الفصل العاشر

# معالجة الكلام

# Speech Processin

## (١٠,١) مقدمة

إن تطبيقات المعالجة الرقمية للإشارات قد اتسعت واشتملت على كل تطبيقات الحياة التى نحياها الآن وبالذات بعد ثورة الحاسبات وثورة الاتصالات التى نعيشها الآن. ولذلك فإنه من المستحيل أن نشمل كل تطبيقات معالجة الإشارات في فصل كهذا أو حتى كتاب واحد أو كتب عديدة. إن موضوع معالجة الكلام الذي نحن بصدده في هذا الفصل من المستحيل أيضاً أن نشمله كله في هذا الفصل، لأن كل مجال من مجالات البحث قد أفردت له كتب مخصوصة بكل نقطة من نقاط الحديث في هذا الموضوع، لذلك فإننا سنمر في هذا الفصل على بعض موضوعات معالجة الكلام دون الدخول فى التفاصيل الدقيقة لأي موضوع حتى يتعرف القارئ المبتدئ في معالجة الإشارات عليها ومن يريد الزيادة في أي فرع من هذه الفروع فإن هناك الكثير من الكتب والمراجع الخاصة بكل موضوع والتي يستطيع قراءتها والاستزادة منها.

إن شرائح معالجة الكلام قد رخصت أسعارها بدرجة كبيرة لدرجة أنك من الممكن أن تجدها مختفية embedded في أي نظام أو تطبيق بدءاً من الأجهزة المنزلية مثل:

الغسالات والأفران، وأجهزة التليفونات وغيرها إلى تطبيقات الحاسب مثل: برامج الإملاء للحاسب، وبرامج قراءة النصوص على الحاسب للمكفوفين، والتصفح الصوتي للإنترنت وغير ذلك الكثير من التطبيقات التى يصعب حصرها هنا. إن مجال معالجة الكلام يمكن تقسيمه إلى أقسام عديدة منها:

١- **نماذج إنتاج الصوت وبناؤها إلكترونياً Speech production modeling:** هذا المجال يهتم بدراسة جهاز الصوت في الإنسان وكيفية محاكاته وبناؤه إلكترونياً أو رقمياً لأن هذه النماذج تعد أساساً للكثير من أنظمة معالجة الكلام.

٢- **تشفير الكلام Speech coding:** لنقل إشارة الكلام عبر قنوات الاتصالات أو تخزينها على أوساط التخزين المختلفة لابد من تشفير هذه الإشارة لتقليل عرض مجالها أو مساحة التخزين التي ستشغلها على الوسط. وكل التطبيقات التي تستخدم الصوت مثل أجهزة التليفونات الخلوية، أو التليفونات العادية، وكل تطبيقات الصوت على الحاسب أو الإنترنت كلها تتعامل مع الصوت بعد تشفيره.

٣- **تحويل النصوص إلى صوت Text to speech conversion:** أو تخليق الكلام sound synthesis وهو عملية إنتاج صوت يحاكي الصوت الآدمي من الآلة أو الحاسب لنقل رسالة صوتية من هذه الآلة. هنا يكون دخل النظام عبارة عن نص يكون في العادة مكتوب بشفرات الأسكي٢ ASCII والخرج إشارة صوتية يمكن للإنسان أن يفهمها. مهم جداً هنا أن يكون الصوت الناتج طبيعياً يشبه الصوت الآدمي ومفهوماً أيضاً. من هذه التطبيقات القواميس الناطقة، وقراءة مواعيد رحلات الطيران، وقراءة البريد الإلكتروني أو الرسائل أو حتى ملفات كاملة للتعامل مع المكفوفين.

٤- **التعرف على الكلام Speech recognition:** وهو عملية التعرف على الكلمة واستخلاص المعلومات المصاحبة لها واستخدام ذلك لأداء أهداف معينة، فأنا

مثلا حينما أقول كلمة "يمين" يتعرف عليها الحاسب أو الماكينة وبناءً على ذلك تتحرك السيارة مثلاً ناحية اليمين. يمكن عن طريق خوارزمات التعرف على الكلام وخوارزمات تخليقه بناء نظام للتحاور الآلي مع الماكينة بحيث يستطيع الإنسان أن يكلم هذه الآلة وهي ترد عليه.

