دراسة أداء خوارزميات التعرف على الكلام وتحسين فعالية الخوارزميات ذات دقة التعرف الأدنى

سرار حمود*

(تاريخ الإيداع 6 / 6 / 2021. قُبِل للنشر في 25/ 8 / 2021)

□ ملخّص □

يتضمن هذا البحث مقارنة أداء الخوارزميات الشهيرة باستخلاص السمات الصوتية في التعرف على الكلام وهي MFCC,BFCC, LPCC,PLP وتحسين فعالية الخوارزميات ذات دقة التعرف الأدنى من خلال الدمج بينها للوصول إلى الدقة والزمن الأمثل للتعرف، تم في هذا البحث إدراج أربعة أنظمة للتعرف على الكلام، تختلف عن بعضها البعض بالطرق المستخدمة في مرحلة استخراج السمات، حيث استخدم النظام الأول خوارزمية MFCC والنظام الثاني خوارزمية BFCC والنظام الثالث خوارزمية LPCC والنظام الرابع خوارزمية PLP، وتم استخدام HMM كمصنف. تمت مقارنة أداء هذه الخوارزميات في التعرف وكان التفكير في تحسين الخوارزميات الأقل دقة في التعرف من خلال الدمج بينها، حيث تم تطبيق خوارزمية الدمج على النظامين الأقل دقة في التعرف ومقارنة أداء النظامين المفردين الأقل دقة مع أداء النظام المجمع. كما تمت دراسة تأثير زيادة عدد السمات على نتائج عملية التعرف.

الكلمات المفتاحية: التعرف على الكلام– استخراج السمات – خوارزمية MFCC –خوارزمية BFCC – خوارزمية LPCC– خوارزميةPLP – مصنف HMM.

* مشرف أعمال– قسم هندسة الحاسبات والتحكم الآلي– كلية الهندسة الميكانيكية و الكهربائية– جامعة تشرين– اللاذقية– سورية.

Study of the Performance of Speech Recognition Algorithms and Improving the Efficiency of Less-recognition Algorithms

Serar Hammoud*

(Received 6 / 6 / 2021. Accepted 25 / 8 / 2021)

□ ABSTRACT □

In this research I compare performance of famous algorithms of feature extraction in speech recognition MFCC,BFCC,LPCC and PLP, and improving the performance of less recognition algorithms by combination of them to get better time and recognition, despite of the diversity of these methods. Four systems were used for speech recognition, they differ in the used methods during features extraction, the first system used MFCC algorithm, the second system used LPCC algorithm, the third system used BFCC algorithm,and the fourth system used PLP algorithm .The systems used HMM as classifier.I compare results to improvement performance of less recognition algorithms by the combination algorithm. The results were satisfactory in accordance to the improvement level, the effect of increasing the features number on recognition results were studied .

Keywords: speech recognition, features extraction, BFCC algorithm, LPCC algorithm, PLP algorithm, BFCC algorithm , HMM classifier.

* **Work Supervisor, Department Computer and Automatic Control Engineering, Faculty of Mechanical and Electrical Engineering, Tishreen University, Lattakia, Syria.**

مقدمة:

بدأ اهتمام خبراء الحاسب والباحثين في مجال التعرف على الكلام منذ أكثر من أربعة عقود، وذلك لكي يصل الإنسان إلى مرحلة تجعله قادراً على التخاطب مع الكمبيوتر وإعطائه الأوامر والتعليمات صوتياً وبدون الحاجة إلى الكتابة وغيرها من الطرق، وذلك توفيراً للوقت والجهد. وفي السنوات الأخيرة تطورت نظم التعرف على الكلام تطوراً واضحاً وكبيراً، بحيث أصبحت برامج التعرف الآلي تدخل في أغلب مجالات الحياة، ووصلت إلى دقة مرضية نوعاً ما.

إنّ الكلام عبارة عن سياق من الرموز الصوتية التي تخضع لنظام معين متفق عليه بين أفراد الثقافة الواحدة، فمن خلال عملية الكلام يستطيع الفرد التعبير عن آرائه وأفكاره ومشاعره ونقل المعلومات إلى من حوله من البشر. ولقد حاول الإنسان التواصل مع الحاسب عن طريق الكلام محاكياً بذلك أسهل وأكثر وسائل التواصل الطبيعية بين البشر منذ عشرات القرون، وذلك باستخدام تقنيات التعرف على الكلام وهو تحويل الكلمات المنطوقة إلى دخل يمكن قراءته من قبل الحاسب، وعملية الكلام عملية معقدة تشترك فيها عدة أجهزة عضوية وتمر بمراحل مختلفة وعلى الرغم من أن هذه الأجهزة تقوم بعملية خاصة بها في عملية نطق الكلام إلا أنه لا يمكن لأي جهاز من هذه الأجهزة أن يعمل بشكل منفصل ومستقل عن الأجهزة الأخرى بل لابد لها أن تشترك مع بعضها البعض في إتمام عملية الكلام والتواصل[1].

أجريت العديد من البحوث والدراسات في مجال التعرف على الكلام خلال العقود الماضية، وأحد الأبحاث تناول دراسة التعرف على كف اليد باستخدام الدمج بين خوارزميتي PTA,LDA وحقق الدمج نتائج أفضل بالنسبة لدقة التعرف. وتم في بحث آخر دراسة تحسين أداء الخوارزميات في تمييز الكلام باستخدام الدمج بين خوارزميتي MFCC,PLP وحقق الدمج أيضاً نتائج أفضل. وتناول بحث آخر تحسين نتائج التعرف على الصوت باستخدام تكامل أنظمة مختلفة حيث تم دمج خوارزميتي MFCC,PLP والتي حققت أفضل نتيجة تعرف، وفي هذا البحث تمت مقارنة أداء الخوارزميات LPCC, BFCC, PLP, MFCC، حيث تبين أن خوارزميتي BFCC,LPCC أعطت نتائج أقل دقة بالتعرف لذلك تم استخدام خوارزمية الدمج بينهما بهدف تحسين النتائج، وقد أظهرت النتائج تحسناً ملحوظاً بزيادة دقة التعرف للنظام المجمع. وبالتالي فقد أثبتت الدراسات السابقة أن عملية الدمج تؤدي إلى نتائج أفضل في عملية التعرف ولكن لا يمكن الجزم بوجود خوارزمية أفضل من غيرها لأن دقة التعرف ترتبط بطبيعة اللغة وجنس المتحدث وعمره والكثير من العوامل الأخرى، كما أن المصنف يلعب دور كبير في معدل التعرف وزمنه، وحتى بارامترات المصنف تؤثر بشكل كبير على عملية التعرف لذلك لا يمكن اعتماد نتائج بحث ما كنتائج شاملة وذلك بسبب التنوع الكبير جداً في ظروف عملية التعرف ومحدداتها.

