مجلة جامعة تشرين للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (39) العدد (2) 2017
Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series Vol. (39) No. (2) 2017

اختبار وتقييم العوامل المؤثرة في أنظمة التعرف السمعية لكلمات معزولة في اللغة العربية

الدكتور جعفر محسن الخير*
الدكتورة مريم محمد ساعي**
نور سميع غضبان ***

(تاريخ الإيداع 14 / 2 / 2017. قُبِل للنشر في 5/ 4 / 2017)

□ ملخّص □

تعد تقنيات التعرف على الكلام من أهم التقنيات الحديثة، وقد تم تطوير العديد من الأنظمة المختلفة من حيث الطرائق المستخدمة في استخراج السمات وطرائق التصنيف، لكن مهما كانت الخوارزمية المستخدمة أو طريقة التصنيف فإن في تكنولوجيا معالجة الكلام للتطبيقات الحقيقية يمكن للعديد من الحقائق أن تشوه أو تتلف الكلام، مما يجعل الكلام لا يبدو بالطريقة التي تم تسجيله فيها.

يقترح البحث إنشاء نظام للتعرف على الكلمات المعزولة بالاعتماد على السمات السمعية المستخرجة من فيديوهات منطوقة لكلمات باللغة العربية ، ومن ثم إضافة مكون الطاقة والمشتقات التفاضلية وتحديد عدد قنوات المرشح الأمثلي في مرحلة استخراج السمات لخوارزمية معاملات تردد ميل لزيادة نسبة التعرف. تم استخدام نماذج ماركوف المخفية في مرحلة التصنيف وتحديد عدد الحالات الأمثلي في المصنف.

تم اختبار النظام على 4155 عينة، فبينت النتائج أن السمات المضافة إلى الخوارزمية والاختبارات التجريبية على عدد قنوات المرشح وعدد حالات المصنف مما رفع أداء الخوارزمية وبالتالي معدل التعرف الذي وصل إلى 92%.

ومن أجل أغراض الاختبار والتقييم في أنظمة التعرف على الكلام تم إدخال ضجيج مفتعل ومراقبة تأثيره على نسب التعرف، لذلك قام البحث بتطبيق هذه الطريقة لأول مرة على قاعدة بيانات منطوقة باللغة العربية. ولتقليل من أثره تم تطبيق إحدى طرق تقدير إشارة الضجيج – في بحثنا قمنا بتطبيق الطريقة الأكثر تطبيقا" بالنسبة لقواعد البيانات في اللغات الأخرى وهي الطرح الطيفي – لتقدير إشارة الضجيج وطرحها من الاشارة المشوبة وذلك ليكون مدخلا" للحد من أثر الضجيج. هذا التقدير قدم تحسين منخفض عندما طبق على ملفات SNR المنخفضة وحسن النتائج مع ملفات SNR العالية فقط. لكن نتج عن الطرح الطيفي ضجيج موسيقي حيث تمت مراقبته برفع عتبة التوهين لتلائم الكلمات المنطوقة باللغة العربية .

الكلمات المفتاحية: التعرف على الكلام، استخراج السمات، خوارزمية معاملات تردد ميل، نماذج ماركوف المخفية، تقدير الاشارة، الطرح الطيفي، نسبة الاشارة الى الضجيج.

* أستاذ مساعد – قسم هندسة الحاسبات والتحكم الألي_ كلية الهندسة الميكانيكية والكهربائية_ جامعة تشرين _ اللاذقية _سورية
**أستاذ مساعد – قسم هندسة الحاسبات والتحكم الألي _ كلية الهندسة الميكانيكية والكهربائية_جامعة تشرين _ اللاذقية _سورية.
*** طالبة دكتوراه – قسم هندسة الحاسبات والتحكم الألي _ كلية الهندسة الميكانيكية والكهربائية_جامعة تشرين_ اللاذقية_ سورية.

مجلة جامعة تشرين للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (39) العدد (2) 2017
Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series Vol. (39) No. (2) 2017

Testing and Evaluation Factors That affecting Audio Speech Recognition Systems For Isolated Arabic Words

Dr. Jaffar Mouhsen Alkheer*
Dr. Mariam Mohamad Saee**
Nour Sami Ghadban***

(Received 14 / 2 / 2017. Accepted 5 / 4 / 2017)

□ ABSTRACT □

The speech recognition is one of the most important techniques of modern techniques, there has been many different systems developed in terms of the methods used in the features extraction and classification methods, But whatever the algorithm used, or the method of classification in speech processing technology in real applications can be for many of the facts that the deformation or damage of speech, making speech does not seem the way you recorded it.

This study proposes design a system to identify isolated words depending on the audio features extracted from videos to the words in Arabic Language and then the energy and Temporal derivative components is added, the optimal number of channels in the filterbank determined in extracting features of the method Mel Frequency Cepstral Coefficient (MFCC) stage, it was used hidden Markov models HMM as classification and determine the optimal number of cases in classifier **.**

The system was tested on 4155 samples. The results showed that the added features of the algorithm and experimental tests on the number of channel of filter bank and the number of cases in classifier, raising the performance of the algorithm and thus identify which reached 92% rate.

And for testing and evaluation purposes in the speech recognition systems have been introduced artificially noise and monitor its impact on the recognition ratios, so the research application of this method for the first time on the data base pronunciations in Arabic.

To minimize the effect of the noise signal we have to applied method of estimating - in our research we have implemented the most way application "For databases in other languages, a Spectral subtraction - to estimate the noise signal and subtracted from the reference vestiges so as to have access" to reduce the impact of noise.

This estimate gave a low improve when applied to low SNR and good results with only a high SNR files. But the resulting spectral subtraction musical noise, where it was monitored to raise the threshold of attenuation to match spoken words in Arabic

Key words: Speech recognition, features extraction, MFCC, Spectral subtraction, Markov Hidden models, SNR .

* **Assistant Professor in computer and control engineering in Tishreen University, Lattakia, Syria**
** **Assistant Professor in computer and control engineering in Tishreen University, Lattakia, Syria**
*** **Postgraduate student in computer and control engineering in Tishreen University, Lattakia, Syria**

مقدمة:

بدأ اهتمام خبراء الحاسب والباحثين بالتعرف على الكلام منذ أكثر من أربعة عقود، وذلك لكي يصل الإنسان إلى مرحلة تجعله قادرا على التخاطب مع الكمبيوتر وإعطاءه الأوامر والتعليمات صوتيا وبدون الحاجة إلى الكتابة وغيرها من الطرق، وذلك توفيرا للوقت والجهد.

