الجمهورية العربية السورية
جامعة دمشق
كلية الهندسة المعلوماتية
قسم الذكاء الصنعي ومعالجة اللغات الطبيعية

نظام تحسين البث المعتمد على الذكاء الصنعي
AI Based Streaming Enhancing System

مشروع أُعدّ لنيل الإجازة في الهندسة المعلوماتية

إعداد الطلاب

احمد عبد الفتاح **احمد الطحان**

هبة قزويني سوس

بإشراف

د. مدحت الصوص

2021-2022

ملخص تجريدي Abstract

قمنا في هذا المشروع بإنجاز نظام لتحسين البث بالزمن الحقيقي يتضمن نظام لتنقية الصوت من الضجيج مما يحسن أداء تطبيقات معالجة الكلام، ونظام لتحقيق ضغط موجه بناءً على الحركة أو على الأشخاص مما يساعد في توجيه كم البيانات إلى الأجزاء المهمة من الفيديو فقط. رأينا أيضاً كيف من الممكن أن تعطي الطرق الرياضية الخاصة بإزالة الضجيج والقائمة على تقدير طيف الضجيج نتائج مقبولة في ظروف الضجيج المتوسطة كما رأينا كيف من الممكن أن تساهم عمليات المعالجة المسبقة على الفيديو قبل ضغطه في تقليل حجمه. أخيراً توفر البيئة المقترحة قابلية عالية لإعادة الاستخدام مستقبلاً عن طريق تعريف أنظمة جديدة لمعالجة الصوت أو الفيديو.

جدول المحتويات Contents

الفصل الأول: مقدمة Introduction

يلعب الذكاء الصناعي اليوم دوراً كبيراً في معالجة العديد من المهام بطريقة أكثر تكيفاً وتخصيصاً كما يسهل توظيفه في العمليات التي تتطلب بداهة عالية في المعالجة وسرعة أكبر في اتخاذ القرار في الحصول على نتائج مقاربة لنتائج البشر نذكر أيضاً أنه ساهم كذلك في تسهيل التواصل بين البشر والحاسوب حيث جعل هذا التواصل أكثر مرونة فأصبح البشر يشعرون براحة أكبر عند استخدام الحاسوب.

يعد التواصل المرئي بين البشر من أغنى وسائل التواصل بالمعنى وأكثرها قدرة على نقل التصور الصحيح بين الملقي والمتلقي إلا أن نقل عبر قنوات الاتصال المختلفة كان له العديد من المشاكل من أبرزها التباين الكبير في جودة البنية التحتية لقنوات الاتصال بين المتلقين، واختلاف اللغة المحكية بين الملقي والمتلقي لذا كان لا بد من أيجاد أساليب أكثر ذكاءً في استقبال هذه البيانات تضمن تكوين تصور مطابق للتصور الموجود ببال الملقي وتساهم في تقليل متطلبات استقبالها.

نهدف في هذا المشروع إلى تسهيل هذه العملية حيث سنعمل على تحقيق ضغط موجه للفيديو مما يؤدي إلى كلفة ومتطلبات أقل في استقبال البيانات عند المتلقي كما سنعمل على تحسين جودة الصوت المنقول وإزالة ما يشوه هذه الإشارة الصوتية من ضجيج أخيراً نهدف إلى تحقيق ترجمة للكلام المحكي من قبل الملقي إلى نص بلغة المتلقي كل هذه الميزات يجب تحقيقها بالزمن الحقيقي لكي يكون لهذا النظام توظيف واقعي خصوصاً في تطبيقات التواصل المباشر والفوري وتعد المعالجة بالزمن الحقيقي أبرز قيمة هندسية مضافة للمشروع المذكور.

يعتبر المشروع مستند على مفاهيم من عدة تخصصات حيث نجد في القسم الخاص بالضغط الموجه للفيديو مفاهيم بالرؤية الحاسوبية ونظرية المعلومات وفي القسم الخاص بتحسين جودة الصوت مفاهيم معالجة الإشارة أما في القسم الخاص بالترجمة مفاهيم خاصة بمعالجة اللغات الطبيعية يحيط بكل ذلك مفاهيم تقنية عديدة حول برتوكولات البث عبر الشبكة.

يظهر هذا التقرير كل الخطوات المتبعة لتحقيق النظام حيث بدأنا بالفصل الأول بذكر دراسة مرجعية موسعة تظهر أبرز المنهجيات المتبعة في تحقيق أجزاء النظام وأبرز الأعمال المشابهة لها.

الفصل الثاني: الدراسة المرجعية State of the art

نستعرض في هذا الفصل التعريف العلمي للمسائل المطلوب معالجتها للنظام ونذكر بعدها أبرز المنهجيات المتبعة لتحقيقها؛ حيث نبدأ أولاً بتعريف مشكلة الضوضاء في الإشارات الرقمية وأنواعه ثم تعريف لتقنية الضغط الموجه للفيديو وأنواع بروتوكولات البث عبر الشبكة وأخيراً نذكر أبرز المنهجيات المتبعة لتحقيق كل منها يمكن للقارئ تجاوز هذا الفصل في حال عدم اهتمامه بالأعمال المنجزة مسبقاً واكتفائه بقراءة ما يهم.

2.1 الضوضاء في الإشارات الرقمية The noise in digital signals

2.1.1 تعريف

يمكن تعريف الضوضاء بأنها إشارة غير مرغوب فيها تتداخل مع اتصال أو قياس إشارة أخرى، يعتبر الضوضاء في حد ذاته إشارة تنقل المعلومات المتعلقة بمصدر الضوضاء، كما تعتبر العوامل الرئيسية التي تحد من قدرة نقل البيانات في الاتصالات والدقة في أنظمة القياس لذلك يعتبر إخماد الضوضاء وإزالة التشويه من القياسات من أهم مسائل معالجة الإشارة، يفيد حل هذه المسألة في التحسين على العديد من التطبيقات مثل أنظمة الاتصالات الخليوية، التعرف على الكلام، استعادة التسجيلات القديمة، المؤتمرات عن بعد، معالجة الصور، معالجة الإشارات الطبية والقياسات الحيوية، ومعالجة إشارات الرادارات والحساسات. يكمن التحدي في هذه المسألة في إزالة الضوضاء من الإشارة التي تحوي ضوضاء بدون التأثير على الإشارة الأصلية.

يمكن تعريف مسألة إزالة الضوضاء بشكل عام كمسألة تقدير التوزيع احتمالي شرطي[1] كالآتي:

$$P(x|\tilde{x}) \quad ; x, \tilde{x} \in \mathbb{R}^n$$

حيث x العينة النقية و $\tilde{x}$ العينة الصاخبة

نعرف في هذا سياق معالجة الإشارة مسألتين مهمتين[2] :

A. إلغاء الضوضاء Noise Cancellation

يتم عمل إلغاء الضجيج عند عملية القياس للإشارة وذلك بفرض كنا على علم بمصدر الضوضاء، على سبيل المثال في حال كنا نسجل صوت ونعلم مصدر الضجيج فيمكننا تركيب مسجلين واحد عند مصدر الضوضاء وواحد عند المتحدث ونحاول فصل الإشارة المسجلة من المسجل الأول عن الإشارة المسجلة من المسجل الثاني مع الأخذ بعين الاعتبار التخامد الحاصل بالضجيج وتأخيره، غالباً ما تعالج هذه المسألة على مستوى العتاد الصلب Hardware لجهاز التسجيل أو أثناء المعالجة حيث أصبحت مسجلات الصوت والميكروفونات حديثاً تأتي مدعمة بهذه الميزة بشكل مسبق، تعتبر هذه المسألة خارج سياق المشروع المذكور.

B. إخماد الضوضاء Noise Reduction

تتم عند عملية معالجة إشارة مقاسة مسبقاً حيث لا يمكننا إلا الوصول للإشارة الصاخبة (التي تم قياسها سابقاً وتحوي على ضوضاء) عندها لا نعلم مصدر الضوضاء وطبيعة الإشارة الخاصة فيه لذلك نعمل على اخمادها، تعتبر هذه المسألة أعقد من المسألة السابقة وهي المسألة المطروحة في مشروعنا لذا فالمسألة المطروحة بحالتنا هي مسألة إخماد ضجيج كوننا نتعامل مع إشارة صوتية مسجلة مسبقاً.

يمكن تعريف الإشارة الرقمية الصاخبة في فضاء الزمن من الطول L بالعبارة التالية:

$$y[k] = x[k] + n[k]$$

حيث y تمثل الإشارة الصاخبة (التي تحوي ضوضاء) وx تمثل الإشارة الأصلية النقية وn تمثل الضوضاء في الإشارة و k=0,1,..L دليل الزمن بافتراض استطعنا الحصول على تقريب للضوضاء $\hat{n}$ فيمكنا الحصول على تقريب للإشارة الأصلية (النقية) $\hat{x}$ كالتالي:

$$\hat{x}[k] = y[k] - \hat{n}[k]$$

نؤكد هنا أن الإشارة النقية وكذلك الضجيج من نفس طول الإشارة الصاخبة L يعبر عن ذلك بالعلاقة $y, x, n \in \mathbb{R}^L$ تعتبر عملية إزالة الضوضاء في فضاء الزمن عملية غير مجدية لذا غالباً ما نلجأ للمعالجة بفضاء التردد الزمني فيمكن تعريف الإشارة الرقمية الصاخبة فضاء التردد الزمني بعد استخدام تحويل فوريه قصير الزمن STFT لتحويل الإشارة من فضاء الزمن إلى فضاء التردد الزمني بالعبارة التالية:

$$Y(m,n) = X(m,n) + N(m,n)$$

حيث m تمثل دليل التردد وn تمثل دليل الكتلة

من خلال تطبيق تابع خاص f لتقليل الضوضاء على كتلة الإشارة الصاخبة Y، يمكن تقدير كتلة STFT النقية $\hat{X}$ كالتالي:

$$\hat{X}(m,n) = f\big(Y(m,n)\big)$$

يمكن استبدال التابع السابق f بمرشح خطي H فيمكن كتابة العلاقة السابقة كالتالي:

$$\hat{X}(m,n) = H(m,n)\ \ Y(m,n)$$

تقنياً غالباً ما يتم تقسيم طرق تقدير الضوضاء إلى نوعين طرق تقدير قطرية Diagonal estimation وطرق تقدير غير قطرية non-Diagonal estimation حيث تقوم طرق تقدير الضوضاء القطرية بمعالجة كل كتلة بشكل مستقل وتقدير الضوضاء فيها ومن سيئات هذه الطريقة هو توليد ما يسمى بالضوضاء الموسيقية لذا فتعتبر فعاليتها محدودة. تقوم طرق المعالجة غير القطرية بحل هذه المشكلة لذا تعتبر طرق أكثر فعالية[3].

2.1.2 أنواع الضوضاء

يمكن تصنيف الضوضاء بناءً على مصدرها إلى عدة أصناف [2]:

A. الضوضاء الصوتية (الضجيج) Acoustic noise

تنشأ من تحرك المصادر أو اهتزازها أو تصادمها وهو أكثر نوع مألوف من الضوضاء. تعتبر هذه الضوضاء موجودة بدرجات مختلفة في البيئات اليومية. تتولد عادةً من مصادر مثل السيارات المتحركة ومكيفات الهواء ومراوح الكمبيوتر، حركة المرور، الأشخاص الذين يتحدثون في الخلفية، الرياح، المطر ،...إلخ.

B. الضوضاء الحرارية وضوضاء الإطلاق Thermal noise and shot noise

تحدث الضوضاء الحرارية نتيجة الحركة العشوائية للجسيمات المنشّطة حرارياً في الموصل الكهربائي وتعتبر موجودة في أي موصل كهربائي بدون تطبيق جهد عليه. أما بالنسبة لضوضاء الإطلاق فهي تنشأ من تقلبات عشوائية للتيار الكهربائي حيث تحدث بسبب حقيقة أن التيار يتم بواسطة شحنات منفصلة (أي الإلكترونات) مع تقلبات عشوائية وأوقات وصول عشوائية.

C. الضوضاء الكهرومغناطيسية Electromagnetic noise

موجود في جميع الترددات ولا سيما في نطاق الترددات الراديوية (نطاق kHz إلى GHz) حيث تحدث الاتصالات عن بعد. جميع الاجهزة الكهربائية مثل أجهزة الإرسال والاستقبال الإذاعية والتلفزيونية، تولد ضوضاء كهرومغناطيسية.

D. الضوضاء الكهربائية Electrostatic noise

تنتج عن وجود جهد مع أو بدون تدفق تيار.

E. تشوهات القناة، الصدى، والتلاشي Channel distortions, echo and fading

بسبب الخصائص غير المثالية لقنوات الاتصال. القنوات الراديوية، مثل تلك الموجودة على ترددات GHz التي يستخدمها مشغلو الهواتف المحمولة الخلوية، حساسة بشكل خاص لخصائص الانتشار لبيئة القناة وتلاشي الإشارات

F. ضوضاء المعالجة Processing noise

الضوضاء الناتجة عن المعالجة الرقمية للإشارات، على سبيل المثال الضوضاء الكمومية في التشفير الرقمي لإشارات الكلام أو الصور، أو حزم البيانات المفقودة في أنظمة اتصالات البيانات الرقمية.

كما يمكن اعتمادًا على طيف التردد أو خصائص الوقت، تصنيف الضوضاء كما يلي[2]:

A. الضوضاء البيضاء White noise

ضوضاء عشوائية بحتة لها طيف طاقة مسطح. تحتوي الضوضاء البيضاء نظريًا على جميع الترددات بنفس الشدة.