أهمية البحث وأهدافه:

لقد تم تطوير العديد من الأساليب من أجل إنجاز التعرف على الكلام، والتي تختلف عن بعضها بالطرق المستخدمة في استخراج السمات. حيث أن طرق استخراج السمات المعتمدة للتعرف على الكلام تختلف عن بعضها البعض بالسمات التي تعتمدها، ولا يمكن الجزم أي منها هي الأفضل، لذلك كان التفكير باجراء مقارنة لفعالية الخوارزميات الشهيرة باستخلاص السمات في التعرف على الكلام، وبالتالي فإن الهدف هو دراسة فعالية خوارزميات استخلاص السمات في التعرف على الكلام، وتحسين نتائج الخوارزميات الأقل دقة في التعرف من خلال الدمج، حيث تم تطبيق خوارزمية الدمج على النظامين الأقل دقة في التعرف ومقارنة نتائج النظامين المفردين الأقل دقة مع نتائج النظام المجمع.[2][3]

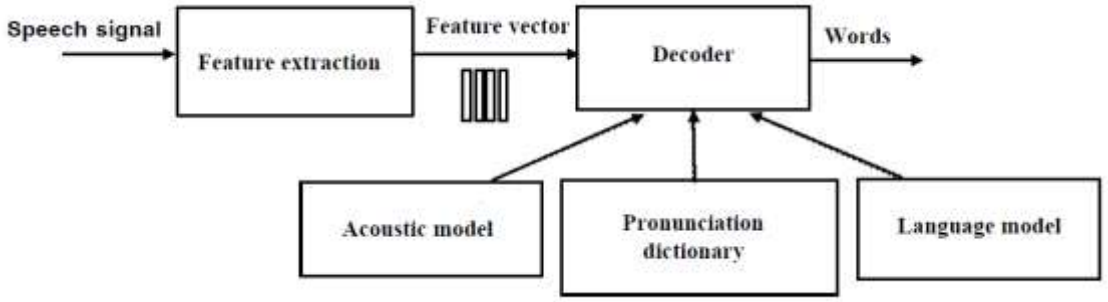
طرائق البحث ومواده:

يتكون نظام التعرف على الكلام من ثلاث مراحل أساسية مبينة في الشكل(1) الذي يمثل هيكلية نظام التعرف على الكلام المعتمد على نموذج ماركوف [1]:

1- مرحلة استخراج السمات: يتم في هذه المرحلة تحويل إشارة الكلام إلى تسلسل من أشعة السمة التي تمثل المعلومات في الكلام المنطوق. يتم في هذه المرحلة تقليل أبعاد إشارة الكلام الأصلية وإعداد هذه الإشارة في صيغة المتطلبات الأساسية لمرحلة التصنيف التالية. من الخصائص الهامة لمرحلة استخراج السمات هو كبت المعلومات التي ليس لها أهمية من أجل تصنيف صحيح مثل المعلومات حول المتحدث أو المعلومات التي تخص قناة النقل.

2- مرحلة التصنيف: وظيفة المصنف هو إيجاد الرسم التخطيطي بين تسلسل أشعة السمة وبين عنصر الكلام المتعرف عملياً، والمصنف المستخدم في البحث هو HMM من أشهر المصنفات المستخدمة في مجال التعرف على الكلام.

3- نماذج اللغة: وظيفة نماذج اللغة هو اختيار الفرضيات التي هي على الأرجح التسلسل الصحيح لعناصر الكلام للغة معطاة .



الشكل (1) هيكلية نظام التعرف على الكلام المعتمد على HMM [1].

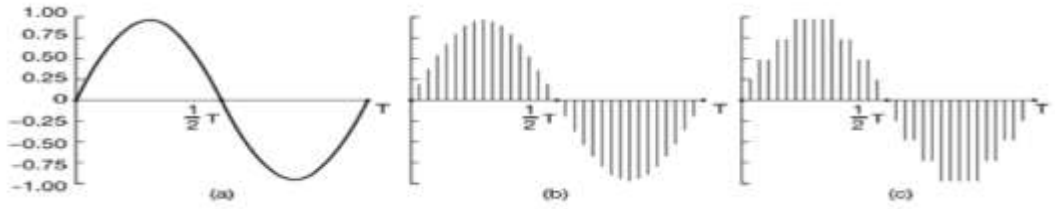
1- مراحل نظم التعرف على الكلام:

أ- تحويل الإشارة الصوتية إلى إشارة رقمية [4]:

إن الإشارة الصوتية تحول إلى إشارة تمثيلية كهربائية عبر المايكروفون، والتي يجري لها عملية الرقمنة (digitization) عن طريق المحول ADC من تمثيلي إلى رقمي، ويقوم هذا المحول بخطوتين أساسيتين هما: التقطيع (sampling)، والتكميم (Quantization) .

التقطيع: هي عملية أخذ قيم الإشارة في نقاط معينة من الزمن، يحدث التقطيع عادة بفواصل زمنية متساوية، تكرار أخذ العينات يسمى معدل التقطيع ويقدر بالهرتز. ومن أجل استعادة الإشارة الأصلية من الإشارة المقطعة يجب اختيار معدل التقطيع بحيث يكون أكبر من ضعفي أكبر تردد للإشارة الأصلية وهذا ما يقوم به المحول ADC والشكل (2)يوضح إشارة جيبية مقطعة.

التكميم: المقصود بعملية التكميم هو رقمنة قيم العينات المأخوذة في المرحلة السابقة بشكل تقريبي (أي تقريبها إلى أقرب مستوى كم). حيث أن الإشارة التي تم تقطيعها في المرحلة السابقة فقد تم إيجاد مطال الإشارة الرقمية من أجل عدد محدود من العينات وتم تحديد قيم الإشارة لبقية العينات بالتقريب للقيم السابقة. وبعد عملية التكميم يتم ترميز مطالات العينات المأخوذة من الإشارة الأساسية بأرقام ثنائية وذلك باستخدام المرمز الذي يخزن القيم الثنائية إما ب 8 أو 16 خانة. كما في الشكل(2).



الشكل (2) (a) إشارة جيبية، (b) إشارة جيبية مقطعة، (c)إشارة جيبية مقطعة ومكممة [4].

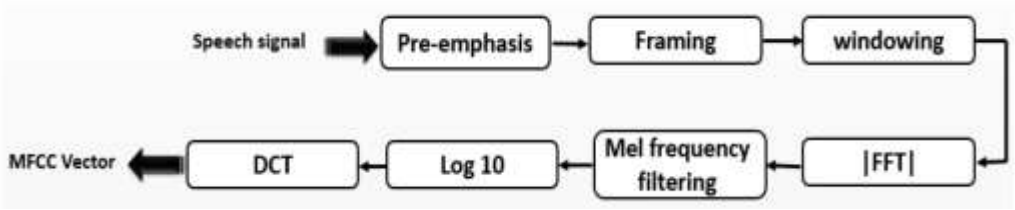
ب- استخراج السمات [12] ،[11]:

تنوعت الطرق المستخدمة في استخراج السمات، ولقد تم في هذا البحث استخدام كل من BFCC, MFCC, PLP, LPCC في مرحلة استخراج السمات مع المصنف HMM .

