K-mean Clustering and Arabic Vowels Formants Based Speaker Identification System

Mikhled Alfaouri  
Communications and Electronics Department, Philadelphia  
E-mail: malfaouri@hotmail.com

Khaled Daqrouq  
Communications and Electronics Department, Philadelphia  
E-mail: haleddaq@yahoo.com

Jamal Al-Nabulsi  
University, Jordan 19392, Biomedical Engineering Department  
Hashemite University, Jordan  
E-mail: jian@hu.edu.jo

Abstract

This paper introduces and addresses the proposes of a new approach for speaker feature extraction based on experimental and theoretical approached, where the formants of Arabic Vowels are proposed to distinguish the speaker features from each other. Discrete Wavelet Transform (DWT) in conjunction with Algorithmic Power Spectrum Density (PSD) is used to illustrate the distinguisher of different speaker formants. This approach provides a more efficient method in speaker recognition rate, i.e., higher accuracy. K-means clustering (KC) and Root Mean Square Difference Similarity Measure (RDSM) are used for features classification. Instead the conventional method extracts the features from one word or more. In this paper the authors proposed a new method to utilize the Arabic Vowels. Ultimately, the attained results by the presented method showed considered a performance in classification, which reaches about 94% in classification rate. As a result of DWT utilization, the system works with considerable capability of features tracking even with 0dB SNR.

Keywords: Speaker Identification, Arabic Vowels, Wavelets, Clustering.

1. Introduction

Based on the resonant frequencies, speech synthesis of the vocal track during voiced speech that recognized as formants are commonly used to be valuable features for both automatic speech recognition and speaker recognition (Malkin et al.; Huang et al., 2000). Furthermore, projects such as the UW Vocal Joystick, University of Washington explored the use of formants for 1-D and 2-D continuous motion control. The Vocal Joystick project is creating a new vocal interface to allow people, especially individuals with motor impairments, to utilize many aspects of their voice to easily interact with computers or other devices (Malkin et al., 2007).

More adequate description of formant is described as a function of the supralaryngeal vocal tract. The air in the oral cavity, oropharynx, laryngopharynx, and in many phonemes the nasal cavities
and nasopharynx vibrates at a range of frequencies in response to the vibratory movement of the vocal folds and air passing during the glottis. These resonant frequencies rely on the size and shape of the vocal tract and its constrictions, in addition to tongue and lip positions, which can change the functional length of the vocal tract (Gelfer and Mikos, 2007). Vocal tract resonances are often studied in terms of vowel formant frequencies. Because the male vocal tract is about 15% longer than the female vocal tract, the speech signal taken from men are predicted to have lower formant frequencies than those considered characteristic of women (Bachorowski and Owren, 1999). Average Vowel Formants (AVF) (in Hertz) for Male Subjects, Transgender Subjects, and Female Subjects are extracted in (Gelfer and Mikos, 2007), for F1 of /i/ vowel (in instance) as 283.16, 272.40 and 323.01, respectively, for F2 as 2200.71, 2365.44, and 2614.15 respectively, and for F3 as 2770.79, 3138.77, and 3230.26 respectively. A fast Fourier transform (FFT) was conducted with a single 1024-point Hamming window and Linear Predictive Coding (LPC). A frequently cited acoustic parameter as a cue to body size is mean fundamental frequency (F0) of voice as defined by (Darwin, 1871).

Experimental literature proposes solution for formant tracking, many are discussed in (Malkin et al., 2007), such as (Xia and Espy-Wilson, 2000), uses LPC spectral analysis to estimate potential formant frequencies. There have also been other types of formant trackers such as HMM-based methods (Acero, 1999) approaches using nonlinear predictors (Deng et al., 2003), and a current one using a Kalman filtering framework (Deng et al., 2004) to name a few.

It should be noted that (Acero, 1999 and Deng et al., 2004) aimed to model the vocal track resonant frequencies in general during both voiced and unvoiced speech segments. The paper of (Malkin et al., 2007) presented a graphical model for the use with formant tracking, inspired by the Graphical model approach to pitch tracking.

**Literature Survey of Speaker Recognition**

The applications of speech signal processing such as speech recognition and speaker identification have been drastically increased in recent years, because of its non-contact characteristic and speaker identification system that can be utilized to suspect identification. The realization of speaker identification can be divided into two main parts: First is the feature extraction and the second is the speaker classification based on the extracted features (Wu & Lin, 2009; Sarikaya, Pellom, & Hansen, 1998).

