Speech Recognition Approach From Midi File

Khadhim Mahdi Almusawi *Mohammed abedul Hadi Daykh* *Raid Luaibi Lafta*
Thi-Qar university

Abstract

In this paper we explore a new technique for speech recognition from recorded human speech on midi file .The microphone is used to record the statement in the computer .The speech features are extracted from input track using a proposed algorithm corresponding to the melody to be used in the recognition algorithm.

The aim is to build a reliable and efficient large scale system that collect thousands of sound as a Midi data, finally the system will returns list of the wave files most similar to the humming and recognize the input sound.

In this paper we record one statement in available language of 200 different person from our city (alnassiryha in south of Iraq) and store as a system data base. Promising result were obtained testing different Midi files as experiments.

Keywords—**speech recognition, speech feature, Melody, midi file**

اسلوب تمييز الاصـوات مـــن ملـــف المـــيديا

كاظم مهدي هاشم محمد عبد الهادي دايخ رائد لعيبي لفته

الملـخـص

في هذا البحث نعرض تقنية جديدة لتمييز كلام الانسان المسجل على ملف الميديا .تم استخدام المايكرفون لتسجيل جملة محددة على الحاسبة وتم استخلاص الملامح من المقاطع الداخلة باستخدام الخوارزمية المقترحة لغرض استخدامها في خوارزمية التمييز .ان الهدف هو بناء نظام واسع وكفوء الذي يجمع عدد كبير من الاصوات ويخزن على ملف من نوع ميديا ويعطينا قائمة من ملفات الاشخاص المشابهة للصوت الداخل لغرض تمييزه .في هذا البحث سجلنا عبارة واحدة باللغة المحلية ل ٢٠٠ شخص في مدينة (الناصرية –جنوب العراق)وتم خزنها لتكون قاعدة بيانات النظام .ثم التحقق من صحة النتائج باستخدام تجارب مختلفة .

الكلمات المفتاحية :.- تمييز الاصوات – ملامح الصوت نغمات الصوت – ملف الميديا

1. Introductions

Speech recognition is an interaction concept in which the identity of a person has to be revealed fast and orderly from a given human speech input using a large database of known midi files. In short, it tries to detect the pitches and melody in a human speech and compares these pitches with symbolic representations of the known melodies. [1]

The Goal of this work is to identify the person using his speech .The speech is recorded on The midi file .The literature concerned this topic is quite poor [2]. Several papers seek to extract the melodic line from audio file [3,4] .

The aim of our work is not to extract a monophonic line from a polyphonic score , but to identify the human speech using important characteristics of the input Midi file .A user's record speech is also transcribed to a melody, and melody features are extracted. The retrievals done by searching for similar occurrences of the user speech in the database. The important characteristics of midi file which are used in this paper (pitch ,velocity ,…..) .[5] figure below.

ONSET (BEATS)	DURATION (BEATS)	MIDI channel	MIDI PITCH	VELOCITY	ONSET (SEC)	DURATION (SEC)

The first column indicates the onset of the notes in beats (based on ticks per quarternote) and the second column the duration of the notes in these same beat-values. The third column denotes the MIDI channel (1-16), and the fourth the MIDI pitch, where middle C (C4) is 60. The fifth column is the velocity describing how fast the key of the note is pressed, in other words, how loud the note is played (0-127). The last two columns correspond to the first two (onset in beats, duration in beats) except that seconds are used instead of beats.[5]

In this paper we use two columns from midi file array instead off all columns that working in early papers.

2. File format:

2.1. Midi file format

The Standard MIDI File SMF is a file format specifically designed to store the data that a sequencer records and plays (whether that sequencer be software or hardware based). This format stores the standard MIDI messages (i.e., status bytes with appropriate data bytes) plus a time-stamp for each message (i.e., a series of bytes that represent how many clock pulses to wait before playing the event). The format allows saving information about tempo, pulses per