وفي السنوات الأخيرة تطورت نظم التعرف على الكلام تطورا" واضحا" وكبيرا"، بحيث أصبحت برامج التعرف الآلي تدخل في أغلب مجالات الحياة، ووصلت إلى دقة مرضية نوعا" ما [1].

أكدت التطبيقات الحقيقية لتكنولوجيا معالجة الكلام أنه يمكن للعديد من الحقائق أن تشوه أو تتلف الكلام، مما لا يمكن التعرف على الكلام لأنه لا يبدو كعينات التدريب، ولذلك في نظم التعرف الآلي على الكلام وأثناء نطق المستخدم فإنه لا يمكن التعرف على الكلمة المنطوقة بسبب وجود هذه الحقائق التي تشوه الكلام فتؤدي إلى تدني مستوى دقة النظام في التعرف على الكلام المنطوق [13] .

تم في هذا البحث تقديم طريقة جديدة لتحسين أداء الأنظمة السمعية البصرية بالاعتماد على السمات السمعية، عن طريق زيادة معدل التعرف بتحسين عملية استخراج السمات السمعية من فيديوهات لكلمات معزولة منطوقة باللغة العربية، وذلك عن طريق إضافة معاملات في مرحلة استخراج السمات السمعية ومعرفة العدد الأمثلي لكل من قنوات المرشح وعدد حالات المصنف.

كما تم التطرق في البحث إلى دراسة لأهم العوامل التي تؤثر على انخفاض نسبة التعرف على الكلام وهو الضجيج ومراقبة تأثيره على نسبة التعرف، وتطبيق الطرح الطيفي لتقدير إشارة الضجيج وطرحها من الإشارة المشوبة وذلك ليكون مدخلا" للحد من أثر الضجيج .

أهمية البحث وأهدافه:

تم تطوير العديد من الأساليب من أجل إنجاز التعرف على الكلام، والتي تختلف عن بعضها بالطرائق المستخدمة في استخراج السمات (feature extraction methods) وطرائق التصنيف (classification methods) التي تعتمدها، ولكن عندما تخضع هذه الأنظمة للضجيج مهما كانت الطريقة المستخدمة في استخراج السمات أو التصنيف سوف تنخفض نسبة التعرف على الكلام [2].

تقوم مرحلة استخراج السمات بالتعبير عن الصوت بواسطة مجموعة من السمات ولذلك فإن هذه السمات تلعب دورا" أساسيا" في تحديد نسبة التعرف، وذلك لأن مرحلة استخراج السمات تؤدي إلى الاحتفاظ ببعض المعلومات من إشارة الكلام وخسارة معلومات أخرى. بما أن طرائق استخراج السمات المستخدمة في التعرف على الكلام تختلف من حيث السمات التي تعتمدها، بالتالي فإن لكل طريقة نسبة تعرف صحيحة محدودة وغير كاملة. ويمكن القول أن بزيادة عدد السمات تزداد المعلومات المعبرة عن إشارة الصوت وبالتالي تكون عملية التعرف على الصوت أسهل ونسبة الخطأ أقل. لذلك قام البحث بإضافة مكون الطاقة والمشتقات التفاضلية من أجل تحسين عملية التعرف، مما يؤدي إلى تحسين استخراج السمات السمعية في الأنظمة السمعية للناطقين باللغة العربية من خلال تحسين خوارزمية استخراج السمات السمعية بإضافة معاملات إلى خوارزمية معاملات تردد الميل MFCC المستخدمة في مرحلة استخراج السمات[19] .

كما تم إجراء اختبارات تجريبية بالاعتماد على Gunter and Bunke [21]، حيث أن هذه الاختبارات خضعت لها أهم أنظمة التعرف على الكلام في اللغات الأخرى ولم تطبق حتى الآن في اللغة العربية لتحديد العدد الأمثلي لكل من عدد قنوات المرشح وعدد حالات المصنف وقد أثبتت أنها تساهم في زيادة نسبة التعرف على الكلام.

كما قام البحث بدراسة أهم الصعوبات التي تواجه أنظمة التعرف على الكلام وهو الضجيج الذي يؤثر في عملية التعرف، لذلك كان من أهم أهداف الدراسة إضافة ضجيج مفتعل ومراقبة تأثيره على نسبة التعرف على الكلام لأن إدخال الضجيج المفتعل يعتبر آلية من أغراض الاختبار والتقييم في أنظمة التعرف على الكلام لمراقبة نسبة التعرف باختلاف قيمة الضجيج المطبق، وكما تمت دراسة الطرح الطيفي وإدخاله كطريقة في تحسين نتائج التعرف على الكلام وذلك من خلال تقدير إشارة الضجيج وطرحها من الإشارة المشوبة للحصول على الإشارة النظيفة ومراقبة نتائج ذلك على النظام باختلاف قيمة الضجيج المطبق.

طرائق البحث وموارده:

تم الاعتماد على قاعدة بيانات قام بإنشائها د.علاء ساهر[25]. تتضمن القاعدة 10000 مقطع فيديو لكلمات معزولة ومنطوقة باللغة العربية من قبل خمسين متكلم تتراوح أعمارهم بين 18 إلى 60 عاما". تشمل الكلمات المنطوقة الأعداد من صفر إلى عشرة ، تم استخدام خورازمية MFCC بعد التعديل عليها في مرحلة استخراج السمات، كما تم استخدام نماذج ماركوف المخفية في مرحلة التصنيف .

تم استخدام برنامج Adobe Premiere Pro للحد من أي ضجيج غير مفتعل وذلك من أجل أن تكون العينات خالية من الضجيج (نظيفة)، وحيث يمكننا بواسطته أيضا" إدخال الضجيج بحيث تكون عينات الاختبار المضافة لقاعدة البيانات تحوي على الضجيج وبالتالي تحقق سهولة لاختبار آثاره على أداء التعرف.

إنّ بيئة العمل المستخدمة في هذا البحث هي Matlab2014a، وتم استخدام مكتبات منه هي , voicebox) signal processing).

قبل مرحلة تدريب المصنف HMM على الكلمات تم إنشاء ملفات معنونة label files لكل تسجيل، حيث يصبح لكل تسجيل كيان خاص به له نفس الاسم السابق، واستخدم برنامج Praat لتعليم الكيانات الاسم بكل سهولة ومن ثم تم استخدام برنامج # C لتحويل الملفات الناتجة عن هذه المرحلة بصيغة Praat's TextGrid إلى waveform.

