تصنيف الاصوات

امل حجازي، براء الشيخة، سيلفي كبه، شربل قلومة

قسم الذكاء الصنعي، جامعة دمشق

ملخص العمل :

الهدف من المشروع تصنيف الاصوات (كسر الزجاج – اطلاق النار) حيث قمنا بتجهيز البيانات من خلال استخراج الميزات من الصوت باستخدام عدد من السمات الفيزيائية التي تؤثر بالصوت بعد ان قمنا بتقطيع الاصوات المدخلة باستخدام تحويل فورييه ومن ثم تخرين هذه الميزات وتطبيق عمليات pre-processing وذلك لفهم البيانات ومعالجتها ومن ثم تطبيق العديد من النماذج الخاصة بالتعلم التلقائي وتم الحصول على افضل النتائج باستخدام Boosting Classifier model .

الكلمات المفتاحية : ***sounds of breaking glass , shooting sounds***

I. مقدمة:

يستطيع الانسان ان يميزويصنف الاصوات التي يسمعها دون بذل اي جهد يذكر حيث يستطيع التمييز بين صوت رنين الجرس وصوت رنين الهاتف, ولكن يمكن ان يواجه بعض المشاكل في التعرف على الصوت عندما يكون ضعيف او يوجد ضجيج في نفس المكان .

يوجد عدة امور دفعت العلماء والباحثين لمعرفة الطريقة التي يميز بها الانسان الاصوات ,اولا لصبحوا قادرين على معرفة وتشخيص الامراض السمعية للانسان وثانيا لمحاولة بناء الة تستطيع ان تفعل ما يفعله الانسان فيمكن استخدامها في المجال الطبي او لاستخدامها لاغراض امنية كاصدار تنبيهات في حال وجود اصوات غريبة في المنازل او الاماكن العامة كاصوات الصراخ وتكسير الزجاج واطلاق النار وهذا ما سنتطرق اليه في دراستنا.

II. الدراسة المرجعية:

خواص الاشارة الصوتية:

يمكن تصنيف الاصوات الى عدة انواع حيث يعرف الصوت بخصائصه الزمنية temporal properties التي تتعلق بمدة الصوت و سعة الموجة و تغيرها ، والخصائص الطيفية spectral properties تتعلق بمكونات الاشارة الترددية وقوتها.

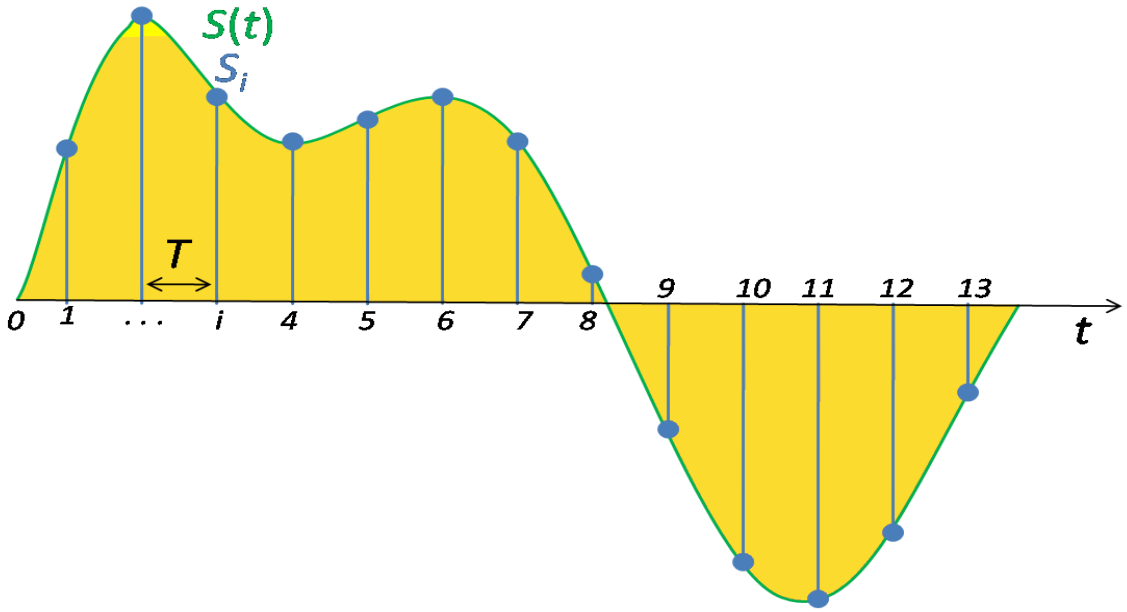
ولان خصائص الاشارة تتغير ببطء نسبيا ،فعند استخراج السمات من الاشارة لا يتم ذلك على كافة الاشارة الصوتية دفعة واحدة و انما بتقسيم الاشارة الى مقاطع صغيرة تدعى windows ، و تعرف هذه الطريقة في التحليل باسم short time analysis و تسمى السمات التي نحصل عليها بهذه الطريقة باسم short-time parameter [1] .

معالجة الاشارة الصوتية :

يعرف الصوت انه ظاهرة فيزيائية معقدة ناتجة عن الحركة ،فكل متحرك يصدر صوت ، و الحركة تعني الاصطدام مع الوسط المحيط (الهواء ،الماء،..)حيث يحدث تغير في جزيئات الوسط المحيط بمصدر الصوت و بالتالي تغير بالضغط و حدوث اهتزاز , و هذه الاهتزازات هي ما نسمعه بالحقيقة .

من اجل التعامل مع الصوت على الحاسوب يجب ربط الصوت بصيغة رياضية ما ،اي باستخدام توابع دخلها و خرجها ارقام لكي نستطيع التعامل معها ،لكن الخرج لايكون واضحا دائما و لايمكن التعبير عن الاصوات (المعقدة منها) بصيغة تابع مباشر مثل f(x)=y وانما نستطيع ايجاد قيم الخرج عند قيم دخل محددة مثل : f(0)=30

و نستطيع الحصول على تابع الصوت بعملية اخذ العينات sampling و التي تعني تسجيل قيم مواقع الحركة في كل لحظة من الزمن ، و بالتالي تابع الصوت هو تابع للزمن قيم خرجه تدعى samples ، و لسنا بحاجة للحصول على تابع مستمر و معرفة قيمة العينة عند كل لحظة زمنية لان هذا مستحيل بل يكفي الحصول على قيم العينات عند قيم محددة للزمن بخطوة ما (مثلا كل ميلي ثانية).



sampling

و تعرف العينة sample على أنها مقدار طويلة الصوت في لحظة زمنية ما .

يعبر محور X عن الزمن و y عن الamplitude أو الطويلة والتي تمثل مقدار ضغط الوسط في هذه اللحظة بحيث تكون القيم الموجبة تعني ضغط الوسط, والسالبة خلخلة، أما الصفرية فتمثل الصمت اي حدوث توازن الضغط .

