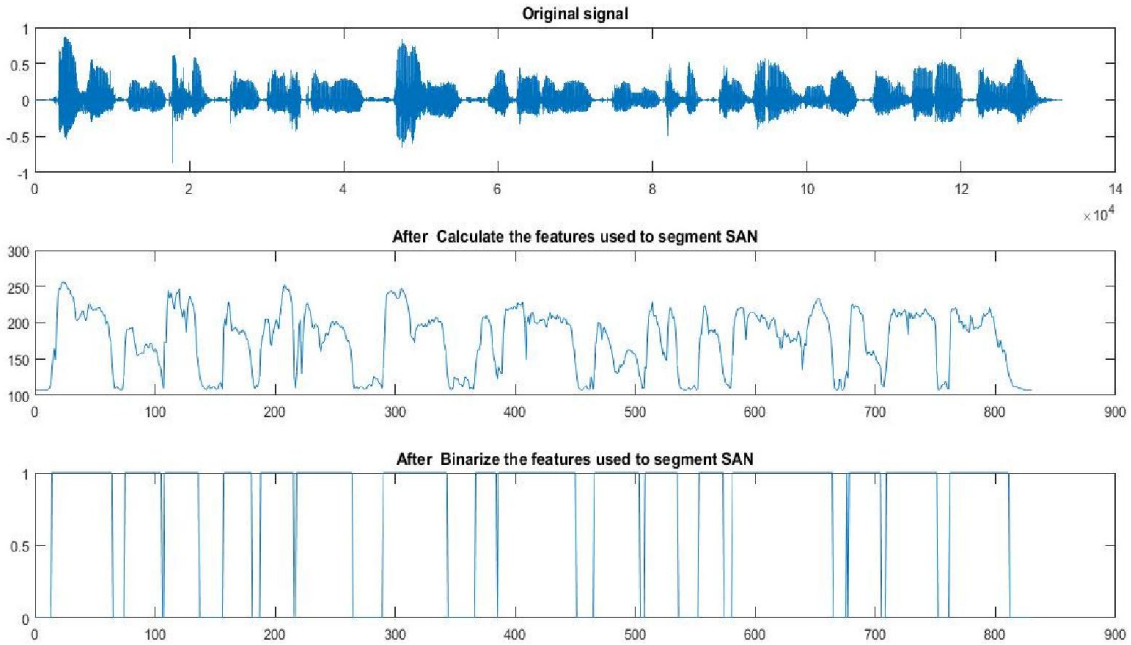


Figure (4.5): Calculate and binarize the features used in segmentation step**.**

The steps in order to facilitate the segmentation process of SAN clip are explained as steps follows:

1- Extract the feature vectors using RASTA-PLP.
2- Calculate the power spectrum of a time series (dB is 10log10).
3- Find the total power value at different frequencies for each frame (the sum of the resulting power values).
4- Normalize the total power values to 100.
5- Binaraize the normalized values depending on a threshold (as the value above or equal to 50 set to one and the value less than 50 set to zero).
6- Locate the indices of the word boundary by taking into account the number of the adjacent frames with zero or one values and the word length,.
7- Extract MFCC feature vectors of each word.

The segmentation idea is illustrated in the figure below, which obtains the segmentation of SAN clip example at word level depending on the steps described above:

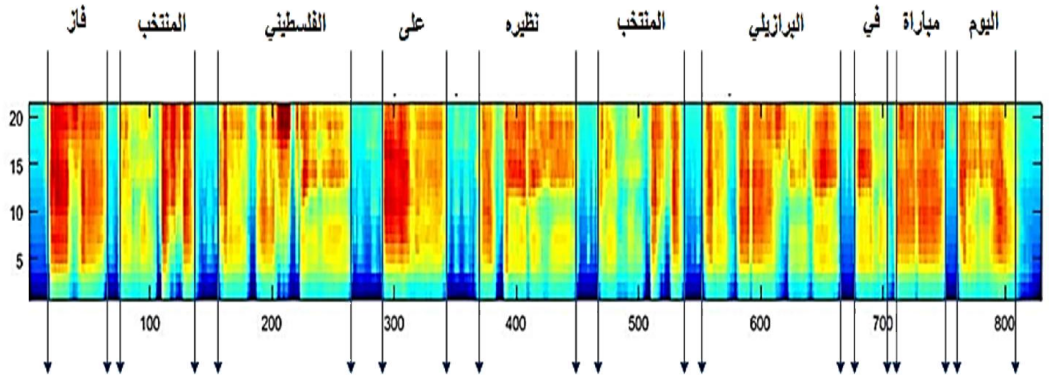


Figure (4.6): Segmentation of SAN clip (example).

4.3 Keywords Extraction

Keywords extraction algorithm consists of three stages, practically similar to existing systems for keywords extraction (Jean-Louis, 2014). Generating candidate keywords is the first stage. These candidate keywords include filtered segmented words (each word in the filtered words storage is considered as a candidate keyword). The second stage involves extracting features for each candidate keyword. These features capture the frequency of a word. The frequency of a word is computed by matching the candidate word with other words to find the distance between them. By using threshold distance, the considered matched words are counted. The final stage is a keyword selector (word ranking). The selector examines the rank of the word *p* for each candidate keyword. The candidate is selected as keyword if *p* is high enough. Figure 4.7 shows the main boxes in keywords extraction method. For each category, SAN clip is segmented using segmentation technique discribed in the previous section. The resulting segmented words, which are stored as features as obtained in the section above, are filtered using discarded words removal processes. Discarded words removal processes are based on a template matching method. The stored discarded words (prepositions, verbs, kāna and her sisters (كان واخواتها) …, etc) features are compared to the resulting segmented

words features in order to eliminate the discarded word using DTW algorithm. Next, filtered words are going through a word frequency counter as shown in the Figure 4.8. with the aim of calculating the word frequency, how often the word *w* appears in SAN clip *n*. The words frequencies are ranked using word ranking process. Finally, the candidate keywords are selectd depending on the word frequency.

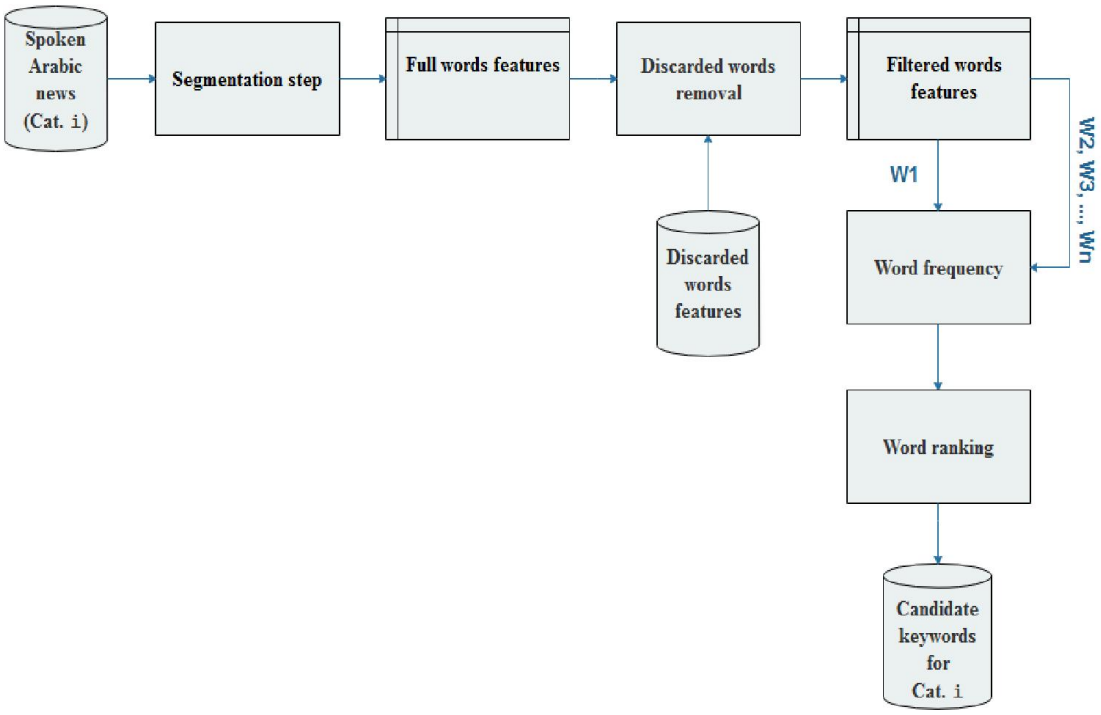


Figure (4.7): Keywords extraction process.

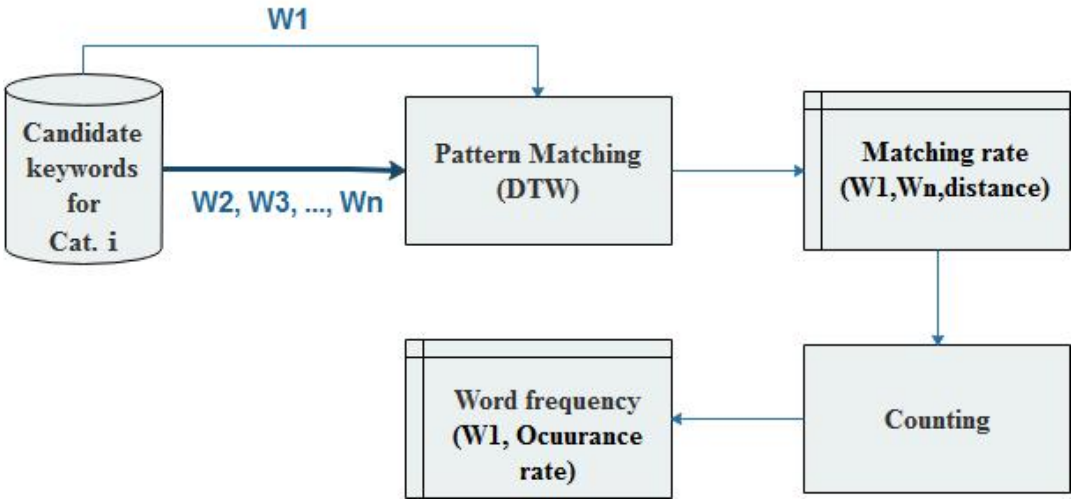


Figure (4.8): Word frequency process.

After extracting candidate keywords for each category, the candidate keywords for categories go through Mutually Exclusive Process (MEP). As shown in the figure below, MEP is used in order to select the keywords for each category. It is used to ensure that keywords in one category cannot be found in any other categories. An additional idea in order not to remove any keywords, which is repeated in another category, probability ratio are assigned to the word depending on the word frequency rate of this word in the categories.

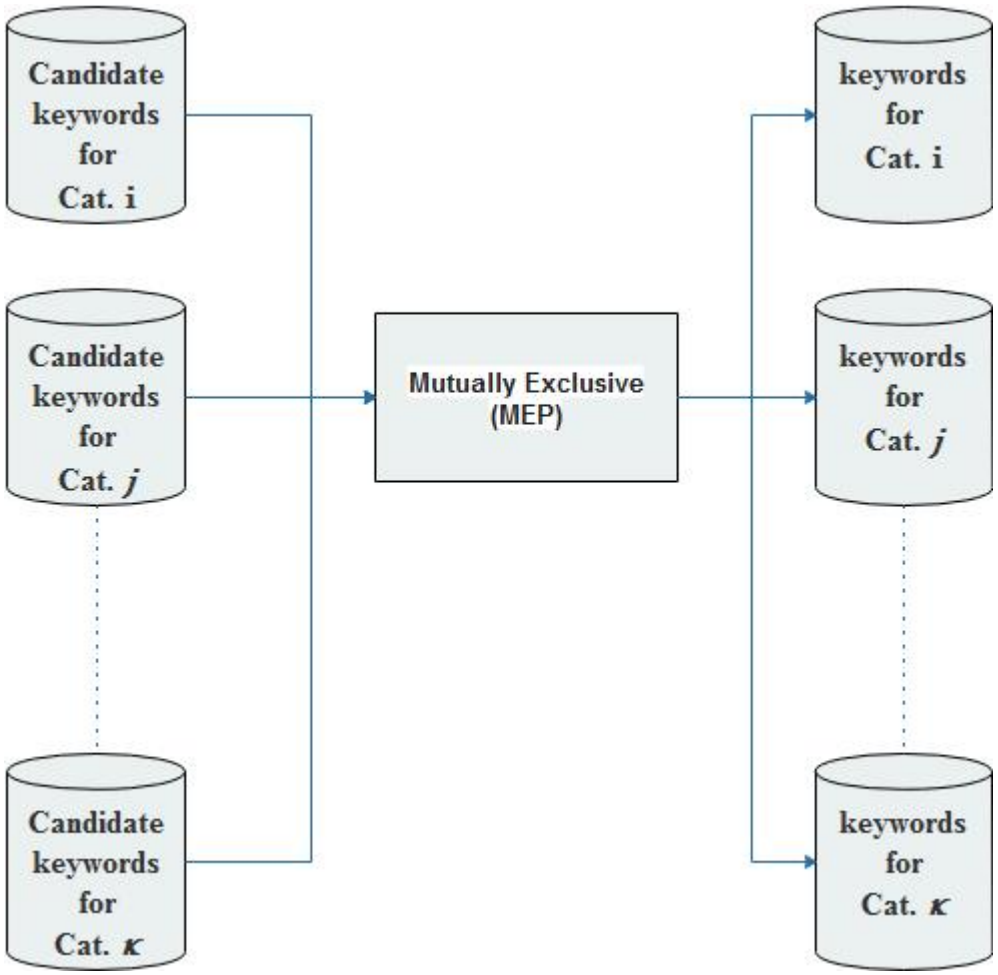


Figure (4.9): Mutually exclusive process.