استخراج السمات وفق خوارزمية معاملات تردد الميل MFCC وتدريب المصنف وفق نماذج ماركوف المخفية HMM :

تعد خوارزمية MFCC (Mel Frequency Cepstral Coeffiecents) من الطرائق السائدة والمهيمنة المستخدمة في استخراج السمات وذلك بسبب حساسية مرشحاتها لخواص إشارة الصوت البشرية [9].

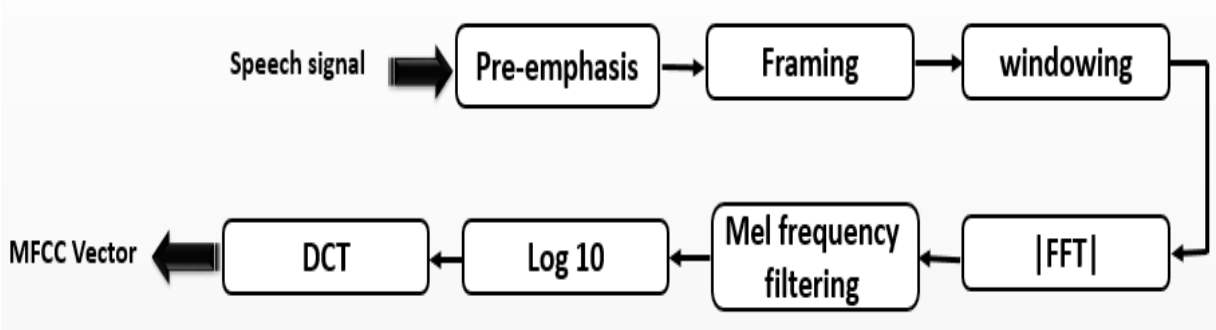
تستخدم معاملات MFCC بشكل كبير في التعرف على الكلام، حيث تم تقديم هذه المعاملات من قبل العالمين Davis Mermelste، في عام 1980 ومازالت متقدمة في هذا المجال منذ ذلك الوقت [20].

إن الأصوات التي تولد من قبل الإنسان يتم ترشيحها حسب شكل المسلك الصوتي (vocal tract)، فإذا تمكنا من تحديد شكل المسلك الصوتي بدقة فإنه يمكن تحديد الصوت (phoneme) الذي يتم إنتاجه. شكل المسلك الصوتي

يتجلى في غلاف طيف طاقة الزمن القصير (short time power spectrum) وهدف MFCC هو تمثيل هذا الغلاف بدقة [5].

3-1مراحل استخراج السمات:

تعتمد خوارزمية MFCC على التغيرات المعروفة في عرض حزمة الترددات للأذن البشرية، حيث أن لمرشحاتها تباعدا خطيا على الترددات المنخفضة ولوغاريتمياً على الترددات المرتفعة. تستخدم هذه الخوارزمية من أجل التقاط الصفات الرئيسية للكلام، حيث تمتلك MFCC تباعدا خطيا على الترددات الأقل من 1000 هرتز وتباعدا لوغاريتمياً على تردد أكبر من 1000 هرتز. خطوات عمل الخوارزمية مبينة في الشكل(1).



الشكل (1) المخطط الصندوقي لعمل الخوارزمية MFCC

- **pre-emphasis**

يتم تطبيق عملية pre-emphasis (وهي عملياً مرشح تردد عالي high pass filter) على الإشارة وذلك من أجل تعويض جزء التردد العالي الذي تم فقده أثناء آلية إنتاج الكلام (زيادة الطاقة النسبية للطيف عالي التردد)، حيث يتم إعادة تقييم كل قيمة في إشارة الكلام باستخدام الصيغة [18]:

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

حيث: s(n): إشارة الكلام

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

a: ثابت تتراوح قيمته بين 0.9 و o.1

- **التأطير (framing)**

إشارة الكلام هي إشارة متغيرة باستمرار لذلك من أجل تبسيط الدراسة نعتبر أنه من أجل نطاق زمني قصير (short time scale) فإن إشارة الصوت لا تتغير كثيراً لهذا السبب يتم تقطيع الإشارة إلى عدد من الإطارات (frames)، 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[10].

- **النوفذة (windowing)**

كل إطار (frame) سوف يخضع لعمية النوفذة (windowing) باستخدام نافذة هامينغ، وذلك من أجل القضاء على الانقطاعات عند الحواف.

تعطى نافذة هامينغ بالعلاقة [18]:

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

حيث w(n) هو مطال العينة الجديد.

n هو ترتيبها في النافذة.

N هو الطول الكلي للنافذة

بعد عملية النوفذة windowing يتم تطبيق تحويل فورييه السريع FFT من أجل كل إطار وذلك من أجل استخراج مركبات التردد للإشارة في مجال الزمن[11].

- **ترشيح الإشارة وفقاً لتردد ميل (Mel frequency filtering)**

تعمل MFCC على ترشيح طيف الإشارة الصوتية (short time power spectrum) عن طريق مجموعة من المرشحات المثلثية (Mel filter bank) (التي صممت كمحاكاة لمرشحات تمرير الحزمة band pass filtering التي تظهر في النظام السمعي)، المتباعدة بانتظام وفقا لمقياس ميل الترددي (Mel frequency scale) الذي يعبر عن علاقة تربط التردد الملاحظ لنغمة صافية (m) (perceived frequency of pure tone) بترددها المقاس الأصلي (f) ويعطى بالعلاقة [18] :

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

يستطيع الإنسان أن يميز التغيرات الصغيرة في الـ pitch (وهي الارتفاع أو الانخفاض النسبي لنغمة (tone) كما تدركها الأذن الذي يعتمد على عدد الاهتزازات التي تنتجها الحبال الصوتية في الثانية) ويميزها بشكل أفضل عند الترددات الصغيرة من الترددات الكبيرة، بالتالي فإن تضمين هذا المقياس يجعل سماتنا أقرب إلى سمع الإنسان[12].

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. نتيجة هذا التحويل يكون قد تم الحصول على شعاع MFCC.

3-2 مكون الطاقة والمشتقات التفاضلية:

تمتلك خوارزمية MFCC عند تطبيقها شعاع سمات ذي 12 معامل هي (Formant, Pitch, Intensity, Spectral Flux, Perceptual Loudness, Centre of Spectral Mass Gravity, Harmonics to Noise Ratio, Zero Crossing Rate, MFCC Coefficient, Frequency, Slope, Bandwidth)، وعند إضافة مكون الطاقة شعاع السمات يصبح ذي 13 معامل وعند إضافة المشتقات التفاضلية يصبح شعاع السمات للخوارزمية (مصفوفة الارتباط في خوارزمية MFCC) ذي 39 معامل.