٥- **التحقق من المتكلم Speaker verification**: وهو عملية التحقق من شخصية المتكلم عن طريق عينة من كلامه ، فعند دخول شخص ما إلى مبنى معين مثلاW يطلب منه أن ينطق اسمه أمام جهاز الحاسب وليكن مثلا "محمد أحمد"، عندها يتحقق الحاسب أن هذا الإسم أو الصوت المنطوق هو فعلاً صوت "محمد أحمد" بناءً على مقارنة هذا الصوت بقاعدة بيانات للأصوات مخزنة عنده. كما نرى فإن تطبيقات هذا النوع من معالجة الصوت كثيرة ومتعددة.

## (١٠,٢) نماذج إنتاج الصوت

هناك طرق عديدة لتصنيف الأصوات وإحدى هذه الطرق للتصنيف هي التصنيف تبعا لطريقة إثارة الجهاز الصوتي لإنتاج هذه الأصوات، وتبعا لذلك يمكن تصنيف الأصوات إلى:

١- **أصوات جهورية Voiced**: يتم إنتاج هذه الأصوات عن طريق إثارة الجهاز الصوتي بنفخات أو دفعات دورية من الهواء تتسبب في ذبذبة الأحبال الصوتية في الحنجرة. هذه الذبذبات تسبب تعديل في انسياب الهواء القادم من الرئتين، ومعدل هذه الذبذبات يتراوح من ٦٠ ذبذبة في الثانية للرجل البالغ إلى ٤٠٠ أو ٥٠٠ ذبذبة في الثانية للنساء أو الأطفال.

٢- **أصوات انفجارية Plosive**: ويتم إنتاجها عن طريق إثارة الجهاز الصوتي بالإطلاق المفاجئ لضغط هواء مثل ب أو ت أو ك.

٣- **الأصوات الأنفية Nasal**: حيث جزءاً أو حتى كل الهواء يسمح له بالمرور فى التجويف الأنفي عن طريق فتح الغشاء الأنفي velum، من هذه الأصوات م أو ن.

٤- **الأصوات الاحتكاكيـــة Fricatives**: تنتج من إثارة الجهاز الصوتي بتدفق توربيني من الهواء يتولد من مرور الهواء في فتحة ضيقة. من هذه الأصوات ف أو س أو ش.

٥- **أصوات احتكاكية جهورية Voiced fricatives**: وتنتج من إثارة الجهاز الصوتي بتدفق توربيني للهواء مع ذبذبات في الأحبال الصوتية مثل v أو z أو zh في كلمة pleasure.

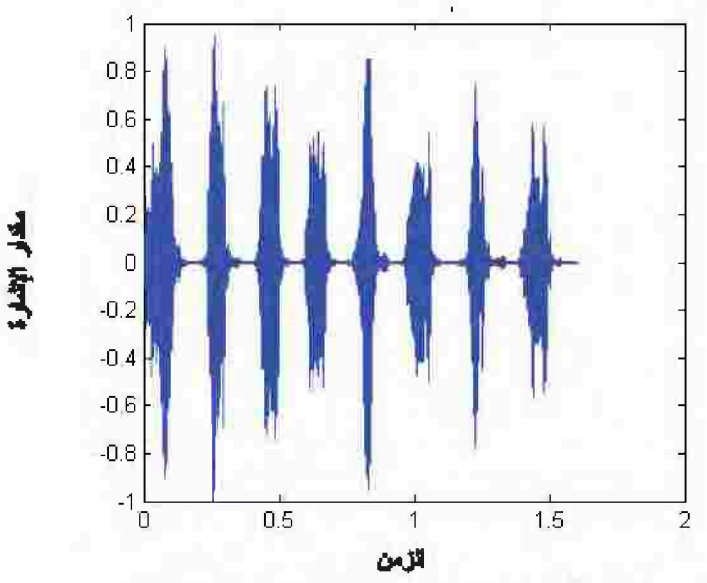
بجانب طريقة إثارة الجهاز الصوتي فإن نوع الصوت الناتج يتوقف أيضاً على حركة اللسان والشفتين والفك السفلي.

