A Comparative Study of Methods for Separating Audio Signals

Riham J. Issa ***Yusra F. Al-Irhaym***
rihamjassim11@gmail.com *yusrafaisalcs@gmail.com*
College of Computer Sciences and Mathematics
University of Mosul, Mosul, Iraq

Received on: 14/01/2020 **Accepted on: 28/03/2020**

ABSTRACT

The process of separating signal from a mixture of signals represents an essential task for many applications including sound signal processing and speech processing systems as well as medical signal processing. In this paper, a review of sound source separation problem has been presented, as well as the methods used to extract features from the audio signal, also, we define the Blind source separation problem and comparing between some of the methods used to solve the problem of source separation.

Keywords: Blind Source Separation, Independent Component Analysis, Deep Neural Networks.

دراسة مقارنة لطرق فصل الإشارات الصوتية

رهام جاسم عيسى ***يسرى فيصل الارحيم***
قسم علوم الحاسوب، كلية علوم الحاسوب والرياضيات
جامعة الموصل، الموصل، العراق

تاريخ استلام البحث: ٢٠٢٠/٠١/١٤ **تاريخ قبول البحث: ٢٠٢٠/٠٣/٢٨**

الملخص

إن عملية فصل الاشارة من خليط من الاشارات تمثل مهمة اساسية لتطبيقات مختلفة منها انظمة معالجة الاشارة الصوتية ومعالجة الكلام فضلا عن معالجة الاشارات الطبية، في هذا البحث تم استعراض مشكلة فصل مصادر الصوت، فضلا عن طرق استخلاص الميزات من الاشارة الصوتية، بالإضافة الى وصف لمشكلة الفصل الاعمى للمصادر ومقارنة بعض الطرق المستخدمة في حل مشكلة فصل المصادر.

الكلمات المفتاحية: الفصل الاعمى للمصادر، تحليل المكونات المستقلة، الشبكات العصبية العميقة.

1. المقدمة

يعد الكلام أحد اساليب التواصل الاساسية حيث يمكن للأذن البشرية إدراك التفاصيل الدقيقة للكلام الذي تتلقاه وبالتالي فإنها قادرة على تمييز متكلم عن الاخر[3]. تعد انظمة السمع والتكلم البشرية انظمة فريدة، اذ تعتبر القدرة على التكلم والسمع أحد الاساسيات اليومية التي يمكن ان يمارسها الانسان والتي لا تتطلب جهدا لأدائها [16]. اذ تحلل الاذن البشرية تغير الضغط في الهواء الى ترددات مختلفة وذلك من خلال عمليه مشابهة لتحويل فورير لإشارة الضغط p(t) حيث تمثل (t) الزمن وتحويل $|p(w)|^2$ اذ تمثل (w) التردد الى الدماغ [14]. عندما يتحدث عدة اشخاص في وقت واحد في غرفة كما في حفل الكوكتيل (cocktail party) فانه بإمكان الاذن البشرية

الاصغاء الى مزيج من الاشارات مرغوبة واخرى غير مرغوبة ، ونظرا لما يتمتع به الانسان من ذكاء فانه بإمكانه تحديد مصادر الصوت النشطة فضلا عن قدرته على التركيز على مصدر صوت واحد وتجاهل بقية الاصوات باعتبارها ضوضاء الخلفية [1]. الا ان المعالجة الحاسوبية للإشارات الصوتية في العديد من تطبيقات معالجة الاشارة تعمل على تحليل ومعالجة الاشارة المعزولة بدقة أكثر من معالجة اشارة مكونة من خليط من الاشارات [17]. ان عملية فصل المصادر الصوتية جزءا مهما في انظمة تحسين واسترجاع الاشارات الصوتية وتعرف على انها عملية استخراج مصادر الاشارات الصوتية الموجودة في خليط من الاشارات الصوتية [28]. وهناك العديد من الطرق لحل مشكلة فصل مصادر الصوت سواء كان التسجيل الصوتي من قناة واحدة او متعدد القنوات[12]. ونظرا لتطور حقل الشبكات الاصطناعية والتعلم العميق، ظهرت عدة معماريات لأنظمة التعلم لغرض فصل مصادر الصوت من قناة واحدة [25]، حيث اظهرت الشبكات الاصطناعية العميقة تميزا في الاداء في مهام فصل المصادر كما في فصل الصوت المتكلم وفصل الصوت الموسيقى [27].

2. استخلاص الميزات (Feature Extraction)

يعرف استخلاص الميزات على انه عملية تحديد معلومات وخصائص ثابتة وكافية في الاشارة لصنف معين ومختلفة بين الاصناف المختلفة لاستخدامها في عملية الفصل[9]، وهناك عدة طرق لاستخلاص الميزات والتي تهيء الاشارة لعملية الفصل منها معاملات التنبؤ الخطي (Linear Prediction Coding (LPC)) ومعاملات درجة النغم (Mel-Frequency Cepstrum Coefficients (MFCC)) وغيرها[5].