quarter note resolution (or resolution expressed in divisions per second, i.e. SMPTE setting), time and key signatures, and names of tracks and patterns. It can store multiple patterns and tracks so that any application can preserve these structures when loading the file. The format was designed to be generic so that any sequencer could read or write such a file without losing the most important data, and exible enough for a particular application to store its own proprietary, extra data in such a way that another application won't be confused when loading the file and can safely ignore this extra stuff that it doesn't need. Think of the MIDI file format as a musical version of an ASCII text file (except that the MIDI file contains binary data too), and the various sequencer programs as text editors all capable of reading that file. But, unlike ASCII, MIDI file format saves data in chunks (i.e., groups of bytes preceded by an ID and size) which can be parsed, loaded, skipped, etc. Therefore, it can be easily extended to include a program's proprietary info. For example, maybe a program wants to save a flag byte that indicates whether the user has turned on an audible metronome click. The program can put this flag byte into a MIDI file in such a way that another application can skip this byte without having to understand what that byte is for. In the future, the MIDI file format can also be extended to include new official chunks that all sequencer programs may elect to load and use. This can be done without making old data files obsolete (i.e., the format is designed to be extensible in a backwardly compatible way). In conclusion, any software that saves or loads MIDI data should use SMF format for its data files. Standard MIDI files provide a common file format used by most musical software and hardware devices to store song information including the title, track names, and most importantly what instruments to use and the sequence of musical events, such as notes and instrument control information needed to play back the song[12].

2.2.WAV file format:

The Wave file format is Windows' native file format for storing digital audio data. It has become one of the most widely supported digital audio file formats on the PC due to the popularity of Windows and the huge number of programs written for the platform. Almost every modern program that can open and/or save digital audio supports this file format, making it both extremely useful and a virtual requirement for software developers to understand. It supports a variety of bit resolutions sample rates, and channels of audio. This format is very popular upon IBM PC (clone) platforms, and is widely used in professional programs that process digital audio waveforms. It takes into account some peculiarities of the Intel CPU such as little Endean byte order. This format uses Microsoft's version of the

Electronic Arts Interchange File Format method for storing data in chunks.

The WAVE format is a subset of RIFF used for storing digital audio. Its form type is WAVE, and it requires two kinds of chunks: the fmt chunk, which describes the sample rate, sample width, etc., and the data chunk, which contains the actual samples. WAVE can also contain any other chunk type allowed by RIFF, including LIST chunks, which are used to contain optional kinds of data such as the copyright date, author's name, etc. Chunks can appear in any order. The WAVE file is thus very powerful, but also not trivial to parse. This subset basically consists of only two chunks, the fmt and data chunks, in that order, with the sample data in PCM format. The WAVE specification supports a number of different compression algorithms. The format tag entry in the fmt chunk indicates the type of compression used. A value of 1 indicates Pulse Code Modulation (PCM), which is a straight uncompressed encoding of the samples. Values other than 1 indicate some form of compression

3.MELODY REPRESENTATION

The fundamental attributes of music are the pitch sequence of notes, rhythm, tempo (slow/fast), dynamics (loud/soft), texture (timbre or voices) and lyrics (if any). It is in these dimensions that we typically distinguish one piece of music from another. Of these descriptors, melody and rhythm are the most distinctive. The melody of a piece of music is a sequence of notes with varying pitch and duration

4. PROPOSED METHODOLOGY

The major algorithmic modules are the extraction of a melody representation from the human speech. and the melodic similarity distance computation. Hence we divide the system in to two layers.

Layer 1:

In this layer the user speech statement .The statement is recorded and stored as a .WAV file. This file is then fed to a Speech Synthesis Software: PRAAT.PRAAT accepts the statement and gives the Pitch and melody of the user speech . Detailed Pitch Analysis along with efficient use of algorithms helps in extracting the notes from the user speech .

Layer 2:

In this layer a Database Schema consisting of different genres of MIDI files is stored. Correspondingly their Pitch and melody Analysis is done using PRAAT and the the extracted notes are stored in various files. Thus, the notes extracted from Layer 1 are evaluated against Layer 2 using an appropriate Pattern Matching Algorithm. The system returns the best matched wave file to the requested speech.

