Smart Voice Search Engine

Shahenda Sarhan
Faculty of Computers and Information,
Mansoura University
Mansoura,Egypt

ABSTRACT
For years and years the search engine researchers focus their efforts on having more accurate and faster search engines. This was more than enough in the past but with the appearance of smart phones the idea of having everything smart became dominant. In this paper we tried to share the dream of having a domain independent search engine and not only an ordinary one but a smart search by voice engine which searches user speech automatically without the user 's request and provide him with evidence on his speech, this engine was called SVSE. Through the paper we will introduce the proposed system in details explaining each part of it and finally discussing the difficulties it faces.

General Terms
natural Language Processing, speech Recognition.

Keywords
speech recognition, speaker recognition, search engine, search by voice.

1. INTRODUCTION
Spoken queries are a natural medium for searching the web in settings where typing on a keyboard is not practical [21]. Also with the users' new expectations about the nature of the services through having constant access to the information and services of the web. Given the nature of delivery devices and the increased range of usage scenarios, speech technology has taken on new importance in accommodating user needs for ubiquitous mobile access any time, any place, any usage scenario as part of any type of activity.

In the last decade many researchers and firms tried to introduce a voice search engine that handles the recognition of spoken search queries accurately. Since (2007) GOOG 411 [5] had been assisting people with obtaining phone numbers, Users who called a toll-free telephone numbers were asked for the city and state of the sought business. Users were able to select the destination by speaking or pressing the number that corresponded to the desired result. In (2010), GOOG-411 shut down its service. While Google did not provide an official reason for the shut down, many believe that Google had simply gathered enough voice samples for its research purposes

With the advent of smart phones with large screens and data connectivity, Google Mobile App (GMA) for IPhone was released on (2008) including a search by voice feature. GMA search by voice extended the paradigm of multi-modal voice search from searching for businesses on maps to searching the entire World Wide Web [5][21].

While on (2009), a multi-modal user interface called Google Maps Navigation with speech or text as the input, and maps with as the output was released in conjunction with Google Android OS 2.0. This version of the system exceeds its previous by adding voice commands, traffic reports, and street view support.

In 2011 many Arabian revolutions were erupted in the Middle East. As a result of these revolutions search engines utilizations especially Google was dramatically increased. Through this time Google represents the outlet of the news to all the Arabian peoples through tabs and smart phones. It Became natural to find yourself in a heated debate on a social network as Facebook or Twitter about politics and certain personalities what they said or even did and you are demanded to provide an evidence (eg: YouTube video file or a picture) on your claims.

From here the researcher sensed the need for a smart voice search engine system for mobile phones. This system works as a personal assistant that can search on Google automatically and did not wait for you to ask, you can think of as a genie that makes your speech become articles and video.

The proposed system needed first to recognize your voice using speaker and speech recognition techniques, and then starts to search for you recognized queries through Google search. That will be described in details in section 4 but first we will discuss the concepts of speaker and speech recognition in next sections.

This paper is organized as follows: Section 2 introduces the concept of speaker recognition in details. While, section 3 discusses the concept of automatic speech recognition, a detailed description of our proposed model SVSE was introduced in section 4. Finally, Section 5 outlines our conclusions and suggested future work.

2. SPEAKER RECOGNITION
Speaker recognition has generally been viewed as a problem of verifying or recognizing a particular speaker in a segment of speech spoken by a single speaker [11][13][16]. But for some applications of interest the problem is to verify or recognize particular speakers in a segment of speech in which multiple speakers are present [1][2][6]. Automatic systems need to be able to segment the speech among the speakers present and/or to determine whether speech by a particular speaker is present and where in the segment this speech occurs.
Speaker recognition process encompasses three terms identification, verification and diarization. In automatic speaker identification (Figure 1), there is no priori identity claim, the system decides who the person is, or the person is known or unknown [2][12].