1.1 العبورات الصفرية (Zero Crossings)

تعتبر احد ابسط طرق استخلاص الميزات في الاشارة الصوتية والتي تستخدم عادة في تطبيقات التعرف على الكلام. حيث يحدث العبور الصفري في اشارات الزمن المتقطع في حالة كون العينات المتعاقبة ذات اشارات جبرية مختلفة، بينما يمثل معدل العبور الصفري مقياسا لعدد مرات التي تمر سعة اشارة الكلام خلال قيمة الصفر في فترة زمنية محددة [2].

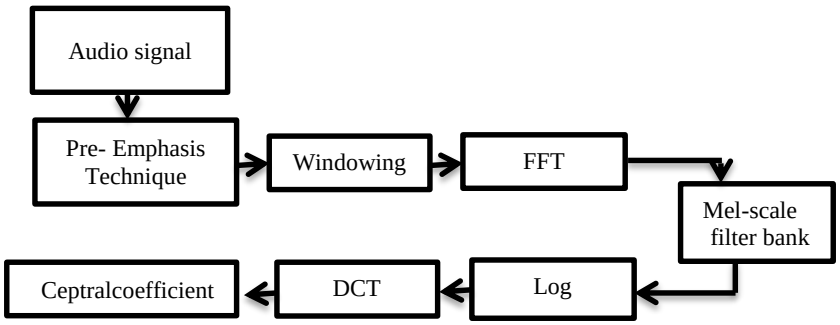
2.2 تحويل فوريير للوقت القصير (Short Time Fourier Transform (STFT))

تستخدم هذه الطريقة لاستخلاص الميزات الزمنية والطيفية من خلال تحويل اشارة الصوت من المجال الزمني الى المجال الترددي، تحتوي الـ(STFT) على مجموعة من الباراميترات التي من الممكن ان تؤثر على حجم ودقة الناتج منها حجم النافذة المستخدمة مع تحويل فوريير السريع (Fast Fourier Transform)[22].

2.3 معاملات درجة النغم (Mel-Frequency Cepstrum Coefficients (MFCC))

تمثل الـ(MFCC) أحد الطرق الاكثر شيوعا في استخلاص ميزات الصوت في مهام التعرف على الكلام فهي تستند الى الادراك البشري للسمع في استخلاص الميزات، اذ تمثل القنوات الصوتية التي تنتج الكلام كمرشحات لإنتاج الكلام، تعمل الـ(MFCC) على ايجاد علاقة بين التردد الحقيقي ودرجة النغم للكلام [24]، حيث يتم تحويل الاشارة المدخلة الى المجال الترددي باستخدام تحويل فوريير، وللحد من تشويه التردد الناتج بسبب التجزئة يتم استخدام نافذة التحليل (Hammimg) قبل تطبيق تحويل فويرير. ومن ثم تحويل التردد من مقياس هرتز الى مقياس ميل باستخدام بنك الترشيح، ومن ثم استخدام تحويل (DCT) لاستخراج متجه الميزات بعد

حساب اللوغارتم لقوة الطيف [5]. والشكل (1) يوضح الخطوات الاساسية لاستخلاص الميزات من الاشارة الصوتية باستخدام (MFCC).



الشكل (1) مخطط لمراحل الاشارة باستخدام (MFCC) [24]

3. الفصل الاعمى للمصادر (Blind Source Separation)

تهدف عملية فصل المصادر الى استرجاع الاشارات الاصلية من خليط من الاشارات، والمثال الاكثر شيوعا هو عملية فصل المصادر في مسألة حفل الكوكتيل (Cocktail Party Problem) والتي يحاول الشخص المستمع الى اتباع صوت متكلم محدد من بين مجموعة من اصوات للمتكلمين في غرفة واحدة وفي نفس الوقت[17].

3.1 تحليل المكونات المستقلة ((Independent Component Analysis (ICA))

تمثل الـ(ICA) أحد الطرق الكلاسيكية واكثرها شهرة في مجال فصل الاشارات من مزيج خطي من مكونات احصائية مستقلة، حيث تستخدم في مجال الفصل الاعمى للمصادر كما تستخدم ايضا في مجال استخلاص الميزات [1]. وتتطلب المعادلة القياسية لـ (ICA) ان يكون عدد اشارات المصدر لا يتجاوز عدد الاشارات، فاذا كان عدد اشارات المصادر أكبر من عدد الاشارات يعرف بالمزيج الناقص او غير المكتمل[13]. كما ان اختلاف الرئيسي بين الـ(ICA) وطرق الـ(BSS) في البيانات (مستمرة، متقطعة) اضافة الى عشوائية المتغيرات، فالمكونات في (ICA) يجب ان تكون مستقلة بينما في (BSS) فان الاستقلالية غير ضرورية [4].