5. WAV to MIDI converted :

To create a MIDI a file for a speech recorded in WAV format must determine pitch, velocity and duration of each note being played and record these parameters into a sequence of MIDI events. The Midi created represents the basic melody and chords of recognized music. The difference between WAV and MIDI formats consists in representation of sound and music. WAV format is digital recording of any sound (including speech) and MIDI format is principally sequence of notes (or MIDI events). Here we have an Output File (.mid) from an Input File (.wav) that contains musical data, and a Tone File (.wav) that consists of monotone data. Illustrated in figur1. The block diagram of this process An advantage of such a structure is also the fact that the query is prepared on the client side of the system. In this case the query is very short. Besides, there is a possibility to evaluate its quality before sending to the server. The system provides for playback of the recognized melody notes in MIDI format. This allows the user to listen to a query and take a decision either to send it to the server or to sing it once again.

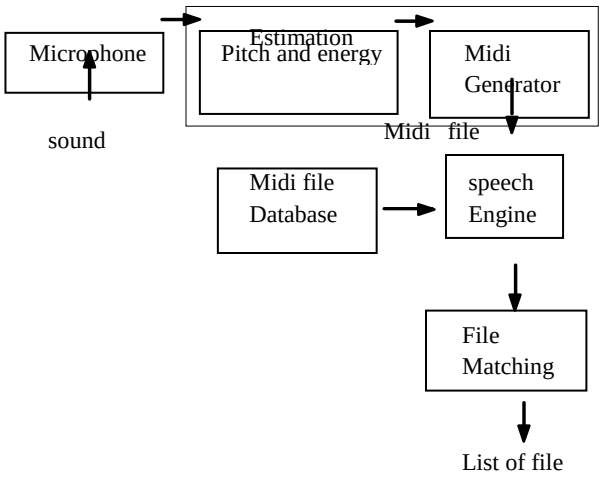

Figure 2 : wave to midi convertor

6.Algorithm for WAV to MIDI:

6.1.Part 1

1. Accept the user query and save it as a .wav file
2. Perform pitch energy estimation on the wav files.
3. From the respective pitch tier calculate the frequencies of various notes at different time intervals.
4. Using the formula

$$p = 69 + 12 \log2(f \div 440)$$

6.2.Part 2

Generation of midi file

1. Declare the data structures for mthd header and the mtrk header according to the midi file format.
2. For the data bytes to be written into the midi file, each event will consist of two events time and message of which the time also called as delta time is a variable length quantity.

3. Delta time can be written using the following pseudo code

pseudo code.

1. unsigned long result
2. unsigned long array [4]
3. count=0
4. if (result less than 128) goto step 9
5. array [count] =((result & 0x7F) | 0x80)
6. count++;
7. shift result by 7 bits right
8. goto step 3
9. array [count]=(result & 0x7F)
10. count++
11. for (i=count to 0)

display array in reverse order.

6.3.Matching process

The database is a set of record sounds indexed by the melody string. Extracting the melody representation from the original soundtrack is a difficult problem. There are many algorithms possible for pattern matching. We convert the wav file to a matrix representation of note events in a MIDI file and the database where our speech are stored is also converted into a matrix representation of note events in a MIDI file. These matrices are then compared using the edit distance algorithm[4]. The edit distance [20] of two matrix, nmat 1 and nmat 2, Calculates the similarity of two matrix in a particular representation.
Output is a value indicating distance between nmat1 and nmat2 under the given representation and metric. Output value is rescaled to (0, 1) if rescale is set to 1.

7. EXPERIMENTAL RESULTS

In order to compare our approach against other techniques for speech recognition , we collected a database of MIDI files for different person for one statement , and recordings of various people sound .We used the algorithm discussed in the papers to compare the human speech to each statement in our database and arrive at a distance between the human speech and each midi files.

Our preliminary results were based on two small databases of MIDI files, one containing 200 statement for various person , and one containing 18 different statement for various persons . Our results were quite promising.

This system could be used to convert a wave files database into a melodic lines database.
The Tables below shows the statistics of the comparing by using one statement for same person and different person the first experiment use all midi file information ,and second experiment use velocity and midi pitch only for same person ,the third and forth experiment use same information (midi pitch ,velocity ,….) for different person .