The general steps used for keywords extraction methodology are summarized as follows:

1- Segment all SAN clips for each category and extract features.
2- Clean the segmented words (candidate keywords) by removing punctuations and stop words.
3- Compute the frequency of the words.
4- Rank the word (candidate keywords).
5- Apply MEP to all categories

Finally, the keywords for categories are stored in the keywords database in a structured tree pattern as shown in Figure 4.10.

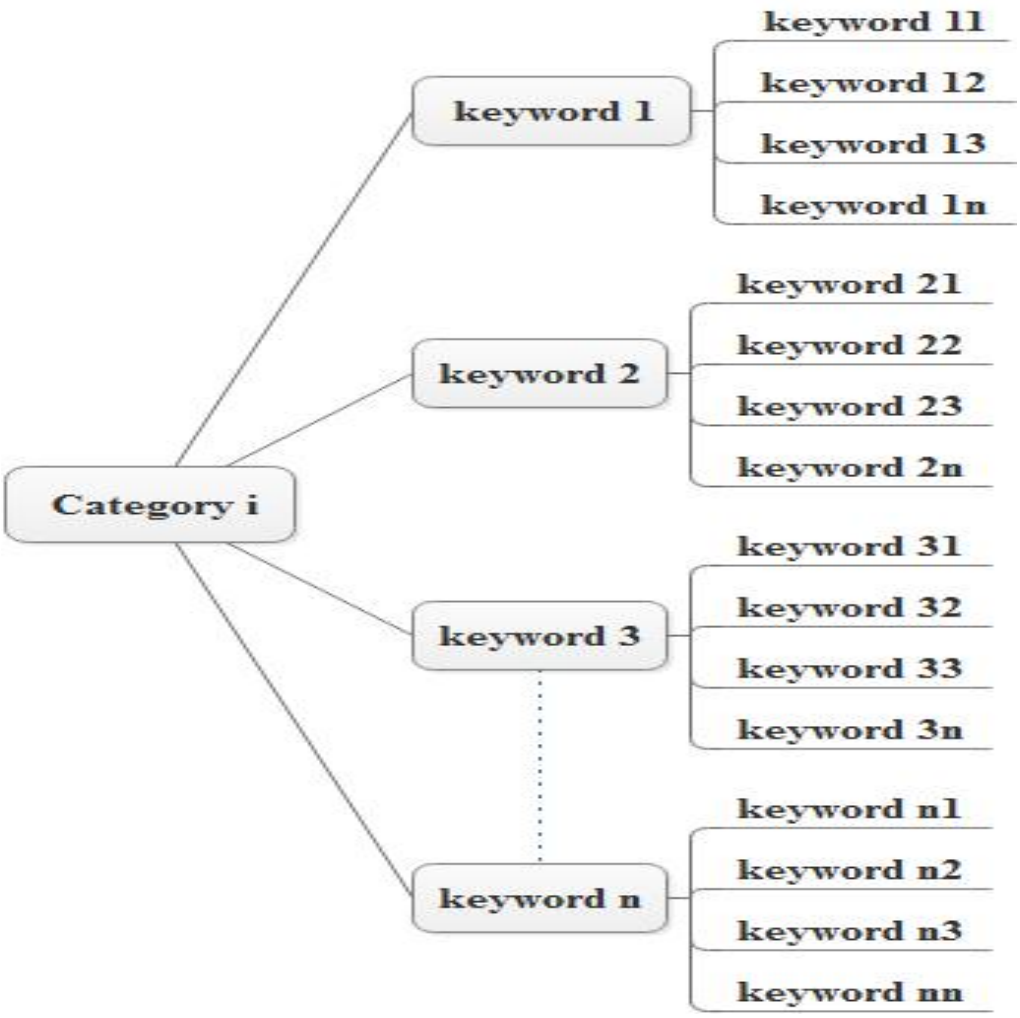


Figure (4.10): Keywords database structure.

4.4 Topic Identification

The last step in the SAN identification system is TID process. The TID process is based on the pattern comparison and scoring techniques. It is known that the pattern recognition technique involves two steps: pattern training and pattern comparison. Pattern comparison or matching is the method followed in order to compare or match speech features directly. The essential feature of this approach is to initiate consistent speech pattern representation for pattern matching from a set of labelled keywords samples. The stochastic approach is used in this thesis. Stochastic approaches are more suitable approaches as they use probabilistic models to deal with undetermined or partial information. Different methods like HMM, DTW, VQ etc., are defined as stochastic approaches. Among all these methods, HMM is the most popular stochastic approach nowadays (Kishori et al., 2015).

With the aim of categorizing a word, the evaluation of a matching ratio or distance between the word and one of keywords has to be calculated. Regarding the other words calculations in all classes, a probable predefined class is identified by discovering the best class matching ratio or distance value of the word. The distance calculation and comparison need an estimation of the correct value to match the extracted features of the words. As discussed in Ch. 3, the feature term refers to a piece of information which can be used for such a purpose. The similarity (or dissimilarity) of two words can be measured via comparing features.

The two main modules form the pattern matching system are the feature extraction and the feature matching. The feature extraction module purpose is to convert the speech signal to some kinds of representations for extra examination and treating. This extracted information is known as a feature vector. In order to complete the process of converting SAN clip signal to feature vector, signal-processing module is used. The commonly used method for extracting feature factor is MFCC. As shown in Figure 4.3, the input is the voice sample (SAN clip) and the output becomes MFCC feature vector after the segmentation step. With the feature matching process as shown in Figure 4.11, the extracted MFCC feature vector from unknown spoken news word sample is compared to acoustic model (keywords). The model succeed when it has a max score (the model with low distance succeed). The output of this system is considered as a matched word. It is notable that, in the

acoustic model, DTW or HMM is used with the purpose of scoring the model by distance or Logliklehood metric.

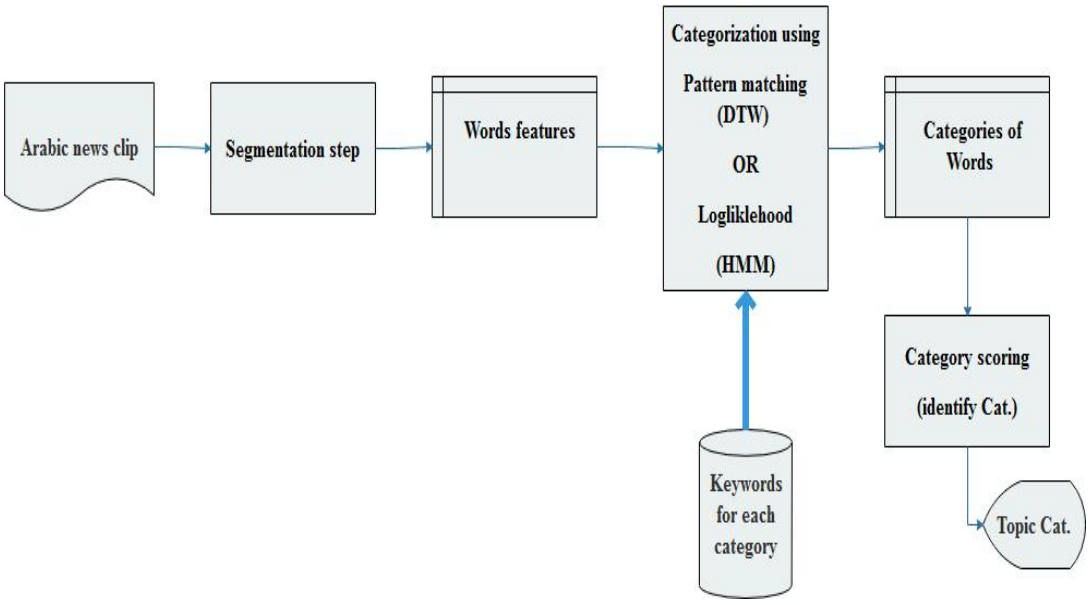


Figure (4.11) : Topic identification process.

After selecting the keywords for each category, the general steps of TID process are shown in Figure 4.11 and summarized as follows:

1- Read the input file (SAN clip **C**) at that time preprocessing and segmentation steps are implemented in order to obtain (**C** = $\mathbf{w_1}$, $\mathbf{w_2}$... $\mathbf{w_n}$) where w_n is the separated word.

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

3- For each word **w**, one of the pattern matching algorithms is applied in order to decide to which category **w** belongs. For example, $\mathbf{w_1}$ is the MFCC features of the first segmented word and $\mathbf{kw_x}$ is the MFCC features of the keywords in the **x** category**.** for example, $\mathbf{kw_s}$ are the MFCC features of the keywords in the sport category sport ($\mathbf{kw_s}$ = $\mathbf{kw_1}$, $\mathbf{kw_2}$, ..., $\mathbf{kw_n}$), noticing that $\mathbf{kw_1}$ is a container of several different samples of $\mathbf{kw_1}$ spoken by different people.

a. By using DTW, the distance between $\mathbf{w_1}$ and each **kw** in the keywords database is calculated to select the least distance. If the least distance exceeds the set threshold, the word is omitted from calculations.
b. By using HMM, the logliklehood is calculated to select the most logliklehood. If the most logliklehood exceeds the predefined threshold, the word is omitted from calculations

For instance, $\mathbf{w_1}$ matches some samples of $\mathbf{kw_1}$ in $\mathbf{kw_s}$ (sport category). In order to select the most suitable pattern, the **kw** with least distance to $\mathbf{w_1}$ 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.

4- After selecting the most suitable representative, which belongs to one of the categories and represents the segmented word $\mathbf{w_1}$, the selected category is saved in the categories of words box.
5- Repeat the steps above from one to four for all segmented words.
6- Rank the frequency (count) of the categories in the categories of words storage; the maximum category count is selected as the identified topic of this SAN clip.

After the description of TID processes the whole components which are needed for TID system are summarized and shown in Figure 4.12

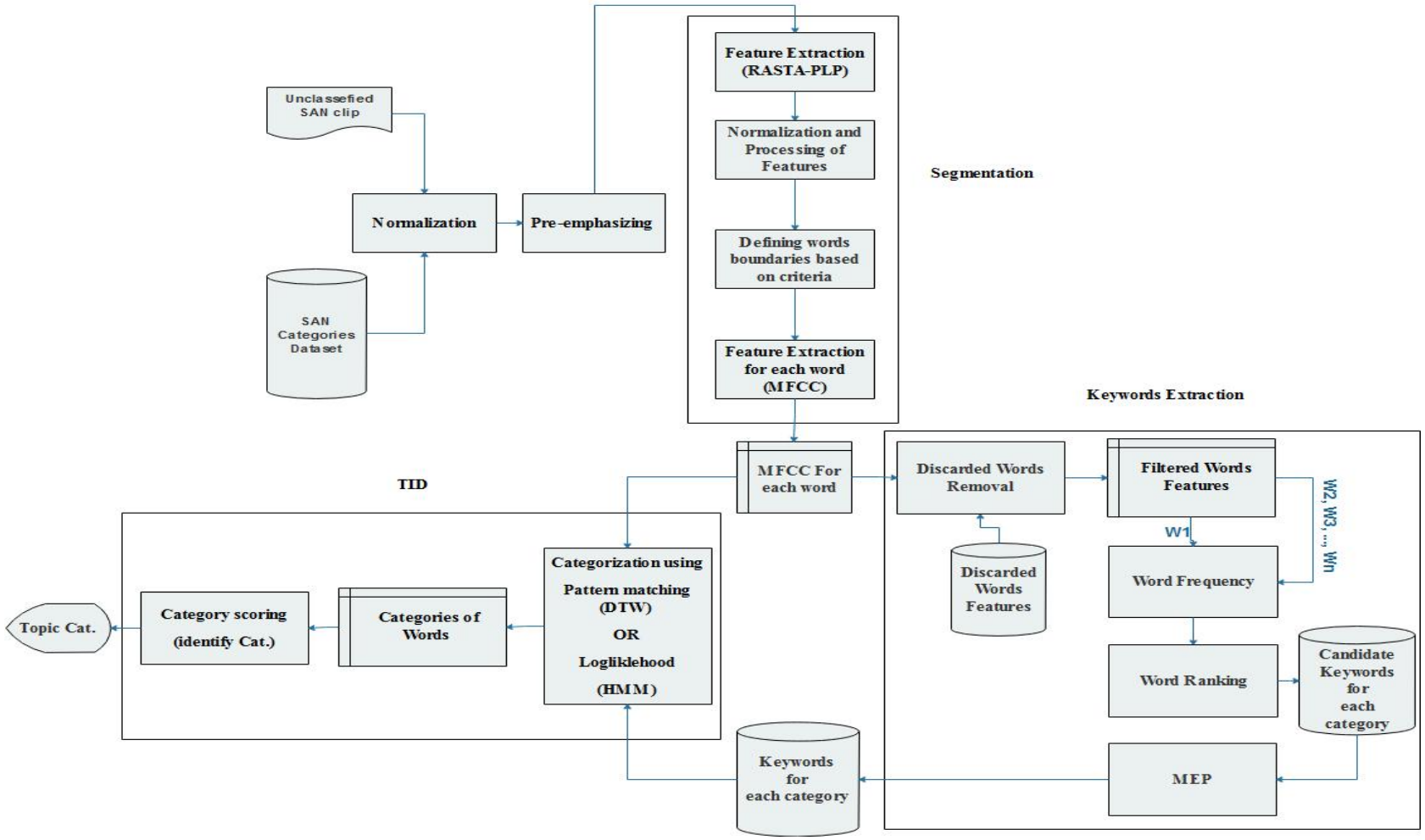


Figure (4.12) : The whole system processes.