3.2 الشبكات العصبية العميقة (Deep Neural Networks)

تستخدم الشبكات العصبية العميقة عناصر بسيطة مستوحاة من الانظمة العصبية البيولوجية لجمع بين طبقات معالجة لا خطية، فهي تتكون من طبقة ادخال وعدد من الطبقات المخفية وطبقة اخراج [8]، وتعد الشبكات العصبية العميقة وخاصة الشبكات العصبية المتكررة (RNNs) من الشبكات الفعالة جدا في عملية فصل مصادر الصوت في حال توفر مجموعة بيانات كبيرة خاضعة للإشراف لتحسين اوزان الشبكة [23].

في هذا الجزء سيتم ايجاز انواع شبكات التعلم العميق الخاضع للإشراف وهي شبكة التغذية الامامية متعددة الطبقات ((Feed-Forward Multilayer Perceptron's (MLPs))، والشبكة العصبية الالتفافية ((Convolutional Neural Networks (CNNs))، والشبكة العصبية التكرارية، Recurrent Neural) (Networks (RNNs)) وبالشبكات التوليدية التنافسية ((Generative Adversarial Networks (GANs)) [7].

3.2.1 الشبكة العصبية ذات التغذية الامامية ((Feed-Forward Neural Networks(FNN)

تستخدم الشبكة العصبية القياسية ذات التغذية الامامية والمتصلة كليا (FNN) لتخمين طيف المصدر، ولغرض استغلال السياقات الزمنية، تستخدم عدد نوافذ متسلسلة كإدخال[18]. ويعد نموذج شبكة التغذية الامامية متعددة الطبقات ((Feed-forward Multilayer perceptron (MLP) من اكثر النماذج شيوعا وابسطها في نماذج التعلم العميق التقليدية، حيث تكون الخلايا العصبية الموجودة في الطبقة l_i متصلة بجميع الخلايا في الطبقة l_{i-1} لكل $i \in [l, L]$ لذا فان هيكلية هذه الشبكة تكون مرتبطة كليا (fully-connected) بالإضافة الى الاوزان في الشبكة ،يمكن تمثيل النموذج العام غير الخطي لمدخلات السلسلة الزمنية X بالمعادلة (1).

$$A_{li} = f(w_{li} * X + b) \dots\dots\dots (1)$$

حيث يمثل w_{li} مجموعة الاوزان مع طول وعدد الابعاد التي تطابق المدخلات $X's$ ، بينما تمثل b الوزن التحيزي، اما A_{li} فهي دالة التفعيل للخلية في الطبقة l_i [9].

عملية التعلم لشبكة الـ(MLP) هي المعالجة التي تكيف الارتباطات الموزونة لغرض الحصول على اقل فرق ما بين اخراج الشبكة والهدف وغالبا ما يتم استخدام خوارزميات الانتشار الخلفي (Back Propagation) التي تعتمد على تقانات تدرج النسب[10]. ولغرض زيادة كفاءة تدريب شبكات (MLP) يتم استخدام تقانات الانتشار الخلفي مع دالة التنشيط (Sigmoid) والتي تعمل على حضر تدريب بعض طبقات الـ(MLP) وتدريب الطبقات المتبقية في نفس الوقت وذلك لتجنب ما يسمى بظاهرة التلاشي التدريجي (Vanishing Gradient Problem) [16]، وتشير الى ان التدرجات المحسوبة من اشارات الخطأ المرتدة من الطبقات العليا الى الطبقات السفلى تبدا بالتناقص او التلاشي تدريجيا، ونتيجة لهذا التلاشي فان الاوزان في الطبقات السفلى لا تعدل اوزانها كثيرا وهذا يؤدي الى ضعف عملية التعلم في الطبقات السفلية خلال عملية التدريب[7].

3.2.2 الشبكة العصبية الالتفافية ((Convolutional Neural Network (CNN)

يعتمد هذا النوع من الشبكات العصبية على التفاف (convolving) مدخلات الشبكة مع نواة قابلة للتعلم. حيث ان الادخال لدالة التنشيط يتمثل بناتج التفاف الادخال الى الطبقة ومجموعة من الباراميترات (w) الخاصة بالطبقة والتي تسمى بالمرشحات او النواة. يتم في الطبقة الالتفافية (Convolutional Layer) حساب قنوات ميزات متعددة قناة من النواة المقابلة لها، اما طبقات التجميع (Pooling Layers) فتستخدم لتبسيط قنوات الميزات المعلمة، وعادة ما تضاف هذه الطبقات فوق الطبقات الالتفافية، ففي الحالات ثنائية الابعاد فان الالتفاف الرياضي بين مصفوفة الاشارة $X \in R^{j \times I}$ ومصفوفة القيم $W \in R^{M \times N}$ حيث ان $M<J$ و $N<I$ يمكن ان تعرف بالمعادلة رقم (2):

$$S_{j.i} = \sum_{m=1}^{M} \sum_{n=1}^{N} x_{j,i} + m, i + n\ \omega_{mn} \dots\dots\dots (2)$$