B. الضوضاء بيضاء محدودة النطاق Band-limited white noise

ضوضاء ذات طيف مسطح وعرض نطاق محدود يغطي عادةً النطاق المحدود للجهاز أو الإشارة محل الاهتمام.

C. الضوضاء ذات النطاق الضيق Narrowband noise

عملية ضوضاء ذات عرض نطاق ضيق مثل "همهمة" 50-60 هرتز من الكهرباء.

D. الضوضاء الملونة Colored noise

ضوضاء غير بيضاء أو أي ضوضاء ذات نطاق عريض يكون طيفها غير مسطح الشكل؛ ومن الأمثلة على ذلك الضوضاء الوردية والضوضاء البنية وضوضاء الانحدار الذاتي.

E. الضوضاء النبضية Impulsive noise

تتكون من نبضات قصيرة المدى ذات سعة عشوائية ولمدة الزمنية عشوائية.

F. نبضات الضوضاء العابرة Transient noise pulses

تتكون من نبضات ضوضاء طويلة الأمد نسبيًا.

2.2 ضغط الفيديو حسب المحتوى Content aware video encoding

2.2.1 تعريف

بشكل عام يعتبر ضغط الفيديو مسؤول عن نوعين من التكرار: التكرار المكاني وهو التكرار بين البكسلات والتكرار الزمني وهو التكرار بين الاطارات. يتم تحقيق ضغط التكرار المكاني من خلال تقنية DCT وتقنيات الترميز الكمي بينما في التكرار الزمني فيتم تحقيق الضغط فيه من خلال تقنيات تقدير الحركة motion estimation.

مع ارتفاع الطلب على تقنيات الاجتماع عبر الإنترنت والبث المباشر؛ أُطلِقت عدة طرائق للبث عبر الإنترنت. من أشهر هذه التقنيات: HTTP Dash، وHTTP Live Stream HLS. إذ يضغط المخدمُ الفيديو إلى عدة دقات مختلفة، ويقدّم الدقة المناسبة للمستخدم عبر ملف Manifest.ومن أجل ضغط هذه الملفات إلى مستويات دقة مختلفة، ولتوفير أماكن الوصول إلى الفيديو عمومًا أُجريت في العقود الثلاثة الأخيرة عدةُ قفزات في مجال ضغط الفيديو ومن أول خوارزميات الـ codec (وهي كلمة مختصرة عن compression وdecompression) المستخدمة بكثرة كانت خوارزمية H.120، ثم طُوّرت إلى الخوارزمية الشهيرة MPEG-2 H.262، ومن ثم H.264/AVC. وقد أُطلِق مؤخرًا (Versatile Video Coding VVC) الذي يفترض أن يخفف الـ Bit Rate (معدل دفق البيانات) بنسبة 30%-50% من الوضع الحالي مع HEVC/H.265 مما يؤثر إيجابًا في حجم الملف النهائي. وعلى نحو منفصل أنتج the Alliance for Open Media خوارزميتهم الخاصة "AV1" وأصدروها بصيغة ملف مفتوح المصدر (Open Source).

2.2.2 بروتوكولات البث عبر الانترنت

معظم ملفات الفيديو غير مصممة للبث فيجب تحويلها إلى ملفات قابلة للبث فيتم تقسيمها إلى اجزاء صغيرة بحيث تصل هذه الاجزاء بالتسلسل ويتم تشغيلها فور وصولها.

للقيام ببث حي نحن بحاجة الى بروتوكول بث معين، تقوم هذه البروتوكولات بتسهيل عملية نقل البيانات من برنامج الى اخر من خلال تقسيم الفيديو الى اجزاء وإرساله الى المشاهد ثم إعادة تجميعه. بحيث يتم نقل ملف الفيديو من والى المشفر ومضيف البث ثم الى مشغل الفيديو حيث تتم مشاهدته.

بروتوكولات البث الاكثر شيوعاً:

A. HTTP Live Streaming (HLS)

أصدرت Apple هذا البروتوكول حيث تدعم كل من متصفحات سطح المكتب وأجهزة التلفزيون الذكية والأجهزة المحمولة التي تعمل بنظام (Android) او نظام HLS (iOS) ومشغلات فيديو HTML5 أيضًا تدعمه بشكل أصلي. يدعم هذا البروتوكول تقنية (adaptive-bitrate) التي توفر أفضل جودة للفيديو وبرنامج ترميز H.265 الذي يوفر ضعف جودة الفيديو بنفس حجم ملف H.264. من نقاط ضعف هذا البروتوكول (high latency) بحيث يكون زمن التأخر مرتفع نسبيا ولكن هناك طرق لتقليله.

B. Real-Time Messaging Protocol (RTMP)

قامت Macromedia بتطوير هذا البروتوكول يُستخدم لإيصال الفيديو فقط حيث يصل الفيديو إلى المنصة التي يقوم المستخدم ببث الفيديو عليها عبر بروتوكول RTMP ثم يصل إلى المشاهد عبر بروتوكول آخر عادةً HLS. يوفر هذا البروتوكول (Low latency) بحيث يكون زمن التأخير منخفض ويوفر هذا للمشاهدين "lag" تأخيرات أقل عند مشاهدة مقاطع الفيديو مع اتصال إنترنت ضعيف مما يسمح لهم باستئناف البث بسرعة بمجرد استقرار اتصالهم بالإنترنت ولكن مشغلات فيديو HTML5 لا تدعم RTMP.

C. Web Real-Time Communications (WebRTC)

من الناحية التقنية WebRTC مشروع للبث وليس بروتوكول للبث ولكن غالبا يتم تجميعها مع البروتوكولات الشائعة للبثوث قامت VoIP بتطوير هذا المشروع. يعتبرWebRTC مهما للبثوث التي تتطلب (real-time latency) زمن تأخر في الوقت

الفعلي. ومن اهم استخدامات WebRTC (Peer-to-peer streaming) التي يطلق عليها (web conferencing).ومن البرامج والتطبيقات التي تستخدم Snapchat WebRTC و Facebook و WhatsApp ومنصات تواصل اجتماعي اخرى تدعم دردشة الفيديو.

D. Secure Reliable Transport (SRT)

يعتبر بروتوكول جديد نسبيا اطلقته Haivision. يتميز SRT بالامان والوثوقية وزمن تأخر منخفض ولكن توجد حاليًا بعض القيود على البث باستخدام SRT نظرًا لأن أجهزة وبرامج البث الأخرى لم يتم تطويرها بعد لدعم هذا البروتوكول.

E. Real-Time Streaming Protocol (RTSP)

يشبه RTSP في بعض النواحي بروتوكول HTTP Live Streaming (HLS)، لا يقوم RTSP بنقل بيانات البث المباشر ليس ما تنجزه RTSP بمفرده حيث تعمل خوادم RTSP غالبًا جنبًا إلى جنب مع بروتوكول النقل في الوقت الحقيقي (RTP) وبروتوكول التحكم في الوقت الحقيقي (RTCP) لنقل البثوث. يتميز RTSP تقنية التخصيص من خلال استخدام بروتوكولات اخرى مثل بروتوكول التحكم في الإرسال (TCP) وبروتوكول مخطط بيانات المستخدم (UDP).يعتبر RTSP من البروتوكولات الاقل شيوعا حيث اغلب مشغلات الفيديو لا تدعمه.

F. Dynamic Adaptive Streaming over HTTP (MPEG-DASH)

يدعم MPEG-DAS تقنية (adaptive-bitrate) التي توفر للمشاهدين أفضل جودة للفيديو يمكن أن تدعمها سرعة اتصالهم بالإنترنت ويميل هذا إلى التذبذب من ثانية إلى ثانية يمكن لـ DASH مواكبة ذلك.لا تدعمه اجهزة Apple ومن ميزاته انه لا يحدد نظام ترميز معين.

يستخدم YouTube بشكل أساسي تنسيقات الفيديو VP9 وH.264, MPEG-4 AVC, AV1 وتقنية (adaptive-bitrate) عبر بروتوكول HTTP Live Streaming (HLS) وللاجهزة المحمولة ترسل خوادم Youtube البيانات عبر RTSP وهو بروتوكول طبقة التطبيق وفي طبقة النقل يستخدم RTSP كلاً من TCP وUDP.

2.3 أعمال ذات صلة Related works

2.3.1 توابع إخماد الضوضاء Noise reduction functions

تم اقتراح العديد من الطرق المحتملة لإيجاد التابع f المعرف بالعلاقة في الأدبيات الخاصة بمعالجة الإشارة. أهم هذه الطرق هي مرشح واينير Wiener filter[4]، الطرح الطيفي spectral subtraction[5]، وفلتر افرايم ومالاه the Ephraim and Malah filter[6]

2.3.1.1 مرشح واينير Wiener filter

يقوم مرشح Wiener بتعديل مطال الإشارة الصاخبة وفقًا لتقدير نسبة الإشارة إلى الضوضاء عند كل تردد. يتطلب ذلك تقدير مصفوفة ترابط بين للإشارة والضوضاء يمكن أن يكتب تابع التعديل كالتالي:

$$H = \begin{cases} \dfrac{|Y|^2 - S_N}{|Y|^2} & ,|Y|^2 > S_N \\ 0 & ,otherwise \end{cases}$$

حيث S_N هو طيف الطاقة الحالي وما يسمى أيضاً ببصمة الضوضاء وتم حذف الدلائل n,m لتبسيط العبارة وتعد هذه الطريقة من طرق التقدير القطرية.

2.3.1.2 الطرح الطيفي *Spectral subtraction*

في الطرح الطيفي، يتم الحصول على تقدير السعة الطيفية للإشارة عن طريق طرح تقدير السعة الطيفية للضوضاء من تلك الخاصة بالإشارة الصاخبة. يكتب تابع التعديل كما يلي:

$$H = \begin{cases} \dfrac{|Y| - \sqrt{S_N}}{|Y|} & ,|Y|^2 > S_N \\ 0 & ,otherwise \end{cases}$$

على الرغم من الطرق السابقة يمكن أن توفر انخفاضًا كبيراً في ضوضاء الخلفية للإشارات الصوتية لكن يوجد العديد من الخصائص غير المواتية في التطبيقات العملية. العيب الرئيسي هو ظهور ما يسمى بالضوضاء موسيقية وتعد هذه الطريقة أيضاً من طرق التقدير القطرية.

2.3.1.3 مرشح افرايم ومالاه *Ephraim and Malah filter*

من المعروف أن طريقة إفرايم ومالاه تنتج قدرًا أقل من الضوضاء الموسيقية نظراً لأن تابع إخماد الضجيج يعتمد إلى حد أقل على التغيرات الزمنية في طيف الوقت القصير للصوت إشارة لذا تعتبر طريقة تقدير غير قطرية.

حساب قاعدة الإخماد عادةً يعتمد على a priori SNR:

$$\xi(m,n) = \frac{\lambda_X(m,n)}{\lambda_N(m,n)}$$

ونعرف الـ a posteriori SNR كالتالي:

$$\gamma(m,n) = \frac{|Y(m,n)|^2}{\lambda_N(m,n)}$$

حيث λ_N , λ_X يمثلان طاقة الكثافة الطيفية للضجيج والإشارة النقية على الترتيب

والشعاع المساعد ρ كالتالي:

$$\rho(m,n) = \frac{\xi(m,n)}{\xi(m,n)+1}\gamma(m,n)$$

فيكون تابع الإخماد كالتالي:

$$H(m,n) = \frac{\sqrt{\pi\,\rho(m,n)}}{2\,\gamma(m,n)}\left[\left(1+p(m,n)\right)I_0\left(\frac{\rho(m,n)}{2}\right)+\rho(m,n)\,I_1\left(\frac{\rho(m,n)}{2}\right)\right]e^{\frac{-\rho(m,n)}{2}}$$

حيث I_i تمثل تابع Bessel المعدل من الرتبة i

تعتبر العلاقة السابقة مكلفة حسابياً لذا تم إيجاد تقريب لها يوفر سرعة أكبر في الحساب وأداء مساوي تقريبا[11] مثل:

$$H(m,n) = \sqrt{\frac{\xi(m,n)}{1+\xi(m,n)}\left(\frac{1+\rho(m,n)}{\gamma(m,n)}\right)}$$

2.3.2 تقدير طيف الضوضاء Noise spectrum estimation

رأينا أن كل التوابع المذكورة سابقاً تحتاج إلى بصمة محسوبة مسبقاً λ_N عادةً ما يتم اختيار هذه البصمة يدوياً في بيئة صامتة وبيئة صاخبة بالكامل لكن هذه الطريقة غير مجدية مع تطبيقات الزمن الحقيقي كونها تتطلب تفاعل مع المستخدم، كما تعمل تلك الطرق بشكل على افتراض أن الضوضاء ثابتة وتم تمييز المقاطع التي تحتوي ضجيج مسبقاً. لذا في حال كانت الضوضاء غير ثابتة في المقطع الصوتي او في حال وجود عدم دقة في تقدير الضوضاء فيمكن أن يؤدي ذلك إلى إخماد مجحف للضجيج أو حذف قيم من الإشارة الأصلية.

يوجد عدة طرق لتقدير طيف الضوضاء بشكل فعال دون تفاعل مع المستخدم حيث يمكن تحديد وتحديث هذا الطيف خلال الوقفات دون تدخل المستخدم عن طريق خوارزميات تحديد النشاط الصوتي [7,8]

كما يوجد بعض الطرق لتقدير الكثافة الطيفية للطاقة في ضوضاء غير الثابتة بدون استخدام كشف النشاط الصوتي [9]. بدلاً من ذلك، يتم تتبع الحدود الدنيا الطيفية في كل نطاق تردد دون أي تمييز بين مرحلتي الكلام وغير الكلام وتستخدم لتقدير طيف الضوضاء التكيفي. على أي حال تعتبر هذه التقديرات غير مناسبة للإشارات الصوتية ذات الأطياف الكثيفة.