Chapter 5
Results and Discussion

Chapter 5

Results and Discussion

This chapter describes the main experimental results of the proposed system. Section 5.1 describes the programming language and tools used in order to develop the proposed algorithms. Section 5.2 describes the dataset used for the keywords extraction and TID steps. The evaluation method is demonstrated in Section 5.3. Finally, Section 5.4 discusses the evaluation results for the proposed methods

5.1 Programming Language and Tools

For the implementation, Matlab programming language is used to implement the proposed methods because it contains many useful toolboxes for speech processing as a VOICEBOX. The version of Matlab that is used in this thesis is Matlab R2016b installed on a server (Dell T 620 Tower) that is 8 core, 2 GHz, 20 MB Cache, RAM 32GB DDR3 in order to accelerate the process and to save time and effort.

5.2 Dataset Description

It is not an obligation that these dataset must contain Arabic news with various topics areas. One of the most appropriate types of such dataset is spoken news. In this case, there is a speaker who reads the newsletter which talks about a well-known topic. These news may be collected from TV-news, radio, …, etc. the dataset for Arabic broadcast news is difficult to be found. Therefore, a new Arabic speech dataset is established in this thesis. It contains 240 files for each class (1440 files for all six classes) for the first stage (keywords extraction step) and 360 unidentified files for the TID stage.

SAN Dataset

SAN dataset is designed to be comparable to Arabic texts categorization dataset. Experiments are conducted on the widely-used Aljazeera news Arabic dataset (Alj-News). The dataset which is used in this study are constituted from Aljazeera News Arabic categories in addition to the additional weather category. 240 files for each category are used for keywords extraction step, which constitute

the total of 1440 files for 6 classes. 60 files for each class are used for TID step, which constitute the total of 360 files for 6 classes. Regarding that Alj-News is a collection of 1500 Arabic news files obtained from Aljazeera online news agency (Mohamed et al., 2005).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, economy and sport. It is worth mentioning that, regardless of the fact that the number of files in this dataset is small, the diversity among the nature of categories is big. All written texts files are converted to spoken files using special recorder by multiple speakers (exactly 30 speakers). The conversion is made by speakers of various genders, ages and intonations. SAN clips are recorded with specific properties in which the sound format is WAV PCM with bit depth 16 bit at sampling rate 16 kHz single-channel. The SAN dataset is created using MSA. SAN is organized in the way shown below:

Category name _ topic id _ speaker id _ gender_ age.wav

As example: Art_3_S1_F_22y.wav

SAN dataset details are shown in tables below:

Table (5.1): SAN details for keywords extraction step.

Category	Minutes	Words	Almost Speaking Time (hours)
Weather	640	23,965	11
Sport	1488	66,485	25
Science	1638	86,449	27
Politics	1431	72,050	24
Economy	990	54,312	17
Art	1470	67,188	25

Table (5.2): SAN details for TID step.

Category	Minutes	Words	Almost Speaking Time (hours)
Weather	240	10,626	4.0
Sport	344	15,154	5.7
Science	390	16,334	6.5
Politics	387	16,298	6.5
Economy	320	13,900	5.3
Art	403	17,193	6.7

As Table (5.1) and Table (5.2), the proposed system works on 128 hours of spoken Arabic news records (with ~370449 words in all classes) in the keywords extraction step and 35 hours of unclassified spoken Arabic news records (with ~89505 words in all classes) in the TID step. It is notable that speakers are chosen to be different of genders and ages.

5.3 Evaluation Metrics

To approve the accuracy of the used machine learning algorithm, F1-Measure and accuracy metrics are used. F1-Measure is usually used in the field of the information retrieval. The value of F1-Measure is governed by two factors which are precision and recall. Binary classification and information retrieval in the pattern matching system are used. Therefore, the positive predictive value (precision) which is the fraction of related examples among the real retrieved examples and the sensitivity (recall) which is the fraction of related examples that have been retrieved over entire related examples are used to evaluate the binary classification system. Hence, precision and recall are based on relevance measurement ("Precision and recall", 2017). The number of correct outcomes divided by the number of all retrieved outcomes defines a *precision* value. The number of correct outcomes divided by the number of outcomes that should have been retrieved defines a *recall* value.

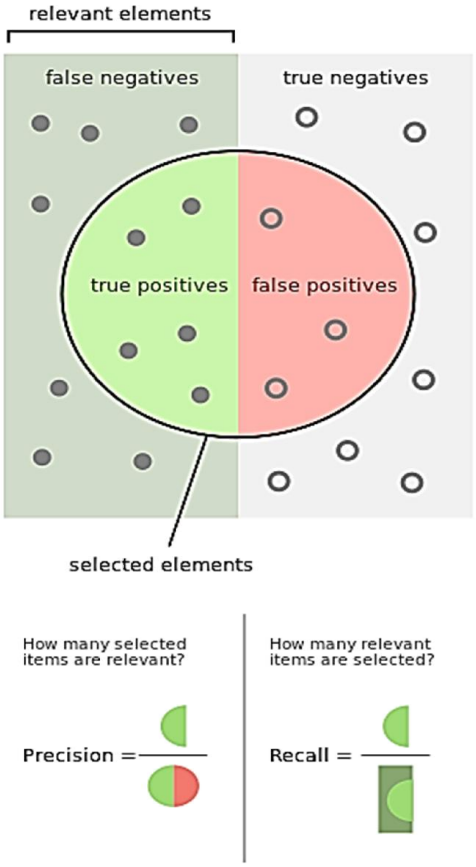


Figure (5.1): Precision and recall description.

Figure 5.1 describes the two factors which are precision and recall, where precision answers the question: how many selected files are relevant? and recall answers the question: how many relevant files are selected? In order to find the values of precision and recall, confusion matrix is constructed. Confusion matrix is a table often used to designate the performance of a classification model (or "classifier") on a set of test data for which the true values are identified. Figure 5.2 shows the confusion matrix, which includes the terms true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) and compares the results of the classifier (identifier) for classification tasks.

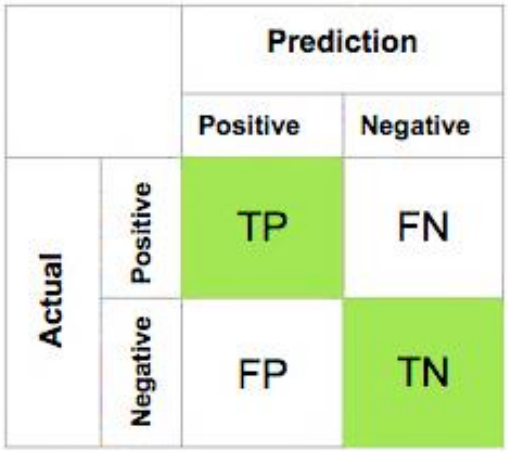


Figure (5.2): Confusion matrix.

Positive and negative expressions denote the classifier's expectation (occasionally identified as the prediction), and the true and false expressions denote if the expectation relates to the observation (occasionally identified as the external judgment). For more clarification:

- False positives (FP), which are files wrongly considered as belonging to the category.
- False negatives (FN), which are files not labeled as belonging to the positive category and they should be part of it.
- True positives (TP), which are files correctly labeled as belonging to the positive category.
- True negatives (TN), which are files considered as belonging to the positive category and should not be part of it.

The F1-Measure is selected to estimate the accuracy of the identification process which is derived from the two factors which are precision and recall. Additionally, the accuracy metric is used to clarify how good predictions are on the average. Nevertheless, f-measure is favoured over accuracy when it is applied to an unbalanced dataset. The equations below show how to calculate precision, recall, F1-Measure and accuracy ("Precision and recall", 2017):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.2}$$

The two measures are often used together in the F1 Score (or F1-Measure) in order to provide a single measurement for SAN system. This is obviously shown in equation 5.3

$$\text{F} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{5.3}$$

$$\text{F} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{5.4}$$

For accuracy metric the equation below is used ("Precision and recall", 2017):