حيث تمثل كل من $x_{j,i}$ و ω_{mn} هي العنصر j, i والعنصر m, n و لـ X و W على التوالي [16]. تقلل شبكات الـ (CNN) عدد الباراميترات بشكل كبير كما تحقق التعميم من خلال تبادل الباراميترات لنموذج الانماط المحلي في الادخال، الا ان الشبكة القياسية تتطلب عمقا كبيرا لتغطية السياقات الطويلة وهذا يزيد من صعوبة عملية التدريب [18]. وعادة ما يتم تدريب شبكة الـ(CNN) باستخدام خوارزمية الانتشار الخلفي الا انها تقلل من مشكلة التلاشي التدريجي، كما ان الخوارزمية المستندة الى التدرجات تدرب الشبكة بشكل كامل لتقليل معيار الخطأ بصورة مباشرة يجعل هذا النوع من الشبكات قادرا على تحقيق اوزان مثلى [15].

3.2.3 الشبكة العصبية التكرارية ((Recurrent Neural Networks (RNNs)

تعد الشبكة العصبية التكرارية واحدة من اولى الشبكات القادرة على تذكر مدخلاتها، لأنها تتضمن ذاكرة داخلية ما يجعلها أكثر ملائمة للاستخدام مع البيانات المتسلسلة مثل الكلام واللغة[20]. حيث يتم في كل خطوة زمنية ارسال المدخلات من خلال الشبكة التكرارية، حيث تستلم العقد المدخلات عبر الحواف متكررة، بينما تستلم مدخلات التنشيط من متجه المدخلات الحالي ومن العقد المخفية في الحالة السابقة للشبكة، و يتم حساب الاخراج من الحالة المخفية في الخطوة الزمنية الحالية، ومن خلال الربط المتكرر فان متجه الادخال السابق في الخطوة الزمنية السابقة يؤثر على ناتج الاخراج الحالي في الخطوة الزمنية الحالية[8]. ويمكن تمثيل الشبكة التكرارية ذات الطبقة الواحدة بالمعادلة(3):

$$h^{(n)} = \phi(Wx^{(n)} + Vh^{(n-1)} + b \dots\dots\dots (3)$$

وتمثل $\phi(.)$ دالة التنشيط بينما b فهي المتجه التحيزي، اما V وW فتمثل قيم المصفوفات بينما $h^{(n)}$ فهو اخراج الشبكة في الزمن n [16]. ولغرض التخلص من مشكلة التلاشي التدريجي، تستخدم معماريات مختلفة لشبكات الـ(RNN) تدعى بـ (Gated-RNNs) ومن هذه المعماريات ما يسمى بـ (Long Short-Term Memory (LSTM)-RNN) والتي تعد اكثر تعقيدا حسابيا مقارنة بـ(RNN) القياسية الا انها سهلة التدريب وتعمل بشكل افضل اذا كانت (N) كبيرة. كما تعد هذه البنية هي أكثر المعماريات شيوعا والتي تستخدم في تطبيقات معالجة الصوت مثل تمييز الصوت وتحسين وفصل الصوت[16].

3.3.3 الشبكات الخصومية ((Generative Adversarial Networks (GANs)

في السنوات الاخيرة قدمت هذه الشبكة لحل مشكلة فصل المصادر، حيث يفرض في هذه الطرق ان مرشحات المزج في اشارة من القناة الواحدة معلومة[21]، وتمثل هذه الشبكات اطارا لشبكتين فرعيتين وهما (Generator and Discriminator(G and D)) حيث تكون عملية التعلم تنافسية[26]. حيث تدرب شبكة التوليد الـ(Generator) لإنتاج عينات من هدف موزع معين. ولتحقيق ذلك، تستخدم شبكة التمييز (Discriminator) للتمييز بين العينات الحقيقية من مجموعة البيانات والعينات المزيفة الناتجة من شبكة المولد [6]. ويمكن ان يكون تدريب الشبكتين معا وذلك باستخدام المعادلة رقم (4)

$$\min_{G}\max_{D} V_{cGAN}(G,D) = E_{x,\sim P_{data(x)}}[\log D\,(x)] + E_{z}\sim P_{G(z)}\big[\log(1 - D\big(G(z)\big)\big] \dots\dots(4)$$

اذ تمثل x عينات بيانات حقيقية تم اخذها من التوزيع P_{data}، بينما $G(z)$ تعتمد على البيانات الاصطناعية يتم اخذها من التوزيع P_G [29].