خوارزمية MFCC (Mel Frequency Cepstral Coeffiecents) [5] ،[6] ، [7] ،[8]

إن خوارزمية MFCC من الطرق السائدة المستخدمة في استخراج السمات وذلك بسبب حساسية مرشحاتها لخواص إشارة الصوت البشرية، تستخدم معاملات MFCC بشكلٍ كبيرٍ في التعرف على الكلام وما زالت متقدمة في هذا المجال. إن الأصوات التي تولد من قبل الإنسان يتم ترشيحها حسب شكل المسلك الصوتي، فإذا تمكنا من تحديد شكل المسلك الصوتي بدقة فإنه يمكن تحديد الصوت الذي يتم إنتاجه.

تعتمد MFCC على التغيرات المعروفة في عرض حزمة الترددات للأذن البشرية، حيث أن لمرشحاتها تباعداً خطياً على الترددات المنخفضة ولوغاريتمياً على الترددات المرتفعة وهي تستخدم من أجل التقاط الصفات الرئيسية للكلام، حيث تمتلك MFCC تباعداً خطياً على الترددات الأقل من 1000 هرتز وتباعداً لوغاريتمياً على تردد أكبر من 1000 هرتز. خطوات عمل الخوارزمية مبينة في الشكل(3).



الشكل (3) المخطط الصندوقي لعمل الخوارزمية MFCC [7] .

يتم تطبيق عملية pre-emphasis وهي عملياً مرشح تردد عالي يقوم بتمرير الإشارات الكهربائية ذات الترددات الواقعة فوق تردد معلوم يسمى تردد القطع f_c للمرشح حيث $f_c=1/(2\pi RC)$، وذلك من أجل تعويض جزء التردد العالي الذي تم فقده أثناء آلية إنتاج الكلام، حيث يتم إعادة تقييم كل قيمة في إشارة الكلام باستخدام الصيغة التالية:

$$s2(n) = s(n) - a*s(n-1) \quad (1)$$

حيث s(n): إشارة الكلام.

s2(n): إشارة الخرج بعد عملية pre-emphasis .

a: ثابت تتراوح قيمته بين 0.9 و 0.1 .

إشارة الكلام هي إشارة متغيرة باستمرار لذلك من أجل تبسيط الدراسة نعتبر أنه من أجل نطاق زمني قصير فإن إشارة الصوت لا تتغير كثيراً، لهذا السبب يتم تقطيع الإشارة إلى عدد من الإطارات، زمن كل إطار من 20 إلى 40 ميلي ثانية

مع وجود تداخل اختياري يساوي إلى نصف أو ثلث حجم الإطار وذلك من أجل تسهيل الانتقال من إطار إلى آخر، سيخضع كل إطار لعملية النوفذة باستخدام نافذة هامينغ وذلك من أجل القضاء على الانقطاعات عند الحواف. تعطى نافذة هامينغ بالعلاقة التالية :

$$w(n)=0.54-0.46\cos(2\pi\frac{n}{N}),\quad 0\le n\le N \qquad (2)$$

حيث w(n) هو مطال العينة الجديد و n هو ترتيبها في النافذة و N هو الطول الكمي للنافذة.

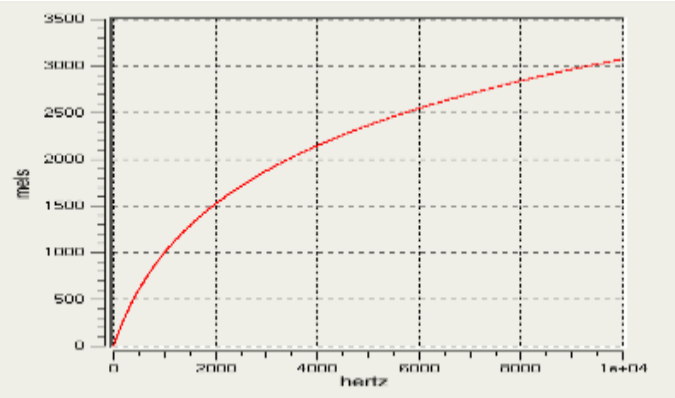
بعد عملية النوفذة يتم تطبيق تحويل فورييه السريع FFT من أجل كل إطار وذلك من أجل استخراج مركبات التردد للإشارة في مجال الزمن.

ترشيح الإشارة وفقاً لتردد ميل(Mel frequency filtering)

تعمل MFCC على ترشيح طيف الإشارة الصوتية عن طريق مجموعة من المرشحات المثلثية المتباعدة بانتظام وفقاً لمقياس ميل الترددي (الشكل(4)) الذي يعبر عن علاقة تربط التردد الملاحظ لنغمة صافية (m) بترددها المقاس الأصلي (f) ويعطى بالعلاقة التالية :

$$m=2595\log(\frac{f}{700}+1) \qquad (3)$$

يستطيع الإنسان أن يميز التغيرات الصغيرة في الـ pitch (وهي الارتفاع أو الانخفاض النسبي لنغمة كما تدركها الأذن، والتي تعتمد على عدد الاهتزازات التي تنتجها الحبال الصوتية في الثانية) وبشكل أفضل عند الترددات الصغيرة من الترددات الكبيرة، بالتالي فإن تضمين هذا المقياس يجعل السمات المستخلصة أقرب إلى سمع الإنسان.



الشكل (4) نطاق ميلmel scale [7] .

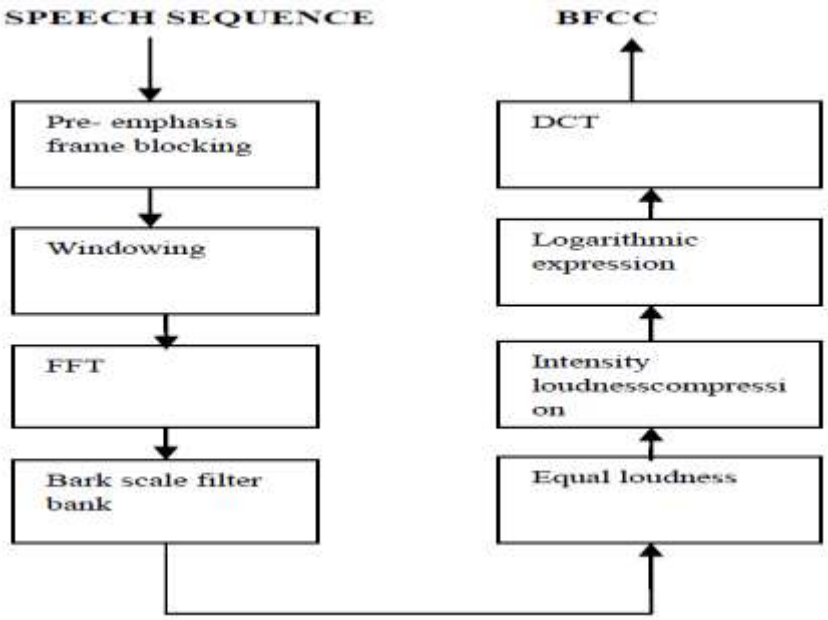
يتم بعد ذلك حساب اللوغاريتم لطيف مجال ميل (Mel scale spectrum)، ومن ثم يستخدم تحويل جيب التمام المتقطع DCT لإعادة تحويل طيف مجال ميل اللوغاريتمي إلى مجال الزمن. وبنتيجة هذا التحويل يتم الحصول على شعاع MFCC .