For decades, speaker identification systems (SI) has been under study by a large number of researches (Avci D. et al., 2009). From commercial pointview, speaker identification system is a technology with potentially large market due to the applications of broadly ranges from automation of operator-assisted service to speech-to-text aiding system for hearing impaired individuals (Avci D. et al., 2009; Reynolds et al., 2000).

In general, a speaker identification system can be implemented by observing the voiced/unvoiced components or through analyzing the energy distribution of utterances. Some digital signal processing methods, such as linear predictive coding technique (Avci D. et al, 2009; Adami & Barone, 2001; Tajima, Port, & Dalby, 1997), Mel frequency cepstral coefficients (MFCCs) (Wu & Lin, 2009; Mashao & Skosan, 2006; Sroka & Braida, 2005; Kandera, Arai, Hermansky, & Pavel, 1999), discrete wavelet transform (DWT) (Fonseca, Guido, Scalassara, Maciel, & Pereira, 2007) and wavelet packet transform.

Because of its suitability for analyzing non-stationary signals, DWT has become a powerful alternative to the Fourier methods in many speech/speaker identification applications (Avci D. et al., 2009; Buckheit and Donoho, 1995; Coifman and Wickerhauser, 1992; Evangelista, 1993, 1994; Ha et al., 2005; Kadambe and Boudreaux-Bartels, 1992; Kadambe and Srinivasan, 1994; Maes, 1994; Mallat, 1989; Reynolds et al., 2000; Saito, 1994; Szu et al., 1992; Visser et al., 2003; Westfried and Wickerhauser, 1993). The main advantage of wavelets is that they have a varying window size, being wide for slow frequencies and narrow for the fast bones, thus leading to an optimal time–frequency resolution in all frequency ranges. Furthermore, owing to the fact that windows are adapted to the
transients of each scale, wavelets lack of the requirement of stationary (Avci D. et al., 2009; Coifman and Wickerhauser, 1992; Evangelista, 1993, 1994; Ha et al., 2005; Kadambe and Boudreaux-Bartels, 1992; Kadambe and Srinivasan, 1994; Maes, 1994; Mallat, 1989; Reynolds et al., 2000; Saito, 1994; Szu et al., 1992; Visser et al., 2003).

Artificial neural network models (ANN) have been effectively used for classification and control in several disciplines. Norton and Zahorian, (2003) have developed an ANN based techniques for speaker verification. Zaki et al., (1996) have used a cascade neural network for speaker recognition. Mak and Kung, (2000) used radial basis function networks for speaker verification. Homayounpour et al., (1995) compared ANN to second order statistical techniques for speaker verification. Reliability is a major problem related to neural network models, and reliability is essential for the success of intelligent systems incorporated in speaker verification systems or other fraud preventing technologies. Reddy et al., (1995) and Das et al., (2001) have developed committee neural networks (CNN) to improve the reliability of ANN based classification systems. The committee opinion was based on a majority voting. The question remains if CNN can be developed for speaker verification. The specific aim of the presented study in (Reddy N.P &. Buch O.A, 2003) was to address this question by developing and evaluating CNN for speech based, text dependent, and verification of the speaker from imposters.

This paper presents a new approach of speaker identification utilizing the first five formants of Arabic vowels (FFAV). K-means clustering (KC) and Root Mean Square Difference Similarity Measure (RDSM) are proposed for features classification. The remainder of the paper is organized as follows: In section 2 Arabic Language in Brief. In section 3 the proposed method is described. The experimental results and discussion in section 4 followed in section 5 by conclusions.

2. Arabic Language in Brief

Recently, Arabic language became one of the most important and broadly spoken languages in the world, with an expected number of 350 millions speakers distributed all over the world and mainly covering 22 Arabic countries. Arabic is Semitic language that characterizes by the existence of particular consonants like pharyngeal, glottal and emphatic consonants. Furthermore, it presents some phonetics and morpho-syntactic particularities. The morpho-syntactic structure built, around pattern roots (CVCVCV, CVCCVC, etc.) (Zitouni I. and Sarikaya R., 2009).

The Arabic alphabet consists of 28 letters that can be extended to a set of 90 by additional shapes, marks, and vowels. The 28 letters represent the consonants and long vowels such as ڥ and ڦ (both pronounced as/a:/), ڥ (pronounced as/i:/), and ڥ (pronounced as/u:/). The short vowels and certain other phonetic information such as consonant doubling (shadda) are not represented by letters, but by diacritics. A diacritic is a short stroke placed above or below the consonant. Table 1 shows the complete set of Arabic diacritics. We split the Arabic diacritics into three sets: short vowels, doubled case endings, and syllabification marks. Short vowels are written as symbols either above or below the letter in text with diacritics, and dropped all together in text without diacritics. We find three short vowels: fatha: it represents the /a/ sound and is an oblique dash over a letter, damma: it represents the /u/ sound and has shape of a comma over a letter and kasra: it represents the /i/ sound and is an oblique dash under a letter as explained in Table 1 (Zitouni I. and Sarikaya R., 2009).