4. قواعد البيانات المستخدمة في عملية الفصل

تتكون قاعدة البيانات من 1000 مقطع غنائي بأصوات مختلفة (ذكر_انثى) تتراوح مدتها ما بين 4−13 ثانية ومعدل عينة 16,000 هرتز. اما قاعدة البيانات (DSD100) فتتضمن 100 مقطع موسيقي بتردد عينة 44,100 هرتز، مأخوذة من جزء من مهمة فرعية تدعى MUS لتقييم فصل الاشارات الصوتية يحتوي كل مقطع على اربعة مصادر صوتية مختلفة (الصوت الغنائي، الغيتار، الطبل، واصوات اخرى)، كما يبلغ متوسط مدة كل مقطع موسيقي 4 دقائق و10 ثوان [29].

5. مقارنة نتائج طرق الفصل باستخدام الشبكات العصبية العميقة

إن عمليات فصل الصوت باستخدام قناة واحدة وباستخدام معماريات مختلفة للشبكات العصبية العميقة اصبحت موضع اهتمام العديد من الباحثين، حيث قدمت العديد من الطرق لتحقيق افضل نتائج في عمليات الفصل، وفي هذا الجزء سيتم مناقشة نتائج فصل مصادر الصوت الغنائي عن الخلفية الموسيقية باستخدام الشبكات العصبية العميقة لثلاث شبكات عصبية وهي (SVSGAN) و(MMDenseLSTM) و (Stacked Hourglass Model) وفقا لبعض المقاييس وهي نسبة الاشارة الى التشوه (Signal-to-Distortion Ratio (SDR)) ونسبة الاشارة الى التداخل (Signal to Interference Ratio (SIR)) ونسبة الاشارة الناتجة من عملية الفصل بصورة عامة (Signal to Artifacts Ratio (SAR)).

ومن ابرز الهيكليات المستخدمة في فصل الصوت الغنائي عن الخلفية الموسيقية باستخدام قناة واحدة هي شبكة (SVSGAN) التي قدمها الباحث Zhe-Cheng Fan واخرون في سنة 2017 وتتألف من شبكتين من الشبكات العصبية التقليدية العميقة، الاولى شبكة التوليد G مدخلاتها اطياف الخليط اذ تعمل على توليد اطياف الصوت الغنائي واطياف الموسيقى الخلفية والثانية شبكة التوليف D التي تعمل على تمييز الصوت النظيف من بين الاطياف المتولدة، كما تم اعتماد حجم الاطياف كميزة ويمكن حسابها باستخدام تحويل فورير القصير ومن ثم استخدام مرشح التردد الزمني (Ideal Ratio Mask IRM) لتحسين ناتج عملية الفصل، وقد تم استخدام شبكة (GAN) الشرطية التي تسمح للمولد G(z,y) بتوليف بيانات واقعية تحت سيطرة (y) كما تسمح بالسيطرة على ناتج النموذج التوليفي D المتمثل بـ D(x,y) بواسطة متجه السياق (y) كما في معادلة رقم (5) التي تمثل دالة الهدف، واظهرت النتائج ان استخدام هذه الهيكلية قد حقق تحسينا في اداء عملية الفصل [29].

$$\min_{G}\max_{D} V_{cGAN}(G,D) = E_{z,s_c\sim P_{data(z,s_c)}}[\log D\,(s_c,z)] + E_z\sim P_{G(z)}[\log(1-D(G(z),z))] \quad \ldots\ldots(5)$$

حيث ان (s_c) عبارة عن سلسلة من (y_1) و (y_2) بينما ناتج G(z) عبارة عن الاطياف التي تم تخمينها بواسطة الا طياف المدخلة (z) اما شبكة التوليف (D) فيتم السيطرة عليها بواسطة قيم المتغيرات لأطياف الادخال (z).

بينما قدم الباحث Naoya Takahashi واخرون في سنة 2018 طريقة تهدف لفصل الصوت من خلال دمج الشبكة العصبية الالتفافية والشبكة العصبية التكرارية باستخدام معماريتين مختلفتين وهما معمارية الـ (Multi-scale Multi- Band DenseNet(MMDenseNet)) ومن ثم دمج dense block مع شبكة اخرى وهي (Long short term memory (LSTM)) حيث تم استخدام تحويل فورير القصير للحصول على حجم الطيف لإشارة الخليط، وقد تم تدريب الشبكة لتخمين طيف المصدر وذلك من خلال تقليل نسبة متوسط مربع الخطأ (Mean square Error MSE) وقد اظهرت هذه الطريقة ان استخدام معمارية (MMDenseLSTM) وبالرغم من قلة عدد الباراميترات المستخدمة تفوقت على نوع اخر من الشبكات المدمجة وهو (MMDenseBLSTM)، كما تفوقت ايضا عن استخدام المرشح الثنائي ((Ideal Binary Mask (IBM) في فصل الصوت الغنائي عن الموسيقى في شبكة مدربة على 900 مقطع غنائي[18].