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5.5}$$

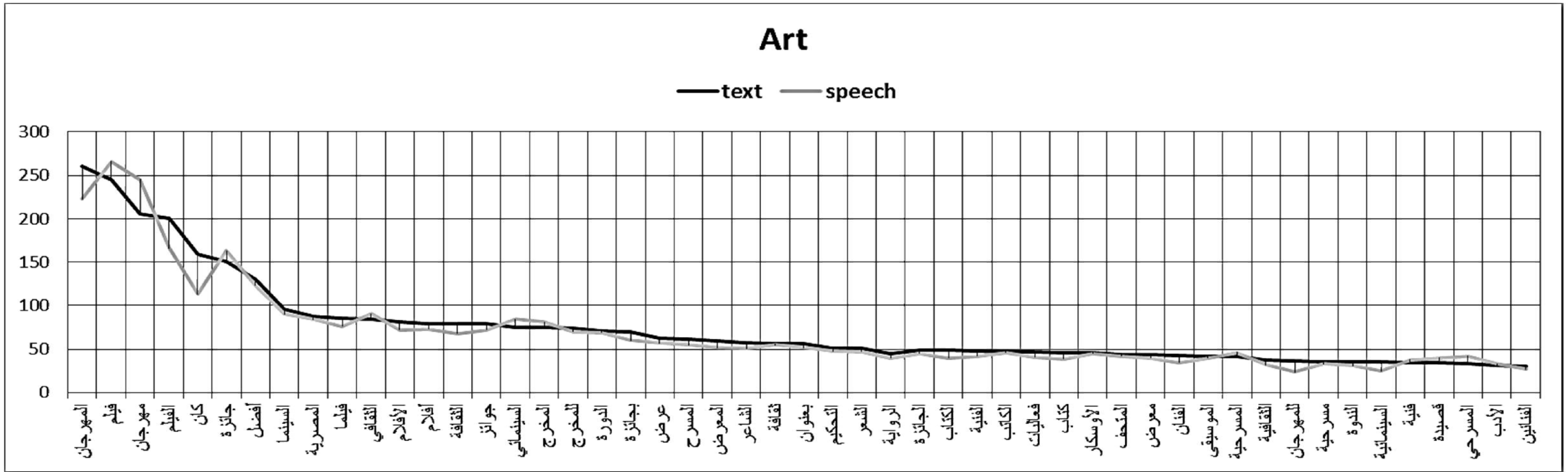
5.4 Evaluation and Results

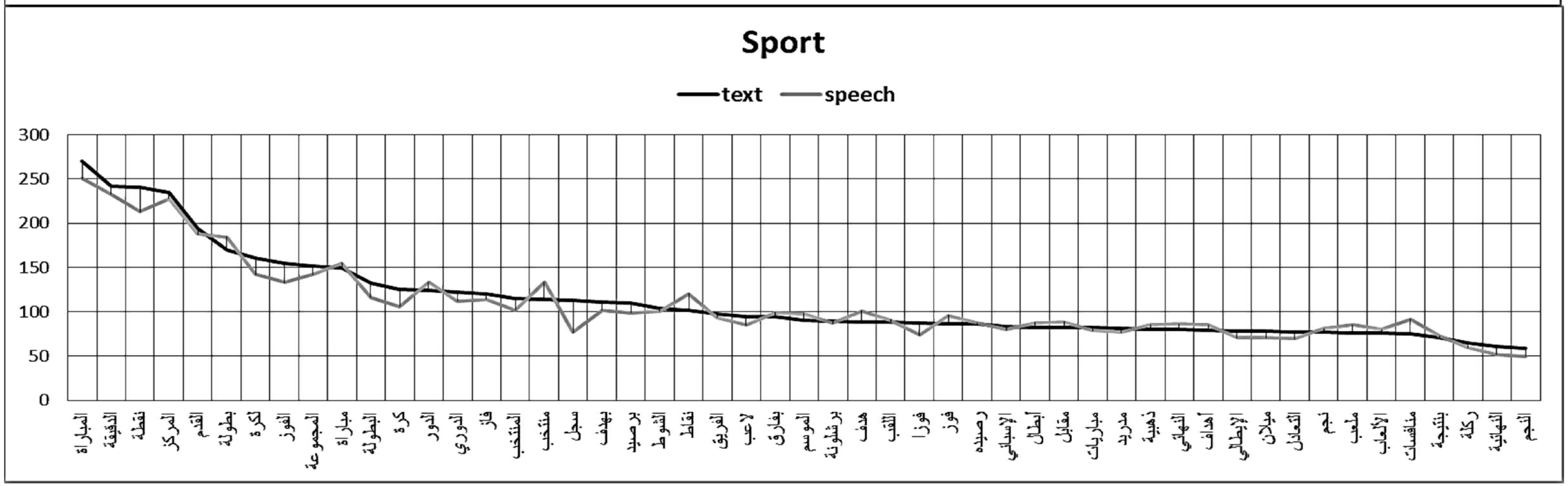
5.4.1 Keywords Extraction Evaluation

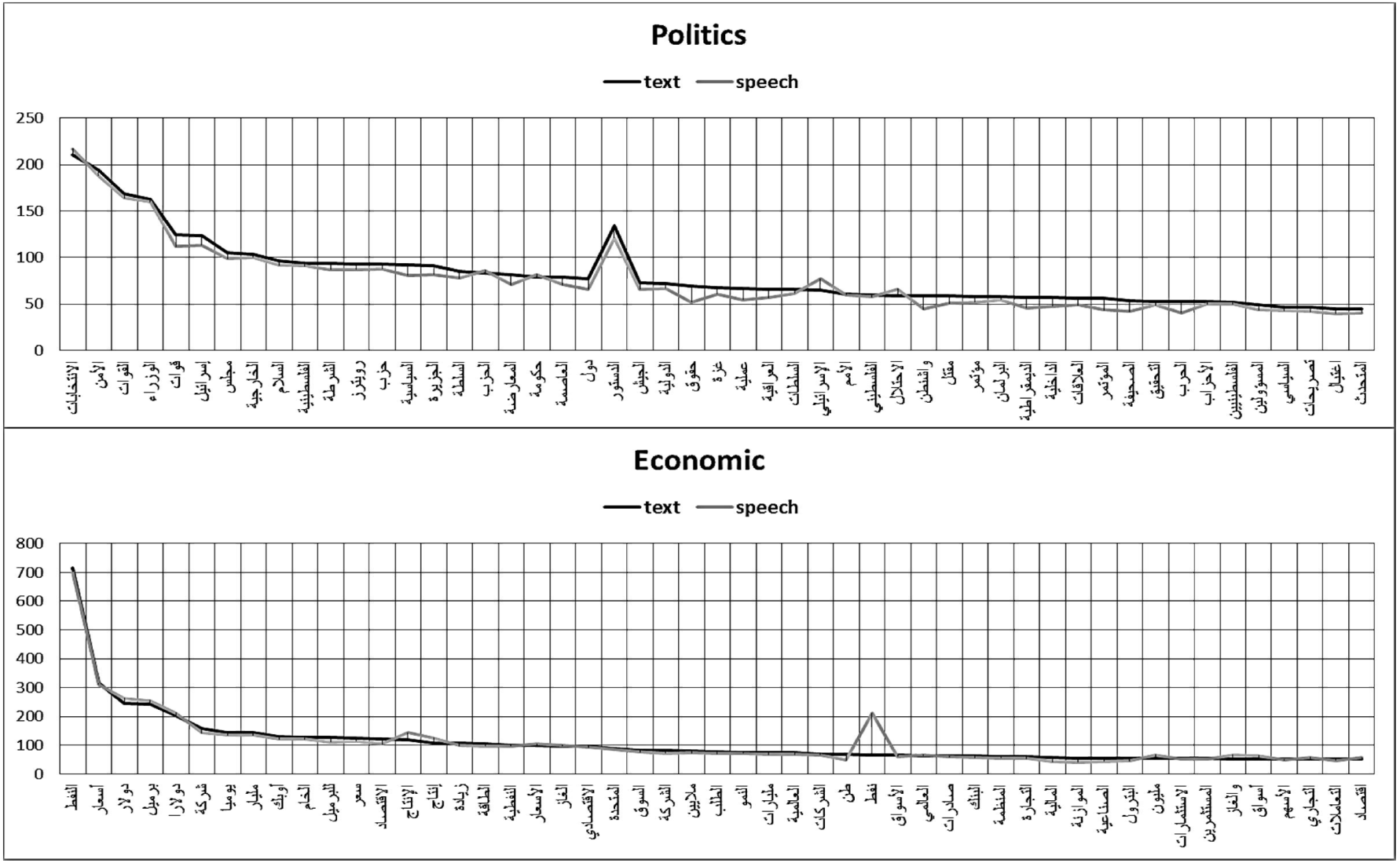
The frequency of spoken word is selected to be the critical line with the aim of scoring the proposed spoken keywords extraction method. In parallel, ("Online word counter", 2017) is used as a word counter tool for text processing methods in order to compare the word frequency values for written and spoken texts. The written texts word frequency results from an online word counter, which is based on written texts, are compared to spoken word frequency method results to evaluate the keywords extraction method.

The top 51 keywords (shown in appendix section) are selected as samples used in the system. The word frequency for each keyword resulting from the DTW matching algorithm applied to spoken files is compared to the keyword frequency resulting from online word counter system ("Online word counter", 2017) for written texts.

The keywords in each class and the word frequency value for each keyword are shown in the figure below; the figure compares the results by using DTW matching method and the online word counter system. The results show good convergence with some notable differences which are due to some challenges of using speech directly as shown and discussed below.


Art
text
speech
300
250
200
150
100
50
0
المهرجان
فيلم
مهرجان
الفيلم
كان
جائزة
أفضل
السينما
المصرية
فيلما
الثقافي
الأفلام
أفلام
الثقافة
جوائز
السينمائي
المخرج
للمخرج
الدورة
بجائزة
عرض
المسرح
المعرض
الشاعر
ثقافة
بعنوان
التحكيم
الشعر
الرواية
الجائزة
الكتاب
الفنية
الكاتب
فعاليات
كتاب
الأوسكار
المتحف
معرض
الفنان
الموسيقى
المسرحية
الثقافية
للمهرجان
مسرحية
الندوة
السينمائية
فنية
قصيدة
المسرحي
الأدب
الفنانين


Sport
text
speech
300
250
200
150
100
50
0
المباراة
الدقيقة
نقطة
المركز
القدم
بطولة
الكرة
الفوز
المجموعة
مباراة
البطولة
كرة
الدور
الدوري
فاز
المنتخب
منتخب
سجل
بهدف
برصيد
الشوط
نقاط
الفريق
لاعب
بفارق
الموسم
برشلونة
هدف
اللقب
فوزا
فوز
رصيده
الإسباني
أبطال
مقابل
مباريات
مدريد
ذهبية
النهائي
أهداف
الإيطالي
ميلان
التعادل
نجم
ملعب
الألعاب
منافسات
بنتيجة
ركلة
النهائية
النجم


Politics
text
speech
250
200
150
100
50
0
الانتخابات
الأمن
القوات
الوزراء
قوات
إسرائيل
مجلس
الخارجية
السلام
الفلسطينية
الشرطة
رويترز
حزب
السلطة
الجزيرة
الحزب
المعارضة
حكومة
العاصمة
دول
الدستور
الجيش
الدولية
حقوق
غزة
عملية
العراقية
السلطات
الإسرائيلي
الأمم
الفلسطيني
الاحتلال
واشنطن
مقتل
مؤتمر
البرلمان
الديمقراطية
الداخلية
العلاقات
المؤتمر
الصحيفة
التحقيق
الحرب
الأحزاب
الفلسطينيين
المسؤولين
السياسي
تصريحات
اغتيال
المتحدث
Economic
text
speech
800
700
600
500
400
300
200
100
0
النفط
أسعار
دولار
برميل
دولارا
شركة
يوميا
مليار
أونك
الخام
للبرميل
سعر
الاقتصاد
الإنتاج
إنتاج
زيادة
الطاقة
النفطية
الأسعار
الغاز
الاقتصادي
المتحدة
السوق
الشركة
ملايين
الطلب
النمو
مليارات
العالمية
الشركات
طن
نفط
الأسواق
العالمي
صادرات
البنك
المنظمة
التجارة
المالية
الموازنة
الصناعية
البترول
مليون
الاستثمارات
المستثمرين
والغاز
أسواق
الأسهم
التجاري
التعاملات
اقتصاد

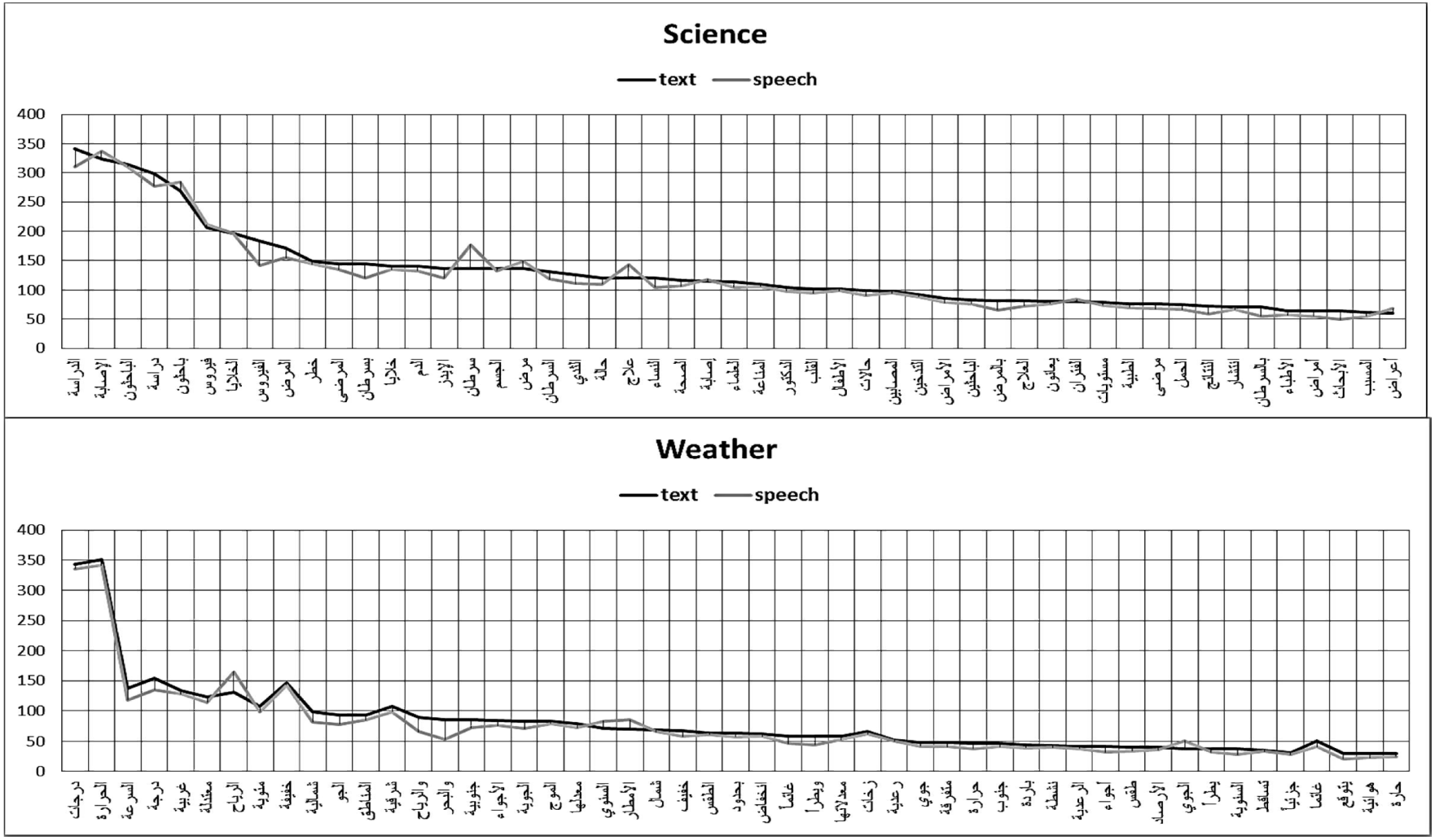


Figure (5.3): Comparing word frequency using text and speech methodology.