خوارزمية BFCC (Bark Frequency Cepstral Coeffiecents) [11]،[8]،[7]،[6]:

إن خوارزمية BFCC إحدى الطرق المستخدمة في استخراج السمات، الشكل(5) يظهر المخطط الصندوقي لخوارزمية BFCC. يتم تطبيق عملية pre-emphasis، التأطير، والنوفذة نفسها المطبقة فيMFCC. بعد عملية النوفذة يتم تطبيق تحويل فورييه السريع FFT من أجل كل إطار وذلك من أجل الحصول على طيف الاشارة، ومن ثم استخدام مرشحات تعتمد على نطاق Bark الذي يربط بين التردد الملاحظ والتردد الخطي الأصلي ويعطى بالعلاقة:

$$f_{bark}=6\ln[\frac{f}{600}+[(f/600)^2+1]^{0.5}] \quad (4)$$

بعد ذلك يطبق على ناتج المرشح منحني قياس ارتفاع الصوت equal-loudness curve، والتي تقارب الحساسية غير المتساوية لسمع الانسان عند ترددات مختلفة.

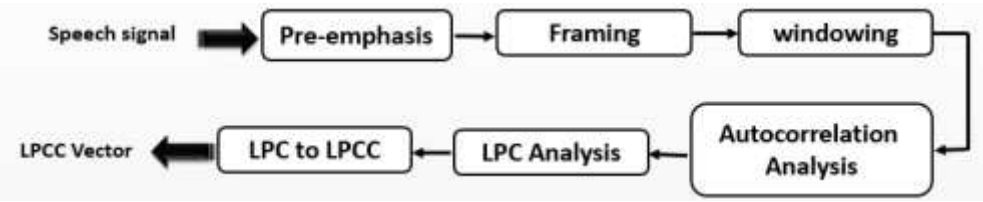


الشكل (5) المخطط الصندوقي لعمل الخوارزميةBFCC [8].

يتم بعد ذلك عملية ضغط لطيف الكلام، ويتم تنفيذ هذه العملية وفق قانونintensity-loudness power الذي يدمج العلاقة غير الخطية بين الكثافة والشدة الملاحظة لها، يتم في هذه المرحلة تقليل التغيرات الديناميكية وتسطيح قمم الطيف، حيث أن الطيف الناتج مسطح أكثر مع قمم أقل وضوحاً. يتم بعد ذلك حساب اللوغاريتم ومن ثم يستخدم تحويل جيب التمام المتقطع DCT لإعادة تحويل طيف مجال اللوغاريتمي إلى مجال الزمن، وبذلك يكون قد تم الحصول على شعاع BFCC.

خوارزمية LPCC [6] ،[9] :

يبين الشكل(6) المخطط الصندوقي لعمل خوارزمية LPCC ، فمن خلال تقليص مجموع مربعات الاختلافات (على فترة زمنية محدودة) بين عينات الكلام الفعلية وقيم التنبؤ الخطية، سوف يتم تحديد مجموعة فريدة من البارامترات (معاملات التنبؤ الخطية)، هذه المعاملات تشكل أساسا لتحليلات التنبؤ الخطي للكلام.



الشكل (6) المخطط الصندوقي لعمل الخوارزميةLPCC [6].

في الواقع إن عوامل التنبؤ الفعلية لا تستخدم في التعرف على الكلام لأنها نموذجية تظهر التباين العالي، لذلك يتم تحويل معاملات التنبؤ هذه إلى مجموعة أقوى من البارامترات هيCepstral Coefficients بواسطة المعادلات الرياضية التالية:

$$C_{LPC}(m)=\begin{cases} -a(m)-\sum_{i=1}^{m-1}\left(1-\frac{i}{m}\right)a(i)\ \mathrm{CLPC}(m-i), 1\le m\le P \\ -\sum_{i=1}^{m-1}\left(1-\frac{i}{m}\right)a(i)\ \mathrm{CLPC}(m-i), m>P \end{cases} \quad (5)$$

حيث a(m) تعبر عن معاملات التنبؤ الخطي LPC coefficients.

P يعبر عن ترتيب النموذج model order.

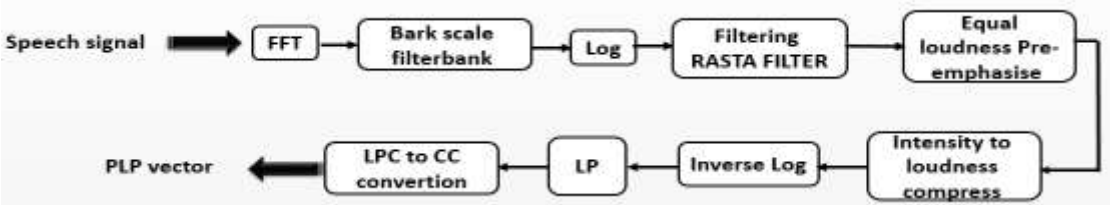
تجري عمليات pre-emphasis، التأطير ،النوفذة نفسها المطبقة فيBFCC.

خوارزمية PLP [13] :

تعتمد هذه التقنية على psychophysics of hearing (العلاقة بين مؤثر فيزيائي والإدراكات المؤثرة)، تقوم باستبعاد المعلومات التي ليس لها صلة بالكلام وبالتالي تحسين عملية التعرف حيث أن خصائصها الطيفية تم تحويلها لتتناسب مع خصائص النظام السمعي عند الإنسان، حيث أن PLP تقارب ثلاث جوانب رئيسية:

- The critical-band resolution curve.
- The equal loudness curve.
- The intensity-loudness power-law relation.

المخطط الصندوقي لعمل الخوارزمية مبين في الشكل(7).

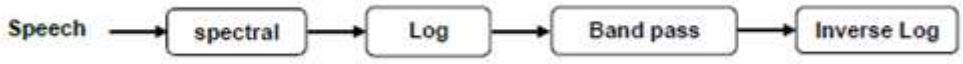


الشكل (7) المخطط الصندوقي لعمل الخوارزمية PLP [13] .

بعد معالجة إشارة الكلام يتم حساب تحويل فورييه السريع FFT للحصول على طيف الإشارة (power spectrum)، ومن ثم يتم استخدام مرشحات بشكل شبه منحرف تعتمد على نطاق bark الذي يعطى بالعلاقة:

$$\Omega(w)=6\ln\{w/1200\pi+[(w/1200\pi)^2+1]^{0.5}\} \quad (6)$$

من أجل دمج طيف الطاقة power spectrum داخل حزم حرجة متداخلة overlapping critical bands ، يتم بعد ذلك الترشيح بواسطة مرشح RASTA، الشكل(8)، وهي تقنية تعتمد على ترشيح تمرير الحزمة (band pass time filtering) المطبقة على لوغاريتم الطيف الممثل للكلام،



الشكل(8) مرشح RASTA [13].