نوع اخر من الشبكات العصبية قدمها الباحث Sungheon Park واخرون في 2018 وهي الشبكة العصبية الالتفافية وبالأخص هيكلية (Stacked hourglass network) في فصل مصدر الموسيقى، وهي شبكة مشابهه لشبكة الـ(U-Net) حيث تم في هذه الطريقة اعتماد طيف الصورة مختلف الابعاد لتوليد مرشح لكل مصدر

موسيقي ومن ثم تحسين المرشح الذي تم تخمينه من خلال تمريره في (stacked hourglass model). حيث يمكن وباستخدام شبكة واحدة فصل العديد من المصادر الموسيقية وقد حققت هذه الطريقة نتائج تنافسية قابلة للمقارنة مع احدث اساليب فصل الاصوات الموسيقية المتعددة ومهام فصل الصوت الغنائي[27]، والجدول (1) يوضح الفرق بين المقاييس باستخدام الشبكات الثلاث وباستخدام قواعد بيانات مختلفة.

جدول (1) نتائج (SIR-SDR-SAR) للشبكات العصبية

SAR db	SIR db	SDR db	الصوت	قاعدة البيانات	نوع الشبكة	رقم المصدر
10.14	9.84	6.57	Vocal+ background music	MIR-1K	DNN(Baseline)	[29]
10.32	9.86	6.69	Vocal+ background music	MIR-1K	(SVSGAN)	[29]
-	-	3.73	Bass	DSD100	(MMDenseLSTM)	[18]
		5.46	Drums			
		6.31	Vocal			
		4.33	Others			
		12.37	Accompaniments			
12.53	16.01	10.51	Singing voice	MIR-1K	4 stacked hourglass networks	[27]
12.36	14.24	9.88	Accompaniments			
-	-	1.77	Bass	DSD100		
-	-	4.11	Drums			
-	-	5.16	Vocals			
-	-	2.36	Other			
-	-	3.91	Bass	DSD100	MMDenseNet	[19]
-	-	5.37	Drums			
-	-	6.00	Vocals			
-	-	3.81	Other			
		12.10	Accompaniments			

من خلال الجدول (1) نلاحظ ان شبكة (SVSGAN) المستخدمة لفصل الصوت الغنائي عن الموسيقى وعلى فرض ان خليط الاشارة مكون من جزأين الاول هو الصوت والاخر هو الموسيقى، اظهرت تحسنا في اداء عملية الفصل حيث تم مقارنتها بالشبكة العصبية العميقة التقليدية تتكون من ثلاث طبقات مخفية بالإضافة الى 1024 خلية في كل طبقة، يمكن ملاحظة ان تطبيق الطريقتين على قاعدة البيانات MIR-1K اظهرت زيادة نسبة الاشارة الى التشوه (SDR) باستخدام الـ (SVSGAN) بمقدار 0.12db، كما ازدادت نسبة الاشارة الى التداخل (SIR) بمقدار 0.02 db وهي نسبة قليلة وقد لا تكون واضحة، اما نسبة (SAR) فقد ازدادت بمقدار 0.18db.

اما نتائج تدريب شبكة (4-stacked hourglass network) باستخدام قاعدة البيانات MIR-1K ومن خلال حساب الوسيط لكل من المقاييس نجد ان فصل الصوت الغنائي حقق اعلى نسب في المقاييس الثلاث بينما الصوت الموسيقي فقد حقق اعلى نسب بمقياسي (SDR,SIR)، وبذلك فانه كلما ازداد عدد نماذج stacked hourglass المستخدمة كلما كانت النتائج افضل، اما عند تدريب الشبكة باستخدام قاعدة البيانات DSD100 نلاحظ ان نسبة الفصل جيدة لفصل الصوت الغنائي وصوت الطبل بينما تكون النتائج اقل جودة في فصل صوت البيس كيتار وبقية الاصوات وذلك بسبب تقارب صوت البيس كيتار من صوت الكيتار العادي وبالتالي يؤثر في عملية التدريب عندما يتم تدريب الاصوات بنفس الشبكة[28–27]. اما شبكة (MMDenseLSTM) المستخدمة لفصل اصوات الادوات الموسيقية نجد ان الطريقة المقترحة تحقق تفوقا واضحا في فصل المصادر بالرغم من قلة عدد الباراميترات المستخدمة اذا ما تم مقارنة هذه الطريقة مع (MMDenseNet)، نجد نسبة نسبه الـSDR

لاشارة الصوت الغنائي ازدادت بمقدار 0.31db عند استخدام MMdenseLSTM، كما ازدادت بمقدار 0.09db بالنسبة لصوت الطبل اما بقية الاصوات الموسيقية 0.52db وهي نسبة عالية جدا، اما .Acco فكانت نسبة الفرق بمقدار 0.63db .