لتحسين خوارزمية استخراج السمات في التعرف على الكلام تم إضافة مكون الطاقة، حيث تعتبر السمة المفيدة التي تستخرج من إشارة الصوت هي قياس الطاقة في إطار الكلام، ويمكن حساب الطاقة بالعلاقة [14]:

$$E = \sum_{n=0}^{N-1} x(n)^2 \qquad (4)$$

يتم احتساب سمات الكلام التي تم النظر فيها حتى الآن من نافذة قصيرة الزمن من الكلام – عادة حول 20ms تصف سمات ثابتة– ولكن يعطي معدل التغير في سمات الكلام معلومات هامة عن أصوات الكلام. ولهذا

يمكن إدراج هذا التغيير عن طريق المشتقات التفاضلية أي حساب سمات السرعة والتسارع، ويشار إلى هذه أيضا" بدلتا السمات delta feature ودلتا دلتا السمات *delta–delta* features .

1. السرعة delta features.
2. التسارع delta–delta features.

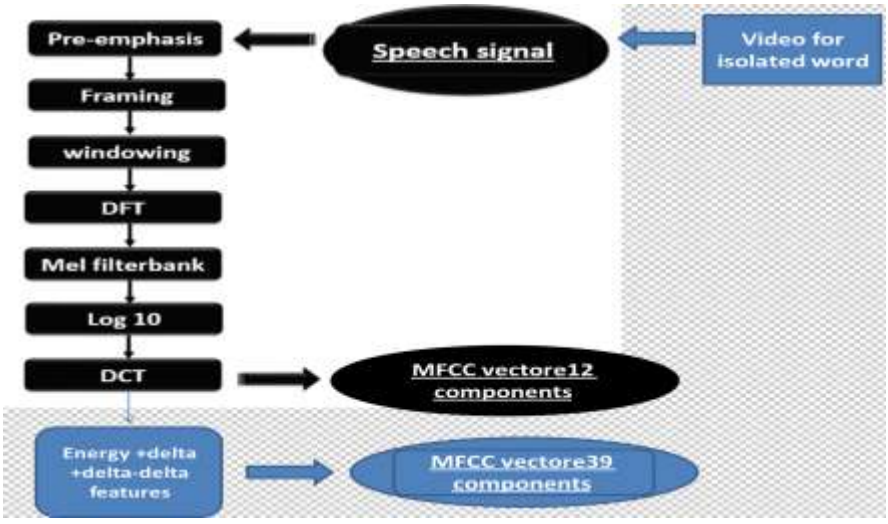
يقيس شعاع السرعة معدل التغير في الأشعة الثابتة، يمكن حساب ذلك عن طريق قياس الفرق بين الإطارات القادمة والسابقة ويعطى بالعلاقة [14]:

$$\Delta \mathbf{c}_n = \mathbf{c}_{n+1} - \mathbf{c}_{n-1} \qquad (5)$$

يقاس التسارع من السمات الثابتة ويتم احتسابها على معدل التغير في شعاع السرعة، يمكن حساب ذلك عن طريق قياس الفرق بين سرعة الإطارات القادمة والسابقة ويعطى بالعلاقة [14]:

$$\Delta\Delta \mathbf{c}_n = \Delta \mathbf{c}_{n+1} - \Delta \mathbf{c}_{n-1} \qquad (6)$$

يبين الشكل (2) المخطط الصندوقي لمراحل استخراج السمات وفق الخوارزمية MFCC والتعديلات التي تمت عليها بإضافة مكون الطاقة والمشتقات التفاضلية:



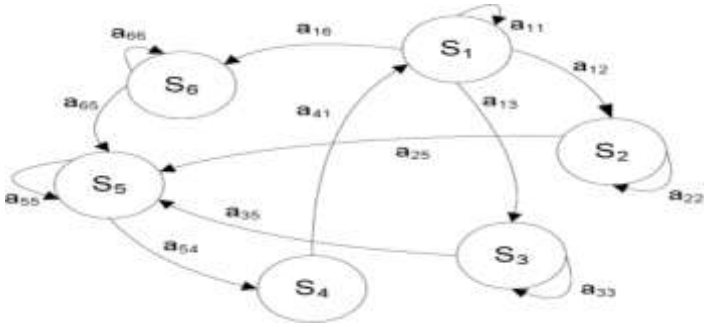
الشكل (2) المخطط الصندوقي للتعديلات على خوارزمية MFCC.

3-3 تدريب المصنف:

يمكن أن تكون النماذج الرياضية محددة(Deterministic) أو عشوائية (Stochastic). ومع ذلك فإن في عدة حالات اجتماعية وحياتية هناك ظواهر عشوائية (وهي ظواهر ذات سلوك غير قطعي لا يمكن السيطرة عليها بشكل تام أو التنبؤ بسلوكها المستقبلي بشكل مؤكد ويطلق عليها مصطلح العمليات العشوائية[16].

فيصبح النموذج العشوائي هو الأكثر ملاءمة لتمثيلها.

المنظومة الموضحة بالشكل(3) يمكن أن توصف خلال أي فترة زمنية، كأن تكون موصوفة في واحدة من مجموعة الحالات المتقطعة(N) (Discrete states) هي : S1،S2،....،SN [7].



الشكل(3) سلسلة ماركوف ل 6 حالات مع انتقالاتها

وخلال 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 بـ (t=1,2,.....)، ويرمز للحالة الحقيقية خلال الزمن (t) ب(Qt). إن وصف الاحتمالية بصورة كاملة للمنظومة أعلاه يتطلب وصف الحالة الحالية عند الزمن (t)، فضلاً عن كل الحالات السابقة لها. فينظر إلى سلسلة ماركوف كنوع من مخطط الاحتمالات (Probabilistic Graphical Model) أو طريقة لتمثيل الفرضيات الاحتمالية [6].

4- الكلام:

الكلام هو عبارة عن سياق من الرموز الصوتية التي تخضع لنظام معين متفق عليه بين أفراد الثقافة الواحدة وهو من صور اللغة وأكثر خصوصية منها والأداء الفعلي والأسلوب الأكثر شيوعاً لها للتواصل بين البشر، فهو الجانب المنطوق والمسموع من اللغة. فمن خلال عملية الكلام يستطيع الفرد التعبير عن آرائه وأفكاره ومشاعره ونقل المعلومات. ولقد حاول الإنسان التواصل مع الحاسب عن طريق الكلام محاكيا بذلك أسهل وأكثر وسائل التواصل الطبيعية بين البشر منذ عشرات القرون. ولقد أثار موضوع الواجهات التخاطبية بين الإنسان والآلة اهتمام المهندسين وعلماء الكلام (اللغة) معا"، وذلك لأن هذه الواجهات تسمح لعامة الناس أن يتواصلوا مع الحاسب حتى لو كانوا لا يعرفون أي شيء عن أساسيات استخدامه وإدارته وذلك باستخدام تقنيات التعرف على الكلام [3].

