Phonemes of Arabic

Authors: Mohamed Alshaari, Hussien ElHarati, and Veton Kepuska.

Terms:

- V – Vowel.
- C – Consonant.
- CVCVCV – ConsonantVowel-ConsonantVowel-ConsonentVowel words.
- Hamza letter – first Arabic letter used to start the pronunciation of consonant letter.
- (a, i, u) – symbols of short vowels.
- (a:, i:, u:) – symbols of short vowels long vowels.
- (C-a, C-i, C-u) – Consonant followed by one of the short vowels.
- 1, 2, … – letters or words numbers.
- A, B, C – represent amount of repeating of each person.

Expected use:

This data can be used for linguistics, Arabic speech recognition, and identification. This data can also be used to extract certain sound features such as formants and MFCC features.

Collections Procedure:

This data was collected from Native Arabic speaking adult males. These files were recorded and saved as wav audio file format. The files were recorded as wav files to retain and keep the accuracy of the measurements because wav files are raw and uncompressed. This data was recorded over a period of approximately 3 months by collecting volunteers from a cultural center with many Arab attendees.

Data Format Specific Details:

This data was collected from 19 Native Arabic speaking adult males. There are 3 short vowels in Arabic language (a, i, u), 3 long vowels (a:, i:, u:), and 28 consonants which the volunteers recorded each of those and repeated them 3 times. The volunteers also recorded 24 Arabic words with CVCVCV pattern and repeated each word 3 times.

The Isolated data directory contains data from 19 speakers, each speaker recorded the following: (24 CVCVCV words, 28 C-a, 28 C-i, 28 C-u, 28 Hamza letter-a-Consonant,3 a, 3 i, 3 u, 3 a:, 3 i: and 3 u:) each of those was repeated 3 times for a total of 8379 audio files.

P.S. (a, i, u) and (a:, i:, u:) were not recorded directly by the volunteers but extracted from the audio files.

Content:

- 1368 (CVCVCV)- Arabic words
- 4788 (CV) isolated Arabic Alphabet followed by short vowels
- 1539 (CVC) Arabic consonants beginning with the Hamza letter
- 171 (VV) Arabic long vowels
- 513 (V) Arabic short vowels

Specifications:

- File Type: wav
- Bit Depth: 32 bits
- Sample Rate: 48000
- Date Source: microphone speech
- Language: Arabic
- Application: speech recognition

Volunteer Nationalities:

AA	Libyan
AK	Libyan
AM	Egyptian
AO	Egyptian
AR	Lebanese
AS	Libyan
AT	Egyptian
BA	Saudi
HR	Syrian
IB	Moroccan
MA	Libyan
MB	Iraqi
ME	Egyptian
MK	Libyan
MZ	Libyan
OA	Saudi
OH	Lebanese
OT	Egyptian
TH	Syrian

File names explained:

In all the files names, the upper case letters correspond to the repetition of each sound, for example if the files has “A” in the name, that means that file has the first repetition of that sound, the same logic goes for “B” and “C”. In Arabic, there are three short vowels which are represented by “a”, “i”, and “u” in the file names. Also in Arabic, there are three long vowels represented by “aa”, “ii”, and “uu” in the files names.

For the specific folders:

1. For the folder titled “isolated_CVCVCV”, the numbers in the file names correspond to the numbers shown in the table below:

1	فَعَلَ	7	فَعِلَ	13	فَعُلَ	19	فُعِلَ
2	رَفَعَ	8	بَخِلَ	14	بَلُغَ	20	ذُكِرَ
3	ذَكَرَ	9	عَمِلَ	15	صَلُحَ	21	جُمِعَ
4	ذَهَبَ	10	حَفِظَ	16	سَهُلَ	22	خُلِقَ
5	شَرَحَ	11	سَمِعَ	17	كَبُرَ	23	كُتِبَ
6	كَتَبَ	12	فَرِحَ	18	كَرُمَ	24	حُشِرَ

2. For the folder titled: “isolated-ArabicAlphbet-with-a-i-u” the numbers correspond to the alphabet letters shown in the table below:

2	ب	**9**	ذ	**16**	ط	**23**	ل
3	ت	**10**	ر	**17**	ظ	**24**	م
4	ث	**11**	ز	**18**	ع	**25**	ن
5	ج	**12**	س	**19**	غ	**26**	ه
6	ح	**13**	ش	**20**	ف	**27**	و
7	خ	**14**	ص	**21**	ق	**28**	ي

3. For the folder titled: “isolated-ArabicConsonant-BeginWith-Hamza-and-a” the numbers follow the same logic from the table above.
4. For the folder titled: “isolated-LongVowels-aa-ii-uu” just has the recordings of the long vowels explained in the paragraph above.
5. For the folder titled: “isolated-Vowels-a-i-u” contains vowels extracted from the words with the numbers corresponding to the table below:

1	رَفَعَ	4	حَفِظَ	7	سَهُلَ
2	كَتَبَ	5	ذُكِرَ	8	كَبُرَ
3	فَعَلَ	6	جُمِعَ	9	بَلُغَ