Tables blow shows the result of our experiments

Table(No.1)

The result of comparing using all matrix information in midi file for same person

		statement	time	Result Comparing by using all Colum	
First file	1010.mid	alnassirya in south of iraq	0.01	0.343	Same person
Second file	10.mid		0.01		
First file	1.mid	alnassirya in south of iraq	0.01	0.7102	Same person
Second file	10.mid		0.01		
First file	3.mid	alnassirya in south of iraq	0.03	0.3925	Same person
Second file	4.mid		0.02		
First file	6.mid	alnassirya in south of iraq	0.02	0.3004	Same person
Second file	7.mid		0.01		
First file	88.mid	alnassirya in south of iraq	0.02	0.9957	Same person
Second file	99.mid		0.03		
First file	Moh.mid	alnassirya in south of iraq	0.03	0.4311	Same person
Second file	Moh1.mid	alnassirya in south of iraq	0.03		
First file	Moh2.mid	alnassirya in south of iraq	0.03	0.6969	Same person
Second file	Moh1.mid		0.02		
First file	90.mid	alnassirya in south of iraq	0.02	0.9898	Same person
Second file	99.mid		0.02		
First file	1313.mid	alnassirya in south of iraq	0.02	0.8998	Same person
Second file	1111.mid		0.03		

Table(No.2)

Result of comparing using (midi pitch ,velocity)for midi file matrix for same person

		statement	time	Result Comparing by using velocity& midi pith	
First file	1010.mid	alnassirya in south of iraq	0.01	0.3303	Same person
Second file	10.mid		0.01		
First file	1.mid	alnassirya in south of iraq	0.01	0.7100	Same person
Second file	10.mid		0.01		
First file	3.mid	alnassirya in south of iraq	0.03	0.2820	Same person
Second file	4.mid		0.02		
First file	6.mid	alnassirya in south of iraq	0.02	0.3001	Same person
Second file	7.mid		0.01		
First file	88.mid	alnassirya in south of iraq	0.02	0.911	Same person
Second file	99.mid		0.03		
First file	Moh.mid	alnassirya in south of iraq	0.03	0.4311	Same person
Second file	Moh1.mid	alnassirya in south of iraq	0.03		
First file	Moh2.mid	alnassirya in south of iraq	0.03	0.6111	Same person
Second file	Moh1.mid		0.02		
First file	90.mid	alnassirya in south of iraq	0.02	0.9999	Same

Second file	99.mid		0.02		person
First file	1313.mid	alnassirya in south of iraq	0.02	0.8810	Same person
Second file	1111.mid		0.03		

Table(No.3)

The result of comparing using all matrix information in midi file for **different person**					
		statement	time	Result Comparing by using all Colum	
First file	Moh2.mid	alnassirya in south of iraq	0.01	1.6594	Deferent person
Second file	9.mid		0.01		
First file	9.mid	alnassirya in south of iraq	0.01	1.9953	Deferent person
Second file	99.mid		0.01		
First file	9.mid	alnassirya in south of iraq	0.03	1.9896	Deferent person
Second file	1010.mid		0.02		
First file	1010.mid	alnassirya in south of iraq	0.02	1.9954	Deferent person
Second file	1111.mid		0.01		
First file	1010.mid	alnassirya in south of iraq	0.02	1.8536	Deferent person
Second file	1313.mid		0.03		
First file	1313.mid	alnassirya in south of iraq	0.03	1.2222	Deferent person
Second file	88.mid	alnassirya in south of iraq	0.03		
First file	7.mid	alnassirya in south of iraq	0.03	1.9941	Deferent person
Second file	1.mid		0.02		
First file	1.mid	alnassirya in south of iraq	0.02	1.4834	Deferent person
Second file	2.mid		0.02		
First file	1.mid	alnassirya in south of iraq	0.02	1.9706	Deferent person
Second file	77.mid		0.03		

Table(No.4)