<table>
<thead>
<tr>
<th>Long Vowel Name</th>
<th>Connected with letter 'پ' (sounds B)</th>
<th>Pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alef</td>
<td>يا بَّ</td>
<td>/baa/</td>
</tr>
<tr>
<td>Waw</td>
<td>ﯾ ﯾ</td>
<td>/buu/</td>
</tr>
<tr>
<td>Yaa'</td>
<td>ﯾ</td>
<td>/pii/</td>
</tr>
</tbody>
</table>
3. Proposed Method

The authors used the resonant frequencies of the vocal track during voiced speech to recognize as formants, which are distinguishable for each person proposed as speaker features. Thus, the first five formants extracted from one of Arabic vowels speech signal are used as the unique features for the speaker; here is an investigation of a new speaker identification system that based on distinct acoustic features contained in logarithmic spectrum taken from speech signals. Features are classified by K-means clustering algorithm. More specifically, this system consists of three main stages (see Fig.1):

**Figure 1:** Block diagram of proposed method

---

**DWT for Filtration**

To expand the PSD capability of formants illustration and speaker features extraction, the authors proposed the DWT approximation coefficients $a_j$ of multiple scales as shown in (Fig.2). Detail coefficients $d_j$ is ignored.

$$a_{j+1}(t) = \sum_{m} h(m-2t)a_j(m) \quad (1)$$

Where, the set of numbers $a_j(m)$ represents the down sampled approximation of the signal at the resolution $2^{-j}$ shown in Fig.1, $h(n)$ is the coefficient of the linear combination that approximates the wavelet scaled version function $\phi(x)$ as found in (Mallat S., 1989):

$$\phi(x/2) = 2^{j/2} \sum_{m} h(m)\phi(x-m) \quad (2)$$

$$h(m) = \frac{1}{2^{j/2}} \int \phi(x/2)\phi(x-m)dx \quad (3)$$

**Figure 2:** DWT coefficients generation

---
Feature Extraction

In this stage, the speech signal is formed to fill the input data matrix to the PSD, which is assists greatly for speaker identification in utterance-dependent system. Fig.3 illustrates the extracted features (first five formants) for same speaker. We can notice that each speaker has distinct features. The five formants picking algorithm is explained via flowing two steps:

1. PSD (is denoted as $P_{xx}$) that is estimated using the Yule-Walker AR method that is also called windowed method. This method fits the AR linear prediction filter model to the signal by minimizing the forward prediction error in the least squares sense. This formulation leads to the Yule-Walker equations, which are solved by the Levinson-Durbin recursion. The spectral-estimate-returned by method is the squared magnitude of the frequency response of this AR model (Marple, S.L 1987). Then $M$ vector of the speaker's formants is represented by logarithmic scale:

$$F_{Pxx}(i) = \sum_{i=1}^{M} 10 \log_{10}(P_{xx}) \text{ dB} \quad (4)$$

With normalized frequency axis ($f = 0:1/128:1$). $F_{Pxx}(i)$ Returns $P_{xx}$ containing sufficient features to demonstrate the speaker identity, as shown in Fig1

2. Picking the first 5 picks by the local maxima concept (Fig. 2):
   - getting the difference:
     $$D = P_{xx}(i+1) - P_{xx}(i)$$
   - Picking the index where the difference goes from + to -; this is the local maxima.
   - Normalize the index; $I = \frac{I}{128}$.

Figure 3: First 5 formants of Arabic Vowels by PSD estimated via Yule-Walker AR method of four utterances for same person a. Person one b. Person two.

Classification

In this paper, K-means clustering algorithm is introduced for classification problem. Clustering in N dimensional Euclidean space $R^N$ is the process of partitioning a given set of n points into a number, say K, of clusters based on some similarity metric which establishes a rule for assigning patterns to the domain of a particular cluster centre (centroid) (Fig.4). Let the set of n points $\{x_1, x_2, ..., x_n\}$ be
K-mean Clustering and Arabic Vowels Formants based Speaker Identification System

represented by the set $S$, and the $K$ clusters is represented by $C_1, C_2, ..., C_K$ (Bandyopadhyay S., Maulik U., 2002). Then

$$C_i \neq \emptyset \text{ for } i = 1, ..., K,$$

$$C_i \cap C_j = \emptyset \text{ for } i = 1, ..., K, j = 1, ..., K,$$

Where $j \neq i$,

$$\bigcup_{i=1}^{K} C_i = S.$$  

K-Means [17, 20] is one of the widely used clustering techniques, which is an iterative hill climbing algorithm. It consists of the following steps:

1. Choosing $K$ initial cluster centroids $z_1, z_2, ..., z_K$, randomly from the $n$ points $\{x_1, x_2, ..., x_n\}$.
2. Assigning point $x_i, i = 1, 2, ..., K$ to cluster $C_j, j \in \{1, 2, ..., K\}$ where $\|x_i - z_j\| \leq \|x_i - z_p\|$, $p = 1, 2, ..., k$, and $j \neq p$.
3. Calculating new cluster centroids: $z^*_1, z^*_2, ..., z^*_K$, where $z^*_i = \frac{1}{n_i} \sum_{z_j \in C_i} x_j, i = 1, 2, ..., K$, where $n_i$ is the number of elements belonging to cluster $C_i$.
4. If $z^*_i = z_i \forall i = 1, 2, ..., K$ then end. Otherwise continue from 2.

**Figure 4:** K-Means data clustering with $K=4$

K-means is a popular clustering algorithm that has been used in a variety of application disciplines, such as image clustering and information retrieval, as well as speech and speaker identification. Different types of clustering algorithms that are based on K-means, are mentioned in (Marroquin, J. & Girosi F, 1993; Bellot P. & El-Beze M, 1999; Wagsta Kiri & Cardie Claire, 2001; Rougui J.E. et al., 2007), such as the modified version for background knowledge, a genetic algorithm, and the syllable contour that is classified into several linear loci that serve as candidates for the tone-nucleus using segmental K-means segmentation algorithm. Moreover, some work involved with solving the problem of slow speaker identification for large population systems.

Here is an investigation of a new speaker identification system that based on distinct acoustic features contained in logarithmic spectrum taken from wavelet transform of speech signals. Features are classified by K-means algorithm. More specifically, this system consists of two main stages: Speaker Features Classification is performed using K-means classifier. The discrimination process is based on finding out the average sums of point-to-centroid distances $D$ in the 1-by-$K$ cluster vector.
The input matrix to the K-means classifier is composed of four columns features matrix belongs to impostor speaker. Each column contains first five speaker formants of four different realization of the same speaker. The sums of point-to-centroid distances are averaged for each \( J \) sub-signal separately. The resulted distances \( D \) are then compared with models features distances stored in the system over \( J \) sub-signals. On other words, the previously mentioned process is repeated for each different model, to find out the model that has the nearest \( D \) to the impostor's \( D \) as shown in Fig.5.

Figure 5: Four Distances \( D \) for Different five Speakers models

Alternative Classification Method

The features may be distinguished based on calculating maximum similarity between input features and the model speech signals features stored in the system memory. Verification between input features vector, and the models features is accomplished by Root Mean Square Difference Similarity Measure (RDSM):

\[
RDSM_n = 100 - \sqrt{\sum_{n=1}^{N} \frac{(\bar{F}_{in} - \bar{F}_{mn})^2}{\bar{F}_{in}^2}} \times 100
\]  \( \text{(6)} \)

Where \( n=1,2,\ldots,N \), \( N \) is referred to the number of stored models, \( \bar{F}_{in} \) is input features vector and \( \bar{F}_{mn} \) model features vector. This measure is calculated for each features vector for input digit at each time, one of stored models. So that, \( N \) magnitudes are achieved for all stored models.

\[
RDSM_n = [RDSM_1, RDSM_2, \ldots, RDSM_N]
\]  \( \text{(7)} \)

The classifier decision is taken by determining the maximum of all \( RDSM_n \) calculated for all models stored in the system

\[
\max_{1 \leq m \leq N} RDSM_n \text{ for } n=1,2,\ldots,N
\]  \( \text{(8)} \)

5. Results and Discussion

In this research paper, speech signals were recorded via PC-sound card, with a spectral frequency of 4000 Hz and sampling frequency of 8000 Hz. For each speaker, the Arabic Vowels: long vowels such as \( \text{ش} \) and \( \text{ث} \) (both pronounced as/\( a:/\)), \( \text{ش} \) (pronounced as/\( i:/\)), and \( \text{ش} \) (pronounced as/\( u:/\)) (tabulated in
Tab.1) were recorded 20 times. 3 females, and four years old along with 14 males participated in speech vowels recording. The recording process was provided in normal university office conditions.

Fig. 5 illustrates the results of investigation of efficacy of the first five formants of Arabic vowels speakers' models discrimination. The figure illustrates four distances $D$ for five different speakers. We can see the clear discrimination between these speakers.