6. الاستنتاجات

تختلف الطرق المستخدمة في عمليات فصل المصادر الصوتية في مدى قوتها في تحقيق عملية الفصل كما تختلف بمدى ملائمتها لإداء المهام المطلوبة، وفي مهمة فصل الاصوات الموسيقية، تم اجراء مقارنة لنتائج ثلاث طرق للفصل بالاعتماد على مقاييس اهمها نسبة التشوه ونسبة التداخل في الاشارة التي تم استرجاعها، وكما موضح في الجدول (1) فان استخدام الشبكة العصبية الالتفافية العميقة (4-stacked hourglass network) (المصممة لتخمين حالة الانسان في الصور الرقمية الملونة) لفصل الصوت الغنائي عن الموسيقي اعطت نتائج افضل من استخدام الشبكة العصبية (SVSGAN) حيث حققت زيادة في النسب (SDR) بمقدار 3.82db و(SIR) بمقدار 4.64db اما (SAR) بمقدار 2.21db وهي نسب عالية ما يجعلها اكثر ملائمة في عملية فصل الصوت الغنائي عن الخلفية الموسيقية، اما في فصل اصوات الادوات الموسيقية نلاحظ ان الشبكة العصبية (MMDenseLSTM) حققت تفوقا كبيرا مقارنة بشبكة (4-stacked hourglass network) و(MMDenseNET) في فصل الاصوات الموسيقية ماعدا اشارة البيس كيتار. ومن هنا نستنتج ان شبكة (4-stacked hourglass network) تحقق نتائج أفضل في مهام فصل الصوت الغنائي عن الموسيقى، بينما شبكة (MMDenseLSTM) فهي تحقق نتائج فصل أفضل بشكل عام من الطرق الاخرى.

المصادر

[1] Abouzid Houda, Chakkor Otman, 2017, "A Novel Method Based on Gaussianity and Sparsity for Signal Separation Algorithms", International Journal of Electrical and Computer Engineering (IJECE), Vol.7, No.4.

[2] Bachu R.G., Kopparthi S., Adapa B., Barkana B.D., 2014, "Separation Of Voiced and Unvoiced Using Zero Crossing Rate And Energy of The Speech Signal", Bulletin of The Polish Academy of Sciences, Vol. 62, No. 3.

[3] Bhavana V.S, Pradip K. Das, 2019, "Speaker Verification Using Simple Temporal Features and Pitch Synchronous Cepstral Coefficients", Arxiv:1908.05553v1.

[4] Birmingham Hang Guan, Anand Rangarajan, 2018, "Signals as Parametric Curves: Application to Independent Component Analysis and Blind Source Separation", http://Arxiv.Org/Abs/1807.03442.

[5] C. Sunitha, E. Chandra, 2015, "Speaker Recognition Using Mfcc And Improved Weighted Vector Quantization Algorithm", International Journal Of Engineering and Technology (IJET), ISSN: 0975-4024, Vol. 7, No .5.

[6] Daniel Stoller Daniel, Sebastian Ewert, Simon Dixon, 2019, "Training Generative Adversarial Networks from Incomplete Observations Using Factorised Discriminators", International Conference on Learning Representations (ICLR), Arxiv:1905.12660v1.

[7] Deliang Wang, Jitong Chen, 2018, "Supervised Speech Separation Based on Deep Learning: An Overview , IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), Vol. 26, No. 10.

[8] Faisal Mohammad, Ki-Boem Lee, Young-Chon Kim, 2018, "Short Term Load Forecasting Using Deep Neural Networks", International Symposium on Information Technology Convergence (ISITC), arXiv:1811.03242 .

[9] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, Pierre-Alain Muller, 2018, "Deep Learning For Time Series Classification: A Review", Data Mining And Knowledge Discovery, Issue 4, https://Doi.Org/10.1007/S10618-019-00619-1 .

[10] Hassan Ramchoun, Mohammed Amine Janati Idrissi, Youssef Ghanou, Mohamed Ettaouil, 2016, "Multilayer Perceptron: Architecture Optimization and Training", International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 4, No.1.

[11] Imad Rida, 2018, "Feature Extraction for Temporal Signal Recognition: An Overview", Computer Science, Engineering, Arxiv:1812.01780.

[12] Joonas Nikunen, Aleksandr Diment, Tuomas Virtanen, 2017, "Separation of Moving Sound Sources Using Multichannel NMF and Acoustic Tracking", IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 26, pp. 281–295.