يه وآخرون. قدم خوارزمية لتقدير مستوى الضوضاء الملونة في الإشارات الصوتية بناءً على الافتراضات أ) أن غلاف الضوضاء يتغير ببطء مع التردد وذلك ب) مطال قمم الضوضاء تخضع لتوزيع رايلي. يمكن بعد ذلك استخدام مستوى الضوضاء المشتق لإخماد الضوضاء.

كما يوجد بعد الطرق التي تجمع بين الـ Maximum Likelihood Estimation والـ Minimum Mean Squared Error Estimator لتقدير الطيف الخاص بالضجيج كما في [21].

2.3.3 Adaptive Noise Reduction for Real-time

قام وينزرييل وآخرون (2010) في [10] باقتراح طريقة لإخماد الضوضاء الثابتة وغير الثابتة من الإشارة الصوتية في الزمن الحقيقي بالاعتماد على إضافة عدة مراحل كالتالي حيث نتعامل مع الإشارة الصوتية ككتل و يتم معالجة كل كتلة على حدا عن طريق تحويل فورييه قصير الزمن STFT من أجل كل كتلة ومن ثم تقدير طيف الضوضاء اللحظي باستخدام نموذج انحدار تلقائي Autoregressive Model وإيجاد العوامل المناسبة لهذا النموذج باستخدام أحد نماذج البرمجة الخطية التي تسمى عودية ليفنسون-دوربين Levinson-Durbin recursion ومن ثم تعديل تابع الإخماد للمحافظة على العابرات في الإشارة وهي عبارة عن مناطق في الإشارة الصوتية تحوي على ترددات عالية تبدو كالضوضاء لكنها فعلياً من الإشارة الأصلية النقية تشبه العابرات في الصوت الحواف في الصور، أخيراً يتم استخدام تقريب لمرشيح افرايم ومالاه Ephraim and Malah filter مع النتائج المستخرجة من المراحل السابقة وتطبيقه على الكتلة الحالية وبعدها يتم تطبيق تحويل فورييه العكسي للعودة لفضاء الزمن مع القيام بتنعيم لتابع التخميد عند كل كتلة

2.3.4 Audio Signal Denoising Algorithm by Adaptive Block Thresholding using STFT

قام دوتا وآخرون (2017) في [12] باقتراح طريقة لإزالة الضوضاء الغاوسية البيضاء من الإشارة الصوتية حيث نتعامل مع الإشارة الصوتية ككتل ثم تتم معالجة كل كتلة من البيانات على حدة. حيث يتم التركيز على أن المهمة المهمة هي اختيار

طول الكتلة المناسب. يتم تجزئة الإشارة إلى كتل، بطول مثالي، ثم يتم تقليل الضوضاء في مجال STFT عن طريق تحديد معاملات STFT. عندما يتم إزالة الضوضاء من كل كتلة عن طريق أخذ حجم النافذة الأمثل يتم استنتاج أن الخوارزمية القائمة على STFT المقترحة متفوقة من حيث جودة الإشارة منزوعة الضوضاء & وقت التنفيذ. لوحظ أن تابع العتبة من النوع الصلب Hard Thresholding المتكيف مع STFT يعطي أفضل نسبة SNR للإشارة الصوتية. وخلص أيضًا إلى أن الخوارزمية المقترحة تؤدي أداءً أفضل من الخوارزميات الأخرى فيما يتعلق بنسبة الإشارة إلى الضوضاء (SNR) ووقت التنفيذ.

2.3.5 إعادة توجيه الصور حسب البروز Saliency-driven image retargeting

من المناهج المقترحة لضغط الفيديو تقنية (saliency-driven image retargeting) إعادة توجيه الصور حسب البروز. حيث يتم اولا استخراج (saliency map) خريطة البروز (وهي الاجزاء البارزة والمهمة من الصورة) من اطارات الفيديو اما تلقائيا او عن طريق المستخدم ثم يتم إجراء تحجيم غير خطي للصورة الذي يقوم بتعيين عدد بكسلات أعلى للمناطق البارزة في الصورة وعدد أقل من البكسلات للمناطق غير البارزة. يعقب هذه العملية نهج متعدد الدقة يتم فيه ضغط نطاقات مختلفة من الصورة بنسب مختلفة باستخدام خوارزميات الضغط الحالية وعملية حساب البروز لا تستغرق الا بضعة اجزاء من الالف من الثانية فإنها لا تزيد من وقت المعالجة بشكل كبير. تم افتراح هذه التقنية في [2] حيث تمت مقارنة قيم PSNR و SSIM لـ CCSIR بالصور المضغوطة عن طريق(JPEG 2000) حيث تم حساب متوسط قيم PSNR و SSIM لمجموعة من خمسة مقاطع فيديو.

يتراوح نطاق الضغط من 0.037 bpp إلى 0.857 bpp بمتوسط 0.313 bpp.اعطت كل من PSNR و SSIM في تقنية (CCSIR) اداء أسوأ قليلاً من أداء JPEG 2000 في الصورة الكلية .ومع ذلك فإنه يؤدي اداء أفضل في المناطق البارزة من الصور بتحجيم معتدل العوامل.

2.3.6 الكشف عن البروز الثابتة والديناميكية Static and Dynamic Saliency Detection

تم العمل على تطوير هذا المنهج في [3] التي اقترحت تقنية Static and Dynamic Saliency Detection حيث استخدمت ترميز الفيديو عالي الكفاءة (HEVC) الذي يتفوق بشكل كبير على المعايير السابقة H.264 / AVC من حيث معدل بتات التشفير وجودة الفيديو. ومع ذلك، فإنه لا يأخذ في الاعتبار النظام البصري البشري (HVS) ، حيث يولي الناس مزيدًا من الاهتمام لمناطق محددة والأشياء المتحركة. في هذا البحث تم اقتراح مخطط مدرك للتحكم في معدل ترميز HEVC استنادًا إلى الكشف عن البروز الثابتة والديناميكية. تتكون الإستراتيجية المقترحة بشكل أساسي من ثلاث تقنيات، كشف البروز الثابتة، كشف البروز الديناميكي، وتخصيص معدل البتات التكيفي. أولاً، تم تدريب نموذج شبكة التفاف عميق (DCN) لاستخراج خريطة البروز الثابتة من خلال تسليط الضوء على المناطق البارزة لغويًا. بالمقارنة مع تقنيات استخراج منطقة الاهتمام التقليدية القائمة على النسيج أو اللون (ROI) ، فإن نموذج هذه الدراسة أكثر انسجامًا مع HVS. ثانيا، نقوم بتطوير تقنية تجزئة الكائن المتحرك لاستخراج المناطق البارزة الديناميكية تلقائيًا لكل إطار. علاوة على ذلك، وفقًا لخريطة البروز الاندماجية، يتم استغلال تقنية التحكم في مستوى البتات لوحدة شجرة التشفير (CTU) لتحقيق تخصيص معدل بتات مرن وقابل للتكيف. ونتيجة لذلك، تم تحسين جودة المناطق البارزة من خلال تخصيص المزيد من البتات، مع تخصيص وحدات بت أقل للمناطق غير البارزة. لقد تحققت من الطريقة المقترحة في كل من مجموعة البيانات الموصى بها -JCT

VC ومجموعة بيانات تتبع العين. تظهر نتائج التجربة أن PSNR للمناطق البارزة يمكن أن يتحسن بمتوسط 1.85 ديسيبل دون إضافة عبء معدل البت مما يحسن التجربة المرئية بشكل كبير.

2.3.7 الطرق المعتمدة على تقدير الحركة Motion estimation based

من المناهج المقترحة ايضا تحقيق ضغط مقاطع الفيديو باستخدام تقنيات تقدير الحركة (Motion estimation) حيث مقطع الفيديو عبارة عن صور متتالية فصورتين او أكثر من الصور المتتالية يجب ان تحوي اختلاف فإن تقدير الحركة هو العملية التي تتبع هذه التغييرات من أجل تقليل التكرار بين الإطارات. مؤخرا هناك العديد من خوارزميات تقدير الحركة تقترح [19] خوارزمية VOS حيث تعتبرها الافضل لتقدير الحركة.

حيث في هذه الخوارزمية يتم تحويل الفيديو أولاً إلى إطارات فردية حيث كل إطار عبارة عن صورة RGB ولكن معالجة RGBصعبة لأن الألوان تأخذ وقت لمعالجتها فتستغرق وقتًا طويلاً لذلك يتم تحويل صور RGB إلى صور من نوع رمادي. ثم يتم إزالة التكرار المكاني الموجود بين بكسلات الاطارات ثم يتم تحقيق الخلفية التكيفية للصور وتحويل الصور لنوع بايناري بعدها سيكون من السهل استخراج مقدمة الصور وطرح خلفيتها عن طريق تعتيم الخلفية باستخدام الفلتر حيث هذه العملية تعمل على تقليل حجم صور الفيديو في النهاية يتم إدخال الفيديو الجديد الى المشفر H.264 الذي يقوم بضغط الفيديو كاملا او حذف التكرار الزمني.

من تجارب هذه الدراسة تم تطبيق خوارزمية VOS على فيديو بحجم 11.7 MB بمدة 12 ثانية تم الحصول على فيديو بحجم 8.61 MB بجودة ممتازة لمقدمة الصور وضعيفة لخلفية الصور وتعد طريقة مناسبة لفيديوهات البث في الوقت الفعلي اما PSNRو SSIM يعطيان اداء متوسط.

2.3.8 تحسين تخصيص البتات Bit assignment improving

تم اقتراح ايضا تقنية تحسين تخصيص البتات في ضغط الفيديو. حيث تحدد هذه التقنية المناطق المهمة (ROI) في الفيديو. فقد اقترحت [16] تقنية استخراج المناطق المهمة لتطبيقات التتبع لان ضغط مقاطع الفيديو يقلل من دقة خوارزميات التتبع. فتم استخراج المناطق المهمة (المراد تتبعها) (ROI) من خلال نموذج غير حدودي قائم على التوزيع الزمني لشدة البكسل. حيث يتم البحث عن البكسلات التي تشهد تغير جذري. وتم تحديد هذه المناطق باستخدام (kurtosis of intensities) درجة تقوس شدة البكسل لكل موضع بمرور الوقت. تشير قيمة درجة التقوس الكبيرة الى تغير كبير بينما تشير القيمة الأقل إلى تغير أصغر. وقد تم اعتماد نهج احتمالي لتحديد العتبة.

وقد تم تجربة هذه التقنية على فيديوهات المرور لمراقبة السير فعند مقارنة نتائج استخراج المناطق المهمة يدويا وباستخدام تقنية درجة التقوس فكانت نتائج هذه التقنية افل بكثير وذلك يعود الى ان تقنية درجة التقوس تركز على المناطق المهمة (ROI) التي تم الإبلاغ عن نشاط او حركة فيها اما المراقب الذي يستخرج المناطق المهمة (ROI) يدويًا يركز على المناطق التي يمكن أن يحدث فيها نشاط اما بالنسبة لقيم PSNRوSSIM يعطيان اداء جيد في المناطق المهمة ولكن تم اكتشاف ان هناك نسبة خطأ في تابع تحديد المناطق المهمة (ROI) وهذا يعتبر من سلبيات هذه التقنية.

2.3.9 تقنية كايف CAVE technology

ومن احدث وافضل تقنيات ضغط الفيديو حسب المحتوى تقنية (CAVE) التي تستفيد من الطريقة المقترحة سابقا استخراج مناطق الاهتمام (ROIs) وتحسن الجودة من خلال تخصيص كميات مختلفة من البتات لمناطق مختلفة في إطارات الفيديو. وتوفير في معدل البت بين 21٪ و46٪ مقابل تشفير HEVC الأساسي. تعتمد هذه التقنية على خطوتين رئيسيتين. أولاً يتم تعيين أوزان للكتل في إطارات الفيديو بناءً على أهمية هذه الكتل من منظور المستخدم. ثم يتم تخصيص البتات للكتل بناءً على أوزانهم مع تلبية متطلبات المستخدم وقت الاستجابة المنخفض للشبكة وتحقيق جودة متسقة للإطارات المشفرة.

استخدمت [18] هذه التقنية في بث العاب الفيديو السحابية وتم اجراء تجارب مكثفة مع العديد من ألعاب الفيديو وقياس العديد من مقاييس الأداء، بما في ذلك SSIM وVMAF وتقلبات الجودة وتوفير معدل البت. حيث تُظهر التجارب أن تقنية (CAVE) تتفوق باستمرار على تقنيات الضغط حسب المحتوى السابقة.

2.3.10 Video Object Segmentation for Content-Aware Video Compression

تم في الورقة البحثة [12] ضغط الفيديو اعتمادا على المحتوى الأكثر أهمية؛ بالبداية يتم تطبيق عملية Background Subtraction بذلك يتم تحديد الحركة في كل إطار، بعد تحديد البكسلات المتحركة في الخلفية بالخطوة السابقة يتم تجميع تلك البكسلات مع بعضها وذلك بتطبيق بعض العمليات المورفولوجية التي من شأنها ملئ الثقوب وتحديد شكل الأغراض المتحركة بشكل واضح و هذه العملية تدعى Blob analysis، في الخطوة اللاحقة يتم تطبيق عملية Color Similarity Calculation و الهدف من هذه الخطوة هو تقدير مسافة كل بكسل عن مقدمة الصورة وعن الخلفية اعتمادا على مقياس التشابه اللوني بين كل بكسل في المنطقة الواقعة ضمن المحدب المغلق للكائن المتحرك والمنطقة المؤكد بكونها من الكائن وبذلك يتم تحديد انتماء كل بكسل من بكسلات الإطار بالنسبة لمقدمة الصورة أو الخلفية بشكل أدق و نهاية يكون المحتوى المهم في الإطار يمثل مقدمة الصورة. لتحقيق عملية الضغط يطبق bilateral filter على خلفية الإطار والذي يمثل الجزء الأقل أهمية فيه، ثم يتم إعادة تشكيل الإطار بشكل كامل عن طريق دمج كلا من مقدمة الصورة والخلفية.