Tab.2 presents the recognition rate of proposed method via five formants of Arabic vowels (FFAV) features. The recognition rate was taken by 400 vowels signals if text-dependent system (vowel dependent system). Classification was accomplished by utilizing two presented in previous section methods: K-means clustering (KC) and Root Mean Square Difference Similarity Measure (RDSM), where KC results were superior.

Table 2: Recognition rate for KC and RDSM in vowel-dependent system

<table>
<thead>
<tr>
<th>Classification</th>
<th>Feature Extraction</th>
<th>Recognition System</th>
<th>NH [%]</th>
<th>Recognition Rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>KC</td>
<td>FFAV</td>
<td>Text-dependent</td>
<td>19</td>
<td>93.12</td>
</tr>
<tr>
<td>RDSM</td>
<td>FFAV</td>
<td>Text-dependent</td>
<td>none</td>
<td>56.82</td>
</tr>
</tbody>
</table>

Table 3: Recognition rate for KC and RDSM in text-independent system

<table>
<thead>
<tr>
<th>Classification</th>
<th>Feature Extraction</th>
<th>Recognition System</th>
<th>NH [%]</th>
<th>Recognition Rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>KC</td>
<td>FFAV</td>
<td>Text-independent</td>
<td>22.1</td>
<td>78.66</td>
</tr>
<tr>
<td>RDSM</td>
<td>FFAV</td>
<td>Text-independent</td>
<td>none</td>
<td>20.18</td>
</tr>
</tbody>
</table>

Both KC and RDSM in text-independent systems are investigated in Tab.3. In this experiment several Arabic vowels are involved where KC was clearly superior.

Table 4 tabulates the results of recognition rate for KC with the most popular feature extraction methods: linear prediction coefficient LPC (Adami & Barone, 2001; Tajima, Port, & Dalby, 1997) and Mel frequency cepstral coefficients MFCC (Mashao & Skosan, 2006; Sroka & Braida, 2005; Kanedera, Arai, Hermansky, & Pavel, 1999). These features were extracted from Arabic vowels. Taking in consideration the results tabulated in table 3 and table 3, The proposed method using FFAV via KC had considerable recognition rate as good as MFCC.

Table 4: Recognition rate for KC with LPC and MFCC

<table>
<thead>
<tr>
<th>Classification</th>
<th>Feature Extraction</th>
<th>Recognition System</th>
<th>NH [%]</th>
<th>Recognition Rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>KC</td>
<td>LPC</td>
<td>vowel-dependent</td>
<td>21.21</td>
<td>93.03</td>
</tr>
<tr>
<td>KC</td>
<td>LPC</td>
<td>vowel-independent</td>
<td>26.80</td>
<td>63</td>
</tr>
<tr>
<td>KC</td>
<td>MFCC</td>
<td>vowel-dependent</td>
<td>17.26</td>
<td>93.74</td>
</tr>
<tr>
<td>KC</td>
<td>MFCC</td>
<td>vowel-independent</td>
<td>20</td>
<td>78.48</td>
</tr>
</tbody>
</table>

In table 2, table 3 and table 4 the concept neighbourhood (NH) result is presented, which is referred to a result of two neighbour speakers. This is observed when $D$ of one impostor are the same or very nearly for two models. This is illustrated in Fig.5 of four speakers that are overlapped with nearby centroids. NH is dealt as limitation of the feature extraction method. From the silhouette plot (Fig.6), you can see that most points in the clusters with low silhouette values, less than 0.8, indicating that they are nearby to points from other clusters. Some points are well-separated from neighboring clusters, with high silhouette values, about 0.6 (see Matlab7 demo of k-means function). If NH is appeared for more than two models the trail is pointed as false.
6. Conclusion
In this Paper, DWT and PSD are proposed for speaker feature extraction. The introduced system worked based on two steps of features extraction the DWT and logarithmic PSD. As a result of DWT utilization, the system works with considerable capability of features tracking even with 0dB SNR. The recognition rate was taken via 400 vowels signals. Classification was accomplished by utilizing two presented methods: K-means clustering and Root Mean Square Difference Similarity Measure. Both Text-dependant system and Text-independent are investigated. More than two thousands trails were done to evaluate the system performance. The results obtained and investigated by the proposed method showed an excellent performance of 94% classification rate in the comparison with the LPC and MFCC.
K-mean Clustering and Arabic Vowels Formants based Speaker Identification System

References


[24] Malkin J., Li X., Bilmes J., A Graphical Model for Formant Tracking, SSLI Lab, Department. Of Electrical Engineering, University of Washington, Seattle, This work was supported by NSF grant ITS-0326382.