هناك أكثر من طريقة لعرض إشارة الصوت حتى يتمكن الباحث من التفريق بين الأصوات المختلفة ومن هذه الطرق:

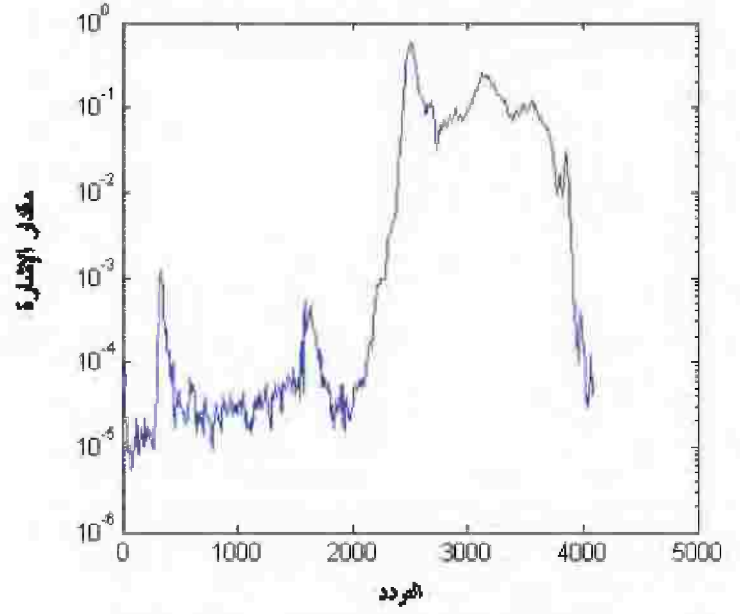
عرض **شكل موجة الضغط Pressure wave form** لأنه كما نعلم أن الأصوات الناتجة تكون عبارة عن موجة من الضغط المنتشرة في الهواء، ويمكن عرض مقدار هذا الضغط مع الزمن كما في الشكل رقم (١٠.١) الذي يعرض مقطعاً من صوت زقزقة العصافير.

الطريقة الثانية للعرض هي عرض طيف الإشارة spectrum وأحياناً يسمى طيف القدرة power spectrum وهو علاقة بين قدرة الإشارة التي هي الضغط والتردد الشكل رقم (١٠.٢) يبين طيف نفس الإشارة السابقة التي هي صوت زقزقة العصافير.

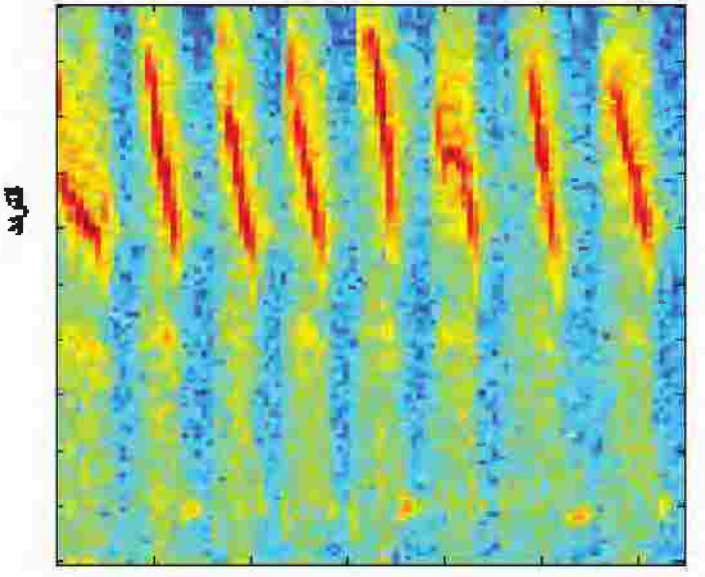
الطريقة الثالثة للعرض هي الأسبكتروجرام spectrogram وهو علاقة بين تردد الإشارة والزمن. الشكل رقم (١٠.٣) يبين الأسبكتروجرام لنفس إشارة الصوت السابقة.




الشكل رقم (١٠.١). عرض إشارة الصوت كموجة ضغط.




الشكل رقم (١٠.٢). طيف إشارة صوتية.