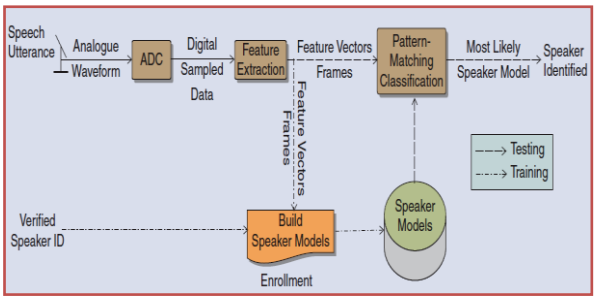

Fig.1. Speaker Identification System. [12]

While automatic speaker verification (Figure 2) involves the use of a machine to verify a person's claimed identity from his voice [6,7,12]. This task is also known as voice verification or authentication, talker verification or authentication, and speaker detection.

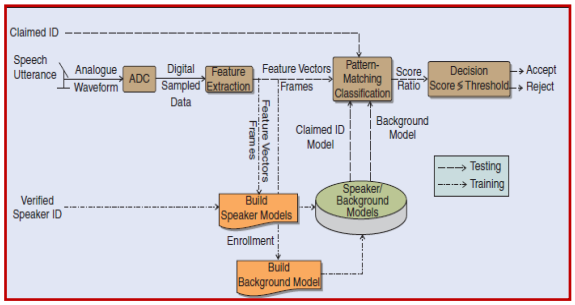

Fig.2. Speaker Verification System. [12]

Finally Speaker diarization techniques [19] are used in multiple-speaker scenarios where speech from the desired speaker is intermixed with other speakers. In this case, it is desired to segregate the speech into segments from the individuals before the recognition process commences. So the goal of this task is to divide the input audio into homogeneous segments and then label them via speaker identity.

According to the constraints placed on the speech used to train and test the system, automatic speaker recognition can be further classified into text-dependent where the recognition system knows the text spoken by the person, either fixed passwords or prompted phrases [6][7]. Or it may be a text-independent recognition in which the recognition system does not know text spoken by person.

Running a speaker recognition system typically involves two phases. In the first phase, a user enrolls by providing voice samples to the system. The system extracts speaker-specific information from the voice samples to build a voice model of the enrolling speaker [6][7]. In the second phase, a user provides a voice sample that is used by the system to measure the similarity of the user's voice to the models of the previously enrolled user and, subsequently, to make a decision.

For speaker recognition, features that exhibit high speaker discrimination power many forms of pattern matching and corresponding models are possible. Pattern-matching methods include Nearest-neighbors models, dynamic time warping (DTW) [15], Gaussian mixture models (GMM) [7], the hidden Markov model (HMM)[14], artificial neural networks, and vector quantization (VQ)[18]. Template models are used in DTW, statistical models are used in HMM, and codebook models are used in VQ.

3. AUTOMATIC SPEECH RECOGNITION

The problem of automatic speech recognition (ASR) has been approached progressively; from a simple machine that responds to a small set of sounds to a sophisticated system that responds to fluently spoken natural language [13] and takes into account the varying statistics of the language where the vocalizations vary in terms of accent, pronunciation, articulation, roughness, nasality, pitch, volume, and speed. And the speech is distorted by a background noise and echoes, electrical characteristics [3][9][10]. All of this affects the accuracy of speech recognition (Figure 3) in addition to the following:

- Vocabulary size and confusability
- Speaker dependence vs. independence
- Isolated, discontinuous, or continuous speech
- Task and language constraints
- Read vs. spontaneous speech
- Adverse conditions

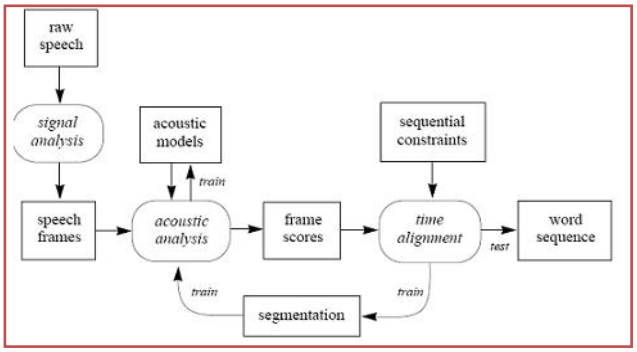

Fig. 3. Speech Recognizer [20]

Many machine learning algorithms can lead to significant advances in ASR. The biggest single advance occurred with the introduction of the expectation-maximization (EM) algorithm for training HMMs [8] . With the EM algorithm, it became possible to develop speech recognition systems for real-world tasks using the richness of GMMs to represent the relationship between HMM states and the acoustic input. In these systems the acoustic input is typically represented by concatenating Mel-frequency cepstral coefficients (MFCCs) [11][17] or perceptual linear predictive coefficients (PLPs) computed from the raw waveform and their first- and second-order temporal differences.