بالاضافة للطويلة يوجد العديد من الخواص والمقادير التي تعبر عن الصوت مثل الloudness و هي واحدة من خواص الصوت الادراكية التي لا يعبر عنها ك soft/loud الصوت فحسب وإنما ايضاً ك high/low، وايضاً يوجد التردد الذي يعبر عن مقدار تكرر الصوت باعتبار الصوت موجة، أو مقدار تكرار الموجة في وحدة الزمن، وتردد الموجة هو معدل تَقَلُّب ضغط الهواء (إلى الداخل والخارج).[2]

خصائص الاشارة الصوتية :

1- Loudness :

ومن أجل تحديد باقي خواص الصوت لا يكفي وجود العينات لوحدها ، فهي لا تعطي معلومات كافية عن الصوت في بعض الأحيان ولابد من وجود طريقة لربط المقادير الفيزيائية للصوت كالتردد والطويلة مع المقادير الادراكية التي نستطيع إدراكها سماعياً لوحدنا كال loudness، حيث ان ال loudness بكافئ الطويلة كمفهوم فيزيائي ، لكننا في العالم الحقيقي عندما نصف الصوت لا نتحدث عن طويلته أو تردده وإنما نتحدث عن كثافة، والكثافة هي عبارة عن مقدار الطاقة في وحدة الوسط ، ونلاحظ أنه مقدار يمكن قياسه بصيغة معينة وبالتالى هو خاصة فيزيائية إدراكية معاً وهو صلة الوصل ما بين مفهوم الطويلة وال loudness ، ويتم إدراك ال loudness بشكل لوغاريتمي مقارنةً بالطويلة اي أن الاحساس بحدوث تغيير في ال loudness يحتاج إلى حدوث تغير كبير في الطويلة حقيقة .

وبشكل مكافئ يوجد ما يسمى pitch و هي خاصية إدراكية مقابلة لمفهوم التردد، فمثلاً ضرب وتر غيتار مرة ثم ضربه مرة أخرى بطريقة أقوى سندركه على سماع صوتين بنفس ال pitch ولكن ب loudnesses مختلفين ، وعلاقة الpitch بالتردد هي أيضاً علاقة لوغارتمية .

2- نوعية الصوت timbre :

وهو عبارة عن كميات الصوت التالية :

- Spectra: مجموعة أشكال الموجات الأبسط التي تشكل الموجة الكلية للصوت ونحصل عليها باستخدام إحدى تحويلات فورييه .
- Envelope: بدء واستمرار واختفاء جزء من الصوت .

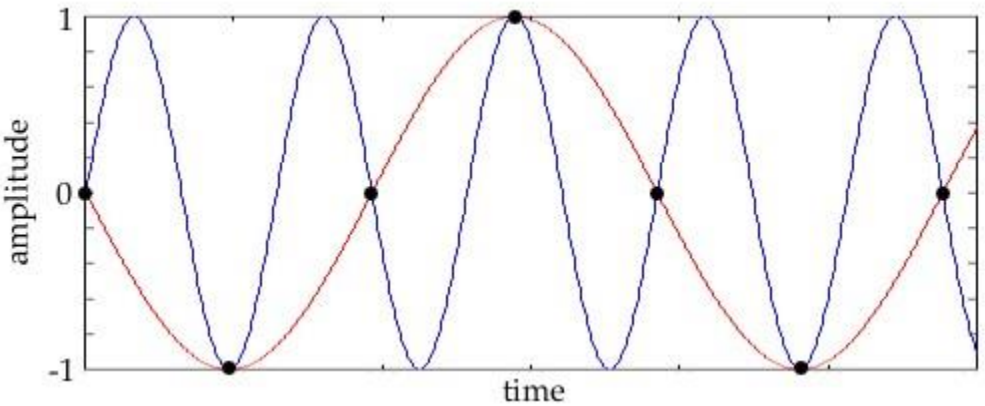
وتدعى أبسط موجة يتألف منها الصوت pure tone هي عبارة عن تابع جيبي وكل صوت هو مزيج من الموجات الجيبية البسيطة والمختلفة والتي نستطيع معرفتها بتحويل فورييه الذي ينتج مجموعة من الثنائيات هي طويلة وطور كل موجة جيبية مؤلف منها الصوت، وتدعى هذه الثنائيات ب spectral histogram .

وبالحصول عليه نحصل على ما يكافئ "صورة" الصوت التي يظهر فيها التردد كسطوع الادنى يقاس بال Pixels والسطوع الأعلى يكافئ التردد الأعلى (سرعة اكبر في تكرار الصوت).[2]

عند إجراء عملية أخذ العينات لتحويل الصوت إلى الصيغة الرقمية يتم بمعدل معين أخذ العينات sampling rate وهو يمثل عدد العينات التي يتم أخذها في وحدة الزمن وزيادته تعني زيادة الدقة، و بحسب نظرية Nyquist sampling فأن معدل أخذ العينات يجب أن يكون ضعفي أعلى تردد للإشارة الصوتية ، اي إذا كان اعلى تردد هو 8000Hz فإن المعدل يجب أن يساوي 16000Hz اي نحتاج أن نأخذ 16000 عينة في الثانية ، وكون أعلى تردد نستطيع سماعه هو 20KHz تقريباً فإن معدل أخذ العينات يجب أن يكون 40KHz حتى وإن كانت الترددات أقل من ذلك بكثير والسبب في ذلك هو وجود ترددات يمكن أن تكون موجودة ولكن لا نستطيع تميزها ، والمتعارف عليه حالياً أن معدل أخذ العينات يساوي إلى 44100Hz .

لو أخذنا معدل أقل من 44KHz مثلاً 2KHz2 فإننا لن نستطيع تمثيل تردد أعلى من 11KHz بناء على نظرية Nyquist وسنحصل على مجموعة من الترددات غير المرغوب بها تدعى aliasing's سببها أننا حصلنا على نفس العينات لموجة جيبية تتغير بالسرعة التي كنا سنحصل عليها لموجة جيبية ذات ترددات منخفضة ، أي تظهر الترددات الأعلى من 11Khz كترددات منخفضة غير متوقعة تمثل شكلاً مستعاراً للترددات الأصلية ونكون في حالة تدعى under sampling.