As shown in Figure 5.3, there are some differences between the word frequency values resulting based on written texts (online word counter) and the word frequency values resulting based on spoken texts (DTW counter) in some words. These differences are caused mainly by the natural of Arabic spoken language, such as the pronunciation of the numbers varies from one person to another person, the hamza, diacritics, vowelization and the conjunctions.

For example the word 'كان' 'kān' in the art class, with sukoon at the end of the word, refers to Cannes Film Festival. The word frequency value based on text counter is 159 words, but the value is reduced to 113 words based on DTW matching counter. This is due to traditional Arabic grammar which has what so-called 'كان وأخواتها' which is known as kāna and her sisters. It has the same word form 'كان' but with different meanings and pronunciations (kāna). Another instance is that the keyword 'كرة' in the sport class varies based on positions or speakers. It can be pronounced as 'كُرَه' 'korah', 'كُرَةْ' 'koratt' or with vowelization as 'كرةُ' 'korato', 'كرةِ' 'korate' or 'كرةَ' 'korata'. So, the word frequency value of 'كرة' using text based counter is 125 words, but it is decreased to 106 words using the speech based counter. also, there is an increase of the word frequency value for 'سرطان' 'saratan' word in the science class by speech based counter. These results are because of the segmentation process and DTW matching system which count the words 'السرطان' 'alsaratan', 'لسرطان' 'lsaratan' or 'بسرطان' 'bsaratan' as the keyword 'سرطان'. Another example of the differences is the keywords 'الجوي' 'aljawwee' and 'الجوية' 'aljawweea' in the weather class. Because of the big similarity of the pronunciations for both keywords, it can be observed that the word frequency value using speech based counter sometimes counts the word 'الجوية' as 'الجوي' and vice versa, and this clearly explains the varity of the results. At the end, the keywords selected based on both techniques, text and speech based, are similar because they have the top score resulting using both techniques.

To measure the accuracy of the keywords extraction system based on the differences between text and speech counter, the accuracy is calculated assuming that:

w_n is the word frequency value of the written keyword(n) resulting from text based counter.

s_n is the word frequency value of the spoken keyword(n) resulting from DTW matching counter.

The accuracy of each class keyword is calculated by averaging the accuracy values of all selected keywords in the class as shown in the next equation:

$$Accuracy(c) = \left(\sum_{n=0}^{k} \left(\frac{\min(w_n, s_n)}{\max(w_n, s_n)} \right) * 100 \right) / k \tag{5.6}$$

Where k is the number of keywords for each class (which is 51 keywords in this thesis.

This equation measures the accuracy based on results differences using both techniques and does not measure the accurate count of each word. The accuracy of each class is shown in the figure below:

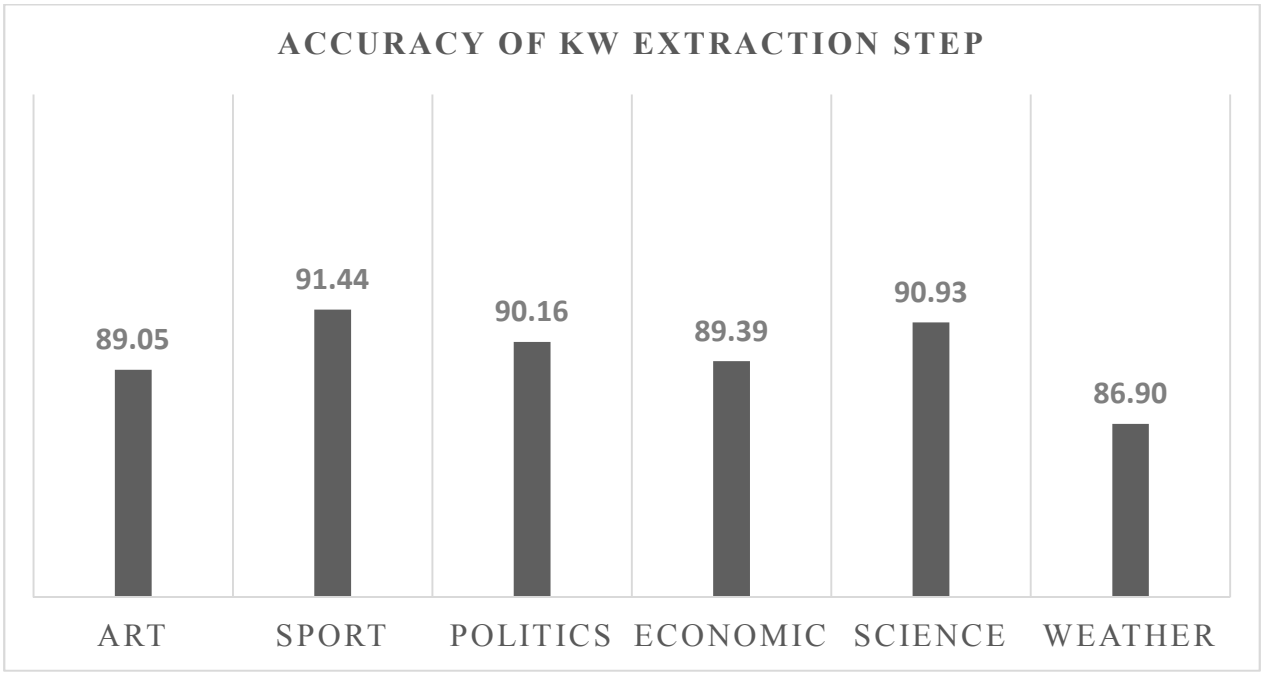


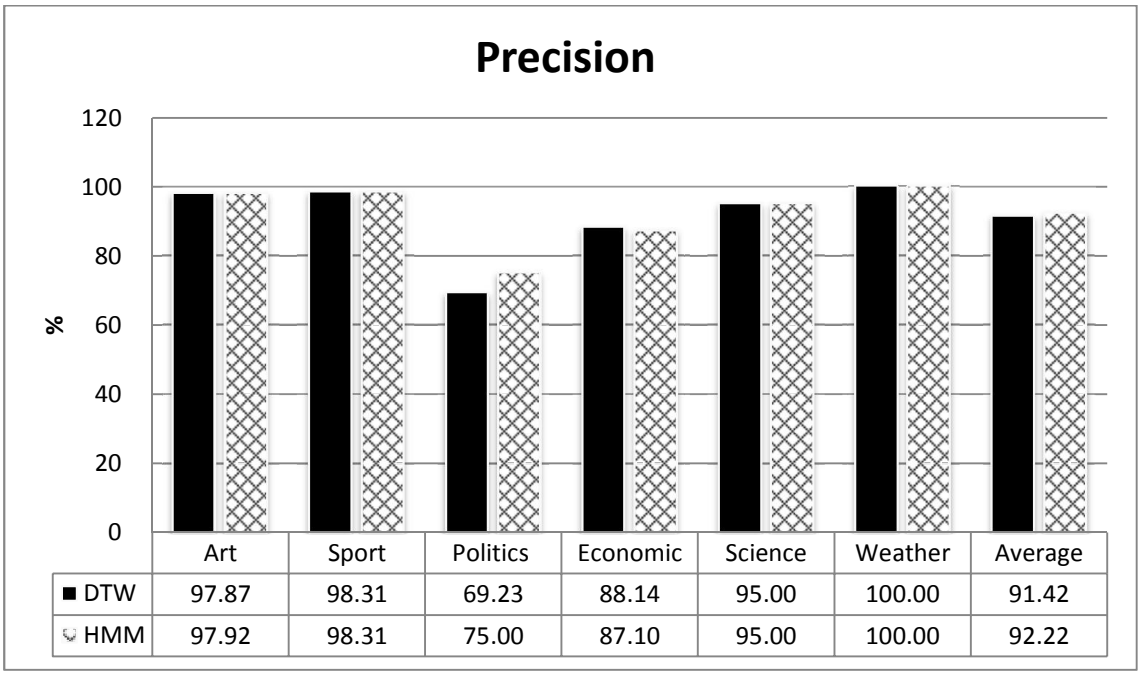
Figure (5.4): Accuracy of keywords extraction step for each class.

As it is shown, the differences in the sport class is the best. This is because that the nature of words related to sports are clear and convergent in spoken texts. Regards the reasons of differences between written and spoken texts results can be better if some steps are followed. Some of the steps are that the pronunciation and the vowelization of the spoken texts should be unified as can as possible, mistakes in pronunciation should be avoided and the pronunciation should be understandable.

5.4.2 TID Evaluation

As discussed in chapter 4, DTW and HMM are the TID techniques which are mainly used in SAN identification system. The pattern matching way, where DTW matching algorithm based on the distance and HMM algorithm based on logliklehood, is followed. The results show encouraging marks for both methods as shown in the figures below. Figure 5.5 shows the evaluation metrics results, the precision, recall, f1-measure and accuracy metrics, of using DTW and HMM methods as identifiers in SAN classification system. The results of applying speech based identification algorithm to Alj-News dataset reveal that it has recorded a good performance accuracy, noticing that results are motivating as the overall F1-measure is 90.26% and 91.36% for DTW and HMM in sequence. However, it is not as better as text based classifiers identification presented previously. For instance, using SVM classifiers on Alj-News Dataset (reduced features) (Chantar and Corne, 2011) shows 93.1% f-measure. Nevertheless, the outcomes of using speech based technique can be enhanced as what is discuss in chapter 6.

Results of using the DTW and HMM as classifiers which are operated on Alj-News dataset expose that the high performance accuracy and speed are scored using HMM classifier as shown in the figures below. The threshold is used to evaluate the resulting logliklehood value using HMM.



	Art	Sport	Politics	Economic	Science	Weather	Average
DTW	97.87	98.31	69.23	88.14	95.00	100.00	91.42
HMM	97.92	98.31	75.00	87.10	95.00	100.00	92.22

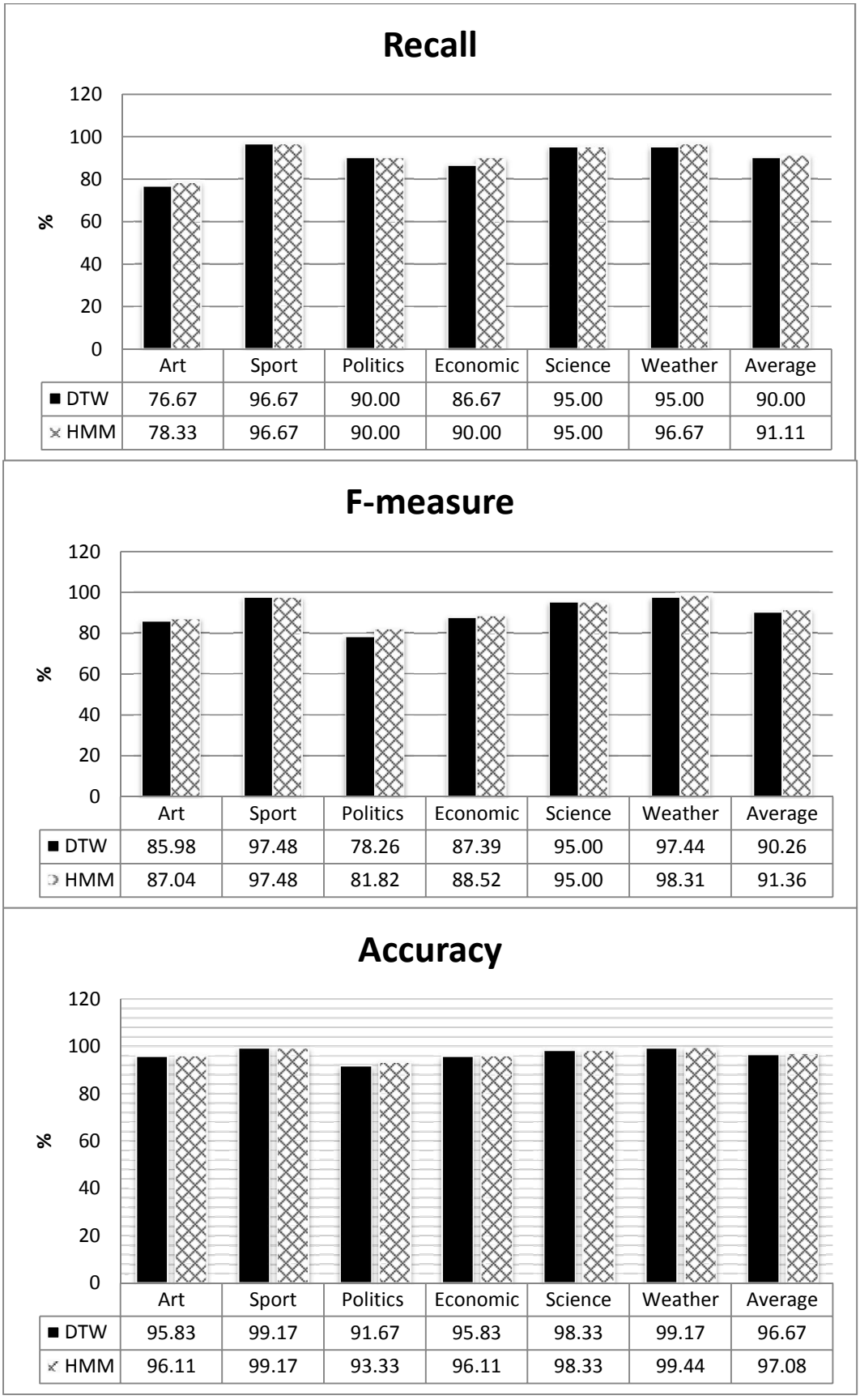


Figure (5.5): Precision, recall, f-measure and accuracy using DTW and HMM method in SAN classification system**.**