4-1 الصعوبات التي تواجه أنظمة التعرف على الكلام [4]:

لقد استحوذت أبحاث الصوت منذ اختراع الحاسوب على اهتمام الكثيرين من العلماء والباحثين والطلاب والهواة، وخاصة بعد ظهور تقنيات الوسائط المتعددة، لما لها من أهمية كبيرة على جميع الأصعدة وفي كثير من نواحي الحياة، وقد تم التوصل إلى الكثير من تطبيقات الصوت من مؤثرات وترشيح وتنقية . . . . الخ.

كانت الحاجة إلى جعل الحاسوب يماثل البشر في استجابته للصوت من إدراك معاني الكلام ومن التفريق بين الأصوات الشغل الشاغل للكثيرين، إلا أن غالب تلك الأبحاث حتى الآن تبوء بالفشل بسبب التعقيدات في تركيب أصوات البشر ، وبسبب الضجيج الآتي من كل جانب .

وقد تبنّت بعض الشركات أمثال IBM و Dragon البحث في هذا المجال فقامت بإنتاج برامج تقوم بمهمة التعرف على الكلام ، ومن هذه البرامج برنامج VoicePad وبرنامج IBM VoiceType وبرنامج Dragon ، لكن تلك البرامج لم تكن تحقق نسبا" عالية" من التعرف، وما يبرر ضعف نسبة التعرف بالرغم من طول زمن التطوير هو الضجيج المرافق للصوت والذي له الكثير من المصادر نذكر أهمها [8]:

- الضجيج الصوتي : له مصادر عديدة – حركة المرور، والناس، الخ
- الضجيج الراديوي.
- تشويه القناة: اختلاف جودة الهاتف.
- صدى الصوت: بين سماعة وميكروفون.
- الترددات العالية ضمن الحاسوب.
- كرت الصوت .

يقدم هذا البحث دراسة عن الضجيج الصوتي الصادر عن الناس ويدعى بضجيج الثرثرة babble noise حيث يعتبر أخطر أنواع الضجيج في أنظمة التعرف على الكلام المطبقة في الزمن الحقيقي. يتم خلق ضجيج الثرثرة بشكل مفتعل، وتبيان نسبة التعرف بعد الحد من هذا الضجيج في محاولة لتحسينها عن طريق تطبيق الطرح الطيفي الذي يقوم بتقليل إشارة الضجيج .

4-2- نسبة الإشارة إلى الضجيج (SNR) Signal-to-noise ratio :

نسبة الإشارة إلى الضجيج هي مقياس لكمية الضجيج الحالي لإشارة، وتقاس في ديسيبل ، وحددت بالعلاقة [26]:

$$SNR(dB) = 10\log_{10}\left(\frac{P_{speech}}{P_{noise}}\right) \qquad (7)$$

طاقة الكلام تقاس بالعلاقة [26]:

$$P_{speech} = \frac{1}{N}\sum_{n=0}^{N-1} x(n)^2 \qquad (8)$$

طاقة الضجيج تقاس بالعلاقة [26]:

$$P_{noise} = \frac{1}{N}\sum_{n=0}^{N-1} d(n)^2 \qquad (9)$$

لذلك يمكن كتابة النسبة بالشكل التالي [26]:

$$SNR(dB)=10\log_{10}\left(\frac{\frac{1}{N}\sum_{n=0}^{N-1} x(n)^2}{\frac{1}{N}\sum_{n=0}^{N-1} d(n)^2}\right)=10\log_{10}\left(\frac{\sum_{n=0}^{N-1} x(n)^2}{\sum_{n=0}^{N-1} d(n)^2}\right) \qquad (10)$$

4-3 آلية توليد نسبة الاشارة الى الضجيج – خلق ضجيج مفتعل – [26]:

كن استخدام المعادلة (10) لخلط الكلام والضجيج لإنشاء SNRخاص ، تعتبر آلية التحكم مفيدة لأغراض الاختبار والتقييم.

وبالنظر إلى كل من الإشارات : الكلام x(n) ، والضجيج d(n) ، نريد أن نمزجها لخلق SNR خاص حيث يتم تغيير قيمة الضجيج المطبق بتغيير قيمة المعامل ألفا الذي يربط بين طاقة الكلام وطاقة الضجيج، لذلك تظهر الحاجة لتوسيع نطاق كل عينة ضجيج عن طريق معامل ألفا **α** للحصول على SNR المطلوب فتصبح العلاقة بالشكل التالي[26] :

$$SNR(dB)=10\log_{10}\left(\frac{\frac{1}{N}\sum_{n=0}^{N-1}x(n)^2}{\frac{1}{N}\sum_{n=0}^{N-1}[\alpha d(n)]^2}\right) \qquad (11)$$

يجب إعادة ترتيب المعادلة السابقة للحصول على معامل الفا **a** بالاستفادة من SNR, P_{speech} , P_{noise} عن طريق الخطوات التالية [26]:

- التقسيم على العدد 10:

$$\frac{SNR}{10}=\log_{10}\left(\frac{\frac{1}{N}\sum_{n=0}^{N-1}x(n)^2}{\frac{1}{N}\sum_{n=0}^{N-1}[\alpha d(n)]^2}\right)$$

- أخذ لوغارتم الطرفين:

$$10^{\frac{SNR}{10}}=\frac{\frac{1}{N}\sum_{n=0}^{N-1}x(n)^2}{\alpha^2\frac{1}{N}\sum_{n=0}^{N-1}d(n)^2}$$

- استبدال القوانين بقيم الطاقة المقابلة:

$$10^{\frac{SNR}{10}}=\frac{P_{speech}}{\alpha^2 P_{noise}}$$

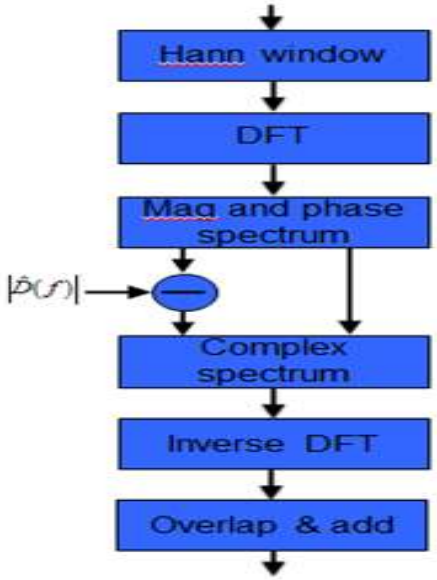
- عزل قيمة ألفا:

$$y(n) = x(n) + \alpha\, d(n) \longrightarrow$$

$$\alpha=\sqrt{\frac{P_{speech}}{P_{noise}}10^{-\frac{SNR}{10}}} \qquad (12)$$

4-4- إزالة الضجيج من الاشارة [27] [22]:

إنّ أكبر تأثير على دقة نظام التعرف على الكلام هو وجود الضجيج ، والذي يسبب أخطاء الحذف لأن النظام غير قادر على استخراج الخطاب الفعلي الممزوج مع الضجيج، ومن هنا علينا أن ننجز إلغاء الضجيج على التسجيلات للحد من أي ضجيج قدر الإمكان. يتم تقدير الضجيج وفق مراحل يبينها المخطط الصندوقي التالي:



الشكل (4) مراحل تقدير الضجيج وإزالته

خرج مراحل هذا المخطط مبين بالخطوات التالية:

- الإشارة بعد النافذة Hann window تكون بالشكل الزمني y(n)
- الإشارة تكون بالشكل الترددي Y(f) بعد تحويل DFT
- الإشارة تكون بالمجال الطيفي الذي يتمثل بالعلاقات [27]:

Magnitude spectrum $|Y(f)| = \sqrt{(Y(f).real)^2 + (Y(f).imag)^2}$

Phase spectrum $\theta_Y(f) = \tan^{-1}\left(\frac{Y(f).imag}{Y(f).real}\right)$

- نقوم بإجراء الطرح الطيفي على الإشارة [27]:

$$\left|\hat{X}(f)\right| = |Y(f)| - \left|\hat{D}(f)\right|$$

- نقوم بتحويل كل من Magnitude و Phase إلى الشكل العقدي [27]:

$$\hat{X}(f).real = \left|\hat{X}(f)\right|\cos(\theta_Y(f))$$

$$\hat{X}(f).imag = \left|\hat{X}(f)\right|\sin(\theta_Y(f))$$

- يحتاج تعزيز قوة تحويل الطيف magnitude spectrum لعودة الإشارة الى المجال الزمني ويتم باستخدام DFT المعكوس، لذلك يتم تطبيق معكوس DFT للطيف المعقد.

- نطبق التداخل والجمع على جميع نوافذ الإشارة لتحسين حجم الطيف magnitude spectrum في الطور الصاخب، وهذا لن يؤثر على الكلام لأن الأذن ليست حساسة لهذه المرحلة.

النتائج والمناقشة:

يبين الشكل المخطط الصندوقي لمراحل العمل في الشكل (5) :

الشكل (5) المخطط الصندوقي لمراحل العمل.

طيلة فترة البحث تم إجراء ما يقارب المئة اختبار من أجل ضبط دقيق لاستخراج السمات واختيار أفضل نموذج HMM وتحقيق دقة التعرف على أعلى مستوى ممكن. في النهاية ساعدت نتائج الاختبار ببناء نظام التعرف على الكلام الذي يحقق 92٪ من الدقة.

تمت المقارنة مع الدراسة [25] وذلك لأن البحث المقدم اعتمد على نفس قاعدة البيانات في هذه الدراسة، وهذا التوحيد يعتبر أساس نظم التعرف على الكلام في مقارنة النتائج.

تمت الدراسة المقترحة على 10000 عينة صوتية (كلامية) دون تحديد فئة عمرية، بيئة العمل المستخدمة Matlab2014a، تم استخدام مكتبات(voicebox , signal processing) .

تمت مرحلة التدريب باستخدام نماذج ماركوف المخفية HMM على مجموعة عينات التدريب لقاعدة البيانات .

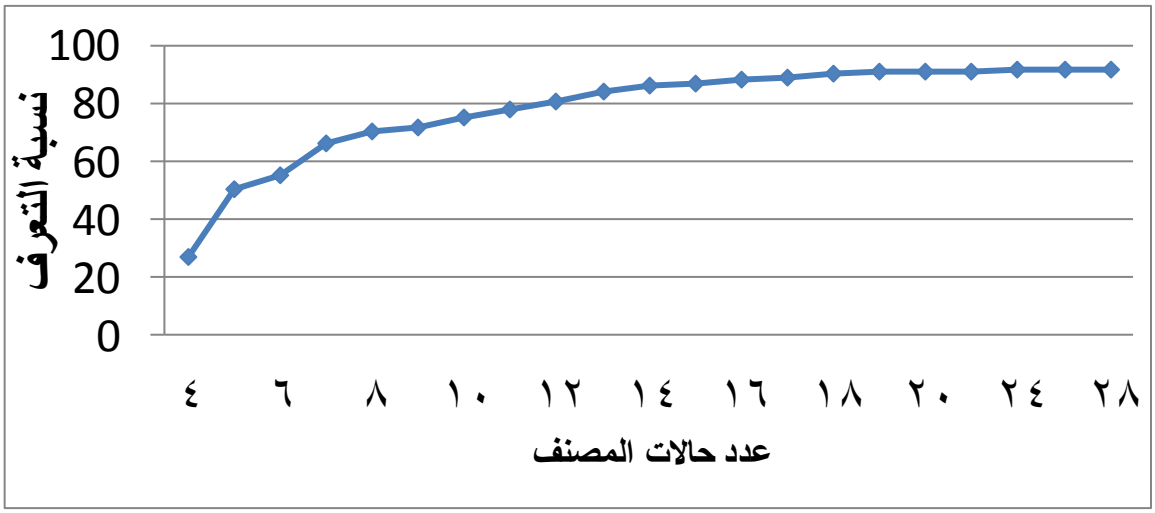
أضاف البحث عن الدراسة التي تمت المقارنة بها والتي تتبع لنفس ظروف غجراء البحث بسبب اعتماد نفس قاعدة البيانات بأنه تمت زيادة عدد المعاملات في مصفوفة الارتباط في الخوارزمية المدروسة، وتم اختيار عدد قنوات المرشح وعدد حالات المصنف من خلال الاختبارات التجريبية من أجل تحديد العدد الأمثل من الحالات، وقد أجريت

الاختبارات على 8 قيم لعدد قنوات المرشح وجد أن العدد الأمثلي هو 100 قناة ، وعدد حالات المصنف المثالي 24 حالة.

الجدول(1) يبين قيم نسبة التعرف الناتجة بوجود المعاملات وعدم وجودها.