This non-adaptive but highly engineered preprocessing of the waveform is designed to discard the large amount of information in waveforms that is considered to be irrelevant for discrimination and to express the remaining information in a form that facilitates discrimination with GMM-HMMs.

Google was one of the pioneers in speech recognition using it to design one of the first search by Voice engine to work in parallel with its text search engine. The goal of the Google search by Voice was to recognize any spoken search query and to handle anything that Google Search can handle. This makes it a considerably more challenging recognition problem, because the vocabulary and complexity of the queries will be so large.

In this study we will introduce a model we termed as smart voice search engine (SVSE). The SVSE system will automatically recognize your voice and on behalf of you will analyze your speech and search for articles and video files supporting your point of view. That, what we will try to explain in the next section, clarifying the system major parts.

4. SMART VOICE SEARCH ENGINE MODEL (SVSE)

As usual if you want to search on the internet you just need to open one of the popular search engines as GOOGLE or YAHOO and type your query in the search box and press enter and you will find what you want in your hands. Even if you are bored of writing or have a problem with your keyboard or mobile touch screen you can search by your voice through GOOGLE search by voice.

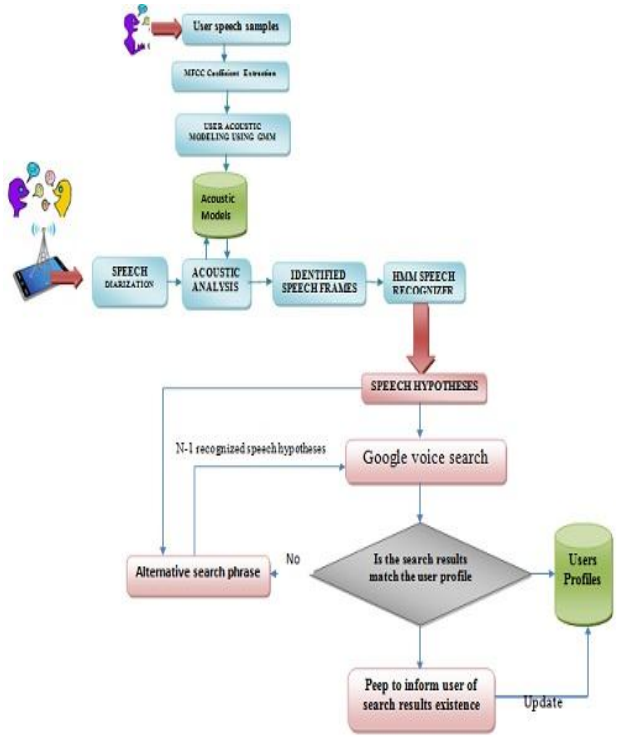

Fig.4. Smart Voice Search Engine.

This can be helpful if you are quietly sitting in your office room and know what you need but what you will do if you are in the middle of a heated debate. As mentioned before the raging events that affect the Middle East based on the revolutions in many Arabian countries can simply hurl you in the middle of a conversation on your opinion of what is happening. In that time especially you will mightily need proof on your opinion. These proofs may involve YOUTUBE videos, pictures, articles and so on.

From this the idea of a smart voice search engine (SVSE) arises to simply help you find the answer to your query even if you did not ask. The SVSE (Figure 4) is a system that works automatically on your speech especially if you are in a conversation. The system captures your speech through the mobile or the laptop microphone and analyzed it as shown in Figure (3) which represents a basic block diagram of the proposed SVSE.