The top performance is scored on 'Weather' and ‘Sport’ classes with (100% and 98.31%) precision , (96.67% and 96.67%) recall, (98.31% and 97.48%) F1-measure and (99.44% and 99.17%) accuracy respectively. However, the lowest F-measure performance recorded is 81.82 % on ‘Politics’ class. This is due to the lack of confusion of news included in the weather or the sport classes in contradiction with the news included in the classes as politics, art, economy or science which might be ambiguous or the news in these classes may be intersected. Another reason for the incorrect classification of SAN clip is the shortness of the clip because it reduces the chance of keywords appearance. An example of intersection is shown in the next news identified via TID system as politics where it is in fact in the science class.

جنون البقر يظهر مجددا في اليابان جنون البقر في اليابان من جديد أكدت وزارة الصحة اليابانية أن الحالة السادسة عشرة من مرض جنون البقر ظهرت في البلاد وتم ذبحها على الفور. وقالت وزارة الزراعة في بيان لها إنه ليس هناك خطر للحوم التي تم توزيعها على الأسواق، مؤكدة أن الفحوصات أثبتت إصابة بقرة في جزيرة هوكايدو بأقصى شمال البلاد وأن السلطات فحصت لكل الأبقار في المنطقة. وكانت اليابان التي تعتبر أكبر سوق لصادرات اللحم البقري حظرت استيراد اللحوم الأميركية أواخر العام ٢٠٠٣ بعد ظهور أول حالة جنون بقر بولاية واشنطن، وقدر حجم تكاليف شرائها للحوم الأميركية بحوالي ١,٣ مليار دولار في نفس العام. ووافقت اليابان في أكتوبر/تشرين الأول الماضي على السماح باستئناف استيراد اللحم الأميركي بمجرد الاتفاق على التفاصيل الفنية، الأمر الذي فسره مشرعون أميركيون بأنه تلكؤ من جانب طوكيو داعين إلى اتخاذ إجراء اقتصادي ضدها. وكانت واشنطن قد اقترحت على مجلس سلامة الأغذية الياباني إلغاء فحوصات الماشية التي يقل عمرها عن ٢٠ شهرا لوجود أدلة علمية تشير إلى أن البروتين المتعلق بجنون البقر لا يظهر في الأبقار الصغيرة، وهو الاقتراح الذي دعمته لجنة حكومية يابانية لكن السلطات تركت القرار النهائي لمجلس سلامة الأغذية. يذكر أن وزيرة الخارجية الأميركية كوندوليزا رايس مارست ضغوطا على اليابان لاستئناف الواردات حيث أكدت بطوكيو أن "الوقت حان لحل هذه المشكلة"، في حين رد وزير الخارجية الياباني نوبوتاكا ماتشيمور بأن بلاده لا تستطيع تقديم موعد محدد لواشنطن لإنهاء الحظر

Nine keywords are identified as politics which are (مجلس (2), الخارجية (2), واشنطن (3)), three keywords which are (إصابة, الصحة, حالة) are recognized as science and three keywords which are (دولار, مليار, الأسواق) are classified as economy. Accordingly, improvements are needed for SAN TID system in order to gain the expected improvements and to achieve the reduction of classification error rate. These improvements can be accomplished if the recommendation steps in chapter 6 are followed.

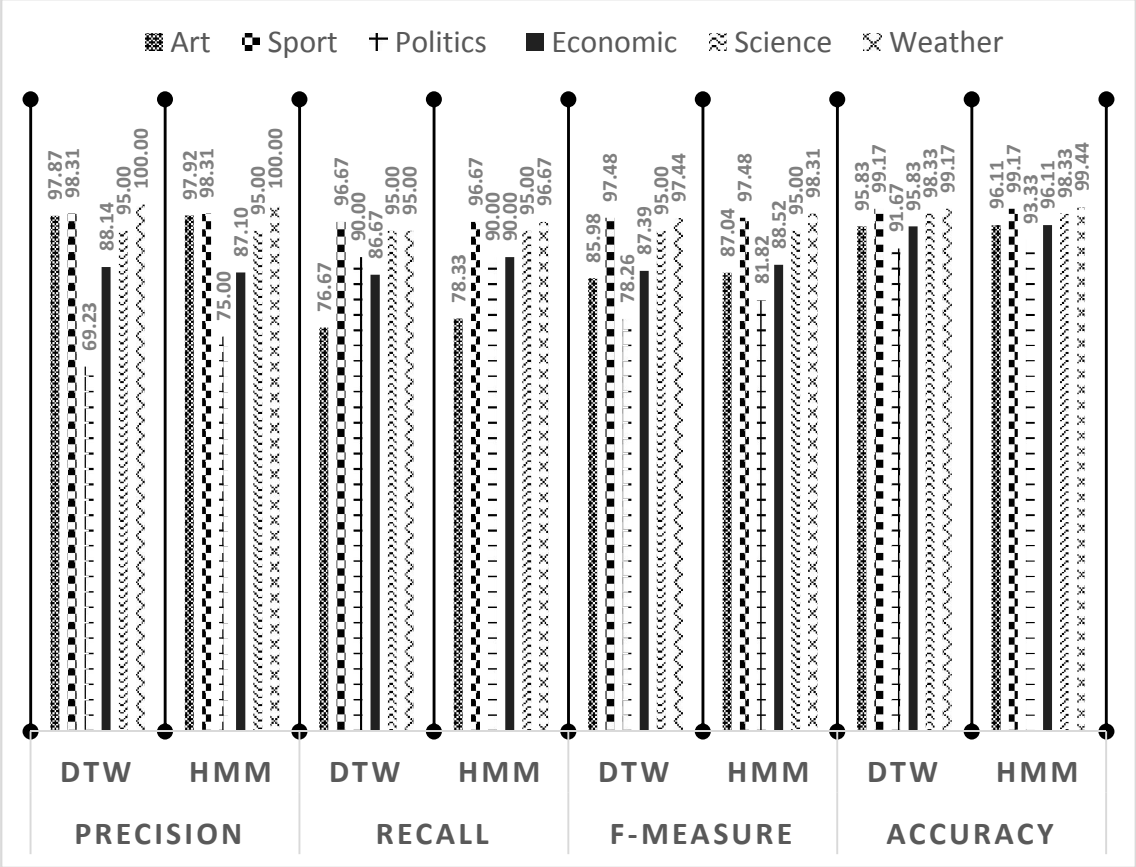


Figure (5.6): Topic identification evaluation summary.

It's is clear from the Figure 5.6, HMM classifier is a competitive algorithm to the best performers in some classes. As seen in politics, using HMM is better than using DTW as F-measure values show. Also, DTW should be used in some suggested enhancement steps which are described in Chapter 6, such as using some different spoken words related to one word.

As a result, the classification F-measure scores which are achieved using speech based techniques are:

a. 90.26% of classification accuracy is achieved by using DTW matching technique which uses a distance as a matching measurement between patterns.
b. 91.36% of classification accuracy is achieved by using HMM matching technique which uses a logliklehood as a matching measurement between patterns.
c. 89.65% of keywords identification accuracy is achieved by using SAN keywords extraction method relatively to the online word counter method as shown in Figure 5.4.

Chapter 6
Conclusions and Recommendations

Chapter 6

Conclusions and Recommendations

This thesis proposes a new methodology in dealing with classification of spoken Arabic texts directly using spoken word features rather than traditional methods which rely on the conversion of the spoken texts to written texts (transcriptions) using ASR model. The proposed approach works well in this area after extracting the experimental results and it can be used in the field of Arabic speech search engines and the classification of the huge amount of spoken Arabic texts into various classes or labels.

First, the keywords extraction system is proposed in order to extract the keywords for each class depending on DTW matching technique where the pattern is the extracted features from the spoken word resulting from applying the segmentation step on the SAN clip. DTW method is used to match (lowest distance under the specific selected threshold) the extracted keyword features with other words features in the SAN clip. The selected keywords, which are the representative of the class, are selected after applying two main steps: eliminating the discarded words and applying the scoring process, which is based on the word frequency in the whole spoken texts.

After the keywords selection is done, an automatic TID system is suggested. The first step in this system is to segment connected spoken texts and extract the features of each separated word. Then, one of the two techniques (DTW or HMM) is used 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 remarkable that choosing the best matching or assigning depends on a specific selected threshold value. Finally, the keywords frequency sorting method is used with the purpose of identifying the class of the SAN clip.

To evaluate TID approach, Arabic news dataset, which are written texts, is converted into spoken files via various speakers reading text clearly in order to create a usable spoken dataset which can be useful in TID area. Aljazeera News dataset, which is widely used in text classification implementations, is used in order to

complete this task. F1-Measure is used as an evaluation metric to estimate the accuracy of the classification process which comes from two factors: precision and recall factors. In addition, the accuracy metric is used in order to obtain how good predictions are on average regardless of the fact that F1-measure is favoured over accuracy when a dataset is unbalanced. The values of F1-Measure for TID system with the DTW classifier is with an average of 90.26% in all classes and 91.36% using the HMM classifier. Moreover, The values of Accuracy metrics are with an average of 96.67% using the DTW classifier and 97.08% using the HMM classifier. The accuracy achieved using SAN keywords extraction method is computed relatively to the online word counter method, which is used with written texts, and it achieves a 89.65% of keywords identification accuracy.

The results are initially good, but there are some problems because dealing with spoken texts is more sophisticated and costly than dealing with written texts. Thus, enhancing the results requires some changes and improvements to be done. Comparing these results with other studies, which have not been applied to spoken Arabic language, is meaningless. Based on the studies which follow the traditional method and other studies working on the speech features directly, we can come to conclusion that it is preferable to benefit from the powerful systems of ASR and reliable text based classification methods. The complexity of dealing with speech features in the classification step is not taken into consideration.

Logically, methodology in this thesis is supposed to be more acceptable than the traditional method because it directly processes the speech without using ASR model and the text processing methodologies. Enhancing results and performance can be accomplished by creating a powerful toolbox via a professional programming team based on this thesis methodology. Speech features, in addition, can be supportive to traditional methods in order to generate more accurate and useful results. Also, working with speech features can be used in speakers and speech search engines which retrieve a clip based on topics, speakers or any other options.

Working on Arabic language TID, which is only based on speech features, is a new technique, so it requires progressive improvements in the future. Some improvements might be borrowed from algorithms applied to text based algorithm. While preparing this thesis, some issues are observed and can be used in order to enhance the overall results and performance of TID system based on speech.

The system has five main processes: pre-processing, Speech segmentation, feature extraction, keywords extraction and TID processes. Any enhancement for one of them can affect the overall results. Some recommendations are mentioned in the pursuit of improving the overall results. In the pre-processing step, more working on speech and speaker normalization may be needed. There is extensive variation in speech. So, listeners agree in their perception of vowels. Some influencing parameters and instruments for vowel normalization are: context, formant ratio, F0, visual information and auditory gestalts. "Vocal tract normalization theories consider that listeners perceptually evaluate vowels on a talker specific coordinate system." (Johnson 2004).

In the segmentation step, feature vectors are extracted depending on the algorithms, such as MFCC, PLP, RAST-PLP and LPC. More than one feature vector can be mixed in order to identify the word boundaries more accurately. The concept of probability may be used. For example, the word 'استقرحصاء' can be 'استقر' and 'حصاء' or 'استقرحصاء'. Therefore, extracting all possible words and providing weights for each one can be useful in this case.