Result of comparing using (midi pitch ,velocity)for midi file matrix for different person

		statement	time	Result Comparing by velocity & midi pitch	
First file	1010.mid	alnassirya in south of iraq	0.01	1.6294	Deferent person
Second file	10.mid		0.01		
First file	1.mid	alnassirya in south of iraq	0.01	1.9153	Deferent person
Second file	10.mid		0.01		
First file	3.mid	alnassirya in south of iraq	0.03	1.7896	Deferent person
Second file	4.mid		0.02		
First file	6.mid	alnassirya in south of iraq	0.02	1.8954	Deferent person
Second file	7.mid		0.01		
First file	88.mid	alnassirya in south of iraq	0.02	1.8536	Deferent person
Second file	99.mid		0.03		
First file	Moh.mid	alnassirya in south of iraq	0.03	1.4222	Deferent person
Second file	Moh1.mid	alnassirya in south of iraq	0.03		
First file	Moh2.mid	alnassirya in south of iraq	0.03	1.8941	Deferent person
Second file	Moh1.mid		0.02		
First file	90.mid	alnassirya in south of iraq	0.02	1.3834	Deferent person
Second file	99.mid		0.02		
First file	1313.mid	alnassirya in south of iraq	0.02	1.9106	Deferent person
Second file	1111.mid		0.03		

Table(No.)

Result of comparing using (Velocity) for midi file matrix for same person					
		statement	time	Result Comparing by using velocity	
First file	1010.mid	alnassirya in south of iraq	0.01	0.3431	Same person
Second file	10.mid		0.01		
First file	1.mid	alnassirya in south of iraq	0.01	0.7002	Same person
Second file	10.mid		0.01		
First file	3.mid	alnassirya in south of iraq	0.03	0.3025	Same person
Second file	4.mid		0.02		
First file	6.mid	alnassirya in south of iraq	0.02	0.3000	Same person
Second file	7.mid		0.01		
First file	88.mid	alnassirya in south of iraq	0.02	0.9157	Same person
Second file	99.mid		0.03		
First file	Moh.mid	alnassirya in south of iraq	0.03	0.4211	Same person
Second file	Moh1.mid	alnassirya in south of iraq	0.03		
First file	Moh2.mid	alnassirya in south of iraq	0.03	0.6369	Same person
Second file	Moh1.mid		0.02		
First file	90.mid	alnassirya in south of iraq	0.02	0.9798	Same person
Second file	99.mid		0.02		
First file	1313.mid	alnassirya in south of iraq	0.02	0.8998	Same person
Second file	1111.mid		0.03		

Table(No.6)

Result of comparing using (Velocity) for midi file matrix for different person

		statement	time	Result Comparing by using velocity	
First file	1010.mid	alnassirya in south of iraq	0.01	1.6294	Deferent person
Second file	10.mid		0.01		
First file	1.mid	alnassirya in south of iraq	0.01	1.9153	Deferent person
Second file	10.mid		0.01		
First file	3.mid	alnassirya in south of iraq	0.03	1.7896	Deferent person
Second file	4.mid		0.02		
First file	6.mid	alnassirya in south of iraq	0.02	1.8954	Deferent person
Second file	7.mid		0.01		
First file	88.mid	alnassirya in south of iraq	0.02	1.8536	Deferent person
Second file	99.mid		0.03		

First file	Moh.mid	alnassirya in south of iraq	0.03	1.4222	Deferent person
Second file	Moh1.mid	alnassirya in south of iraq	0.03		
First file	Moh2.mid	alnassirya in south of iraq	0.03	1.8941	Deferent person
Second file	Moh1.mid		0.02		
First file	90.mid	alnassirya in south of iraq	0.02	1.3834	Deferent person
Second file	99.mid		0.02		
First file	1313.mid	alnassirya in south of iraq	0.02	1.9106	Deferent person
Second file	1111.mid		0.03		

Comparing Chart

This figure explain comparing result with different case (same person ,different person) for all experiment