The proposed model simply involves three main phases

- The Speaker recognition phase
- The automatic Speech Recognition
- The search engine

4.1 The Speaker Recognition Phase

As mentioned before the speaker recognition process involves:

1. In this step the user through the installation of the system record samples of his speech on the system to help the system verify his voice later.
2. Then the system starts to extract the acoustics features of the user from the samples and build a model for the speaker using Standard mel-frequency cepstral coefficient (MFCC) features with Gaussian mixture model (GMM) recognizer are used for speaker identification

$$P(\vec{x} \mid \lambda_j) = \sum_{i=1}^{M} gN(\vec{x}; \vec{\mu}_1; \sum_i) \qquad (1)$$

where g_i are the *mixture weights* satisfying $\sum_{i=1}^{M} g_i = 1$. The $N(\vec{x}; \vec{\mu}_1; \sum_i)$ are the individual component densities, which for a *D*-variate Gaussian function are of the form:

$$N(\vec{x}; \vec{\mu} I; \sum_i) = \frac{1}{(2\pi)^{D/2} \mid \sum_i \mid^{1/2}} e^{\frac{-1}{2} (\vec{x_i} - \vec{\mu_i}) \sum_i^{-1} (\vec{x_i} - \vec{\mu_i})} \quad (2)$$

After initializing the system, it will monitor you and as well as you starts a debates with someone that lasts more than 5 minutes the diariazation step is initialized.

3. Through the diariazation step the system uses the initialized speaker model in step 2 to help in assigning the corresponding frames [4] for each speaker with the help of a Hidden Markov Model. These identified frames present the input of the next phase.

4.2 The Automatic Speech recognition phase

Through this phase the identified speech frames represent the input of the automatic speech recognition engine which is based on HMM. These frames is analyzed and compared to the acoustic models[3,11] to reach the set of recognized word sequences, where w1: L = w1,...,wL which is most likely to have generated Y(where Y represent sequence of fixed size acoustic vectors $Y_{1:T} = y_1, \ldots, y_T$)

$$\hat{w} = argw\ max\ \{P(w|Y)\}. \ (3)$$

However, since P(w|Y) is difficult to model directly, Bayes' Rule is used to transform (3) into the equivalent problem of finding:

$$\hat{w} = argw\ max\{p(Y|w)P(w)\}. \ (4)$$

The likelihood p(Y |w) is determined by an acoustic model and the prior P(w) is determined by a language model.

4.3 Search Engine phase

In this phase the recognized word sequence $w_{1:L} = w_1,...,w_L$ will be searched for on Google search and the search results will be checked to find out if they match the user profile or not. If they did not match an alternative word sequence will be check and so on to find the right one then the system will peep to inform the user of the results.

4.4 System Implementation and Discussion

As mentioned before the MFCC features with GMM were used in building the speaker model while the HMM was used through the rest of the system. The acoustic models were taken from the VoxForge[1] Acoustic Models. A prototype was build using Java programming language through the Android Software Development Kit.

The system performance was evaluated based on the similarity of the search results and the user preferences recorded in its profile using the Nearest-neighbor algorithm. These evaluation results were presented in Table 1 involves two cases the first at 10 minutes and the second at 20 minutes. Each case was evaluated depending on the gender and number of the debate participants.

Table I: SVSE Accuracy results depending on debaters' gender and numbers and debate duration

Duration (min)	# of speakers	Speakers genders	recognition accuracy based on user profile
≤10	2	males	65%
		females	62.17%
		mixed	71.89%
	3	males	62.48%
		females	61.32%
		mixed	69.53%
	4	males	57.01%
		females	52.97%
		mixed	67.54%
	5	males	55.89%
		females	51.08%
		mixed	65.82%
>10	2	males	69%
		females	67.55%
		mixed	74.01%
	3	males	65.25%
		females	64.32%
		mixed	72.53%
	4	males	61.92%
		females	57.4%
		mixed	70.81%
	5	males	60.01%
		females	55.22%
		mixed	69.55%

1 http://www.voxforge.org

The results in Table 1 indicate that the system achieves accepted accuracy up to 74% if the debate duration exceeds 10 minutes and the debate participants were males and females while, the minimum accuracy was achieved during a debate between 5 females.

From Table 1 we can also notice that the system accuracy for mixed genders is higher than for same sex, but only a little higher this is perhaps due to the easier association of speech frames with speakers when the genders differ. Also the speech diarization for female speakers is apparently more difficult than males. Finally as the debate lasts longer the system accuracy gets higher as the system is well trained through time.