2.3.11 Microsoft azure speech translation

وهي تتيح خدمة ترجمة الكلام لأكثر من 88 لغة باستخدام تقنيات التعلم الالي (machine learning). يتم تقديم هذه الخدمة مجانا لمجموع حوالي 5 ساعات من الملفات الصوتية المترجمة كل شهر وبعد ذلك يترتب على المستخدمين دفع المال للحصول على هذه الخدمة.

2.3.12 Google speech translation

يقدم مترجم غوغل الترجمة الآلية للكلام speech-to-text-translated وذلك لحوالي 100 لغة مختلفة حيث يتميز أيضاً بقدرة عالية على مستوى الترجمة السياقية. ومن حيث التسعير يقدم Google API ما يقارب الساعة الواحدة مجانا في كل شهر ويمكن للمستخدم بعد ذلك بتجديد استخدامه مقابل مبلغا ماليا.

2.3.12 Amazon AWS speech translation

يعد من أحد أشهر التطبيقات التي تقدم خدمة ترجمة الكلام وتمتاز بدقة عالية. تتيح عملية الترجمة لأكثر من 50 لغة وتقدم ساعة واحدة مجانا لملفات الترجمة الصوتية

2.4 طرق تقييم النظام System Evaluation Methods

2.4.1 نسبة الإشارة للضوضاء Singal to noise ratio

يستخدم هذا المعيار لتقييم نظم إزالة الضوضاء بفرض x هي الإشارة النقية والـ y هي الإشارة الصاخبة فيمكن تعريف الـ SNR كالتالي:

$$SNR(dB) = 10\ log_{10}\left[\frac{\sum_{i=0}^{n} x_i^2}{\sum_{i=0}^{n}(y_i - x_i)^2}\right]$$

كوننا لا نعرف الإشارة النقية x فنقوم بإضافة ضوضاء غاوسية بيضاء للصوت المراد إزالة الضوضاء منه ونحسب الـ SNR بعد إزالة الضوضاء منه

2.4.2 جذر متوسط مربع الخطأ Root mean squared error

يعتبر معيار أداء آخر لنظم إزالة الضوضاء بفرض x هي الإشارة النقية والـ y هي الإشارة الصاخبة فيمكن تعريف الـ RMSE كالتالي:

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n}(y_i - x_i)^2}{n}}$$

كوننا لا نعرف الإشارة النقية x فنقوم بإضافة ضوضاء غاوسية بيضاء للصوت المراد إزالة الضوضاء منه ونحسب الـSNR بعد إزالة الضوضاء منه

2.4.3 مقياس التقييم الإدراكي لقياس جودة الصوت Perceptual Evaluation of Speech Quality

يعمل على قياس جودة الصوت ويأخذ بعين الاعتبار كل من الخصائص التالية: حدة الصوت، حجم المكالمة، الضوضاء في الخلفية، زمن الانتقال المتغير أو التأخر في الصوت، والتداخل في الصوت. يقارن هذا المقياس الصوت المتنبأ به (خرج النظام الخاص بإزالة الضجيج) بالنسبة للصوت الأصلي، يعد هذا المقياس أكثر دقة من الطرق الأخرى لقياس جودة الصوت، يرجع مقياس PESQ قيما من -0.5 إلى 4.5 حيث تشير القيم الأعلى جودة أفضل.

2.4.4 معدل البت Bit rate

في دفق الفيديو، يشير معدل البت إلى عدد البتات التي يتم نقلها أو معالجتها في وحدة زمنية معينة. يُقاس معدل البت عادةً بالبت في الثانية (بت / ثانية) في الفيديو، يستوعب معدل البت الأعلى جودة صورة أعلى في إخراج الفيديو. معدل البت للفيديو عالي الدقة، على سبيل المثال، عادة ما يكون في النطاق من 5 إلى 20 ميجابت في الثانية.

2.4.5 ذروة نسبة الإشارة للضوضاء Peak signal to noise ratio

نسبة ذروة الإشارة إلى الضوضاء (PSNR) هي مصطلح هندسي للنسبة بين أقصى طاقة ممكنة للإشارة وقوة الضوضاء الفاسدة التي تؤثر على دقة تمثيلها، أسهل تعريف لـ PSNR عبر (MSE). حيث بالنظر إلى صورة أحادية اللون خالية من الضوضاء I أبعادها MxN و K هو الـ noisy approximation للصورة I، يتم تعريف MSE على أنه:

$$MSE = \frac{1}{m.n}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2$$

يتم تعريف PSNR بالديسيبل كالآتي:

$$PSNR = 10\,.\,log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

$$= 20\,.\,log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right)$$

$$= 20\,.\,log_{10}(MAX_I) - 10.\,log_{10}(MSE)$$

هنا، MAX I هو أقصى قيمة بكسل ممكنة للصورة. عندما يتم تمثيل وحدات البكسل باستخدام 8 بت لكل عينة، يكون هذا 255. بشكل عام، عندما يتم تمثيل العينات باستخدام PCM الخطي مع B بت لكل عينة، يكون MAX I هو 2^B-1.

2.4.6 مؤشر التشابه الكمي Structural Similarity Index

مقياس مؤشر التشابه الهيكلي (SSIM) هو طريقة للتنبؤ بالجودة المتصورة للتلفزيون الرقمي والصور السينمائية، بالإضافة إلى أنواع أخرى من الصور ومقاطع الفيديو الرقمية. يستخدم SSIM لقياس التشابه بين صورتين. مؤشر SSIM هو مقياس مرجعي كامل؛ بمعنى آخر، يعتمد قياس جودة الصورة أو التنبؤ بها على صورة أولية غير مضغوطة أو خالية من التشويه كمرجع.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

μ_x هو متوسط x، μ_y هو متوسط y، σ_x^2 هو تباين x، σ_y^2 هو تباين y،و σ_{xy} هو التغاير x و y

$c_1 = (k_1L)^2, c_2 = (k_2L)$ متغيرين لتثبيت القسمة ذات المقام الضعيف ويكون L هو النطاق الديناميكي لقيم البكسل عادة ما يكون $(2^{\#\ bits\ per\ pexil} - 1)$

بشكل افتراضي $k_1 = 0.01$ و $k_2 = 0.03$

$$l(x,y) = \frac{(2\mu_x\mu_y + c_1)}{(\mu_x^2 + \mu_y^2 + c_1)}$$

$$c_{(x,y)} = \frac{(2\sigma_{xy} + c_2)}{(\sigma_x^2 + \sigma_y^2 + c_2)}$$

$$s_{(x,y)} = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}$$

حيث:

$$c_3 = {}^{c_2}/_2$$

فيكون:

$$SSIM_{(x,y)} = [\,l(x,y)^\alpha\,.\,c(x,y)^\beta\,.\,s(x,y)^\gamma]$$

2.5 ملخص Summary

بالنسبة للضغط الموجع للفيديو نستنتج حسب اراء الباحثين والدراسات ان مناهج ضغط مقاطع الفيديو حسب المحتوى تنقسم الى ترميز الفيديو القائم على التركيب (Texture-Based Video Coding) وهذا المنهج يعتمد على تحديد الاجزاء المهمة من الفيديو ويتم ذلك اما من خلال تحديد خريطة البروز (saliency-driven image retargeting)، من حيث الجودة اعطت حسب مقياسي (PSNR وSSIM) قيم جيدة في المناطق البارزة وقيم اقل في المناطق غير البارزة واعطت معدل بت جيد نوع وهي تقنية فعالة لتطبيقات البث في الوقت الفعلي. وتم تطويره هذه التقنية لتحسين النتائج في تقنية (Static and Dynamic Saliency Detection) واعطت نتائج افضل من حيث الجودة حسب مقياسي (PSNR و SSIM) ومعدل البت ولكن من سلبيات هذه التقنية الاختلاف الواضح في جودة ودقة الصورة بين المناطق البارزة وغير البارزة حيث يكون الانتقال بين المناطق البارزة وغير البارزة غير سلس.

الطريقة الثانية تحديد المناطق المهمة في الفيديو (ROI) وهناك طرق عديدة لتحديد هذه المناطق منها تابع درجة التقوس لتغير البكسل وهذه التقنية تستخدم في تطبيقات التتبع حيث يتم تتبع المناطق المهمة (ROI) وتعطي هذه الطريقة نتائج جيدة في تحديد المناطق المهمة او المراد تتبعها في الجودة حيث اعطى مقياسي (PSNR وSSIM) في المناطق المهمة (ROI) قيم جيدة. واخيرا تقنية (CAVE) التي تعتمد على تخصيص البتات للكتل في إطارات الفيديو بناءً على أوزانهم تُستخدم هذه التقنية في ضغط فيديوهات الالعاب السحابية وتم اجراء تجارب على عدة العاب ومقارنة النتائج مع نتائج ترميز HEVC الأساسي فتم توفير في معدل البت بين 21٪ و 46٪ مقابل ترميز HEVC الأساسي واعطى مقياسي (PSNRوSSIM) ايضا قيم افضل من تشفير HEVC وتعتبر تقنية فعالة لتطبيقات البث في الوقت الفعلي.

اما المنهج الاخر المعتمد لضغط الفيديو حسب المحتوى هو ترميز الفيديو القائم على الحركة (Motion-Based Video Coding) يتم في هذه المنهج تجزئة صور الفيديو الى مقدمة الصورة وخلفية الصور وتعريض الخلفية لفلاتر لتقليل حجمها ونتائج هذه التقنية كانت جيدة حيث اعطت من حيث الجودة حسب مقياسي (PSNR وSSIM) قيم عالية في مقدمة صورة الفيديو وقيم اقل في خلفيتها وبمعدل بت جيد وتعتبر جيدة للبث في الوقت الفعلي.

أما بالنسبة للتحسين على الصوت وتنقيته فنجد أن الأساليب المجدية هي تلك التي تعتمد على العمل في فضاء التردد الزمني ونخص بالذكر الطرق غير القطرية لكونها تعطينا أكبر نسبة SNR عند التجريب كما تعمل على إنتاج ضوضاء موسيقية أقل في الإشارة الصوتية الناتجة على عكس طرق التقدير القطرية. بناءً على ما ورد بالأعمال السابقة نجد أن الخوارزمية المقترحة في [10] أعطت نتائج أفضل من الطرق التي تعتمد على بصمة ضوضاء محسوبة مسبقاً.

الفصل الثالث: الدراسة التحليلية والتصميمية Analysis and Design

3.1 المتطلبات الوظيفية Functional Requirements

- تحقيق ضغط موجه للبث بناءً على المحتو ى والحركة فيه؛
- تحقيق تحسين لجودة الصوت وإزالة الضجيج منه؛
- تحقيق ترجمة فورية للبث بين اللغة العربية واللغة الإنكليزية؛
- تحقيق المتطلبات السابقة بالزمن الحقيقي.

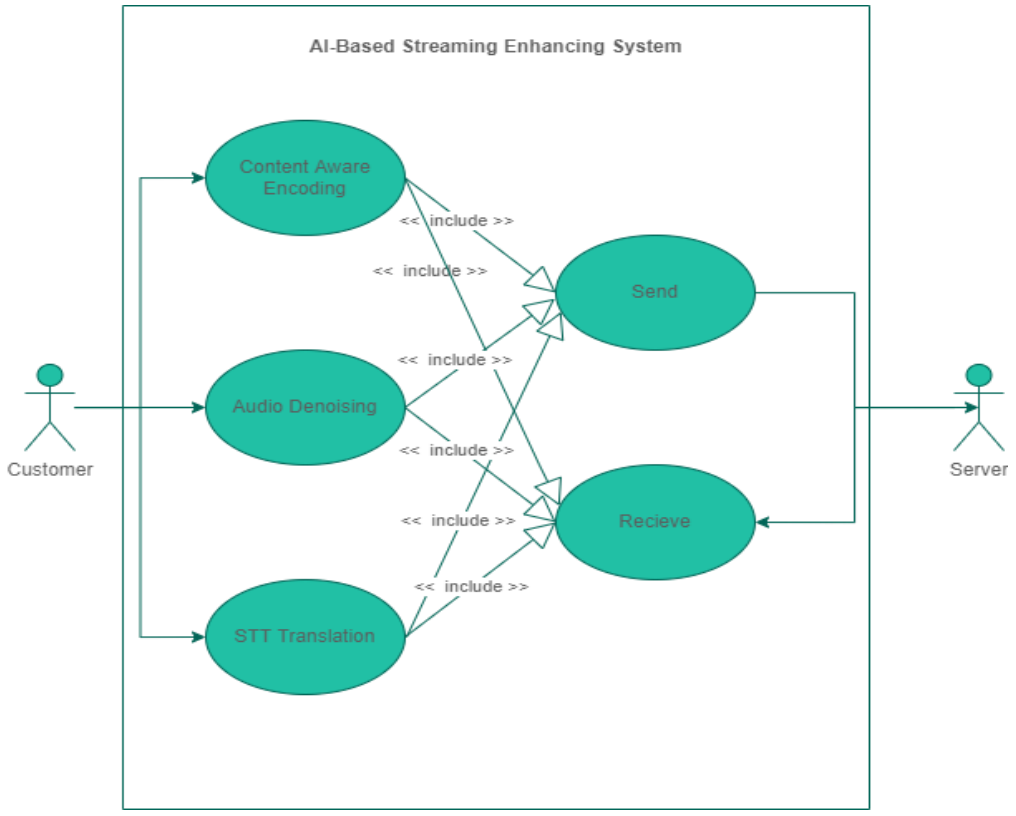
3.2 المتطلبات غير الوظيفية Non-Functional Requirements

- الأمان: وذلك حسب الطلب كون النظام من الممكن أن يستخدم في تطبيقات تتطلب السرية؛
- سهولة الصيانة: وذلك عن طريق فصل النظام الكلي لأنظمة جزئية منفصلة مما يؤدي إلى فصل عالي للمهام وسهولة في تحديد الأعطال وصيانتها مستقبلاً.