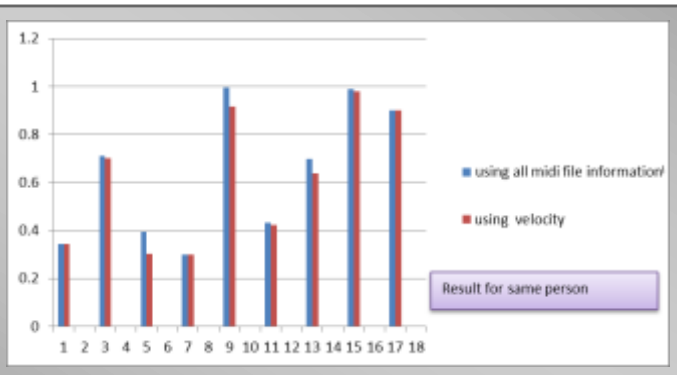

Figure 3

Figure 4

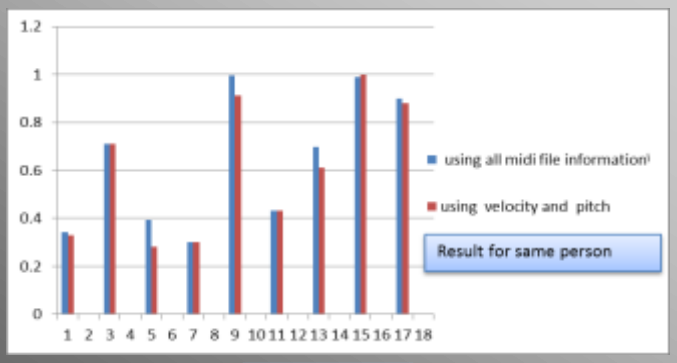

Figure 5

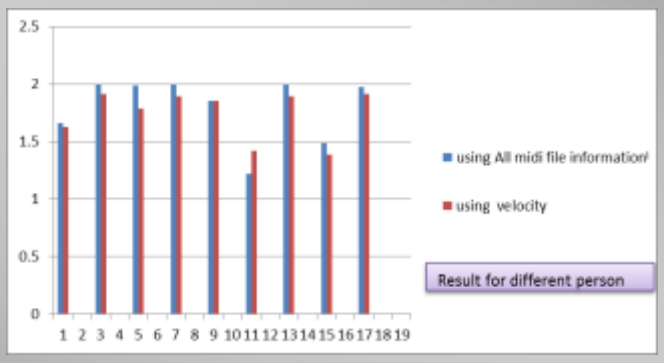

Figure 6

8.Conclusion

In this paper we built a identify human speech engine based on the characteristic of midi file The wave format can be transferred to midi format .we can identify any person using midi file The experiment yielded promising results using different data bases .the result will be compared with other system and they are more efficient .The experiment show that enough training data of each style is needed in order to successfully identify the human speech . We hope we can resolve some other problems of this area in the future research work.

9. REFERENCES

[1] M. A. Raju, B. Sundaram, and P. Rao, TANSEN: "AQuery-By- Humming based Music Retrieval System", In Proc. National Conference on Communications (NCC), 2003.

[2] Alan V.Oppenheim and Ronald W. Schafer, "Discrete-time Signal Processing", Prentice Hall, Fourth Edition,2005.

[3] Yashwant Kanetkar, Let Us C, BPB Publication, Fourth Edition.

[4] Brian W. Kernighan and Dennis Ritchie," The ANSI C programming language", Tata McGraw-Hill, Fourth Edition,2005.

[5] MIDI Toolbox version 1.0, http://www.jyu.fi/musica/miditoolbox

[6] R.J. McNab, L. A. Smith, D. Bainbridge and I.H. Witten.
The New Zealand Digital Library MELody inDEX (MELDEX). D-Lib Magazine, May 1997.

[7] A. Kornstädt. "Themefinder: A web-based melodic search tool." Computing in Musicology, v11, pp. 231-36, 1998.

[8] D. Huron et. Al. Themefinder (website). http://www.themefinder.org/

[9] R. Typke. Tuneserver(website). http://wwwipd.ira.uka.de/tuneserver/

[10] http://www.cs.cornell.edu/Info/Faculty/bsmith/"query-by- humming.html". (Accessed on 13 January 2008)

[11] http://www.praat.org. (Accessed on 14 January 2008)

[12] http://www.sloud.com/download/Sloud QBH SearchMusic. PDF. (Accessed on 05 February 2008)