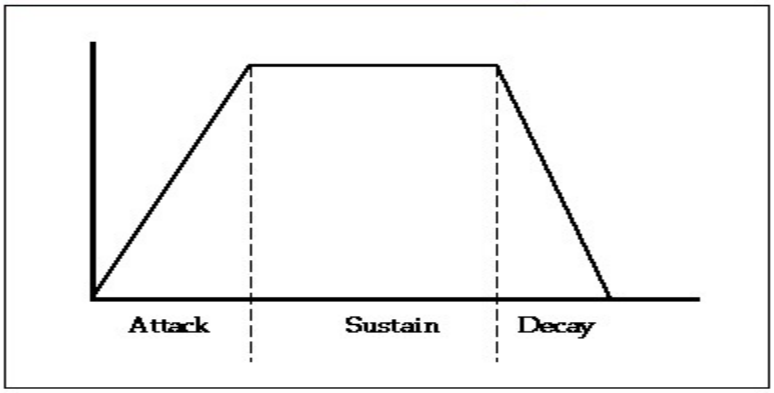
يمكن التخلص من هذه المشكلة عن طريق استخدام فلتر يدعى anti-aliasing filter الذي يسمح فقط بمرور الترددات الأقل من التردد الحدي .[2].



aliasing samples

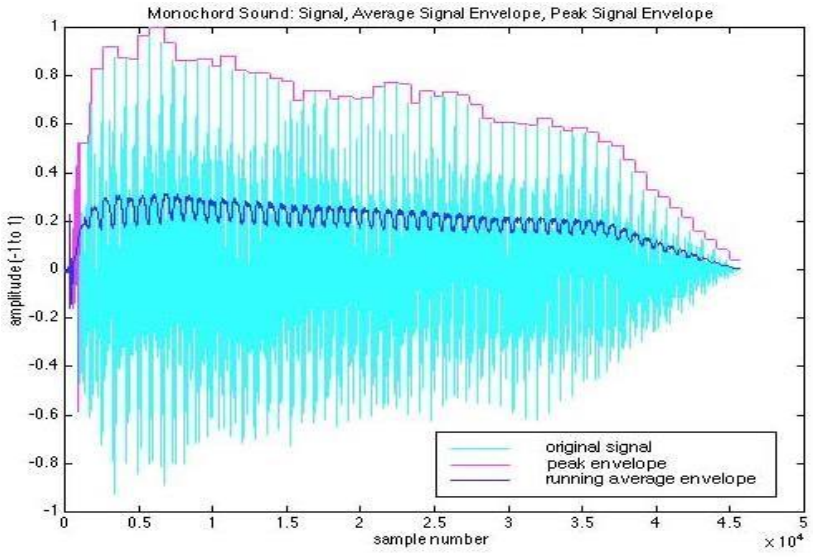
3- المجال الزمني time domain :

يعبر المجال الزمني للصوت عن طويلة الاشارة في لحظات زمنية معينة، ومن خلاله نستطيع معرفة لحظة بدء attack ، واستمرار steady-state ، وانتهاء decay الاشارة (كما موضح بالشكل)، وتسمى لحظة البداية و الانتهاء أحياناً ب transients لأنهما تحدَدان مرة واحدة على مستوى الصوت ولا تستمران وهما مهمان جداً لوصف تقلبات الصوت .



بدء وانتهاء إشارة

من المعلومات الهامة التي نستطيع الحصول عليها في المجال الزمني :



معلومات يمكن استخراجها زمني

Peaks: تمثل اللون البنفسجي وتصف الطويلات الاعلى للصوت وتشابه إلى حد ما القول أنها المجال الأعلى للطويلات .

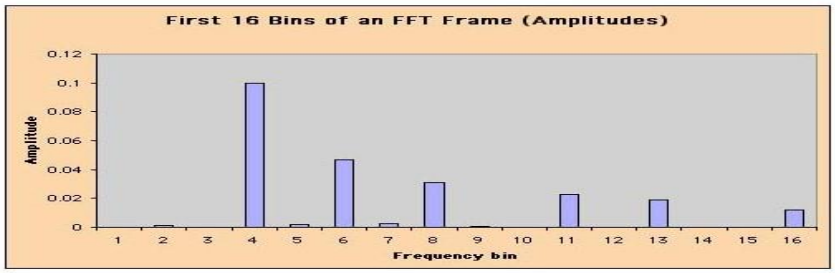
RMS:هي معدل طاقة الاشارة وتشابه القيم الممثلة باللون الأزرق ،هي معدل القيم المطلقة للعينات المسؤولة عن تنعيم الصوت .

Phasor: يشبه العجلة التي تدور عكس عقارب الساعة ومن خلاله نستطيع تحديد الزاوية التي تصنعها الاشارة بعد دورة كاملة في وحدة الزمن ، أما الphasor فهو 360 درجة في الثانية أو 2π راديان (السرعة الزاوية) ويمكن التعبير عنه من خلال تابع زمني : $\theta(t) = 2\pi$

و طول السهم في العجلة يمثل الطويلة amplitude الاشارة ، وبتغيره تتغير قيمة ال phasor. [2]

4- المجال الترددي frequency domain :

يتم تحويل من المجال الزمني إلى الترددي باستخدام تحويل فوريه الذي يحول العينات إلى fourier coefficients التي تشكل بمجموعها ما يسمى spectrum أو مخطط "صورة" الصوت المؤلف من الترددات أو bins ممثلة على محور ال x كما في الشكل أدناه ، أما محور Y يمثل قوة أو طاقة الاشارة .



Bin Frequency(hz)	Amplitude	Phase
0-22	0.00003	
22-	0.0015	
44-	0.0001	
66-	0.1	
88-	0.002	
110-	0.047	
132-	0.0023	
154-	0.031	
176-	0.0005	
198-	0.00026	
220-	0.023	
242-	0.0001	
264-	0.019	
286-	0.00013	
318-	0.00005	
340-	0.0123	

تحويل إلى مجال ترددي

حصلنا على bin بعرض 22 لأن حجم ال frame المأخوذة تحوي 24 عينة ، ومعدل أخذ العينات يساوي 44100Hz وبالتالي التردد الأعلى للإشارة هو 22050Hz وبتقسيمه على حجم ال frame نحصل على 22، ونلاحظ ايضاً أن قيم الطويلة واقعة ضمن المجال [1 – 0] ومجموعها جميعها يساوي الواحد [2].

5- تحويل فورييه Fourier transform :

حسب نظرية فورييه فإن الموجة المنتظمة يمكن وصفها كمجموع غير منتهٍ من الموجات الجيبية وترددات هذه الموجات هي أعداد صحيحة من تردد الموجة تدعى harmonic .

أي لو كان لدينا صوت تردده 440Hz فإنه وحسب نظرية فورييه مؤلف من مجموعات موجات ترددتها 440Hz, 880, 1320,.. أي 1, 2, 3,.. من مقدار التردد الاساسي .

باختصار يمكن القول أن سلسلة فورييه هي عبارة عن مجموعة المركبات البسيطة للموجة الأصلية و يكون لكل منها طويلة وتردد وطور مختلف، و في حال لم تكن الموجة دورية فإن ترددات الموجات البسيطة لن يكون من مضاعفات التردد الاساسي.