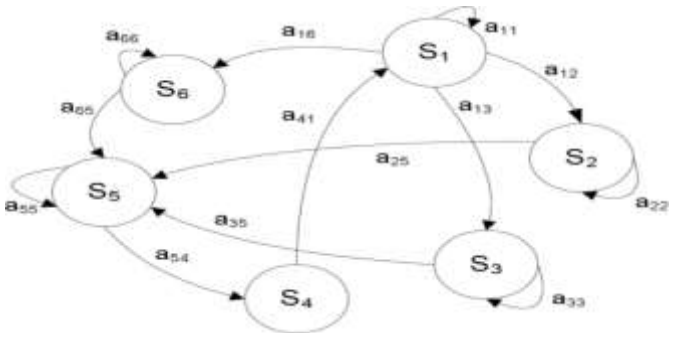
تتم بعدها عملية pre-emphasis لطاقة الطيف (power spectrum) بواسطة منحني قياس ارتفاع الصوت (equal-loudness curve)، والتي تقارب الحساسية غير المتساوية لسمع الإنسان عند ترددات مختلفة، عند حوالي 40db، حيث أن كل معامل لطاقة الطيف power spectrum coefficient يتم ضربه بالوزن Eالذي يعطي بالعلاقة:

$$E(w)=(w^2+56.8*10^6)w^4 / (w^2+6.3*10^6)^2(w^2+0.38*10^9) \quad (7)$$

والذي يسبب انخفاضاً في الحساسية في نطاق التردد العالي. تتم بعد ذلك عملية ضغط لطيف الكلام ويتم تنفيذ هذه العملية وفق(power-law of hearing) الذي يمزج العلاقة غير الخطية بين الكثافة intensity والشدة loudness الملاحظة لها، حيث يتم في هذه المرحلة تقليل التغيرات الديناميكية وتسطيح قمم الطيف حيث أن الطيف الناتج مسطح أكثر(smoother) مع قمم (peaks) أقل وضوحاً. يتم بعد ذلك حساب المعاملات التنبؤية، وتحويلها إلى معاملات cepstral ومن ثم الحصول على شعاع PLP.

سلاسل ماركوف Markuv Chain [10] :

يمكن أن تكون النماذج الرياضية محددة (Deterministic) أو عشوائية (Stochastic)، إلا أنه في عدة حالات اجتماعية وحياتية هناك ظواهر تصادفية (وهي ظواهر ذات سلوك غير قطعي لا يمكن السيطرة عليها بشكل تام أو التنبؤ بسلوكها المستقبلي بشكل مؤكد ويطلق عليها مصطلح العمليات التصادفية)، فيصبح النموذج التصادفي هو الأكثر ملاءمة لتمثيلها. المنظومة الموضحة بالشكل(9) يمكن أن توصف خلال أي فترة زمنية، كأن تكون موصوفة في واحدة من مجموعة الحالات المتقطعة (N) (S1, S2,…., SN).



الشكل (9) سلسلة ماركوف لستة حالات مع انتقالاتها [10].

من خلال تلك الأنظمة المتقطعة تخضع المنظومة إلى تغيرات في الحالة (من الممكن الرجوع إلى الحالة نفسها) وفقاً لمجموعة من الاحتمالات المرتبطة بالحالة. ويرمز إلى الزمن المرتبط بتغير الحالة، ويرمز للحالة الحقيقية خلال الزمن (t) بالرمز (Q_t). إن وصف الاحتمالية بصورة كاملة للمنظومة المبينة في الشكل(9) يتطلب وصف الحالة الحالية عند الزمن (t)، فضلاً عن كل الحالات السابقة لها. ينظر إلى سلسلة ماركوف كنوع من مخطط الاحتمالات أو طريق لتمثيل الفرضيات الاحتمالية. وسلسلة ماركوف محددة بالمكونات التالية:

1 –مجموعة N من الحالات و تمثل بـ $Q=\{q_1,q_2,\ldots q_N\}$

2–المصفوفة الاحتمالية الانتقالية A وتمثل بالمصفوفة :

$$A=\begin{pmatrix} a_{11} & \cdots & a_{1N} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ a_{N1} & \cdots & a_{NN} \end{pmatrix}$$

حيث أن كل a_{ij} تمثل احتمالية الانتقال من الحالة i إلى الحالة j بحيث تحقق الشرط التالي:

$$\sum_{j=1}^{N} a_{ij} = 1 \qquad \forall i$$

3- حالات خاصة هي حالة البداية q_0 وحالة النهاية q_f التي لا ترتبط مع أية مشاهدات (Observations)

4- التوزيع الاحتمالي الابتدائي على الحالات (Initial probability distribution)

$$\pi = \pi_1, \pi_2, \ldots, \pi_N$$

$$\sum_{i=1}^{N} \pi_i = 1$$

وكذلك

وتكون الاحتمالية التي تبدأ بها سلسلة ماركوف عند الحالة i في بعض الحالات $\pi_i = 0$ يعني لا يمكن أن تكون الحالة ابتدائية. وتعرف فرضية ماركوف بالعلاقة التالية:

$$P(q_i \backslash q_1 \ldots q_{i-1}) = P(q_i \backslash q_{i-1})$$

حيث أن:

$P(q_i \backslash q_1 \ldots q_{i-1})$ تمثل احتمالية حدوث الحالة q_i عند توفر الحالات $q_1 \ldots q_{i-1}$

$P(q_i \backslash q_{i-1})$ تمثل احتمالية حدوث الحالة q_i عند توفر الحالة q_{i-1} التي تسبقها فقط.

نموذج ماركوف المخفي Hidden Markuv Model :

نموذج ماركوف المخفي (HMM) عبارة عن نظام محطات الآلة المحدودة (finite state machine) القادر على توليد مشاهدات باحتمالية انتقال الحالة عند الزمن t التي تعتمد فقط على الحالة السابقة لها عند الزمن t-1 علماً أن تسلسل الحالة التي تنتج المشاهدة المعطاة مجهول، لذا ففي نموذج ماركوف المخفي تكون الحالة ليست مرئية، لذلك سمي بنموذج ماركوف المخفي، والانتقالات بين الحالات تحكمها مجموعة من الاحتمالات إضافة إلى احتمالات الانتقال من حالة معينة والتي يمكن أن تنتج نتيجة أو مشاهدة وحسب توزيع الاحتمالية المرتبط بتلك الحالة. والاختلاف بين نموذج ماركوف المخفي ونموذج ماركوف هو وجود الاحتمالات الإضافية، ويمثل هذا الجزء المخفي للنموذج ويرتبط بالمشاهدة الناتجة من كل حالة، فنموذج ماركوف الخفي هو نموذج تصادفي قادر على التصنيف الإحصائي.

مكونات بناءHMM [1]،[10]:

تعرف عناصر نموذج ماركوف المخفي HMM كالتالي:

N:عدد الحالات في النموذج، ويمكن تمثيل فضاء الحالة (S) كما يلي:

$$S=\{S_1,S_2,\ldots.S_N\}$$

حيث يرمز للحالة عند الزمن (t) بـ (q_t) .