In the keywords extraction step, some features, which are used in text methodologies as rooting and stemming, can be used in order to enhance keywords findings and selections. Some steps of Khoja Arabic root extractor method (Khoja & Garside, 1999) seems useful as it suggests:

1) Format the word by removing any punctuation, diacritics and non-letter characters (it might not be applicable with directly speech methodology).
2) Stop words are discarded.
3) The definite article should be excluded, such as. ال وال بال كال فال : .
4) The special prefix (و) is eliminated.
5) If the last letter is a shadda, duplicate the letter (it might not be applicable with directly speech methodology).
6) Replace أ إ آ with ا (it might not be applicable with directly speech methodology).
7) Prefixes should be excluded للـ لـ سـ ف .
8) Eliminate Suffixes, such as كن هما كما :.
9) The outcome should be matched against a list of defined Patterns, such as فاعل افعل : تفعيل فعال.

10) all occurrences of Hamza should be replaced, such as ء ؤ ئ : with ا (it might not be applicable with directly speech methodology).
11) Two letter roots are checked to see if they should contain a double character; if so, the character is added to the root (it might not be applicable with directly speech methodology).

In keywords scoring methodology, the position and the coverage of the word in both clip, which includes the word, and all related class clips might be taken into account. In TID process, some suggested steps can make notable improvements, but they may include some risks. As an alternative of rooting and stemming processes, a list of patterns for each keyword can be created. This may work well with DTW matching method as it depends on measuring distance between patterns. Nevertheless, it might be more difficult to be applicable to HMM method which depends on logliklehood (for created models) regardless of the fact that it requires more additional processing time and resources. Using a large number of keywords in addition to using a weighting technique can be useful. Mutually exclusive technique, discussed in chapter 4 section 4.3, can be replaced with a weighting technique. The suggested weighting technique depends on providing a weight for each keyword depending on the word coverage, which is the count of the word appearance in the clip, in the related class and in the other classes.

Finally, this system is suggested to eliminate the step of converting the spoken texts to written texts and to 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 used in this system as a search engine for spoken clips.

The References List

Aguilo, M., Butko, T., Temko, A. & Nadeu, C. (2009) A hierarchical architecture for audio segmentation in a broadcast news Task, pp. 17–20

Al-Harbii, S., Almuharreb, A., Al-Thubaiity, A., Khorsheed, M. & Al-Rajeh, A. (2008) Automatic arabic text classification. In: JADT; 08, France, pp. 77–83.

Arabic phonology. (2017, June 12) Retrieved From Wikipedia, https://en.wikipedia.org/wiki/Arabic_phonology

Belfield, W. & Gish, H. (2003) A topic classification system based on parametric trajectory mixture models," in Proc. Interspeech, Geneva.

Bhandari G.M., Kawitkar R.S. & Borawake M.P. (2014) Audio segmentation for speech recognition using segment features. vol 249. Springer, Cham.

Bishop, C.M. (2007) Pattern recognition and machine learning. information science and statistics, ISSN 1613-9011, Springer.

Caranica, A., Cucu, H. & Buzo A. (2016) Exploring an unsupervised, language independent, spoken document retrieval system, Content-Based Multimedia Indexing (CBMI), **doi:** 10.1109/CBMI.2016.7500262

Carbonell, J., Yang, Y., Lafferty, J., Brown, R., Pierce, T. & Liu, X. (1999) CMU report on TDT-2: Segmentation, detection and tracking, Proceedings of DARPA Broadcast News Workshop 1999. Herndon, VA, USA

Cerisara, C. (2009) Automatic discovery of topics and acoustic morphemes from speech. Computer Speech and Language 23(2), pp. 220–239. doi: 10.1016/j.csl.2008.06.004

Cettolo, M., Vescovi, M. & Rizzi, R. (2005) Evaluation of BIC-based algorithms for audio segmentation, Computer Speech and Language, vol. 19, no. 2, pp. 147–170.

Chantar, H.K. & Corne, D.W. (2011) Feature subset selection for Arabic document categorization using BPSO-KNN, IEEE, pp. 546–551,10.1109/NaBIC.2011.6089647

Chelba, C., Silva, J., & Acero, A. (2007) Soft indexing of speech content for search in spoken documents. Comput. Speech Lang., 21:458–478.

Chelba, C., Hazen, T.J., & Saraclar, M. (2008) Retrieval and browsing of spoken content. IEEE Signal Processing Magazine 25(3), pp. 39–49, DOI: 10.1109/MSP.2008.917992

Cheng, O., Abdulla, W. & Salcic, Z. (2005) Performance evaluation of frontend algorithms for robust speech recognition, Proc. ISSPA, pp. 711-714.

Cieri, C., Miller, D. & Walker, K. (2003) From Switchboard to Fisher: telephone collection protocols, their uses and yields, Proceedings of Interspeech. Geneva, Switzerland.

Complete list of Arabic speaking countries 2017 (2017), Retrieved from: http://istizada.com/complete-list-of-arabic-speaking-countries-2014/.

Dredze, M., Jansen, A. , Coppersmith, G. & Church, K. (2010) "NLP on spoken documents without ASR," in Proc. of EMNLP.

Dynamic time warping, (2017) Retrieved From Wikipedia, https://en.wikipedia.org/wiki/Dynamic_time_warping

Eriksson, L. (1989) Algorithms for automatic segmentation of speech, Lund University, Dept. of Linguistics, Working Papers (35), pp. 53-61

Fiscus, J. (2004) Results of the 2003 Topic detection and tracking evaluation, Proceedings of the 4th International Conference on Language Resouces and Evaluation (LREC). Lisbon, Portugal

Fiscus, J., Doddington, G., Garofolo, J. & Martin, A. (1999) Topic detection and tracking evaluation (TDT2), Proceedings of DARPA Broadcast News Workshop 1999 . Herndon, VA, USA

Flamary, F., Anguera, X. & Oliver, N. (2011) Spoken WordCloud: Clustering Recurrent Patterns in Speech, International Workshop on Content-Based Multimedia Indexing, pp. 133–138.

Forward-Backward, (2017) Retrieved From http://curtis.ml.cmu.edu/w/courses/index.php/Forward-Backward

Garofolo, J., Fiscus, J. & Ajot, J. (2008) The rich transcription 2007 meeting recognition evaluation. Multimodal Technologies for Perception of Humans, Lecture Notes in Computer Science, Springer Berlin / Heidelberg. pp. 373–389.

Giannakopoulos, Th. & Pikrakis, A., (2014) Introduction to audio analysis: a MATLAB approach, Academic Press, © 2014 Elsevier Ltd.

Gish, H., Siu, MH., Chan, A. & Belfield, W. (2009) Unsupervised training of an HMM-based speech recognizer for topic classification, Proceedings of Interspeech . Brighton, UK

Gold, K. & Scassellati, B. (2006) Audio speech segmentation without language-specific knowledge. In: Cognitive Science, pp. 1370–1375 (2006)

Hagen A., Connors D.A. & Pellm B.L. (2003) The analysis and design of architecture systems for speech recognition on modern handheld-computing devices. Proceedings of the 1st IEEE/ACM/IFIP international conference on hardware/software design and system synthesis, pp. 65-70

Harwath, D., Hazen, T. & Glass, J. (2013) Zero resource spoken audio corpus analysis, IEEE ICASSP, pp. 8555–8559.

Hazen, T. J., (2011) Topic identification, Chapter 12 in Tur, G., De Mori, R. (Editors), Spoken Language Understanding: Systems for Extracting Semantic Information from Speech, John Wiley & Sons, West Sussex, United Kingdom

Hazen, T. J., (Nov, 2011) 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.

Hazen, T., Richardson, F., & Margolis, A. (2007) Topic identification from audio recordings using word and phone recognition lattices, Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding . Kyoto, Japan.

Hazen, T., Siu, M., Gish, H., Lowe, S., & Chan, A. (2011) Topic modeling for spoken documents using only phonetic information, Published in: Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop, doi: 10.1109/ASRU.2011.6163964,

Hidden Markov models – forward & Viterbi algorithm, (2014) Retrieved From http://gekkoquant.com/2014/05/26/hidden-markov-models-forward-viterbi-algorithm-part-2-of-4/

Hidden Markov model, (2017) Retrieved From Wikipedia, https://en.wikipedia.org/wiki/Hidden_Markov_model

Hermansky, H. (1990) Perceptual linear predictive (PLP) analysis of speech, in J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738-1752.

Hermansky, H. & Morgan, N. (1994), "RASTA processing of speech", IEEE Trans. on Speech and Audio Proc., vol. 2, no. 4, pp. 578-589.

Hulth, A. (2003) Improved automatic keywords extraction given more linguistic knowledge. In Proceedings of the 2003 Conference on Emprical Methods in Natural Language Processing, Sapporo, Japan

Ishizuka K. & Nakatani T. (2006) A feature extraction method using subband based periodicity and aperiodicity decomposition with noise robust frontend processing for automatic speech recognition. Speech Communication, Vol. 48, Issue 11, pp. 1447-1457

Johnson, K. (2004) Speaker normalization in speech perception. Ohio State University

Katakis, I., Tsoumakas, G. & Vlahavas, I. (2005) On the utility of incremental feature selection for the classification of textual data streams, Proceedings of the Panhellenic Conference on Informatics . Volas, Greece.

Kesiraju, S., Pappagari, R., Ondel, L. & Burget, L. (2017) Topic identification of spoken documents using unsupervised acoustic unit discovery, in Proc. ICASSP.

Khoja, Sh. & Garside, R. (1999) Stemming Arabic text. computer science department, Lancaster University , Lancaster, UK, http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps,

Farghaly, A. & Shaalan, K., (2009) Arabic natural language processing: challenges and solutions, ACM Trans. Asian Lang. Inf. Process., 8(2009), pp. 1-22, https://doi.org/10.1145/1644879.1644881

Kishori, R., Ghule, R. & Deshmukh, R. (2015) Feature extraction techniques for speech recognition: A review, International Journal of Scientific & Engineering Research, Volume 6, Issue 5.

Knill, K. & Young, S. (1997) Hidden Markov models in speech and language processing., Corpus-Based Methods in language and speech processing. Kluwer Academic Publishers, pp. 27-68.

Kuhn, R., Nowell, P. & Drouin, C. (1997) Approaches to phoneme-based topic spotting: An experimental comparison, in Proc. ICASSP, Munich,

Larsen, E., Cristia, A. & Dupoux, E. (2017). Relating unsupervised word segmentation to reported vocabulary acquisition. Retrieved from osf.io/86tu3

Liu, C., Trmal, J., Wiesner, M., Harman, C. & Khudanpur, S. (Mar., 2017) Topic identification for speech without ASR, Computation and Language, Publication: eprint arXiv:1703.07476

Liu, C., Yang, J. , Sun, M., Kesiraju, S., Rott, A., Ondel, L., Ghahremani, P., Dehak, N., Burget, L. & Khudanpur, S. (Feb., 2017) An empirical evaluation of zero resource acoustic unit discovery, in Proc. ICASSP.

Lo., Y-Y. & Gauvain, JL. (2003) Tracking topics in broadcast news data, Proceedings of the ISCA Workshop on Multilingual Spoken Document Retrieval . Hong Kong.

Mamou, J., Carmel, D. & Hoory, R. (2006) Spoken document retrieval from call-center conversations. ACM SIGIR conference on Research and development in information retrieval, SIGIR'06, pp. 51–58.

Manning, C. & Schütze, H. (1999) Foundations of statistical natural language processing, MIT Press Cambridge, MA, USA chapter Text Categorization, pp. 575–608

Matlab audio processing examples, last update (2012) http://www.ee.columbia.edu/~dpwe/resources/matlab/

Matsuo, Y. & Ishizuka, M. (2004) Keywords extraction from a single document using word co-ocuurrence statistical information. International Journal on Artificial Intelligence

May, Ch., Ferraro, F., McCree, A., Wintrode, J., Garcia-Romero, D. & Van-Durme, B. (2015) Topic identification and discovery on text and speech, Human Language Technology Center of Excellence, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2377–2387

Meng, X., Lee, K.K. & Xu, Y. (2006) Human driving behavior recognition based on Hidden Markov models. In Proceedings of the 2006 IEEE International Conference on Robotics and Biomimetics, Kunming, China, 17–20 December 2006; pp. 274–279.

McDonough, J., Ng, K., Jeanrenaud, P., Gish, H. & Rohlicek, J.R. (1994) 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, April 1994., doi: 10.1109/ICASSP.1994.389275.

Mohamed, S., Ata,W. & Darwish, N. (2005). A new technique for automatic text categorization for Arabic documents. In Proc. of the 5th IBIMA International Conference on Internet and Information Technology in Modern Organizations Cairo, Egypt.