التعبير الرياضي :

ان أي تابع دوري يمكن كتابته بالشكل :

$$f(t) = A_0 + \sum_{n=1}^{\infty} A_n \sin(2\pi * nwt) + \sum_{n=1}^{\infty} B_n \cos(2\pi * nwt)$$

تدعى العوامل بعد إشارة المجموع ب Fourier coefficients ، ويدعى العامل الأول 0A ب DC offset ويمثل معدل التابع ، تمثل قيم A وB عند n صغرية ترددات صغيرة تدعى low-order Fourier coefficients ، أما عند القيم الكبيرة تمثل high-order Fourier coefficients.

باستخدام هذه النظرية أصبح بالhمكان تحليل analysis اي صوت إلى مركباته spectral components ، وبتحويل فورييه المعاكس يمكن تركيب الصوت الأصلي انطلاقاً من مجموعة مركباته الجيبية وتدعى هذه العملية synthesizing sound . [2]

6- الفلترة filtering :

من عمليات تعديل الصوت التي تسمح أو لا تسمح بمرور بعض ال coefficients ، ونقول عن المكونات التي يسمَح لها بالعبور بأنها pass band .

من اهم أنواع الفلاتر low-pass filter الذي يسمح فقط بمرور المكونات الصغيرة , وعكسه الhigh-pass filter [2].

7- Pre-processing audio :

هي مرحلة وسيطية ما بين معالجة الاشارة واستخراج السمات، ومن أهم العمليات التي يمكن إجراءها ضمنها :

- Normalize: تجرى على العينات بعدة طرق و أبسطها إيجاد القيمة المطلقة لأعلى عينة وتقسيمها على كل العينات، من أجل تجنب وجود clipping اي وجود قيم أعلى من القيم الذي تمثلها أعلى bit depth، مثلاً لو كانت ال bit depth تساوي 16 فإن وجود clipping يعني وجود عينات قيمها أكبر من 216 .
- down Sampling: خفض معدل أخذ العينات عند الحاجة لتقليل حجم البيانات وتسريع عملية استخراج السمات ،هي ضرورية في حال اختلاف معدل أخذ العينات بين ملفات الصوت التي يجب تصنيفها ، وعدم توحيد معدل أخذ العينات سيعطي نتائج غير صحيحة في التصنيف لأن أعلى تردد يمكن تمثيله سيختلف من ملف لآخر وبالتالي السمات غير متكافئة من الناحية النظرية ولا يمكن التصنيف على أساسها ، ولكن ستسبب هذه العملية في المقابل ضياع بعض المعلومات .
- Channels merging: جمع القيم المتقابلة في قنوات الملف الصوتي من أجل تقليل الحجم ويمكن الاستعاضة عنها باختيار قناة واحدة فقط عند الحاجة لتقليل حجم الداتا اللازمة للمعالجة .
- Rectified: عملية حذف القيم السالبة من الاشارة ولها نوعان:

Full-wave rectification -> xfull[n] = |x[n]|.

half-wave rectification -> xhalf[n] = x[n] if x[n] >=0, xhalf[n] =0 otherwise.[5]

استخراج السمات من الملفات الصوتية :

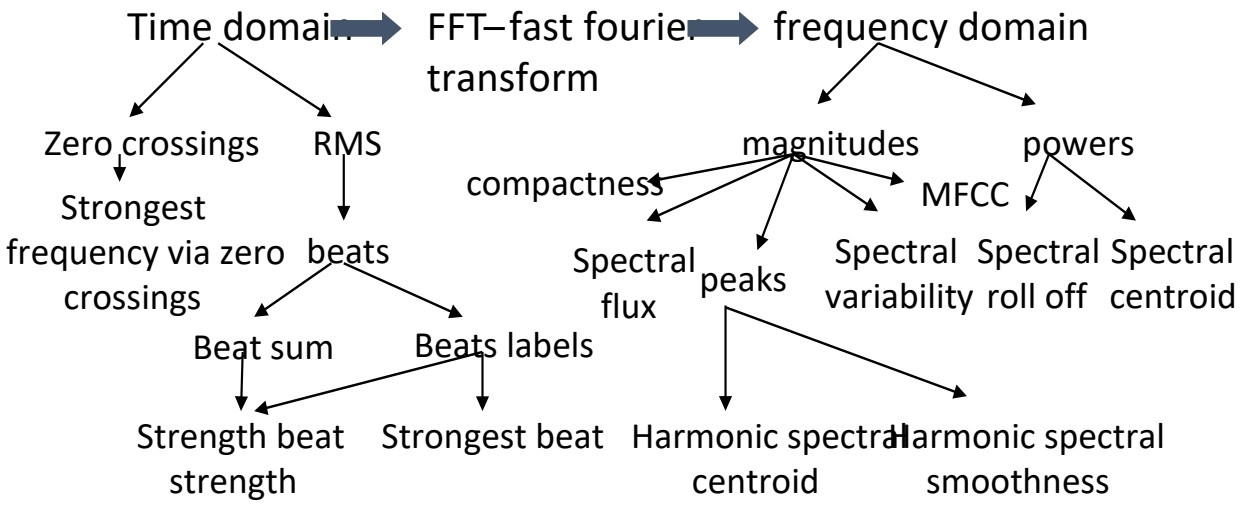
من العمليات الهامة جداً في معالجة الاشارات الصوتية وتتضمن الحصول أو حساب قيم رقمية ما من الصوت يمكن استخدامها لتمييز هذا المقطع الصوتي عن آخر بناءً على فضاء السمات ، إذ أن أشعة السمات التي يتبع كل منها لصوت معين والوجود بنفس المنطقة من هذا الفضاء حتماً ستكون منتمية لنفس الصف أو النمط .

تقسم السمات إلى سمات فيزيائية physical features وسمات إدراكية perceptual features ، الأولى منها تعني الخصائص التي تحسب بعلاقات رياضية ما ولا علاقة لها بآلية إدراك الانسان للصوت ، أما الثانية فهي خصائص يدركها الانسان ويستطيع تميزها أثناء سماعه للصوت .

وتصنف السمات أيضاً ك static وdynamic ، أما الأولى تمثل خصائص تم استخراجها في لحظة زمنية معينة أي يتم الحصول عليها من short-time analysis أما الثانية فيتم استخراجها من أجل تحسين عملية التصنيف .