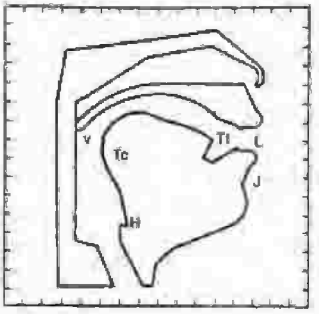



الشكل رقم (١٠.٣). الاسكيروجرام.

هناك طرق عديدة تم استخدامها لتصوير الجهاز الصوتي في الإنسان أثناء السكوت والكلام وذلك للوصول إلى نموذج تقريبي لهذا الجهاز المركب. من هذه الطرق التصوير باستخدام أشعة إكس أو الرنين المغناطيسي MRI أو باستخدام الموجات فوق الصوتية ultrasound حتى إنهم وضعوا ملفات متناهية الصغر على هذه الأجزاء المتحركة بحيث عندما تتحرك هذه الملفات في مجال مغناطيسي فإنه يمكن رسم شكل أو استجابة لهذه الحركة. يبين الشكل رقم (١٠.٤) رسماً تخطيطياً للجهاز الصوتي في الإنسان بناء على طرق التصوير المختلفة السابقة.

إن حجم فراغ الجهاز الصوتي وشكله يتحكم فيه الأجزاء المسؤولة عن النطق وهي اللسان (موضعه ومقدمته ) والفكين والشفتين ومكان عظمة تثبيت اللسان وفتحة التجويف الأنفي، حيث تتحكم كل هذه الأجزاء أو تعدل من حجم هذا الفراغ أثناء

الكلام مما يعطي الكلام الخاصية الفريدة التي تحدد كل شخص على حده.




H = HYOID POSITION عظمة تثبيت اللسان
J = ANGLE OF JAW OPENING زاوية فتح الفك
L = LIP PROTRUSION AND ELEVATION نتوء أو بروز الشفتين
Tc= TONGUE CENTER مركز اللسان
Tt= POSITION OF TONGUE TIP موضع مقدمة اللسان
V = VELUM OPENING فتحة الغشاء الأنفي

**الشكل رقم (١٠،٤). نموذج الجهاز الصوتي في الإنسان.**

لدراسة انتشار الموجات الصوتية في تجويف الجهاز الصوتي فإنه لا بد من وضع بعض الافتراضات لتبسيط التعامل مع هذا الفراغ المعقد، هذه الافتراضات ليس لها تأثير كبير على نتائج الدراسة ولكنها تسهل وبدرجة كبيرة العمليات الحسابية اللازمة لذلك. هذه الافتراضات تتمثل في ثلاثة افتراضات فقط وهي كالتالي:

١- تبسيط الفراغ الصوتي في صورة أنبوبة مستقيمة بدايتها عند فتحة المزمار glottis ونهايتها عند الشفتين، وبتغير نصف قطرها عند مواضع مختلفة وأزمنة مختلفة على حسب الكلام الذي يتم نطقه.

٢- الموجة الصوتية تنتشر في هذا الفراغ في صورة موجة مستعرضة، وهذا يعني أن جميع خواص هذه الموجة تكون ثابتة على أي مقطع عرضي عمودي على الأنبوبة.

٣- الفرض الثالث أن عملية انتشار الموجة الصوتية في هذا الفراغ تكون خطية. وإن كانت هناك بعض الأبحاث التي تفترض عدم الخطية في مناطق معينة وبالذات في

ظل التقدم الهائل في الحاسبات والتي تساعد على سرعة الحساب في ظل عدم الخطية.

بناءً على هذه الافتراضات أمكن عمل نماذج كاملة للجهاز الصوتي يمكن بها تخليق كل أنواع الأصوات وبدقة معقولة وللمزيد من المعلومات في هذا المجال نحيل القارئ إلى مراجع متخصصة في هذا المجال وهى كثيرة وبالذات الفصل رقم ٤٤ في المرجع [Vijay K. Madisetti and Douglas 1999] الذي يحتوي ٢٥ مرجعاً في هذا الموضوع.