3.3 الفاعلون وأهدافهم Stakeholders

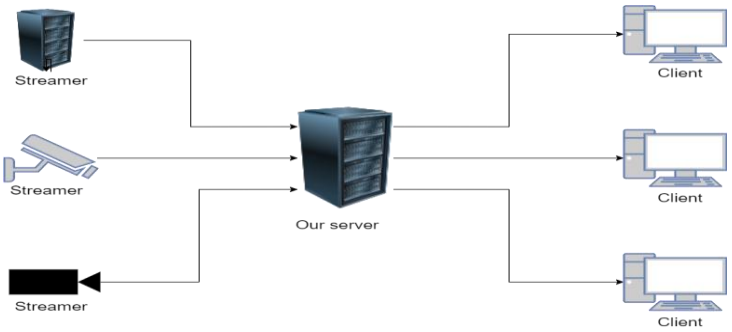
- المستخدمين العاديين الذين يقومون باستخدام واجهة من النظام للحصول على تجربة أفضل في مشاهدة المحتوى؛
- المطورين الذين يمكن أن تفيدهم الميزات الذي يقدمها النظام كخدمة في أنظمتهم

3.4 مخطط حالات الاستخدام Use Case Diagram

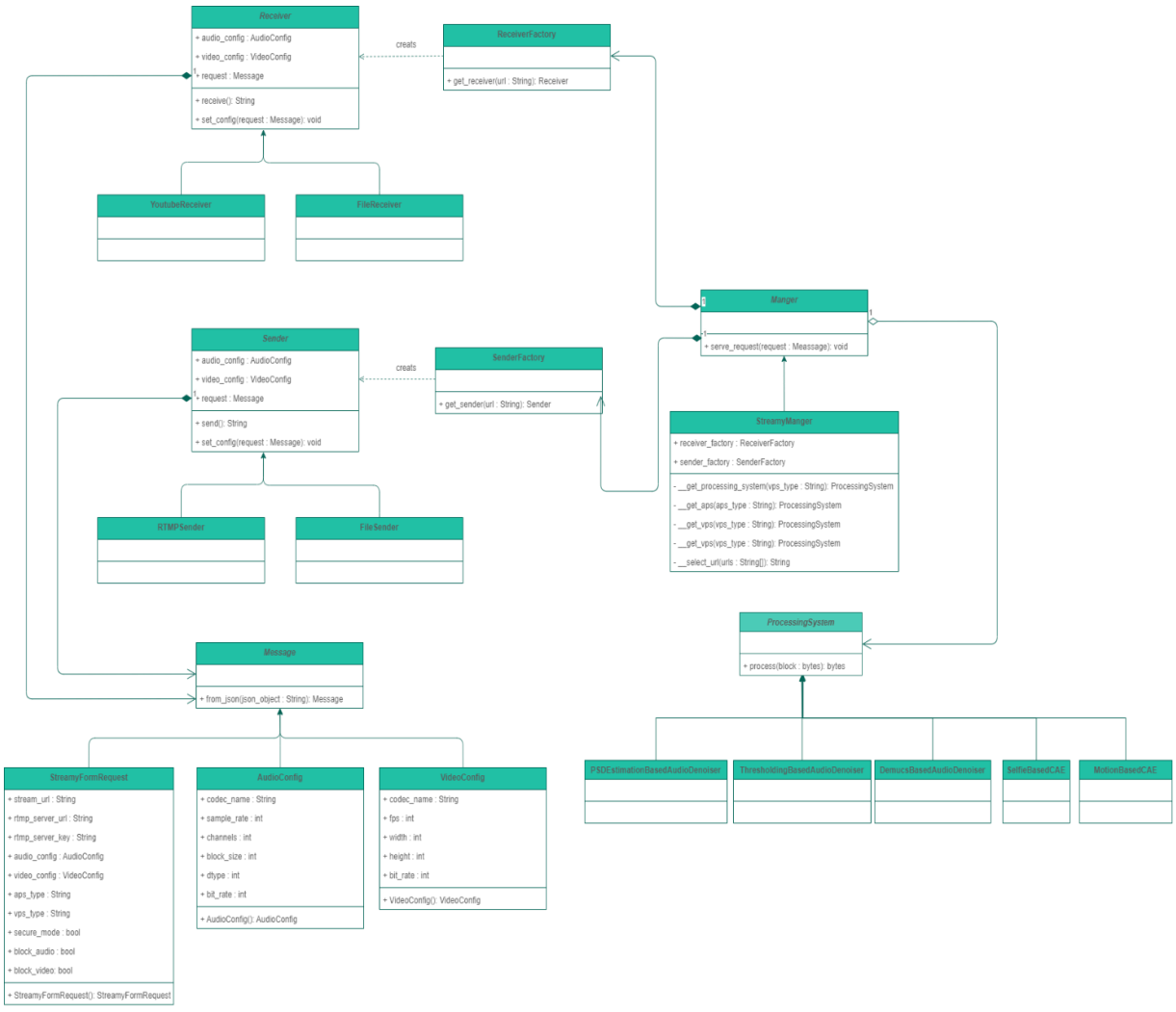


3.5 البنية المعمارية Architecture

ستتمثل بنية المشروع بمعمارية مخدم-مستخدم Client-Server مع Thin Client أي يكون طرف المستخدم متمثل بالأجهزة الذكية والمواقع الالكترونية التي توفر واجهة استخدام فقط وطرف المخدم يحوي على كل العمليات اللازمة لتحقيق عمليات الاستقبال من مصدر البث والمعالجة وإعادة الإرسال إلى المستخدم أو مخدم أخر.



3.6 مخطط الصفوف Class Diagram



الفصل الرابع: النظام المقترح Proposed System

بالاعتماد على خرج الفصل الثاني الخاص بالدراسة المرجعية وخرج الدراسة التحليلية نودر في هذا الفصل المنهجيات المتبعة لتحقيق المتطلبات الوظيفية والتي سيتم اعتمادها في هذا المشروع نبدأ في هذا الفصل بذكر التقسيم الوظيفي للأنظمة الجزئية ضمن النظام المقترح، من ثم نذكر الطرق المتبعة لتحقيق المتطلب الخاص بإزالة الضجيج من الصوت، نذكر أخيراً الطرق التي سنتعبها لتحقيق المتطلب الخاص بالضغط الموجه للفيديو.

4.1 تقسيم الأنظمة الجزئية Subsystems Partition

وظيفيا سيقوم المخدم بثلاث مهام استقبال، معالجة، وإرسال؛ لذا يمكن بناء على التقسيم الوظيفي السابق اقتراح البنية الآتية للنظام:

A. نظام استقبال Receiver System

يتولى هذا النظام مهمة استقبال البث من الانترنت الدخل لهذا النظام رابط لفيديو على الشبكة والخرج هو بث (فيديو مع صوت)

B. نظام معالجة Processing System

الدخل لهذا النظام بث أصلي والخرج بث معالج يتم تقسيمه كذلك الأمر إلى خمسة أجزاء كالتالي:

a. نظام تقسيم Divider System

يتولى هذا النظام مهمة تقسيم البث إلى أقنية منفصلة الدخل لهذا النظام بث أصلي الخرج هو قناة صوت وقناة فيديو

b. نظام معالجة الصوت Audio Processing System

يحقق هذا النظام كل المتطلبات المتعلقة بمعالجة الصوت وتنقيته الدخل فيه قناة صوت أصلية والخرج قناة صوت معالجة

c. نظام معالجة الفيديو Video Processing System

يحقق هذا النظام كل المتطلبات المتعلقة بتحقيق ضغط محسن على الفيديو الدخل فيه قناة فيديو أصلية والخرج قناة فيديو معالجة

d. نظام معالجة الترجمة Translation Processing System

يحقق هذا النظام الترجمة من الصوت للنص باللغتين العربية والإنكليزية الدخل فيه قناة صوت أصلية باللغة أولى والخرج نص مترجم للغة ثانية نذكر هنا أننا لن نقوم بتحقيق هذا النظام لذا سنلجأ للحصول عليه بشكل جاهز

e. نظام تجميع Aggregator System

يتولى هذا النظام مهمة تجميع قناة الفيديو المعالجة وقناة الصوت المعالجة والنص المترجم الدخل فيه الأقنية السابقة والخرج بث معالج

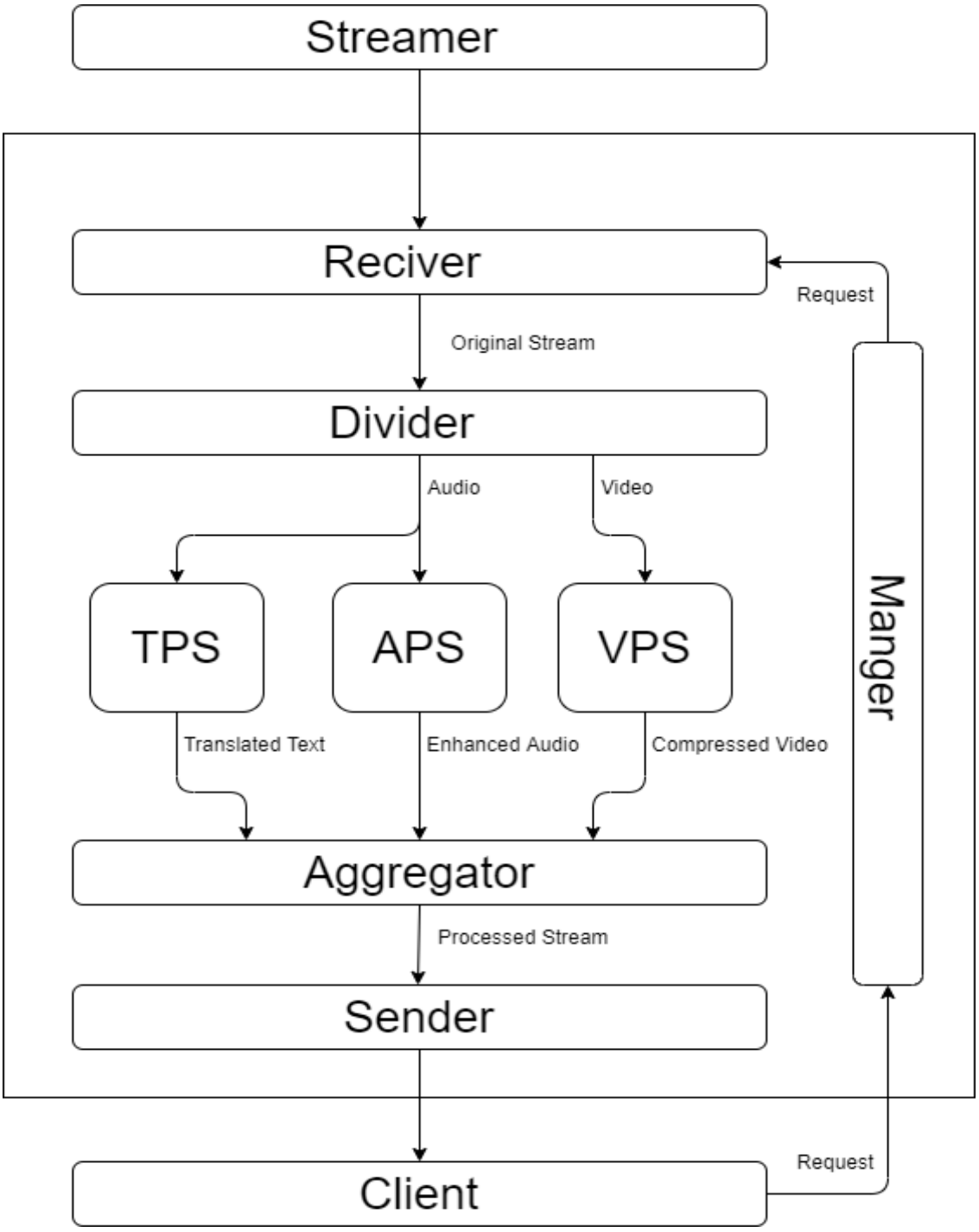
C. نظام إرسال Sending System

يتولى هذا النظام مهمة إعادة إرسال البث المعالج إلى العميل

D. مدير Manger

يتولى هذا النظام استقبال الطلب من المستخدم وتوجيهه للمستقبل وتعريف طريقة الإدارة المناسبة

يظهر المخطط التقسيم الوظيفي للنظام لمقترح:



-المخطط -

4.2 أنظمة معالجة الصوت Audio Processing Systems

4.2.1 نظام إزالة الضجيج المعتمد على العتبة Threshold Based Audio Denoiser

يعتبر هذا النظام بمثابة نموذج أولي لإزالة الضجيج ويعد فعال وسريع جداً في حالة كان الضجيج مستقر حتى درجة معينة وذا مطالات صغيرة نسبياً، يضع هذا النموذج المعادلة التالية:

$$\hat{X}(m,n) = g(Y(m,n))$$

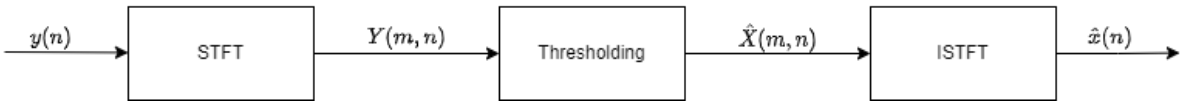
حيث يمثل g(.) تابع العتبة المستخدم وله حالتان ممكن أن يكون تابع عتبة حاد Hard Thresholding كالتالي:

$$g_T(x) = \begin{cases} 0 \ , |x| \leq T \\ x \ , |x| > T \end{cases}$$

سنقوم بتعريف تابع التعتيب كتابع ضبابي Fuzzy Thresholding حيث يهدف ذلك تقليل الخطأ الناتج عن عدم الدقة في تحديد العتبة فيتم أخذ موقع الفرق على منحني غاوس من أجل كل المكونات التي لها مطال أصغر من العتبة المختارة كالتالي:

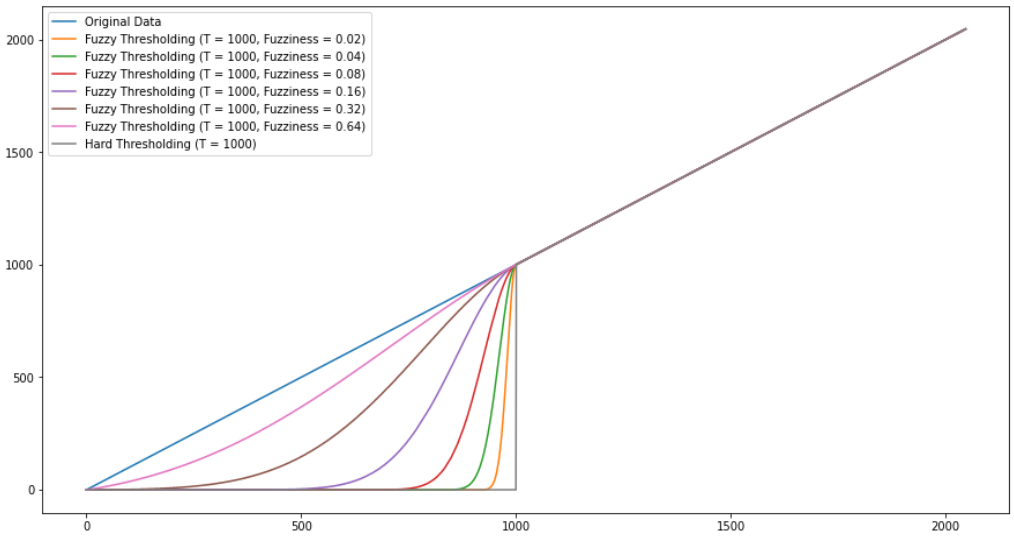
$$g_T(x) = \begin{cases} f(|x| - T) \ , |x| \leq T \\ x \ , |x| > T \end{cases}$$

حيث f(.) تابع غاوس ويمثل الـ standard division الخاص بالتابع الـ f(.) درجة ضبابية التابع g(.) فكلما كانت الـ standard division أكبر يؤدي ذلك إلى ضبابية أكبر في التعتيب يظهر مخطط الفرق بين تابع العتبة القاسي والعتبة يمثل المخطط 1 مراحل عمل هذا النموذج:



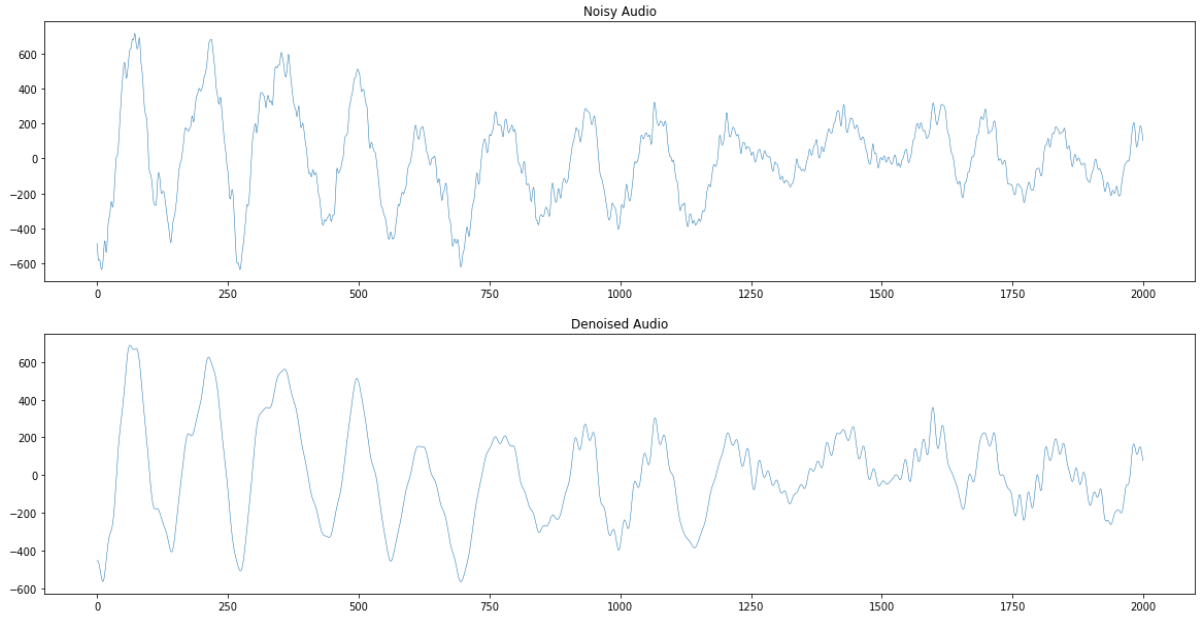
-المخطط 1-

يمثل المخطط 2 مقارنة بين تابع Fuzzy Thresholding وتابع Hard Thresholdingحالة المطالات من 0 إلى 2048 والعتبة T = 1000:



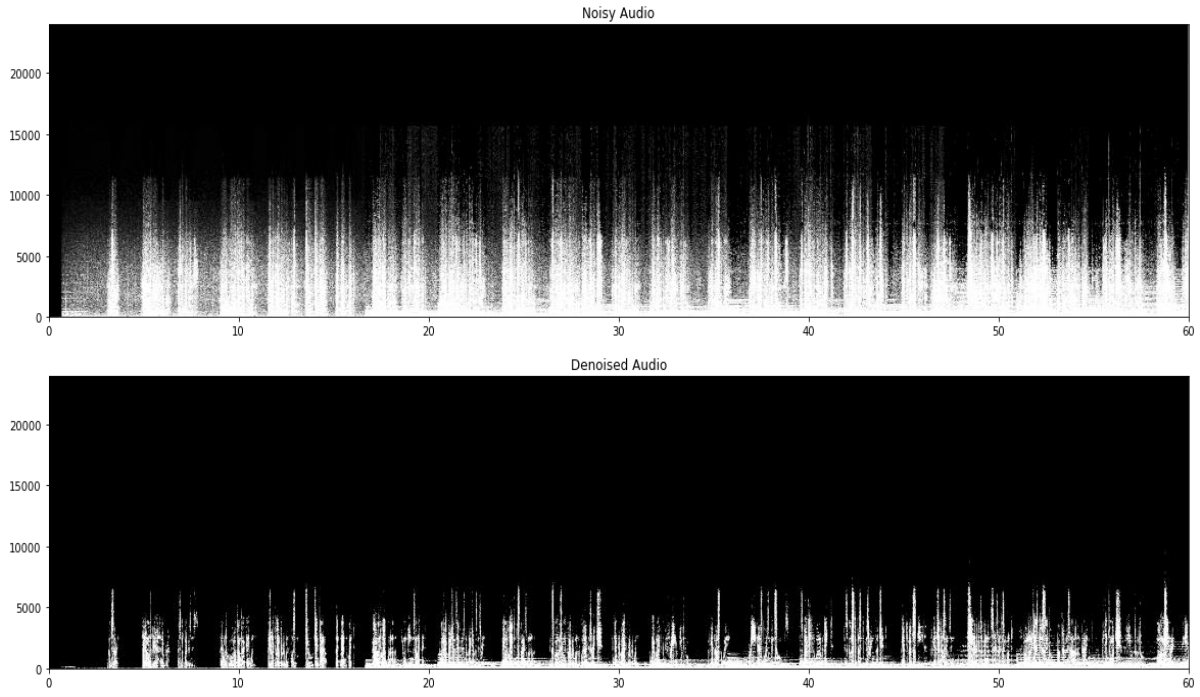
-المخطط 2-

يوضح المخطط 3 مثال لتأثير تابع الـ Hard Thresholding على جزء من مقطع صوتي حيث T=10 في الفضاء الزمني Time Domain:



- المخطط 3-

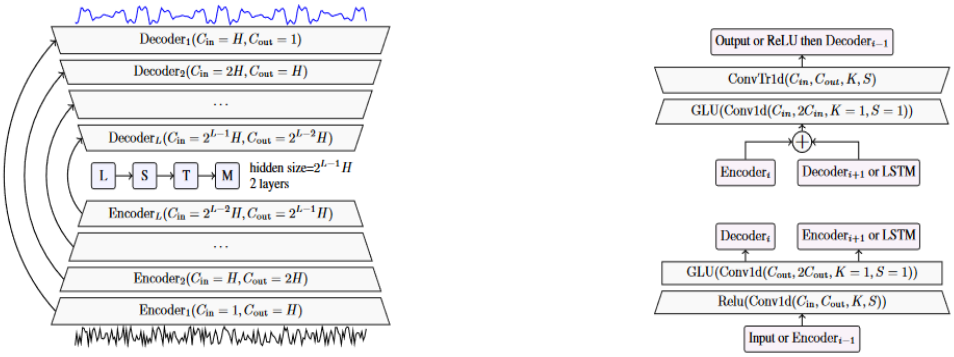
ويوضح المخطط 4 تأثير نفس التابع في فضاء التردد الزماني Time-Frequency Domain يكون الـ Spectrogram لنفس المقطع كالتالي:



-المخطط 4-

4.2.2 نظام إزالة الضجيج المعتمد على المشفرات التلقائية Demucs Based Audio Denoiser

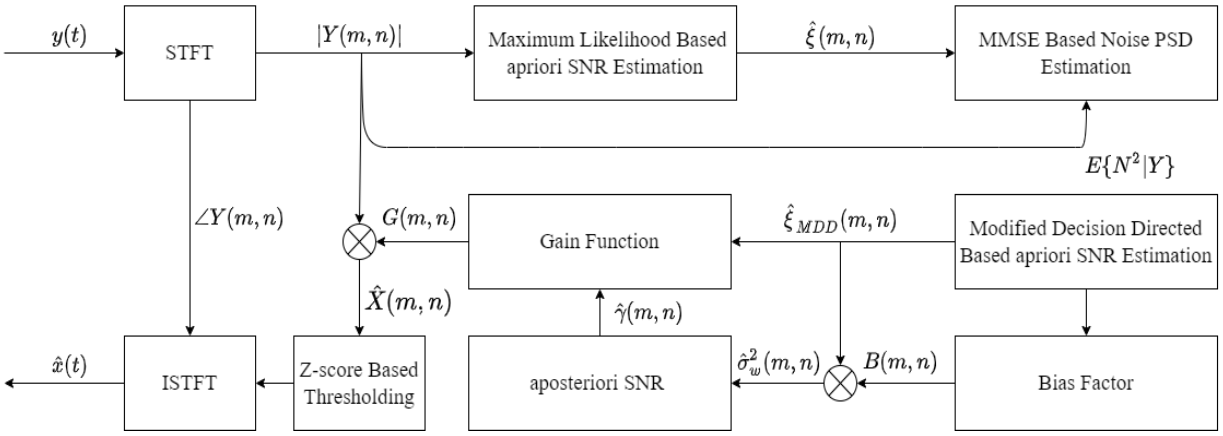
يعتمد هذا النظام على نموذج تعلم عميق Deep Learning Method مدرب مسبقاً مبني باستخدام المشفرات التلقائية Autoencoder وفق معمارية Demucs، الدخل للنموذج بهذه الحالة الإشارة الصوتية الصاخبة بالبعد الزمني والخرج هو الإشارة الصوتية النقية بالبعد الزمني. البنية الأساسية لـ Demucs والمقترحة [22] في مصممة أصلا لفصل المنابع الموسيقية عن الكلام قام الباحثون في [13] بالتعديل عليها لإزالة الضجيج من الصوت بالزمن الحقيقي قمنا في مشروعنا بعمل إرساء لهذا النموذج المدرب مسبقاً بما يناسب متطلباتنا من حيث آلية الدخل والخرج ومجالاته:



-المخطط 5-

4.2.3 نظام إزالة الضجيج المعتمد تقدير الكثافة الطيفية للضجيج PSD Estimation Based Audio Denoiser