ويتم الحصول على السمات بدايةً بتقسيم الاشارة إلى windows مدة كل منها 10-20ms ، وكونها متداخلة ستكون مدة الframes التي ستؤخذ منها أقل وغالباً يتم الحصول عليها بمعدل 100 إطار في الثانية ، وتسمى المسافة الفاصلة بين كل إطار وإطار مجاور (overlapping window) ب hop size وتمون بين 5ms إلى 20ms ، وتتم معالجة كل إطار منها واستخراج السمات الساكنة منه ويتم جمع هذه السمات بشعاع فيكون لكل إطار شعاع السمات الخاص به ، ثم يتم جمع الاطارات مع بعضها البعض للحصول على texture window نحصل منها على السمات الديناميكية ، وغالباً تكون مدة ال texture window من 500ms إلى 1s .[1]

تم تلخيص السمات بالمخطط التالي :



السمات الفيزيائية :

وتدعى أيضا ب low-level parameters ويتم الحصول عليها مباشرة من short-time spectral أي بعد القيام بتحويل فورييه كالتالي ، وبفرض r هي رقم الframe و xr[n] هي العينة رقم n من الاطار(r) عندئذٍ :

$$\boldsymbol{xr[n]; n = 1..N \rightarrow \quad Xr[k]atfreqf[k] ; k = 1..N} :$$

وهذه السمات هي [3] :

A. Zero-crossings rate:

تقيس عدد مرات تغيري الموجة لاشارتها ضمن الframe كالتالي :

$$ZCR_r = \frac{1}{2}\sum_{1}^{N} |sign(x_r(n)) - sign(x_{r-1}(n)|$$

$$\boldsymbol{sign(x) = \begin{cases} 1 ; x \geq 0 \\ -1 ; x < 0 \end{cases}}$$

B. Short-time energy:

هي مجموع مربعات العينات في ال frame الواحدة ، ويستخدم كمؤشر لقوة الاشارة و يعرف :

$$\boldsymbol{E_r = \frac{1}{N}\sum_{1}^{N} |x_r(n)|^2}$$

C. Band-level energy:

يتم الحصول عليها من مجموع مربعات قيم power spectrum الذي يمثل مربعات قيم المخطط الطيفي الحاوي على السعات :

$$\boldsymbol{E_r = \frac{1}{N}\sum_{k=1}^{N/2} (X_r[k].w[k])^2}$$

حيث w[k] تابع من أجل وزن القيم بقيم غير مساوية للصفر .

D. Spectral centroid:

يمثل مركز ثقل المخطط الطيفي للإشارة ، ويقابل السطوع في الصور brightness of sound ويُعرف كالتالي :

$$C_r = \frac{\sum_{k=1}^{N/2} f[k]|X_r[k]|}{\sum_{k=1}^{N/2} |X_r[k]|}$$

E. Spectral roll-off:

وهو تردد القيمة التي تحقق مايلي :

$$\sum_{k=1}^{K} |X_r[k]| \leq 0.85 \sum_{k=1}^{N/2} |X_r[k]|$$

$$R_r = f[k]$$

أي يمثل تردد القيم الطيفية المتراكمة الأولى والتي تمثل 85% من الاشارة .

F. Spectral flux:

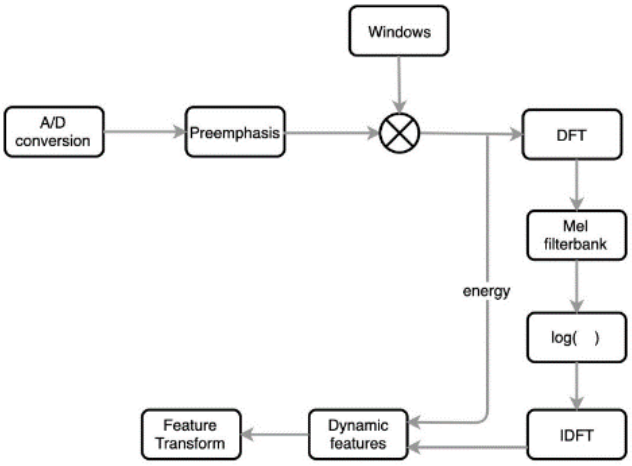
تمثل مربع الفرق بين الframe والframe السابق له ، ويدل على مقدار التغير في الاشارة و يعرف كالتالي :

$$F_r = \sum_{k=1}^{\frac{N}{2}} (|X_r[k]| - |X_{r-1}[k]|)^2$$

C. MFCC:

وهي من اكثر المعاملات استخداما لاستخراج الميزات بسبب شبهها بالادراك السمعي البشري حيث تعد افضل الطرق للتعرف التلقائي على الكلام البشري فوجد ان اول 13 ميزة يتم استخراجهم من خلالها هي الاكثر اهمية.

ويوضح المخطط التالي الخطوات الخاصة ب MFCC:



السمات الإدراكية :

يتم استخراجها في حال كانت مصادر الصوت غير جيدة كبديل عن السمات السابقة ، وهذه السمات هي [3] :

A. Loudness:

تعتمد على كثافة الصوت كما يمكن ان تعتمد ايضاً على مدى وطيف الصوت .

B. Pitch:

تمثل معدل تكرار الصوت ويوجد العديد من الخوارزميات لحسابه أهمها خوارزمية تعرف باسم pitch detection algorithm

II. الطريقة **Methodology**:

- جمع وتقسيم الداتا سيت (data set):

الداتا سيت في مشروعنا عبارة عن مجموعة من الاصوات (كسر الزجاج , اطلاق النار,اصوات أخرى) تم تجميعها وتحويل صيغة الملف الصوتي الى صيغة المقبولة من قبل نظامنا (.wav)

- الخوارزمية العامة والعمليات التي سوف يتم إجراءها على الملف الصوتي [**4**] :

1. تحويل صيغة الملف الصوتي إلى صيغة مقبولة من قبل نظامنا .
2. تقسيم الملف إلى frames .
3. معالجة الملف الصوت pre-processing باستخدام Normalize
4. استخراج السمات :

تم اخذ العينات من خلال تقسيم الاشارة الصوتية الى تحويل فورييه و استخدام عدة سمات وهي (zero-crossing - short-term-energy – short-term-entropy- spectral-centroid - spectral-entropy - spectral-flux - spectral-rolloff - 13 feature MFCC)
حيث تم شرح معادلتها في القسم السابق ثم اخذ المتوسط والانحراف المعياري لجميع feature لكل المقاطع وتخرين النتائج في ملف csv.

5. تقسيم data set الى train و test حيث تم اخذ 40% من الداتا من اجل test و 60% منها train

IV. النماذج الاحصائية:

سوف يتم عرض بعض المعلومات المأخوذة من الداتا وعددها بالإضافة الى بعض النماذج الإحصائية منها :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 540 entries, 0 to 539
Data columns (total 43 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   MFCCs1-deviation                540 non-null    float64
 1   MFCCs1-median                   540 non-null    float64
 2   MFCCs10-deviation               540 non-null    float64
 3   MFCCs10-median                  540 non-null    float64
 4   MFCCs11-deviation               540 non-null    float64
 5   MFCCs11-median                  540 non-null    float64
 6   MFCCs12-deviation               540 non-null    float64
 7   MFCCs12-median                  540 non-null    float64
 8   MFCCs13-deviation               540 non-null    float64
 9   MFCCs13-median                  540 non-null    float64
 10  MFCCs2-deviation                540 non-null    float64
 11  MFCCs2-median                   540 non-null    float64
 12  MFCCs3-deviation                540 non-null    float64
 13  MFCCs3-median                   540 non-null    float64
 14  MFCCs4-deviation                540 non-null    float64
 15  MFCCs4-median                   540 non-null    float64
 16  MFCCs5-deviation                540 non-null    float64
 17  MFCCs5-median                   540 non-null    float64
 18  MFCCs6-deviation                540 non-null    float64
 19  MFCCs6-median                   540 non-null    float64
 20  MFCCs7-deviation                540 non-null    float64
 21  MFCCs7-median                   540 non-null    float64
 22  MFCCs8-deviation                540 non-null    float64
 23  MFCCs8-median                   540 non-null    float64
 24  MFCCs9-deviation                540 non-null    float64
 25  MFCCs9-median                   540 non-null    float64
 26  category                        540 non-null    int64
```