Nöth, E., Harbeck, S., Niemann, H. & Warnke, V. (1997) A frame and segment based approach for topic spotting, in Proc. Eurospeech, Rhodes,

Oelze, I. (2009) Automatic keywords extraction for database search, Ph.D. Thesis, University of Hannover, Hannover.

Online word counter (2017) site: http://countwordsfree.com/

Pallett, D., Fiscus, J., Garofolo, J., Martin, A. & Przybocki, M. (1999) 1998 broadcast news benchmark test results: English and non-English word error rate performance measures, Proceedings of DARPA Broadcast News Workshop 1999 . Herndon, VA, USA.

Precision and recall (2017) Retrieved From https://en.wikipedia.org/wiki/Precision_and_recall

Plas, L., Pallotta, V., Rajman, M. & Ghorbel, H. (2004) Automatic keywords extraction from spoken text. A comparison of two lexical resources: the EDR and WordNet. Proceedings of the 4th International Language Resources and Evaluation, European Language Resource Association.

Prithvi, P. & Kishore-Kumar, T. (2016) Comparative analysis of MFCC, LFCC, RASTA –PLP , IJSER ,Volume 4, Issue 5.

Qaroush, A., Hanani, A., Jaber, B., Karmi, M. & Qamhiyeh, B. (2016) Automatic spoken customer query identification for Arabic language. ICIME 2016, doi:10.1145/3012258.3012261, pp. 41-46

Rabiner, L. (1989) A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE 77.2, pp. 257-286.

Räsänen, O. (2007) Speech segmentation and clustering methods for a new speech recognition architecture, M.Sc Thesis, department of Electrical and Communications Engineering, Helsinki University of Technology, Espoo

Reverse engineering DynamicHedge's Alpha curves (2013) Retrieved From http://qusma.com/2013/12/30/reverse-engineering-dynamichedges-alpha-curves-part-1-3-dynamic-time-warping/

Rosenberg, A. (2016) 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

Rose, R., Chang, E. & Lippman, R. (1991) Techniques for information retrieval from voice messages, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing . Toronto, Ont., Canada

Rybach, D., Gollan, C., Schlüter, R. & Ney, H. Audio segmentation for speech recognition using segment features", in Proc. DARPA.

SaiJayram, A.K.V, Ramasubramanian, V. & Sreenivas, T.V. (2002) Robust parameters for automatic segmentation of speech. Proceedings IEEE International Conference on Acoustics (ICASSP '02), Vol. 1, pp. 513-516.

Saha, G., Chakroborty, S. & Senapati, S. (2005) A new silence removal and end point detection algorithm for speech and speaker recognition applications, in Proc. of Eleventh National Conference on Communications (NCC), IIT- Kharagpur, India, pp. 291-295.

Sakoe, H. & Chiba, S. (1978) Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing. 26 (1): 43–49. doi:10.1109/tassp.

Sandsmark, H. (2012) Spoken document classification of broadcast news, master thesis, department of electronics and telecommunications, Norwegian University of Science and Technology, Norway.

Saxena, S. (2015) Hidden Markov model [ppt], Retrieved from: https://www.slideshare.net/shivangisaxena566/hidden-markov-model-ppt

Sebastiani, F. (2002) Machine learning in automated text categorization. ACM Computing Surveys, 34(1), pp. 1–47, doi:10.1145/505282.505283.

Sharma, M. & Mammone, R. (1996) Blind speech segmentation: automatic segmentation of speech without linguistic knowledge. ICSLP 96. Proceedings. Vol. 2, pp. 1237-1240.

Siegler, M. A., Jain, U., Raj, B., & Stern R. M. (1997) Automatic segmentation, classification and clustering of broadcast news audio," in Proc. DARPA Speech Recognition Workshop, , pp. 97–99.

Siu, M., Gish, H., Chan, A. & Belfield W. (2010) Improved topic classification and keywords discovery using an HMM-based speech recognizer trained without supervision, in Proc. Interspeech, Makuhari.

Siu, M., Gish, H., Chan, A., Belfield W. & Lowe, S. (2014) Unsupervised training of an hmm-based self-organizing unit recognizer with applications to topic classification and keywords discovery, Computer Speech & Language, vol. 28, no. 1, pp. 210–223.

Sivaram, G.S.V.S. & Hermansky, H. (2011) Multilayer perceptron with sparse hidden outputs for phoneme recognition. IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), Prague, pp. 5336–5339

Suzuki, Y., Fukumoto, F. & Sekiguchi, Y. (1998) Keywords extraction of radio news using term weighting with an encyclopedia and newspaper articles. SIGIR.

Theunissen, M.W., Scheffler, K., & du-Preez, J. A. (2001) Phonemebased topic spotting on the Switchboard Corpus," in Proc. Eurospeech, Aalborg.

Tsiporkova E., (2017) "Dynamic time warping algorithm for gene expression time series" Retrieved from: https://www.psb.ugent.be/cbd/papers/gentxwarper/DTWalgorithm.htm

Uchat, N.S. (2012) Hidden Markov model and speech recognition, seminar report, department of computer science and engineering IIT, Mumbai.

Vimala, C. & Radha, V. (2012) A review on speech recognition challenges and approaches', World Computer. Sci. Inf. Technol, 2, (1), pp. 1–7

Wayne, C. (2000) Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation, Proceedings of the 2nd International Conference on Language Resouces and Evaluation (LREC). Athens, Greece.

Wintrode, J. & Kulp, S. (2009) Techniques for rapid and robust topic identification of conversational telephone speech, Proceedings of Interspeech . Brighton, UK

Wintrode, J. & Khudanpur, S. (2014) Limited resource term detection for effective topic identification of speech, IEEE ICASSP, pp. 7118–7122.

Women's health (2013), Retrieved from: http://www.who.int/mediacentre/factsheets/fs334/en/

Wright, J., Carey, M., & Parris, E. (1996) Statistical models for topic identification using phoneme substrings, in Proc. ICASSP, Atlanta.

Appendix 1: Keywords

Keywords used in this thesis:

No.	Politics	Economic	Art	Science	Sport	Weather
1	الانتخابات	النفط	المهرجان	الدراسة	المباراة	درجات
2	الأمن	أسعار	فيلم	الإصابة	الدقيقة	الحرارة
3	القوات	دولار	مهرجان	الباحثون	نقطة	السرعة
4	الوزراء	برميل	الفيلم	دراسة	المركز	درجة
5	قوات	دولارا	كانْ	باحثون	القدم	غربية
6	إسرائيل	شركة	جائزة	فيروس	بطولة	معتدلة
7	مجلس	يوميا	أفضل	الخلايا	لكرة	الرياح
8	الخارجية	مليار	السينما	الفيروس	الفوز	مئوية
9	السلام	أوبك	المصرية	المرض	المجموعة	خفيفة
10	الفلسطينية	الخام	فيلما	خطر	مباراة	شمالية
11	الشرطة	للبرميل	الثقافي	المرضى	البطولة	الجو
12	رويترز	سعر	الأفلام	بسرطان	كرة	المناطق
13	حزب	الاقتصاد	أفلام	خلايا	الدور	شرقية
14	السياسية	الإنتاج	الثقافة	الدم	الدوري	والرياح
15	الجزيرة	إنتاج	جوائز	الإيدز	فاز	والبحر
16	السلطة	زيادة	السينمائي	سرطان	المنتخب	جنوبية
17	الحزب	الطاقة	المخرج	الجسم	منتخب	الأجواء
18	المعارضة	النفطية	للمخرج	مرض	سجل	الجوية
19	حكومة	الأسعار	الدورة	السرطان	بهدف	الموج
20	العاصمة	الغاز	بجائزة	الثدي	برصيد	معدلها
21	دول	الاقتصادي	عرض	حالة	الشوط	السنوي
22	الدستور	المتحدة	المسرح	علاج	نقاط	الأمطار
23	الجيش	السوق	المعرض	النساء	الفريق	شمال

24	الدولية	الشركة	الشاعر	الصحة	لاعب	خفيف
25	حقوق	ملايين	ثقافة	إصابة	بفارق	الطقس
26	غزة	الطلب	بعنوان	العلماء	الموسم	بحدود
27	عملية	النمو	التحكيم	المناعة	برشلونة	انخفاض
28	العراقية	مليارات	الشعر	الدكتور	هدف	غائماً
29	السلطات	العالمية	الرواية	القلب	اللقب	ويطرأ
30	الإسرائيلي	الشركات	الجائزة	الأطفال	فوزا	معدلاتها
31	الأمم	طن	الكتاب	حالات	فوز	زخات
32	الفلسطيني	نفط	الفنية	المصابين	رصيده	رعدية
33	الاحتلال	الأسواق	الكاتب	التدخين	الإسباني	جوي
34	واشنطن	العالمي	فعاليات	الأمراض	أبطال	متفرقة
35	مقتل	صادرات	كتاب	الباحثين	مقابل	حرارة
36	مؤتمر	البنك	الأوسكار	بالمرض	مباريات	جنوب
37	البرلمان	المنظمة	المتحف	العلاج	مدريد	باردة
38	الديمقراطية	التجارة	معرض	يعانون	ذهبية	نشطة
39	الداخلية	المالية	الفنان	الفئران	النهائي	الرعدية
40	العلاقات	الموازنة	الموسيقى	مستويات	أهداف	أجواء
41	المؤتمر	الصناعية	المسرحية	الطبية	الإيطالي	طقس
42	الصحيفة	البترول	الثقافية	مرضى	ميلان	الأرصاد
43	التحقيق	مليون	للمهرجان	الحمل	التعادل	الجوي
44	الحرب	الاستثمارات	مسرحية	النتائج	نجم	يطرأ
45	الأحزاب	المستثمرين	الندوة	انتشار	ملعب	السنوية
46	الفلسطينيين	والغاز	السينمائية	بالسرطان	الألعاب	تساقط
47	المسؤولين	أسواق	فنية	الأطباء	منافسات	جزئياً
48	السياسي	الأسهم	قصيدة	أمراض	بنتيجة	غائما
49	تصريحات	التجاري	المسرحي	الأبحاث	ركلة	يتوقع

50	اغتيال	التعاملات	الأدب	المسبب	النهائية	هوائية
51	المتحدث	اقتصاد	الفنانين	أعراض	النجم	حارة

Appendix 2: Discarded Words

Discarded words:

No.	Word	No.	Word	No.	Word	No.	Word
1	في	21	لا	41	وقد	61	بعض
2	من	22	أو	42	حتى	62	إلا
3	على	23	قد	43	عدد	63	دون
4	أن	24	لم	44	بينما	64	بن
5	إلى	25	قبل	45	أخرى	65	لكن
6	التي	26	الماضي	46	أي	66	مثل
7	الذي	27	أنه	47	هو	67	لدى
8	عن	28	يوم	48	أمام	68	لها
9	مع	29	الأول	49	أيضا	69	عندما
10	بين	30	كان	50	بسبب	70	آخر
11	بعد	31	أكثر	51	الأولى	71	أنها
12	ما	32	غير	52	حول	72	عبر
13	هذه	33	منذ	53	خاصة	73	الثلاثاء
14	هذا	34	ذلك	54	أمس	74	فيما
15	كما	35	الثاني	55	المقبل	75	أكبر
16	العام	36	كل	56	كانت	76	تكون
17	حيث	37	بشكل	57	يمكن	77	بها
18	خلال	38	يكون	58	بأن	78	ثلاثة
19	عام	39	الذين	59	فيها	79	أحد
20	اليوم	40	نحو	60	قال	80	عدم

81	هي	101	تحت	121	يتم	141	إليه
82	الشهر	102	تلك	122	الإثنين	142	حاليا
83	فيه	103	نت	123	إضافة	143	مارس
84	له	104	أقل	124	تم	144	يجب
85	منها	105	ثم	125	فإن	145	سبتمبر
86	حين	106	ضمن	126	الأحد	146	أضاف
87	عليه	107	وذلك	127	الأربعاء	147	خارج
88	إذا	108	الخميس	128	جانب	148	اللاحق
89	مما	109	السبت	129	فقط	149	مايو
90	أعلى	110	أكد	130	فوق	150	التالي
91	عليها	111	بشأن	131	هناك	151	أبريل
92	يذكر	112	حد	132	بما	152	نوفمبر
93	إذ	113	به	133	الآن	153	يونيو
94	أول	114	غدا	134	الجمعة	154	فبراير
95	عند	115	داخل	135	فقد	155	يناير
96	لتصبح	116	ضد	136	ولا	156	أشار
97	مساء	117	بأنه	137	السابق	157	صباحا
98	وهي	118	رغم	138	قالت	158	يوليو
99	الأسبوع	119	عدة	139	حوالي	159	أكتوبر
100	أما	120	وسط	140	لن	160	أغسطس