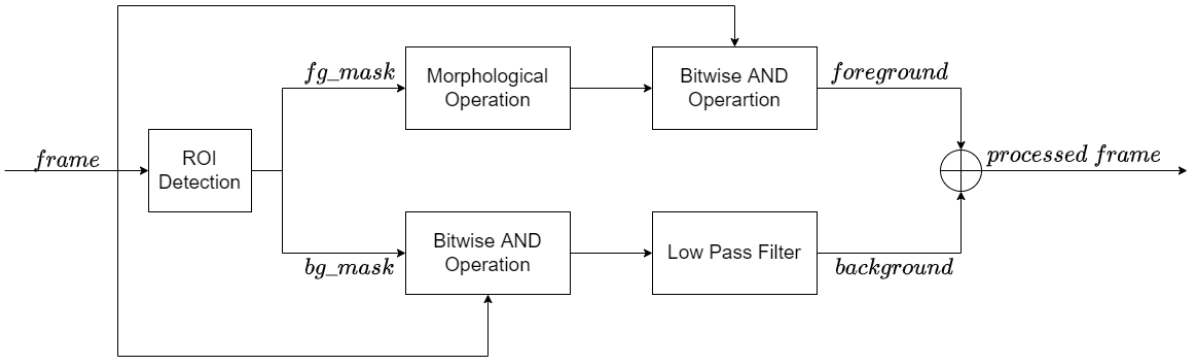
يعتمد هذا النظام على تقدير الكثافة الطيفية للضجيج وذلك بالاعتماد على الطريقة المقترحة في [21] كما قمنا ببعض التعديل فتم استخدام الطريقة المقترحة في [23] لتقدير الـa priori SNR عوضاً عن الطريقة المقترحة بالورقة الأصلية والتي تستخدم الطريقة المقترحة في [6] كما تم عمل تعتيب لمكونات الإشارة قبل الـ ISTFT بناء على الـ z-score الخاص بها فقمنا بحذف كل المكونات التي لها z-score أصغر من -1 باعتبار أنها غالباً ستكون musical noise وتم أخيراً عمل تنعيم لمكون الإشارة الصاخبة قبل حساب الـ a posteriori SNR، ليكون للنظام المقترح الخطوات المذكورة بالمخطط 6:



-المخطط 6-

4.3 أنظمة معالجة الفيديو Video Processing Systems

سنستخدم في نظامنا آلية هجينة من الطريقة الواردة في؛ تعتمد الفكرة بشكل أساسي على عمل عمليات معالجة للإطار قبل ضغطه تعطي أكبر قدر ممكن من الخسارة في الحجم بعد ضغطه. نبدأ أولاً بإدخال الإطار إلى خوارزمية ROI Detection فينتج لدي mask يعبر عن المناطق المهمة بالصورة ومن ثم يتم عمل بعض العمليات المورفولوجية Morphological Operations إن لزم ومن ثم اقتصاص نفس المناطق المحددة بالـ mask من الإطار الأصلي عن طريق عملية Bitwise AND على التوازي يتم تعريف بقية الإطار المحدد من معكوس الـ mask الأصلي كمنطقة أقل أهمية ويتم تمريرها على Low Pass Filter وذلك لكون العين حساسة بشكل أكبر للترددات المنخفضة لذا سيبقي الـ Low Pass Filter أهم المكونات من صورة الخلفية مما يؤدي للحصول على درجة أكبر من الخسارة في الحجم بعد الضغط، أخير يتم دمج ناتج العمليتين السابقتين يوضح المخطط 7 الخطوط العامة لنظام معالجة الفيديو المعني بالضغط الموجه ويبقى هنا الفرق بين الأنظمة بطريقة تحديد المنطقة المهمة لذا قمنا بتعريف طريقتين سنأتي على ذكرهم لاحقاً:



-المخطط 7-

4.3.1 نظام الضغط الموجه المعتمد على الأشخاص Selfie Segmentation Based Content Aware Encoding

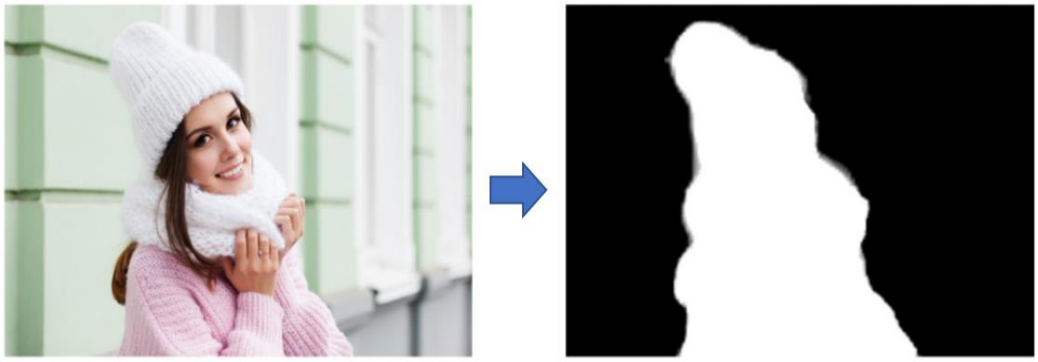
تعتبر MediaPipe إطار عمل يوفر نماذج تعلم آلي جاهزة ومتاحة للعديد من المنصات لمعالجة بيانات السلاسل الزمنية مثل الفيديو بالزمن الحقيقي real time، توفر MediaPipe حل يدعى Selfie Segmentation وهي شبكة عصبية تلافيفية مبينة على معمارية MobileNetV3 يعمل هذا الحل على تحديد الأشخاص ضمن مشهد معين. لذا يمكن اعتبار خوارزمية الـ Selfie Segmentation خوارزمية ROI Detection في البنية المقترحة أعلاه ويكون النظام الناتج في هذه الحالة موجه ومناسب بشكل أكبر لتطبيقات مؤتمرات الفيديو Video Conferencing والمقابلات التلفزيونية. بهدف التحسين سنقوم باستخراج الـ mask الخاص بالشخص مرة واحدة كل n إطار مما يساهم في تحسين السرعة على فرض أن حركة الشخص لن تكون سريعة بشكل كبير جداً. يوجد لدى MediaPipe نموذجين من Selfie Segmentation:

A. النموذج الأساسي General Model

هو شبكة عصبية تلافيفية CNN دخله إطار فيديو أو صورة ممثلة بالصيغة (256 × 256 × 3) وخرجه صورة الشخص البارز ممثلة بالصيغة (256 × 256 × 1) ويعتبر أكثر دقة ويعطي mask أكثر نعومة.

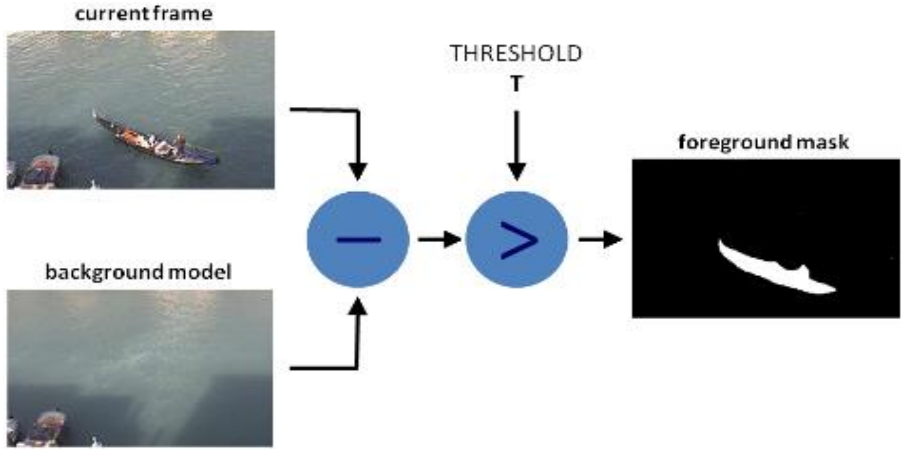
B. النموذج الأفقي Landscape Model

هو ايضا شبكة عصبية تلافيفية CNN دخله إطار فيديو أو صورة ممثلة بالصيغة (144 × 256 × 3) وخرجه صورة الشخص البارز ممثلة بالصيغة (144 × 256 × 1) ويعتبر أسرع من النموذج السابق.



4.3.2 نظام الضغط الموجه المعتمد على فصل الخلفية Background Subtraction Based Content Aware Encoding

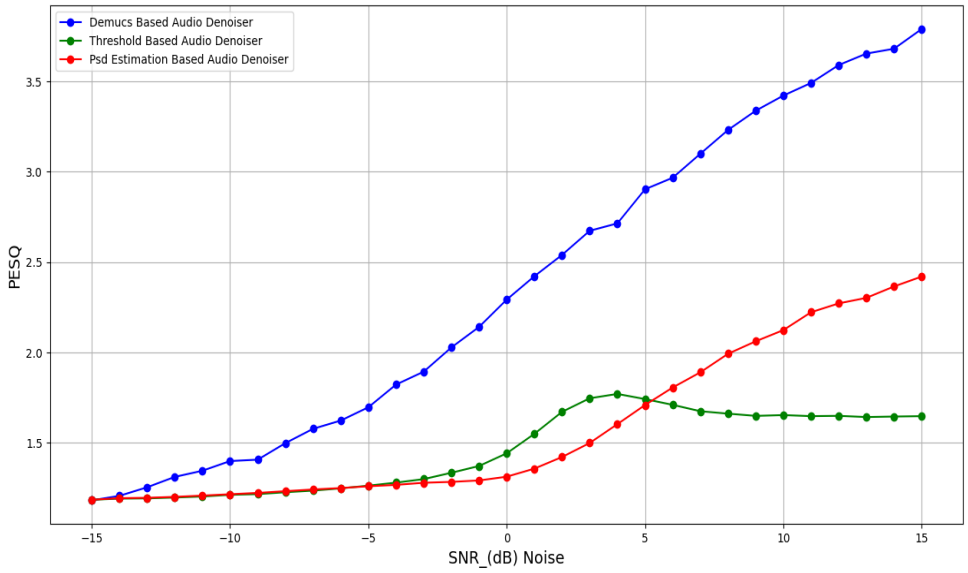
تعد طرح الخلفية Background Subtraction من أشـــهر الطرق الفعالـة لإيجـاد الأغراض المهمـة foreground بالفيديو والتي تعمل في الزمن الحقيقي real time. تتم العملية من خلال مقارنة كل إطار مع الإطار الســابق او مجموعة إطارات ســابقة له وبعدها يتم طرح هذا الإطار الســابق من الإطار الحالي وتكون نتيجة هذا الطرح هو التقاط الحركة (الأغراض المهمة في الإطار) أي يتم بذلك تحديد foreground. ســيتم اســتخدام إحدى خوارزميات فصـــل خلفية كخوارزمية ROI Detection. يوجد عدة طرق لفصـــل الخلفية تعد أكثر تعقيد ودقة مثل MOG ,CNT, KKN ,MOG2



الفصل الخامس: التجارب والتقييم Evaluation and Experiments

5.1 اختبار أنظمة إزالة الضجيج Denoising Systems Testing

يتم عادة اختبار أنظمة إزالة الضجيج عن طريق إضافة ضجيج من نوع White Noise بمقدار SNR معين على ملف صوتي نقي ومن ثم إدخال الملف الصوتي بعد إضافة الضجيج إليه إلى النظام المراد اختباره وحساب التشابه بين خرج هذا النظام والاشارة الأصلية النقية أو حساب مقدار الخسارة بالضجيج. سنقوم في هذا الفصل باختبار كل الأنظمة المقترحة الخاصة بإزالة الضجيج على ضجيج بقيم SNR بالمجال [15 , 15-] وحساب الإشارة الناتجة ثم اجراء عملية التقييم حسب مقياس التقييم الإدراكي لقياس جودة الصوت PESQ بمقارنة الإشارة الناتجة إشارة الصوت الخاصة بملف الصوت النقي بينت التجارب أن عمل تنعيم للتابع G(.) ومكونات الإشارة عبر الزمن من الممكن أن تؤدي إلى تخفيف أثر الضوضاء الموسيقية Musical Noise لذا قمنا بعمل هذا الخطوة في أول نظامين على الترتيب. سنستخدم ملف صوتي نقي باللغة العربية مدته 24 ثانية بقناة واحدة Mono وتردد قطع 16000Hz بالنسبة للبارمترات المستخدمة في النظام الأول Thresholding Based Audio Denoiser تم استخدام حجم نافذة 2048 = window size من نوع Hanning وتداخل overlapping = 1024 وعتبة Threshold = 25 ومعامل ضبابية Fuzziness = 0.03 ومعامل تنعيم Alpha = 0.05، وبالنسبة للنظام الثاني PSD Estimation Based Audio Denoiser فأيضاً تم استخدام حجم نافذة 1024 = window size من نوع Hanning وتداخل overlapping = 512 ومعامل تنعيم Alpha = 0.2 ومعامل موازنة لتقدير نسبة الضجيج DD Weighting Factor = 0.98، أما بالنسبة للنظام الثالث Demucs Based Audio Denoiser تم استخدام حجم نافذة 592 = window size وتداخل overlapping = 336 والنموذج demucs48 وكانت نتائج التقييم النهائية كالتالي ننوه هنا إلى أن درجة الـ SNR الأقل تعني وجود ضجيج أكبر:

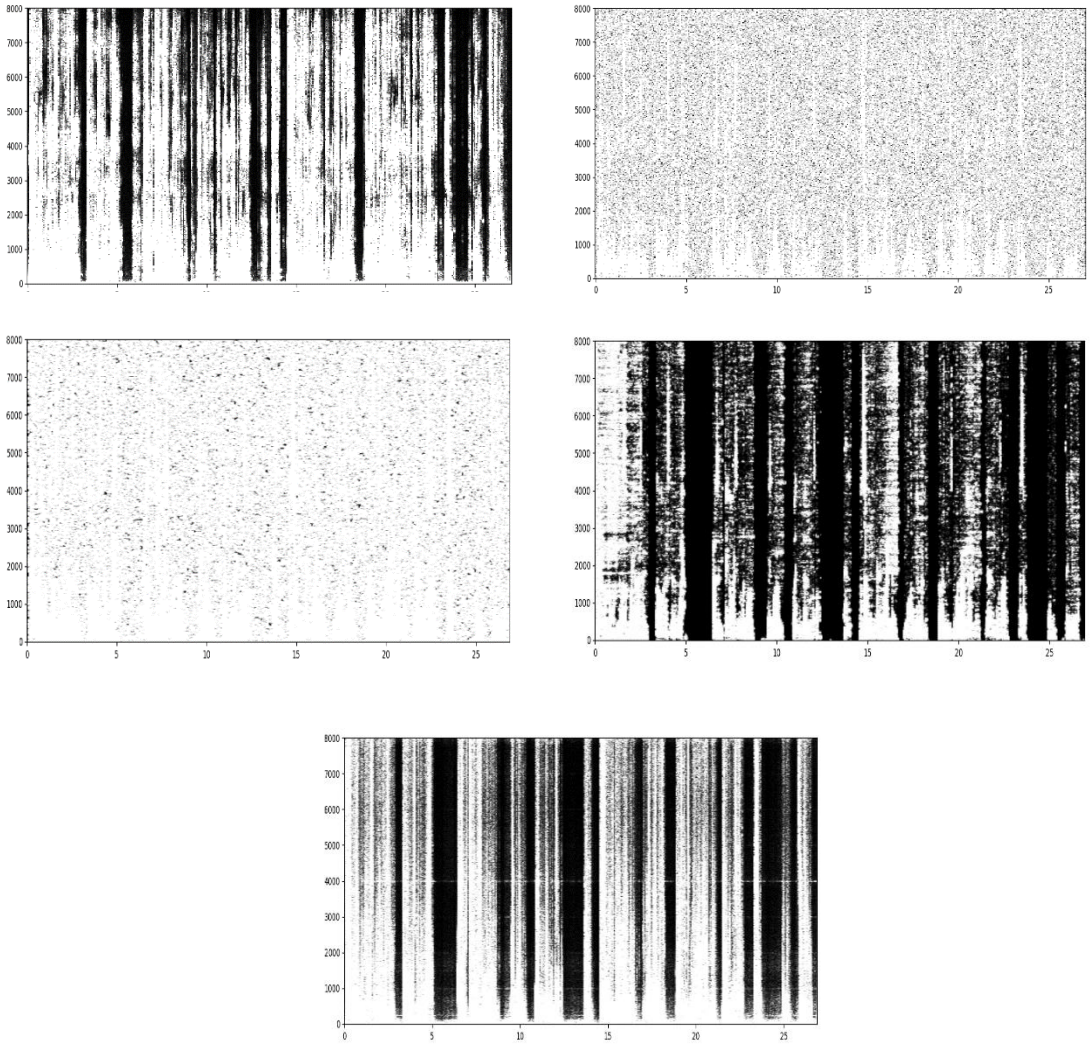


-المخطط-

يظهر الجدول التالي مقارنة في وقت التنفيذ من أجل نفس الملف الصوتي:

System	Execution Time (sec)
Thresholding	0.8
PSD Estimation	1.6
Demucs	14.4

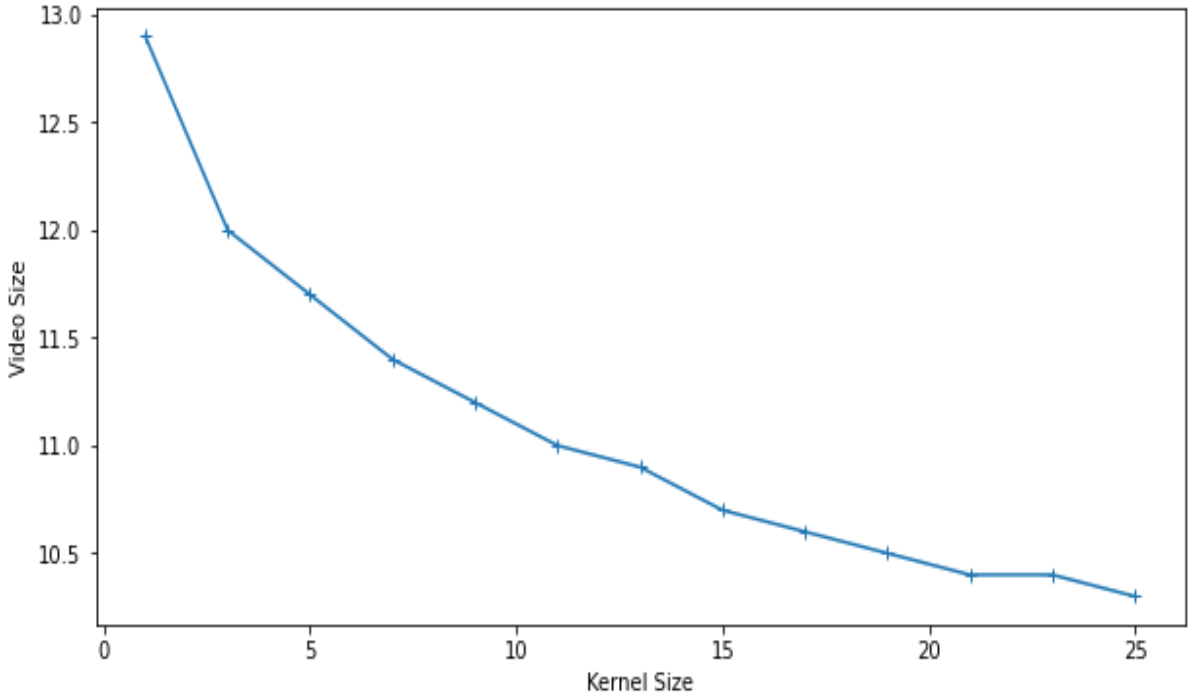
بتحليل النتائج السابقة نلاحظ تفوق التعلم العميق Demucs بشكل عام لكن في حال كانت الضوضاء مستقرة على طول الملف فتعتبر الطريقة الخاصة بالتعتيب Thresholding فعالة جداً وسريعة ونلاحظ أن الدقة التي حصلنا عليها من النموذج الرياضي الخاص بتقدير طيف الضوضاء PSD Estimation تعتبر مقبولة مقارنة بوقت التنفيذ حيث كانت أسرع بتسعة



بالترتيب من اليسار إلى اليمين الإشارة النقية – الإشارة الصاخبة – نتيجة التعتيب – نتيجة الـ PSD – نتيجة Demucs

5.2 اختبار أنظمة الضغط الموجه للفيديو Content Aware Video Encoding Systems Testing

تم اختبار نظام الضغط الموجه المعتمد على تحديد الأشخاص Selfie Based CAE على فيديو بأبعاد 480x854 بعدد إطارات 30 إطار/ثانية وكانت مدة المقطع 3 دقائق وخمسة ثواني بعد إزالة الصوت منه أصبح حجمه 13.6 MB كما تمد استخدام النموذج الأول General Model من Selfie Segmentation أخيراً تم استخدام Gaussian Blur كـ Low Pass Filter وتم تجريب قيم مختلفة من حجم الـ Kernel الخاص بالمرشح وكانت النتائج بعد المعالجة والضغط باستخدام نفس البارمترات وخوارزمية الضغط المستخدمة سابقا لهذا الفيديو كالتالي:



لوحظ أثناء التجريب أن الزيادة بالزمن عند زيادة حجم المرشـــح شـــبه مهملة وذلك نظراً لأبعاد الفيديو الذي تتم معالجته.
عموماً أتضح أيضاً أن حجم المرشح من 11 إلى 15 يعطي أفضل خسارة بالحجم مقارنة بالدقة.
بالنسبة للنظام الموجه المعتمد على فصل الخفية Background Subtraction CAE اتضح أنه يكون أفضل في حالة كانت كاميرة التصوير ثابتة وغير فعالة بشكل كبير عند ما تكون الكاميرة متحركة بشكل كبير أو الـــ Object ذو الأهمية الأكبر ساكن.

الفصل السادس: الخاتمة Conclusion

بالنظر إلى ما تم إنجازه ومطابقته مع المتطلبات الخاصة بالمشروع نجد أننا من أجل أنظمة معالجة الصوت اقترحنا أكثر من نموذج لإزالة الضجيج وقارننا فيما بينها لذا نقترح مستقبلاً عمل إضافات وتحسينات رياضية أخرى على النموذج الخاص بتقدير طاقة الكثافة الطيفية PSD الخاصة بالضجيج، أما بالنسبة لأنظمة معالجة الفيديو فنجد ان الطرق المستندة على اختيار منطقة الاهتمام ROI بشكل مستقل عن الحركة من الممكن أن تعطي نتائج أفضل وذلك كونها تحدد منطقة الاهتمام بغض النظر عن بعد الزمن، أخيراً بالنسبة للنظام الخاص بالترجمة فلم يتسنى لنا تحقيقه بشكل كامل بسبب أيقاف الخدمة الخاصة بالترجمة أثناء العمل على المشروع. بالعودة للمتطلبات غير الوظيفية نجد أن متطلب الأمان محقق وذلك عن طريق استخدام بروتوكول rtmps عوضاً عن rtmp أثناء البث مما يوفر أمان أكبر وكذلك الأمر بالنسبة للاستقبال من مخدم البث الأساسي Streamer وبالنسبة للمتطلب غير الوظيفي الخاص بسهولة التعديل والصيانة فنجد أن البنية المقترحة في مرحلة التصميم سهلة التعديل والإضافة مستقبلاً وتؤمن درجة عالية من الفصل في المهام بين الأنظمة الجزئية.

References المراجع

[1] Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep learning*. MIT press.

[2] Vaseghi, S.V., 2008. *Advanced digital signal processing and noise reduction*. John Wiley & Sons.

[3] Yu, G., Mallat, S. and Bacry, E., 2008. Audio denoising by time-frequency block thresholding. *IEEE Transactions on Signal processing*, *56*(5), pp.1830-1839.

[4] Wiener, N., 1949. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications Application: With Engineering Applications*. MIT press.

[5] Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, *27*(2), pp.113-120.

[6] Ephraim, Y. and Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE transactions on acoustics, speech, and signal processing*, *33*(2), pp.443-445.

[7] Marzinzik, M. and Kollmeier, B., 2002. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Transactions on Speech and Audio Processing*, *10*(2), pp.109-118.

[8] Derakhshan, N., Akbari, A. and Ayatollahi, A., 2009. Noise power spectrum estimation using constrained variance spectral smoothing and minima tracking. *Speech Communication*, *51*(11), pp.1098-1113.

[9] Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on speech and audio processing*, *9*(5), pp.504-512.

[10] Wiesener, C., Flohrer, T., Lerch, A. and Weinzierl, S., 2010, May. Adaptive Noise Reduction for Real-time Applications. In *Audio Engineering Society Convention 128*. Audio Engineering Society.

[11] Wolfe, P.J. and Godsill, S.J., 2003. Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. *EURASIP Journal on Advances in Signal Processing*, *2003*(10), pp.1-9.

[12] Athaley, A. and Dutta, P., 2017. Audio Signal Denoising Algorithm by Adaptive Block Thresholding using STFT. *International Journal of Trend in Scientific Research and Development*, Volume-1(Issue-6), pp.289-300.

[13] Defossez, A., Synnaeve, G. and Adi, Y., 2020. Real time speech enhancement in the waveform domain. *arXiv preprint arXiv:2006.12847*.

[14] Wang, L., Zheng, W., Ma, X. and Lin, S., 2021. Denoising speech based on deep learning and wavelet decomposition. *Scientific Programming*, *2021*.

[15] Zünd, F., Pritch, Y., Sorkine-Hornung, A., Mangold, S. and Gross, T., 2013, September. Content-aware compression using saliency-driven image retargeting. In *2013 IEEE international conference on image processing* (pp. 1845-1849). IEEE.

[16] Reznik, Y.A., Li, X., Lillevold, K.O., Jagannath, A. and Greer, J., 2019, July. Optimal multi-codec adaptive bitrate streaming. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* (pp. 348-353). IEEE.

[17] Soyak, E., Tsaftaris, S.A. and Katsaggelos, A.K., 2010, March. Content-aware H. 264 encoding for traffic video tracking applications. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 730-733). IEEE.

[18] Hegazy, M., Diab, K., Saeedi, M., Ivanovic, B., Amer, I., Liu, Y., Sines, G. and Hefeeda, M., 2019, June. Content-aware video encoding for cloud gaming. In *Proceedings of the 10th ACM multimedia systems conference* (pp. 60-73).

[19] Devi, N.,Thkur ,V. ,2017 ,August. Content Aware Video Compression: An Approach To VOS Algorithm. Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, Volume-4 | Issue-4

[20] Sun, L., Décombas, M. and Lang, J., 2016, June. Video Object Segmentation for Content-Aware Video Compression. In *2016 13th Conference on Computer and Robot Vision (CRV)* (pp. 116-123). IEEE.

[21] Hendriks, R.C., Heusdens, R. and Jensen, J., 2010, March. MMSE based noise PSD tracking with low complexity. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4266-4269). IEEE.

[22] Défossez, A., Usunier, N., Bottou, L. and Bach, F., 2019. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254*.

[23] Yong, P.C., Nordholm, S. and Dam, H.H., 2013. Optimization and evaluation of sigmoid function with a priori SNR estimate for real-time speech enhancement. *Speech communication*, *55*(2), pp.358-376.