```
 26  category                        540 non-null    int64
 27  short-term-energy-deviation     540 non-null    float64
 28  short-term-energy-median        540 non-null    float64
 29  short-term-entropy-deviation    540 non-null    float64
 30  short-term-entropy-median       540 non-null    float64
 31  spectral-centroid1-deviation    540 non-null    float64
 32  spectral-centroid1-median       540 non-null    float64
 33  spectral-centroid2-deviation    540 non-null    float64
 34  spectral-centroid2-median       540 non-null    float64
 35  spectral-entropy-deviation      540 non-null    float64
 36  spectral-entropy-median         540 non-null    float64
 37  spectral-flux-deviation         540 non-null    float64
 38  spectral-flux-median            540 non-null    float64
 39  spectral-rolloff-deviation      540 non-null    float64
 40  spectral-rolloff-median         540 non-null    float64
 41  zero-crossing-deviation         540 non-null    float64
 42  zero-crossing-median            540 non-null    float64
dtypes: float64(42), int64(1)
memory usage: 181.5 KB
```

حيث كما نرى وذكرنا سابقا بوجود حوالي 43 ميزة وعدم وجود قيم Non بالأضافة الى عرض نوع كل ميزة موجودة .

	MFCCs1-deviation	MFCCs1-median	MFCCs10-deviation	MFCCs10-median	MFCCs11-deviation	MFCCs11-median	MFCCs12-deviation	MFCCs12-median	MFCCs13-deviation	MFCCs13-median	...	spectral-centroid2-deviation	spectral-centroid2-median
0	0.077713	0.271746	3.324902e-02	0.155706	0.059562	2.432404e+00	3.225215e-02	0.104898	0.049655	5.717350e-01	...	0.069458	3.204824
1	0.065226	0.194548	8.395482e-03	0.170885	0.032308	1.144038e+00	-8.303745e-02	0.035652	0.037379	-6.303222e-01	...	0.041040	2.561634
2	0.089770	0.269012	-1.105905e-01	1.057548	0.062033	-3.138759e-01	1.396222e-01	0.192089	0.065782	-2.340937e-01	...	0.124805	3.158970
3	0.003935	0.086331	1.460000e-15	0.556781	0.005801	9.370000e-15	2.190000e-15	0.031285	0.000039	3.890000e-15	...	0.008972	3.321059
4	0.057182	0.185445	-1.750000e-14	4.785027	0.035301	7.230000e-14	-3.700000e-14	0.189126	0.090568	2.940000e-14	...	0.097428	3.256161

5 rows × 43 columns

وهنا تم عرض dataframe حيث تحوي على 43 ميزة

IV. عرض المعطيات:

قمنا أولا بقراءة الداتا والعمل على فهمها بشكل جيد لكي نتمكن من الوصول الى افضل طريقة عرض بيانات واخذ اهم الميزات الموجودة بداخلها تم عمل visualization و collaboration للمعطيات للحصول على افضل عرض للبيانات

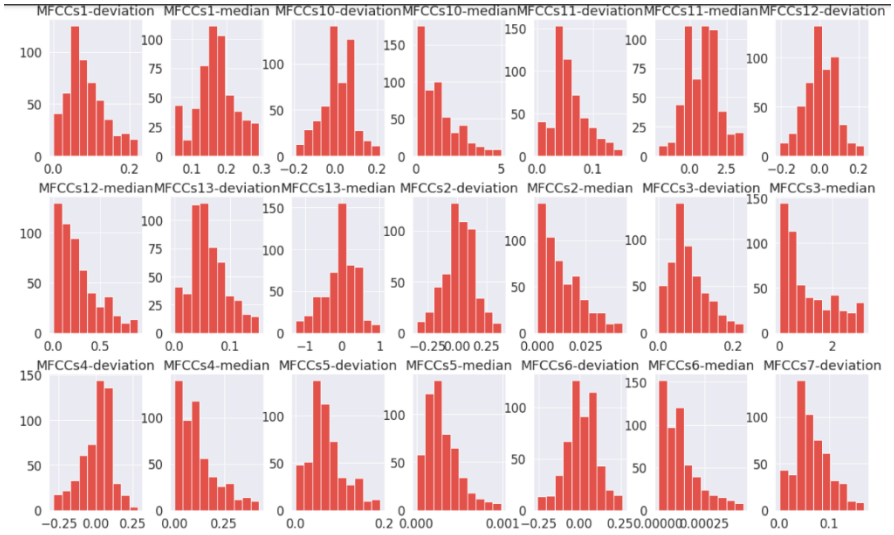
حيث تم التاكد من :

1- عدم وجود حقول فارغة او ان تحوي على قيم nan

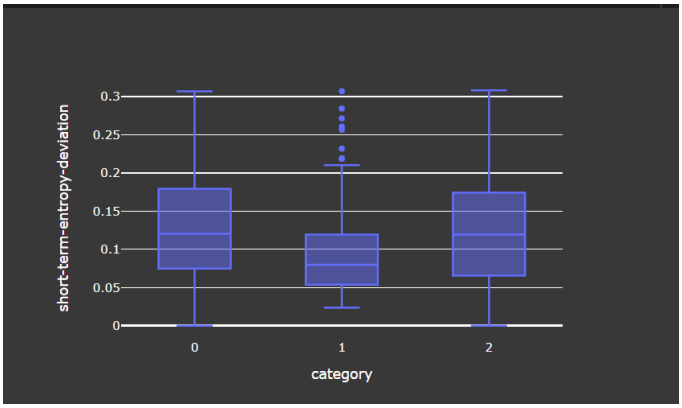
	data_type	missing_val	missing_val_ratio
MFCCs1-deviation	float64	0	0
MFCCs1-median	float64	0	0
MFCCs10-deviation	float64	0	0
MFCCs10-median	float64	0	0
MFCCs11-deviation	float64	0	0
MFCCs11-median	float64	0	0
MFCCs12-deviation	float64	0	0
MFCCs12-median	float64	0	0
MFCCs13-deviation	float64	0	0
MFCCs13-median	float64	0	0
MFCCs2-deviation	float64	0	0
MFCCs2-median	float64	0	0
MFCCs3-deviation	float64	0	0
MFCCs3-median	float64	0	0
MFCCs4-deviation	float64	0	0
MFCCs4-median	float64	0	0

2- جميع البيانات عددية
3- معرفة اذا كان هناك علاقة بين الميزات

سوف نقوم بعرض بعض الأمثلة على تمثيل بعض الميزات الموجودة ضمن الداتا المعطاة حيث تم رسم العلاقة بين الميزات وcategory :

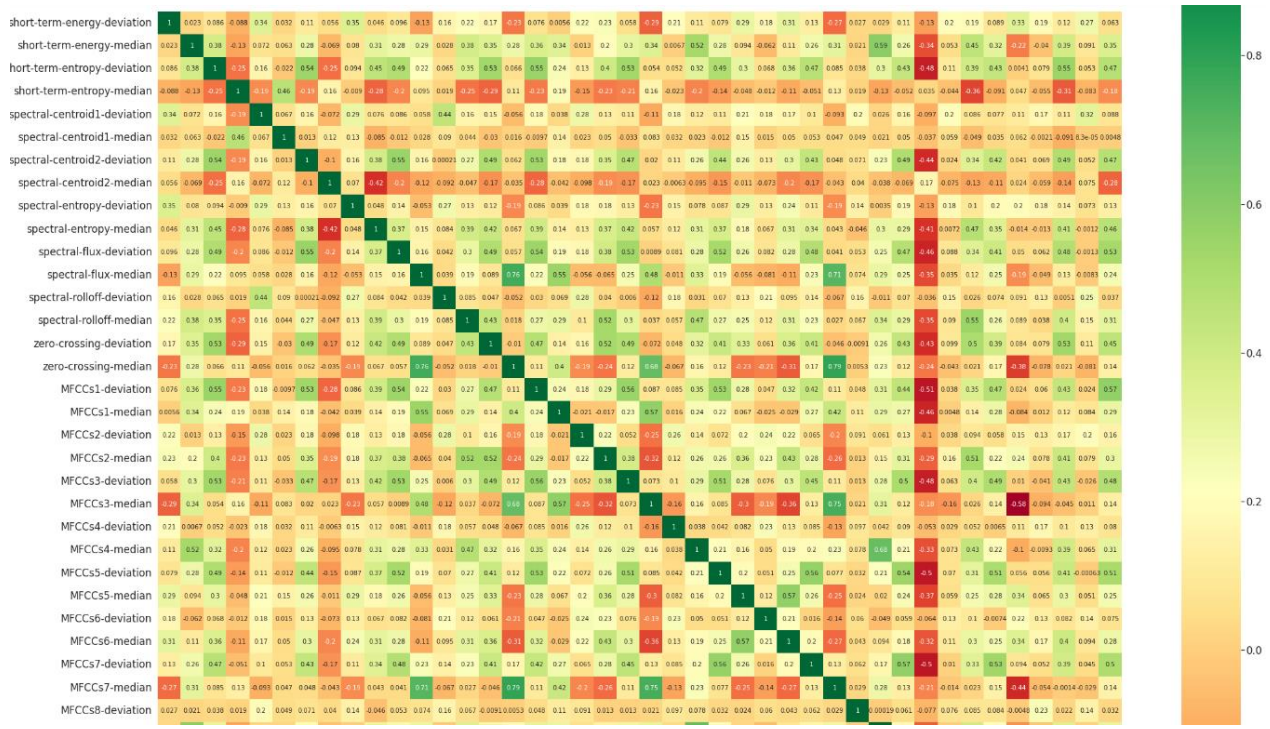


مخططات box plot : تستخدم هذه مخططات لتصور التوزيع الكامل لمتغير كمي واحد على مستويات متعددة لفئتين أو أكثر. حيث يتم من خلال اكتشاف فيما اذا كان هناك قيم شاذة (outliers) وذلك لازالتها من الداتا سوف نقوم بعرض بعض الأمثلة:



تم توزيع مجال قيم الميزة short-term-entropy على category الموجودة وملاحظة وجود قيم شاذة

مخططات correlation: تستخدم لايجاد الترابط بين الميزات وتشير القيم الاكبر من 0.6 الى وجود ترابط بين الميزتين المعروضتين:



V. التجارب:

- التعلم التلقائي المشرف عليه (supervised learning):
- تم الاعتماد في التقييم والمقارنة على معامل f1_score وذلك لان الداتا غير متوازنة

- تم استخدام العديد من النماذج وذلك لتصنيف البيانات الى (اكتشاف الصوت – طلق رصاص – صوت اخر) سوف نقوم بذكرها وتلخيص نتائجها في الجدول التالي:

- Logistic Regression:

أحد المصنفات الخطية البسيطة وله مجموعة hyperparameter من المهم اختيارها بشكل صحيح

1- Penalty: تساهم في ال Regularization وبالتالي تخفض من تأثير المتغيرات الغير مهمة.

2- Max_iter : عدد مرات التكرار يتم زيادها للحصول على نتائج افضل .

نتائج تدريب هذا النوذج على افضل قيم خاصة به حيث تم استخراجها باستخدام Grid Search :