The low performance of the system is also due to the effect of noisy environment especially that the model is text independent which is considered a real problem in the way of the recognition process but we hope to solve this problem soon.

Finally the system performance demonstrated an accepted evaluation results comparing to other search by voice engines as in GOOGLE. For example Figure (5) illustrates that the average percentage of users Queries about news using GOOGLE in the fourth week of January 2014 approached 4% while in SVSE was about 95% which is justified according to system users as the SVSE Unasked analyzed their most interested and talked about topics and not as in google you need to have a clear mind to select your query spoken words and to have the will to open google on you smart phone and search.

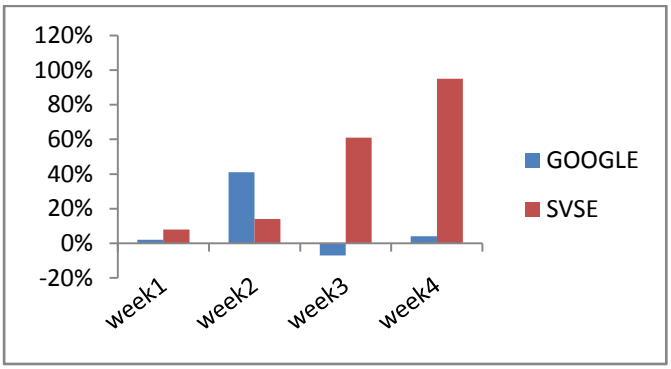

Fig.5. Users' News Queries Percentages in January 2014

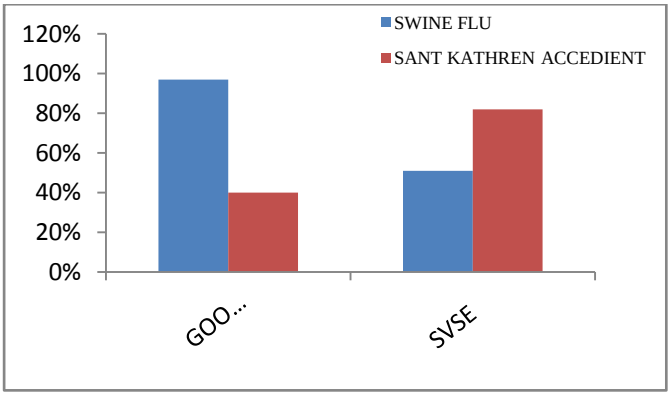

Fig.6. The Most Common Search Queries in February 2014

Figure (6) demonstrated that the most common search on Google was about the Swine Flu while in SVSE was Saint Kathrin accident as SVSE is concerned about being as a personal assistant that can support you during you debate with your colleges or friends or even in your own thoughts. using SVSE you don't need to open your phone and run an application the speak to a microphone you need just to talk and it will do all the work.

5. CONCLUSIONS

For decades web and mobile user used to type a text and press a button to find what they need or even speak after pressing a microphone button. This was really a very good way in the past but with the severe changes in the folk and the lack of time this will not be a useful way. So through this paper the researcher tried to introduce a proposed model of a smart voice search engine called (SVSE) that automatically capture your voice and recognize your speech if you are alone or in a party and search for it even if you did not ask. You can thought of it as your voluntarily search assistant. A prototype of the system was build and achieved an accepted accuracy up to 74% but it still need more training and testing.

In the future, the researcher plan to pursue several future researches on the speech diariazation especially in a multi-model speaker.

6. REFERENCES

[1] Alvin F Martin, Mark A. Przybocki, (2001). "Speaker Recognition in A Multi-Speaker Environment". In: Proc. Eur. Conf. Speech Commun. Technol., Vol. 2, pp.787 - 790.

[2] Cemal Hanilc¸et.al., (2013). "Speaker Identification From Shouted Speech: Analysis And Compensation". In**:** IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 8027 – 8031, 26-31 May 2013. Vancouver, BC, Canada.

[3] Janne Pylkkönen, (2009). "Investigations on Discriminative Training in Large Scale Acoustic Model Estimation". In: the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009), Brighton, UK, pp. 220–223.