M:عدد رموز مشاهدات الحالة الواحدة ويمكن تمثيل رموز المشاهدة الواحدة كما يلي:

$$V = \{v_1,v_2,\ldots.v_M\}$$

$A=\{a_{ij}\}$ التوزيع الاحتمالي للحالة الانتقالية (A) حيث $a_{ij} = p[q_t+1=S_j \backslash q_t=S_j]$, $1 \le i, j \le N$

$B=\{b_j(k)\}$ هو التوزيع الاحتمالي لرمز المشاهدة عن الحالة j

حيث $1 \le k \le M$ ،$1 \le j \le N$ ،$b_j(k)=p[v_k \text{ at } t\backslash q_t=S_j]$ و $\pi= \{\pi_i\}$ توزيع الحالة الابتدائية

حيث $1\le i \le N$ ،$\pi_i = p[q_1 = S_i]$

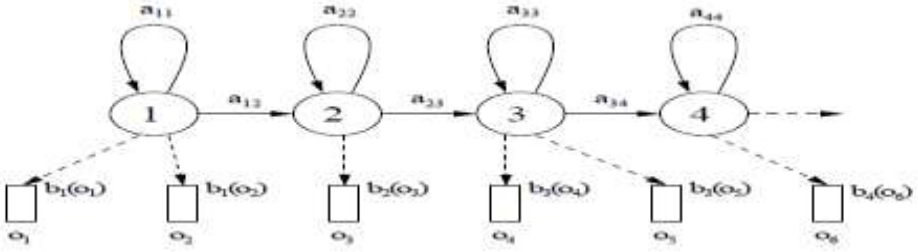
وبإعطاء القيم المناسبة لكل من (π, B, A, M, N) يكون بالإمكان استخدام نموذج ماركوف المخفي كمولد لمتسلسلة المشاهدات O:

$$O = O_1\ O_2\ \ldots\ldots O_T$$

ويمكن أن يمثل نموذج ماركوف المخفي كالتالي: $\lambda = (\pi ،B،A)$

حيث أن: π تمثل احتمالية الحالة الابتدائية و A تمثل مصفوفة احتمالية انتقال الحالة و B تمثل احتمالية مشاهدة الرمز عند الحالة i.

آلية عمل HMM:



الشكل (10) نموذج ماركوف المولد[10].

يتم عند بداية دخول الملاحظات إلى القالب اختيار الحالة الأولى، وهي الأعلى احتمالية في احتمالية الحالة الابتدائية π، فعند دخول الملاحظات يتم حساب احتمالية الانتقال من الحالة الحالية إلى كل الحالات الممكن الانتقال اليها بناء على الملاحظة التي تمت قراءتها ويتم الانتقال إلى الحالة التي لديها احتمالية a_{ij} أعلى من الحالات الأخريات وتستمر هذه العملية لغاية الانتهاء من الملاحظات والحصول على احتمالية انتماء هذه الملاحظة إلى هذا القالب، أي بمعنى آخر احتمالية انتماء هذه الإشارة الصوتية إلى الكلمة $P(O|\pi)$، وبذلك يؤخذ القالب π ذو الاحتمالية الأعلى على أنه هو الكلمة الصحيحة.

خطوات خوارزمية الدمج بين نظامي تعرف:

خطوات خوارزمية الدمج بين نظامي تعرف مبينة في الشكل(11) حيث تبدأ بمرحلة استخراج السمات FE ثم مرحلة التصنيف حيث يدخل شعاع السمات المستخرج من المرحلة الأولى إلى المصنف المستخدم HMM وهو نموذج ماركوف الخفي، خرج المصنف عبارة عن تسلسل من الكلمات واحتمالاتها. يتم اختيار القيمة الاحتمالية الكبرى لتسلسل الاحتمالاتM1 ، ثم يتم اختيار ثاني أكبر قيمة احتمالية لتسلسل الاحتمالاتM2 . ويتم حساب الفرق بين أكبر قيمة احتمالية و ثاني أكبر قيمة احتمالية لنظامي التعرف وقيمة الفرق الأكبر تكون هي ناتج خرج خوارزمية الدمج. حيثRR1(Recognition Rate1) معدل التعرف لنظام التعرف الأول وRR2(Recognition Rate2) معدل التعرف لنظام التعرف الثاني وRR(Recognition Rate) معدل التعرف النهائي الناتج عن تطبيق خوارزمية الدمج، كما أن $P(O|M1)$,$P(O|M2)$,$P(O|M4)$,$P(O|M3)$ هي خرج نموذج ماركوف الخفي.

يبين الشكل(12) مخطط صندوقي لمراحل العمل، حيث تم بناء أربع أنظمة للتعرف على الكلام تختلف فيما بينها بطريقة استخراج السمات، يعتمد النظام الأول على MFCC في استخراج السمات والنظام الثاني على PLP في

استخراج السمات بينما نستخدم في النظام الثالث LPCC في استخراج السمات، وكذلك نستخدم BFCC في النظام الرابع لاستخراج السمات. تم استخدام HMM في مرحلة التصنيف، وتم الاعتماد على حاسوب بنظام تشغيل ويندوز 10 وسرعة المعالج \1\غيغاهرتز وذاكرة الوصول العشوائي \2\غيغابايت وتم استخدام مكتبات معالجة الإشارة signal processingومكتبات الصوت voicebox في برنامج Matlab. تم حساب معدل التعرف لكل نظام على حدى، حيث أن: معدل التعرف= عدد العينات الصحيحة \ عدد العينات الكلية.

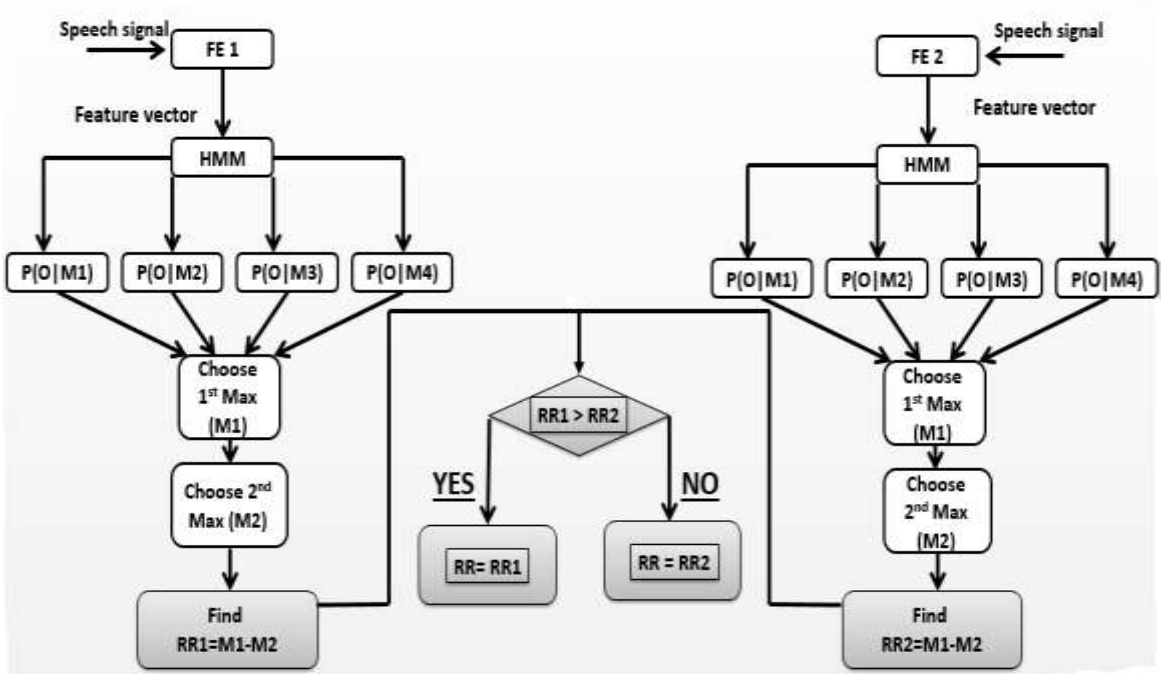
تم تطبيق كل خوارزمية من الخوارزميات السابقة و من ثم دراسة النتيجة الحاصلة حيث تمت الدراسة على ثلاث حالات:

الحالة الأولى: تم تطبيق الخوارزميات LPCC, BFCC, PLP, MFCC بحيث تم الحصول على شعاع من 8 سمات تعبر عن كامل العينة الصوتية .

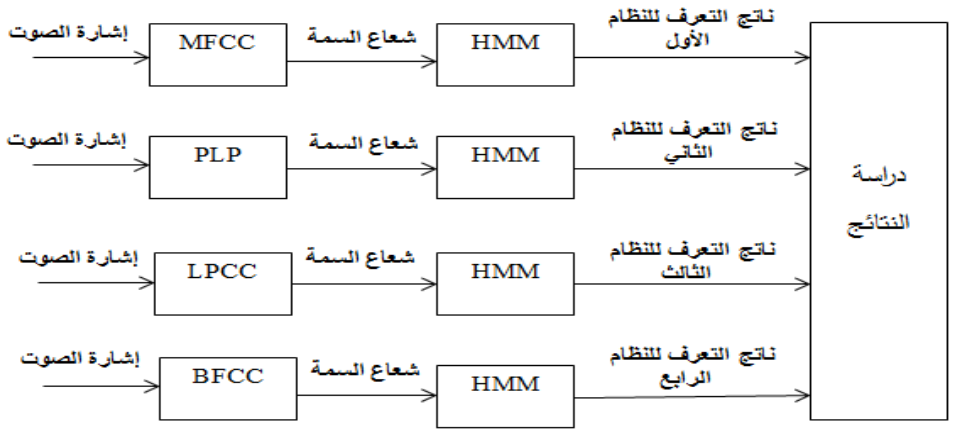
الحالة الثانية: تم تطبيق خوارزميات استخراج السمات الأخيرة LPCC, BFCC, PLP, MFCC بحيث يتم الحصول على 12 من عدد السمات من أجل كل إطار.

الحالة الثالثة: تم تطبيق الخوارزميات LPCC, BFCC, PLP, MFCC لاستخراج السمات بحيث يتم الحصول على 25من عدد السمات من أجل كل إطار.

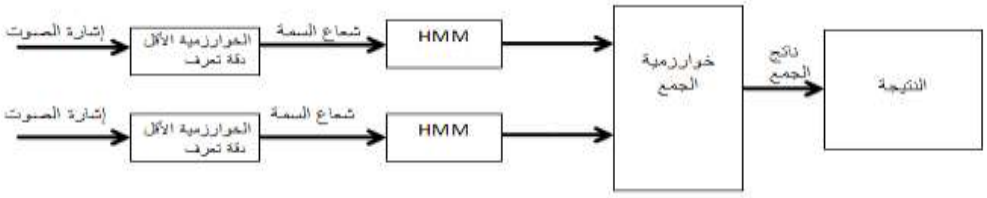
تم تطبيق خوارزمية الدمج على النظامين الأقل دقة تعرف ومقارنة النتيجة الحاصلة مع نتائج تطبيق كل من النظامين على حدى، ومن ثم تمت مقارنة نتائج التعرف.



الشكل (11) ، مخطط خوارزمية الدمج بين نظامي تعرف



أ – اختيار الخوارزمية الأقل دقة

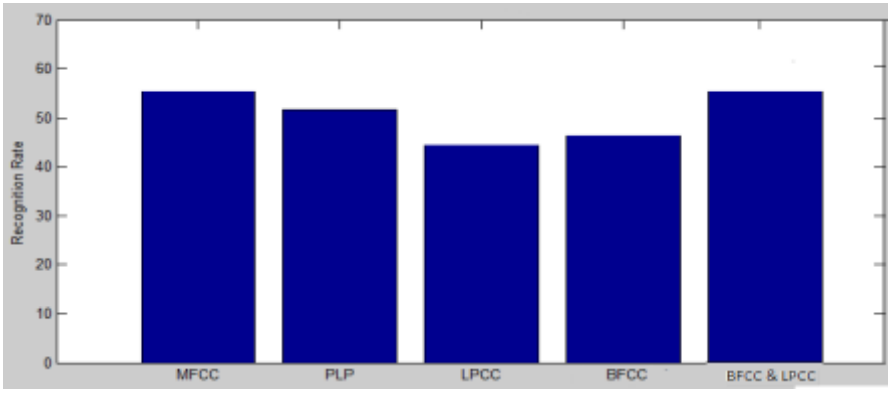


ب – تطبيق خوارزمية الدمج على الخوارزميتين الأقل دقة

الشكل (12) المخطط الصندوقي لمراحل العمل

النتائج والمناقشة:

تم تطبيق الخوارزميات على 150 عينة صوتية حيث عرض **الشكل (13) قيمة معدل التعرف في حال** تطبيق كل خوارزمية في الحالة الأولى.

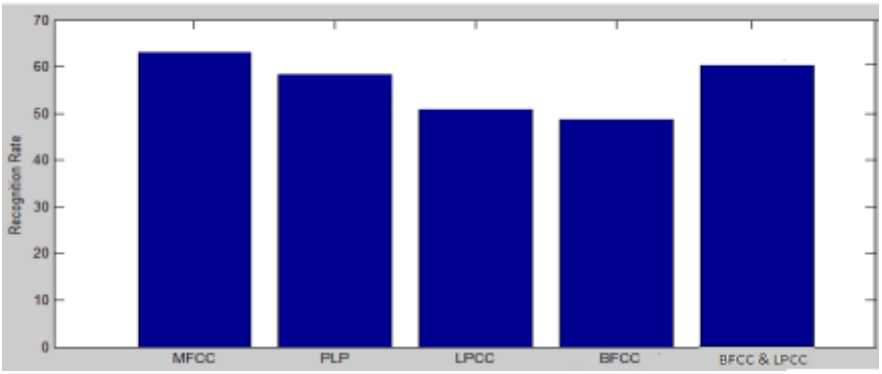


الشكل (13) نتائج حساب معدل التعرف لكل خوارزمية في الحالة الأولى(شعاع السمات 8 سمات).