الخوارزمية	عدد المعاملات في مصفوفة الارتباط بخوارزمية MFCC	عدد القنوات	عدد الحالات	نسبة التعرف
MFCC [25]	12 معامل	لم يذكر	لم يذكر	65%
MFCC بإضافة مكون الطاقة والمشتقات التفاضلية المقترحة	39 معامل	100	24	92%

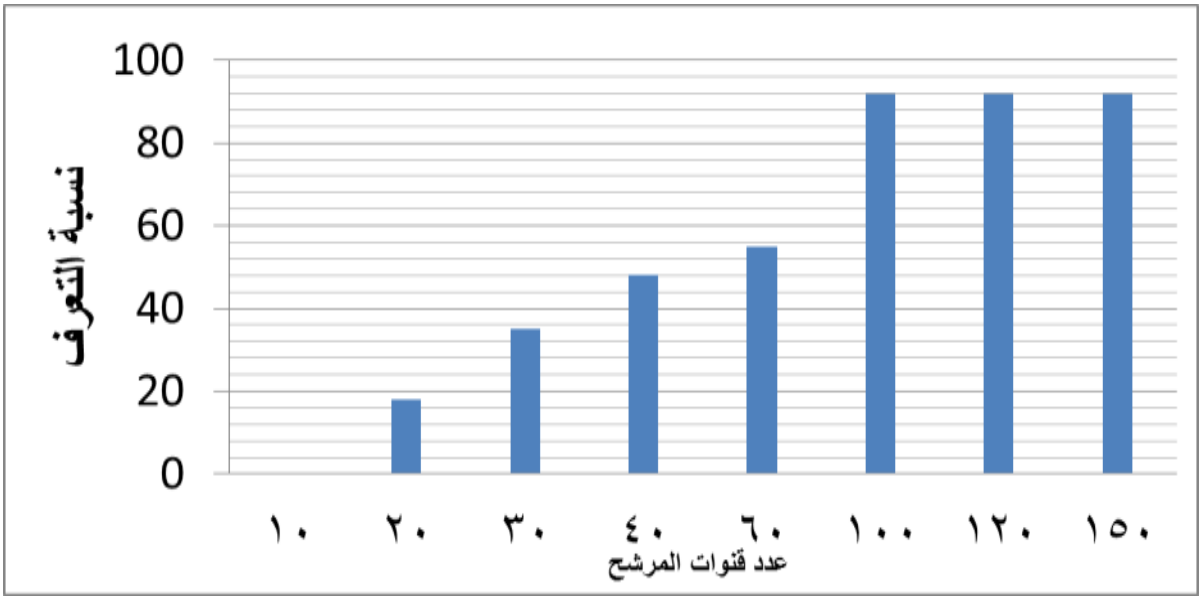
يبين الشكل(6) اختلاف نسبة التعرف باختلاف عدد الحالات في نموذج ماركوف المخفية وأن العدد الأمثلي هو 24 حالة.



الشكل (6) نتائج نسبة التعرف باختلاف عدد الحالات في نموذج ماركوف المخفية.

يمكن للأذن البشرية الكشف عن الاختلافات بين إشارات الترددات المنخفضة أفضل من تغيير النغمة pitch في الطيف الترددي العالي (ميل التردد Cepstral وهو معامل (MFCC))، وبالإضافة إلى ذلك فإن الترددات المنخفضة تحتوي على بيانات الجهاز الصوتي التي نستخدمها للتعرف على الكلام.

لذلك نحن بحاجة إلى مزيد من الدقة في الطيف الترددي المنخفض، وجدنا بأن زيادة عدد القنوات يقدم دقة مضمنة ساعدت في الكشف عن التغيرات في الترددات المنخفضة والتي أدت إلى نتائج أكثر دقة . وبناء على نتائج الاختبار فإن مرشح له 100 قناة أعطى أفضل النتائج لجميع تجاربنا كما هو مبين في الشكل (6) .



الشكل (7) نتائج نسبة التعرف باختلاف عدد القنوات في المرشح.

إن إدخال الضجيج المفتعل لكل من عينات التدريب والاختبار يخفض نسبة التعرف على الكلام، أظهرت النتائج أن الضجيج babble غير ثابت يقلل من أداء التعرف على الكلام بشكل كبير إلى نحو 30٪ في SNR10 كما هو مبين في الشكل (8).

الشكل (8) نتائج نسبة التعرف باختلاف نوع الضجيج المطبق.

زاد تقدير الضجيج أداء التعرف عند تطبيقها على كل من بيانات الاختبار والتدريب المقدمة، وتوصل البحث أنه في حالة بيانات التدريب نظيفة وبيانات الاختبار نظيفة مع تطبيق هذا التقدير حققنا دقة تصل ل 92٪. ولكن عندما يطبق على مزيج من بيانات التدريب نظيفة وبيانات الاختبار ضجيجية نحصل على نتائج متباينة .

هذا التقدير يقدم تحسين منخفض عندما يطبق على ملفات SNR المنخفضة ويحسن النتائج مع ملفات SNR العالية، حيث ارتفاع SNR يحسن هذا التقدير وفعلا" يقوم بتحسين الأداء . دقة التعرف زادت بنسبة 4٪ لبيانات الاختبار ذات الضجيج (babble noise) عند النسبة 10SNR.

الشكل (9) نتائج نسبة التعرف بتطبيق الطرح الطيفي.

تقدير الضجيج يعمل بشكل أفضل على الضجيج الثابت ولكنه عندما طبق على ضجيج الثرثرة قدم ضجيج موسيقي إضافي، وله أثر يذكر للحد من الضجيج الأصلي، ولإتمام الاختبارات تم رفع الحد الأقصى لعتبة التوهين من 0.00 إلى القيمة 0.02.

الاستنتاجات والتوصيات:

تم التوصل في هذا البحث إلى:

- زيادة عدد السمات وتحديد العدد الأمثلي لكل من عدد حالات المصنف وعدد قنوات المرشح لنحصل على نسبة تصل الى 92% في بيئة مثالية.
- دراسة الضجيج لأنه من أهم العوامل التي تخفض هذه النسبة ومراقبة اختلاف نسبة التعرف باختلاف الضجيج المطبق.
- تطبيق الطرح الطيفي أحد الحلول الشائعة لتخفيض أثر الضجيج، لكنه أدى لضجيج موسيقي كان أخطر من الضجيج المعالج فلذلك قمنا برفع قيمة عتبة التوهين لإمكانية إجراء الاختبارات

يمكننا أن نقدم توصيات تتمثل ب:

- إجراء مزيد من التجارب والتقييم مع بارامترات أخرى لخوارزمية السمات التي يمكن أن تؤدي إلى نتائج مثيرة للاهتمام.
- الاختبار على عينات ضجيج أخرى لمراقبة أثرها على نسبة التعرف (factory and machine gun).