[4] Janne Pylkkönen, (2013). "Towards Efficient and Robust Automatic Speech Recognition: Decoding Techniques and Discriminative Training". In: Aalto University publication series Doctoral Dissertations, 44/2013. ISSN: 1799-4942 (electronic), 1799-4934 (printed), 1799-4934 (ISSN-L). Aalto University, Finland.

[5] Johan Schalkwyk,et.al.,(2010). "Google Search by Voice: A Case Study". In: Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics, Springer, pp. 61-90.

[6] Joseph P. Campbell, (1997). "Speaker Recognition: A Tutorial". In: Proceedings of The IEEE, Vol. 85, NO. 9.

[7] Manas A. Pathak, and Bhiksha Raj,(2013)." Privacy-Preserving Speaker Verification and Identification Using Gaussian Mixture Models" In IEEE Transactions on Audio, Speech & Language Processing, Vol.21,No.2,pp. 397-406.

[8] Mark Gales and Steve Young, (2007)." The Application of Hidden Markov Models in Speech Recognition". In: Foundations and Trends in Signal Processing, Vol. 1, No. 3, pp. 195–304, DOI: 10.1561/2000000004.

[9] Mohammed Benzeghiba, et.al. (2007) "Automatic Speech Recognition and Speech Variability: A Review". In: Speech Communication, Vol. 49, pp. 763-786.

[10] Nelson Morgan, (2012). "Deep and Wide: Multiple Layers in Automatic Speech Recognition". In: IEEE Transactions On Audio, Speech, And Language Processing, Vol. 20, No. 1.

[11] Qin Jin, (2007). "Robust Speaker Recognition". In: partial fulfillment of the requirements for the degree of Doctor of Philosophy in Language and Information Technologies, Language Technologies Institute School of Computer Science, Carnegie Mellon University , 5000 Forbes Avenue, Pittsburgh, PA 15213.

[12] Roberto Togneri and Daniel Pullella,(2011). "An Overview of Speaker Identification: Accuracy and Robustness Issues". In: IEEE Circuits And Systems Magazine, Vol.11, No. 2 , pp. 23-61, ISSN : 1531-636X

[13] Sadaoki Furui, (2005). "50 Years of Progress in Speech and Speaker Recognition Research", In: Ecti Transactions on Computer And Information Technology, Vol.1, No.2.

[14] Shaikh Salleh, et.al.,(2000). "Speaker Recognition Based On Hidden Markov Model". In: Natiaonal Conference on Telecommunication Technology 2000, 20th - 21st Nov. 2000, Hyatt Regency Hotel, Johor Bahru.

[15] Svetlana SEG Ă RCEANU & Tiberius ZAHARIA,(2013). "SPEAKER VERIFICATION US ING THE DYNAMIC TIME WARPING". In: U.P.B. Sci. Bull., Series C, Vol. 75, No. 1, ISSN 1454-234x.

[16] Uergen Luettin, (1997)."Visual Speech And Speaker Recognition". Dissertation submitted to the University of Sheffeldfor the degree of Doctor of Philosophy. Department of Computer Science. University of Sheffeld.

[17] Vahid Majidnezhad & Igor Kheidorov , (2012). "A HMM Based Method for Vocal Fold Pathology Diagnosis". In : IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, pp. 135-138, ISSN (Online): 1694-0814.

[18] Wan-Chen Chen , Ching-Tang Hsieh and Chih-Hsu Hsu, (2008). "Robust Speaker Identification System Based on Two-Stage Vector Quantization". In : Tamkang Journal of Science and Engineering, Vol. 11, No. 4, pp. 357-366.

[19] Xavier Anguera, ChuckWooters, Barbara Peskin and Mateu Aguil, (2005). "Robust Speaker Segmentation for Meetings:The ICSI-SRI Spring 2005 Diarization System". In **:** MLMI'05 Proceedings of the Second international conference on Machine Learning for Multimodal Interaction , pp. 402-414.

[20] Speech Recognition, available at: www.learnartificialneuralnetworks.com/speechrecognition.html

[21] WikipediA: WWW.Wikipedia.org