الجدول (1) قيم نسبة التعرف الناتجة لكل خوارزمية للحالة الأولى

الخوارزمية	معدل التعرف	زمن التعرف (بالثانية)
MFCC	54.63 %	1.01
PLP	51.42 %	1.01
LPCC	44.31%	1.0
BFCC	46.36%	1.01
BFCC & LPCC	54.61 %	1.02

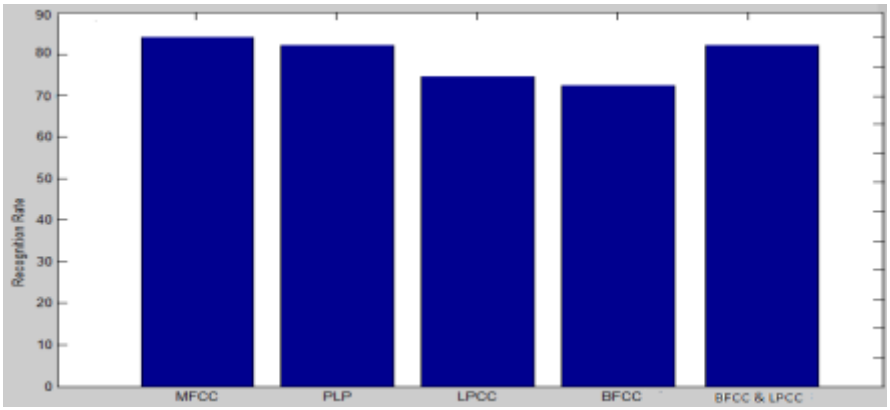
كما ظهر معدل التعرف الناتج من تطبيق كل خوارزمية في الحالة الثانية في الشكل (14).



الشكل (14) نتائج حساب معدل التعرف لكل خوارزمية في الحالة الثانية (12 سمة).

الجدول (2) قيم نسبة التعرف الناتجة لكل خوارزمية للحالة الثانية		
الخوارزمية	معدل التعرف	زمن التعرف (بالثانية)
MFCC	63.42 %	1.86
PLP	59.87 %	1.85
LPCC	52.14 %	1.84
BFCC	50.91 %	1.84
BFCC & LPCC	61.81 %	1.82

كما ظهرت نتائج حساب معدل التعرف عند تطبيق كل خوارزمية في الحالة الثالثة كما هو مبين في الشكل(15).



الشكل(15) نتائج حساب معدل التعرف لكل خوارزمية في الحالة الثالثة (25 سمة).

الجدول (3) قيم نسبة التعرف الناتجة لكل خوارزمية للحالة الثالثة		
الخوارزمية	معدل التعرف	زمن التعرف (بالثانية)
MFCC	84.21 %	4.52
PLP	83.34 %	4.52
LPCC	76.25 %	4.52
BFCC	74.98 %	4.51
BFCC & LPCC	83.63 %	4.61

تبين النتائج زيادة دقة التعرف بالنسبة لخوارزمية MFCC في جميع الحالات. لذلك تم تطبيق خوارزمية الدمج على الخوارزميتين الأقل دقة تعرف LPCC وBFCC بهدف تحسين النتائج. وقد أظهرت النتائج تحسناً ملحوظاً في معدل التعرف للنظام المجمع مقارنة بنتائج كل نظام على حدى.

الاستنتاجات والتوصيات:

نلاحظ من النتائج الحاصلة سابقاً أن خوارزمية MFCC بالنسبة لدقة التعرف تفوقت على خوارزمية PLP بشكل قليل بينما ازداد الفرق بشكل ملحوظ عن خوارزميتي LPCC وBFCC في جميع الحالات. لذلك تم تطبيق خوارزمية الدمج على الخوارزميتين LPCC وBFCC بهدف تحسين النتائج. وقد أظهرت النتائج تحسناً ملحوظاً في معدل التعرف للنظام المجمع مقارنة بنتائج كل نظام على حدى. ومن الملاحظ زيادة دقة التعرف عند زيادة عدد السمات ولكن على حساب زمن التعرف الذين ارتفع بشكل ملحوظ، حيث أنه عند استخلاص عدد أقل من السمات يتم الحصول على نتيجة تعرف مقبولة نوعاً ما بالنسبة للخوارزميات MFCC، PLP، BFCC & LPCC وسريعة التنفيذ من حيث زمن التعرف. أماعند زيادة عدد السمات تزيد نسبة وزمن التعرف. فإذا كان الاهتمام الأكبر حول زمن تنفيذ سريع نسبياً فالأفضل استخدام سمات قليلة مثل تطبيقات الاتصالات الهاتفية وألعاب الكمبيوتر والمحاكاة، أما اذا كان التوجه لاستخدام أنظمة تعرف تهتم بزيادة دقة التعرف بغض النظر عن استهلاك زمن أكبر فالأفضل استخدام سمات كثيرة مثل الأنظمة الأمنية والحربية ووحدات تحكم الحركة الجوية.

أقترح في المرحلة اللاحقة زيادة عدد العينات الصوتية وزيادة عدد التقنيات المستخدمة في مرحلة استخراج السمات وبالتالي زيادة مساحة المقارنة للحصول على نتائج أفضل وأشمل.

References:

[1] LEVENE, M., *Speech Recognition*,Wiley,London, 2020.
[2] HOMAYOON BEIGI, *Fundamentals of speaker Recognition*, Springer Sience,2013.
[3] NEUSTEIN AMY; PATIL HEMANT, *Forensic Speaker Recognition*, Springer, 2014.
[4] DING, C.H. ; BUYYA,R. *MULTIMEDIA-OPERATING-SYSTEMS*, University of Melbourne, Australia, 2017.
[5] PITZ, M.; SCHLUTER, R; NEY, H.; MOLAU, S., *Computing Mel-frequency cepstral coefficients on the power spectrum*, IEEE International Conference, Speech and Signal Processing , Vol. 44, No. 9, 2012.
[6] NAMRATA DAVE, *Feature Extraction Methods LPC, PLP and BFCC In Speech Recognition*, international journal for advance research in engineering and technology, Speech and Signal Processing, , Vol. 97, No. 12, 2015.
[7] HERSH, W., *Improving Speech Recognition Rate Through Analysis Parameters*, Springer, German ,2018,455.
[8] MANNING, C.; RAGHAVAN,P.; SCHÜTZE,H., *A Comparative Study of Speech Recognition*, Cambridge UP, England,2017,533.
[9] MARIUS ZBANCIOC; MIHAELA COSTIN, *Using Neural Networks And LPCC To Improve Speech Recognition*, International IEEE SCS Conference, Speech and Signal Processing, , Vol. 56, No. 11, 2016.
[10] SANDEEP, S. *Speech Recognition Technique* ,Wiley, London,2017,334.

[11] TAABISH GULZAR; ANAND SINGH, *Comparative Analysis of LPCC, MFCC and BFCC for the Recognition*, International Journal of Computer Applications , Volume 101– No.12, September 2014.
[12] SINGH, V., *Classification of Spoken Words using Artificial Neural Networks*, International Journal of Computer Applications, , Vol. 102, No. 17, 2016.
[13] Marius,M., *Speech Techniqe*, Wiley, London, 2019.