- وجدنا حسب الدراسة المقدمة أنه عند الحد من الضجيج في كل أنظمة التعرف على الكلمات باستخدام الطرح الطيفي تؤدي لأنواع ضجيج أخرى كالضجيج الموسيقي، وهذه الأنواع من الضجيج تؤثر على نسبة التعرف في النظام، لذلك يمكن الاتجاه لإدخال عوامل أخرى لتحسين التعرف على الكلام كالأنظمة التي تعتمد على السمات البصرية كإدخال حركة الشفاه التي لا تتأثر بالضجيج السمعي.

المراجع

[1] Marius Zbancioc, Mihaela Costin :*using neural networks to improve speech recognition*, International IEEE SCS Conference, Proceedings, Vol. 1, 2003 EX 720, pp. 445 – 448.

[2] Levy, C., Linares, G., Nocera, P., Bonastre, J.-F. : *Reducing computational and memory cost for cellular phone embedded speech recognition system*, Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on (Volume:5) , pages(309-12(vol.5 , Print ISBN:9-8484-7803-0.

[3] Dimitriadis, Maragos, P. Potamianos*: Robust AM-FM Features for Speech Recognition*, IEEE signal processing letters, VOL. 12, NO. 9, 2005.

[4] Takami Yoshida, Kazuhiro Nakadai, and Hiroshi G. Okuno: *Automatic Speech Recognition Improved by Two-Layered Audio-Visual Integration For Robot Audition*, 9th IEEE-RAS International Conference on Humanoid Robots December 7-10, 2009 Paris, France.

[5] Lavneet Singh, Girija Chetty *A Comparative Study of Recognition of Speech Using Improved MFCC Algorithms and Rasta Filters*, Information Systems, Technology and Management Communications in Computer and Information Science Volume 285, 2012, pp 304-314.

[6] Aleksander Pohl, Bartosz Ziółko *: Using Part of Speech N-Grams for Improving Automatic Speech Recognition of Polish*, Machine Learning and Data Mining in Pattern Recognition Lecture Notes in Computer Science Volume 7988, 2013, pp 492-504.

[7] BEN FRED , Kaïs OUN *: Phoneme Recognition using Hidden Markov Models* , International Journal of Control, Energy and Electrical Engineering (CEEE) , vol.1, pp.57-61, 2014.

[8] Deividas Eringis, Gintautas Tamulevičius*: Improving Speech Recognition Rate through Analysis Parameters*, Electrical, Control and Communication Engineering. Volume 5, Issue 1, Pages 61–66, ISSN (Online) 2255-9159, May 2014.

[9] Annika Hämäläinen, Hugo Meinedo, Michael Tjalve, Thomas Pellegrini, Isabel Trancoso, Miguel Sales Dias*: Improving Speech Recognition through Automatic Selection of Age Group – Specific Acoustic Models*, Computational Processing of the Portuguese Language Lecture Notes in Computer Science Volume 8775, pp 12-23, 2014.

[10] Homayoon Beigi: *Fundamentals of speaker Recognition*-Springer Sience 2011-ISBN: 978-0-387-77591-3.

[11] Neustein, Amy; Patil, Hemant: *Forensic speech Recognition* –A. (Eds) – Springer 2012-ISBN 10: 146140262X/ ISBN 13:9781461402626.

[12] K.K PaliWal: *Advances in speech, Hearing and Language Processing*, Volume1, pages 1-78, 1990, ISBN:1-55938-210-4.

[13] Santosh K.Gaikwad, Bharti W.Gawali, Pravin Yannawar: *A Review on Speech Recognition Technique,* International Journal of Computer Applications (0975 – 8887), Volume 10– No.3, November 2010.

[14] Pitz, M.; Schluter, R; Ney, H. Molau, S., *Computing Mel-frequency cepstral coefficients on the power spectrum*, Print ISBN: 0-7803-7041-4 INSPEC Accession Number: 7120280 Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on (Volume: 1) Page(s): 73 - 76 vol.1

[15] Namrata Dave, *Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition,* international journal for advance research in engineering and technology,Volume 1, Issue VI, July 2013.

[16] H. Hermansky, N. Morgan, A. Bayya, P. Kohn: *RASTA-PLP speech analysis techniqe* , IEEE International Conference on , 1992 , pp: 121-124.

[17] Lukas Burget :*Measurement of complementarity of Recognition Systems* , Springer-Verlag Berlin Heidelberg ,ISBN 3-540-230421,pages(283-288),2004.

[18] Mel Frequency Cepstral Coefficient (MFCC) tutorial. (n.d.). Retrieved November 9, 2015, from http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs.

[19] S. B. Davis and P. Mermelstein, "*Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.Standardised by ETSI Aurora group in 2000

[20] T. W. Parsons, *Voice and Speech Processing*;. New York; McGraw Hill 1986

[21] Gunter, S., & Bunke, H. (2003). *Optimizing the number of states, training iterations and Gaussians in an HMM-based handwritten word recognizer. Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* Retrieved November 9, 2015, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.58.6366&rep=rep1&type=pdf

[22] Noise reduction. (2006). Retrieved November 10, 2015, from http://sound.eti.pg.gda.pl/denoise/noise.html

Mel Frequency Cepstral Coefficient (MFCC) tutorial. (n.d.). Retrieved November 9, 2015, from http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[23]README-HTK-AUDIO.(n.d.).RetrievedNovember9,2015,from http://mi.eng.cam.ac.uk/projects/sacti/corpora/SACTI-1/utterance-audio/README HTKAUDIO.TXT

[24] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., . . . Woodland, P. (2006). The HTK book (Version 3.4 ed.). Cambridge: Entropic Cambridge Research Laboratory.

[25] Alaa Sagheer, *Multimodal Arabic Speech Recognition for Human-Robot Interaction Applications*, Appl. Math. Inf. Sci. 9, No. 6, 2885-2897 (2015)

[26] POLLAK, P. *Estimation methods of speech signal-to-noise ratio*. ´ Acoustic Sheets, c.7 ˇ , 2001 (in Czech).

[27] Karel Palecek and Josef Chaloupka , *Audio-Visual Speech Recognition in Noisy Audio Environments*, IEEE , 2013 36th International Conference onTelecommunications and Signal Processing (TSP),Pages: 484 - 487, DOI: 10.1109/TSP.2013.6613979