```
Confusion Matrix:
[[46  2  7]
 [ 0 74  3]
 [ 6 13 65]]
Classification Report
              precision    recall  f1-score   support

           0       0.88      0.84      0.86        55
           1       0.83      0.96      0.89        77
           2       0.87      0.77      0.82        84

    accuracy                           0.86       216
   macro avg       0.86      0.86      0.86       216
weighted avg       0.86      0.86      0.85       216
```

- Support Vector Machine Classifier:

له مجموعة من hyperparameter :

1- C: يعبر عن مدى سماحية الخطا بتصنيف النقاط ضمن الهامش
2- Kernel :يساهم في القيام بحسابات على أبعاد مختلفة عن ابعاد المسالة العادية مما يساعد في تخفيف التعقيد الذي قد يصل اليه

تم اختيار انواع من kernel المختلفة ومقارنة النتائج بينهم :

1- Linear kernel :

```
Confusion Matrix:
[[47  1  7]
 [ 1 67  9]
 [ 8 13 63]]
Classification Report
              precision    recall  f1-score   support

           0       0.84      0.85      0.85        55
           1       0.83      0.87      0.85        77
           2       0.80      0.75      0.77        84

    accuracy                           0.82       216
   macro avg       0.82      0.82      0.82       216
weighted avg       0.82      0.82      0.82       216
```

2- Rbf kernel :

```
Confusion Matrix:
[[28  0 27]
 [ 0 70  7]
 [ 0  7 77]]
Classification Report
              precision    recall  f1-score   support

           0       1.00      0.51      0.67        55
           1       0.91      0.91      0.91        77
           2       0.69      0.92      0.79        84

    accuracy                           0.81       216
   macro avg       0.87      0.78      0.79       216
weighted avg       0.85      0.81      0.80       216
```

3- Polynomial kernel :