[13] K. Prakash, Hepzibha Rani D, 2015, "Blind Source Separation for Speech Music and Speech Speech Mixtures", International Journal of Computer Applications (0975 – 8887), Vol. 110, No. 12.

[14] Keith Y. Patarroyo, Vladimir Vargas Calderon, 2017, "Pronunciation Recognition of English Phonemes/Ə/,/Æ,/ɑ:/And/^/Using Formants And Mel Frequency Cepstral Coefficients", Arxiv:1702.07071.

[15] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M.S. Nasrin, M. Hasan, B.C. Van Essen, AA.S. Awwal, V.K. Asar, 2019, "A State-of-The-Art Survey on Deep Learning Theory and Architectures", MDPI Electronics (292), Vol. 8, No. 3.

[16] Morten Kolbaek, 2018, "Single-Microphone Speech Enhancement and Separation Using Deep Learning", Ph.D. Thesis, Aalborg University, Denmark.

[17] Ms. Monali R. Pimpale, Prof. Shanthi Therese, Prof. Vinayak Shinde, 2016, "A Survey On: Sound Source Separation Methods", International Journal of Computer Engineering in Research Trends, Vol. 3, pp. 580-584.

[18] Naoya Takahashi, Nabarun Goswami, Yuki Mitsufuji, 2018, "MMDENSELSTM: An Efficient Combination of Convolutional and Recurrent Neural Networks for Audio Source Separation", IEEE, 16th International Workshop on Acoustic Signal Enhancement (IWAENC), DOI: 10.1109/IWAENC.2018.8521383.

[19] Naoya Takahashi And Yuki Mitsufuji., 2017, "Multi-Scale Multi-Band Dense Nets for Audio Source Separation", In Applications of Signal Processing to Audio and Acoustics (Waspaa), IEEE Workshop On, pp. 21–25.

[20] Nilay Ganatra, Atul Patel, 2018, "A Comprehensive Study of Deep Learning Architectures, Applications and Tools", International Journal of Computer Sciences And Engineering, Vol. 6, Issue 12, E-ISSN: 2347-2693.

[21] Qiuqiang Kong, Yong Xu, Wenwu Wang, Philip J.B. Jackson and Mark D. Plumbley, 2019, "Single-Channel Signal Separation and Deconvolution With Generative Adversarial Networks", 28th International Joint Conference On Artificial Intelligence (IJCAI-19), DOI:10.24963/IJSAI.2019/381, pp.2747-2753.

[22] Saber Malekzadeh, Shahla Rezazadeh Azar, Maryam Samami, Maryam Rayegan, 2019, "Classical Music Generation In Distinct Dastgahs with Alimnet Acgan", 27th Iranian Conference on Electrical Engineering, Arxiv:1901.04696 .

[23] Scott Wisdom, Thomas Powers, James Pitton, Les Atlas, 2017, "Deep Recurrent Nmf for Speech Separation by Unfolding Iterative Thresholdin", IEEE Workshop On Applications of Signal Processing to Audio and Acoustics, Doi:10.1109/Waspaa.2017.8170034,E-ISSN: 1947-1629.

[24] Shaik Riyaz, Bathula Lakshmi Bhavani, S.Venkatrama Phani Kumar, 2019, "Automatic Speaker Recognition System in Urdu Using MFCC & HMM", International Journal Of Recent Technology And Engineering (IJRTE), ISSN:2277-3878, Vol. 7, Issue 5S4.

[25] Shrikant Venkataramani, Paris Smaragdis, 2018, "End-To-End Networks For Supervised Single-Channel Speech Separation", IEEE Journal of Selected Topics in Signal Processing (J-STSP), Arxiv:1810.02568.

[26] Steven Spratley, Daniel Beck, And Trevor Cohn, 2019, "A Unified Neural Architecture for Instrumental Audio Tasks", PIEEE, International Conference On Acoustics, Speech, and Signal Processing. DOI: 10.1109/Icassp.2019.8682765.

[27] Sungheon Park, Taehoon Kim, Kyogu Lee, Nojun Kwak, 2018, "Music Source Separation Using Stacked Hourglass Networks", Computer Science, Engineering, ISMIR, Arxiv:1805.08559v2 .

[28] V.S. Narayanaswamy, S. Katoch, J. J. Thiagaraja, H. Song, A. Spanias, 2019, "Audio Source Separation Via Multi-Scale Learning With Dilated Dense U-Nets", Arxiv:1904.04161 .

[29] Zhe-Cheng Fan, Yen-Lin Lai, Jyh-Shing R. Jang, 2017, "SVSGAN: Singing Voice Separation Via Generative Adversarial Network", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), DOI**:** 10.1109/ICASSP.2018.8462091.