```
Confusion Matrix:
[[46  1  8]
 [ 1 70  6]
 [ 3 15 66]]
Classification Report
              precision    recall  f1-score   support

           0       0.92      0.84      0.88        55
           1       0.81      0.91      0.86        77
           2       0.82      0.79      0.80        84

    accuracy                           0.84       216
   macro avg       0.85      0.84      0.85       216
weighted avg       0.85      0.84      0.84       216
```

4- Sigmoid kernel :

```
Confusion Matrix:
[[ 0 24 31]
 [ 0  3 74]
 [ 0 29 55]]
Classification Report
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        55
           1       0.05      0.04      0.05        77
           2       0.34      0.65      0.45        84

    accuracy                           0.27       216
   macro avg       0.13      0.23      0.17       216
weighted avg       0.15      0.27      0.19       216
```

من النتائج نلاحظ ان Polynomial kernel قام بتصنيف القيم افضل من الانواع الاخرى حيث اخطأ بتصنيف قيم الكلاس الاول ب 9 امثلة فقط من اصل 55 أما قيم الكلاس الثاني فقد اخطا ب 7 امثلة من اصل 77 مثال اما الكلاس الثالث فقط أخطا ب 18 مثال من اصل 81 مثال .

- Random Forest Classifier:

تعتبر افضل من decision tree كونها تضيف فكرة ال bagging لها وهو ما يخفف من مشكلتها بالتاثر بالبيانات التي تمثلها وكذلك تقوم بالتخفيف من overfit التي تقوم بها الشجرة الوحيدة .

Hyperparameter المهمة الخاصة بها :

1- N_estimaters : عدد الاشجار والاكبر هو الافضل
2- Max_depth: عمق الشجرة الوحيدة

فكانت نتائج افضل نموذج :

```
array([[47,  1,  7],
       [ 1, 69,  7],
       [ 1, 14, 69]])
```

جدول النتائج للمقارنة بينهم

	MODEL	PARAMETER TUNING	ACCURACY-TRAIN	ACCURACY-TEST	PRECISION	RECALL	F1-SCORE
0	Logistic Regression with feature_selection	C	0.9722	0.8564	0.86	0.86	0.86
1	Logistic Regression with feature_selection	C	0.8703	0.7962	0.80	0.80	0.80
2	KNeighbors Classifier	n_neighbors	0.5216	0.5138	0.51	0.51	0.51
3	Decision Tree	min_samples_leaf	0.898	0.7638	0.76	0.76	0.76
4	SVC-linear	C	0.8672	0.8194	0.82	0.82	0.82
5	SVC- poly	Degree -C-coef0	0.8765	0.8425	0.84	0.84	0.84
6	SVC- rbf	Gamma-C	0.9938	0.8101	0.81	0.81	0.81
7	SVC- sigmoid	C	0.2407	0.2685	0.27	0.27	0.27
8	Random Forest without feature_selection	n_estimators oob_score max_leaf_nodes min_samples_split	0.9475	0.9398	0.91	0.91	0.91
9	Random Forest with feature_selection	n_estimators oob_score max_leaf_nodes min_samples_split	0.956	0.861	0.80	0.80	0.80
10	Boosting without feature_selection	n_estimators learning_rate max_features max_depth	0.9938	0.9861	0.99	0.99	0.99
11	Boosting with feature_selection	n_estimators learning_rate max_features max_depth	1.0	0.8657	0.87	0.87	0.87

- نتائج افضل model :

تم الحصول على افضل النتائج باستخدام Boosting Classifier ولكن من دون feature selection حيث لاحظنا ان قد تاثرت النتائج في حال اخذ كم عينة لان العينات مستقلة وتعطي ميزات مهمة لاستخراج الصوت فيجب استخدامها جميعا وهو نموذج له نفس بنية decision tree وrandom forest بالتالي لها نفس Hyperparameter حيث confusion matrix له :

```
array([[52,  0,  3],
       [ 0, 77,  0],
       [ 0,  0, 84]])
```

نلاحظ انه تم تصنيف classes حيث اخطأ بتصنيف class الاول بثلاثة امثلة فقط و نتائجه ممتازة جدا

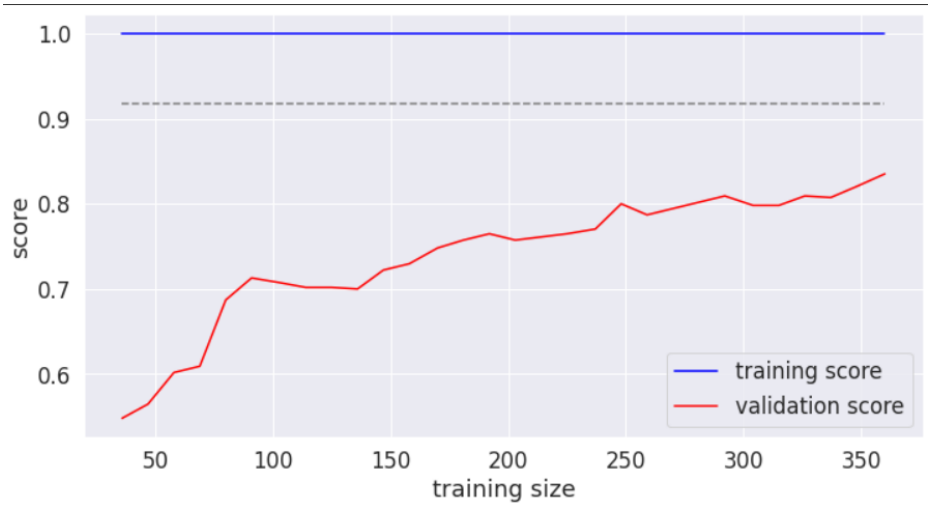
اما Classification Report :

```
Classification Report
              precision    recall  f1-score   support

           0       1.00      0.95      0.97        55
           1       1.00      1.00      1.00        77
           2       0.97      1.00      0.98        84

    accuracy                           0.99       216
   macro avg       0.99      0.98      0.98       216
weighted avg       0.99      0.99      0.99       216
```

اما learning curve الخاص به :



وبالملاحظة نجد ان curve يتجه نحو fit بالتالي فان زيادة الداتا في حالتنا يساعد في تحسين الاداء اكثر .

المراجع :

[1] audio signal processing Chapter in Speech, Audio, Image and Biomedical. P. Rao.

[2] Music and Computers, A Theoretical and Historical Approach, Columbia Phil Burk, Larry Polansky, Douglas Repetto, Mary Roberts.

[3] Music Information Retrieval, University of Victoria, 2014. Tzanetakis G.

[4] Automatic Music Classification with jMIR. Cory McKay.

[5] Music Genre Classification. Michael Haggblade, Yang Hong, Kenny Kao