## (١٠,٣) تشفير الصوت

المقصود بتشفير الصوت هو تحويل الصوت أو تمثيله في صورة مناسبة لنقله عبر قنوات الاتصال أو تخزينه على أوساط التخزين. بالطبع فإن هذه العملية يصاحبها فقد لبعض المعلومات الموجودة في إشارة الصوت والخوارزمات التي تقوم بذلك يطلق عليها lossy algorithms ولحسن الحظ أن الفقد في المعلومات لا يؤثر بدرجة ملحوظة على جودة الصوت عند استرجاعه في مقابل الفائدة العظيمة التي نحصل عليها من هذه الخوارزمات. إن التطبيقات التي تستخدم الصوت في إحدى مراحلها مثل: أجهزة التلفونات وأجهزة المحمول والراديو والتلفزيون كلها لا بد أن تتعامل مع طرق تشفير الصوت.

النقل الرقمي للصوت عبر شبكات التليفونات يبدأ بتحويل إشارة الصوت من الصورة التماثلية أو التناظرية إلى الصورة الرقمية عن طريق العينة sampling بمعدل مقداره ثمانية KHz مثلا ثم التكميم quantization أو تمثيل هذه العينات في ثمانية بتات بت أو تمثيلها بكمية من ٢٥٦ كمية ممكنة باستخدام الثمانية بتات. بعد ذلك يتم إرسال هذه الإشارة بمعدل 64Kbit/s على قنوات النقل. الاتصالات الدولية عبر الكابلات البحرية أو الأقمار الصناعية تستخدم معدل إرسال أقل (32Kbit/s) للتوفير في عرض المجال.

تمر إشارة الصوت بعدة مراحل قبل إطلاقها على قناة الاتصال حيث يتم تحويل الإشارة من الصورة التماثلية أو التناظرية إلى الصورة الرقمية digital form حيث تمثل كل عينة بعدد معين من البتات، بعد ذلك يتم تشفير encryption لهذه الإشارة وإرسالها على القناة بالمعدل المطلوب. عند المستقبل يتم فك شفرة الإشارة decryption وتحويلها إلى الصورة التناظرية مرة أخرى قبل سماعها عن طريق المستقبل.

هناك بعض الخواص المهمة لمشفرات الصوت والتي عن طريقها يتم الحكم بجودة هذا المشفر من عدمه. من هذه الخواص: معدل البتات bit rate، والجودة quality، والتعقيد (ومن ثم التكلفة) complexity، والتأخير delay.

١- **معدل البتات bit rate**: كما ذكرنا سابقاً فإن شبكات التلفونات تستعمل معدل عينة مقداره 8KHz وتمثل كل عينة في ثمانية بتات، ولذلك فإن معدل الإرسال يكون 64Kbit/s. تبعاً لظروف قنوات الاتصال وعرض المجال على كل منها يتم ضغط هذا المعدل ويقاس مقدار هذا الضغط بمقدار الضغط من الكمية القياسية في التلفونات 64Kbit/s. في شبكات الاتصالات الدولية يمكن لعرض المجال أن ينزل من 63Kbit/s إلى 5.3Kbit/s مما يعني توفير كبير في عرض المجال. بعض الاتصالات الخلوية تنزل بعرض المجال حتى 13Kbit/s-3.54Kbit/s. ليس من الضروري أبداً أن يكون عرض المجال ثابتاً دائماً على القناة إذ إنه في أثناء المحادثات يكون فترات كثيرة من السكوت فلماذا يكون معدل التراسل عالي في هذه الفترات، لذلك يمكن جعل معدل التراسل متغير مع طبيعة الإشارة.

٢- **التأخير delay**: عملية التأخير ليس لها أهمية في تطبيقات تخزين الصوت حيث العمل في الزمن الحقيقي ليس ضرورة في هذه الحالة. أما على قنوات الاتصال فإن عملية التشفير يجب ألا تتأخر أو تستغرق وقتاً أكثر من ٣٠٠ مللي ثانية وإلا فإن

عملية الاتصال تكون غير مريحة أو مقبولة. هناك مصادر عديدة لهذا التأخير منها التأخير نتيجة الحسابات أو معالجة الإشارة سواء تكبيرها أو ترشيحها وهذا بالطبع سيعتمد على نوع المعالج processor المستخدم. في العادة تتم عملية التشفير على بلوكات أو إطارات blocks أو frames من إشارة الصوت وهذه العملية تحتاج لوقت أيضاً، كما أن التراسل نفسه على القناة وتعدد الإرسال multiplexing يسبب تأخيراً أيضاً.

٣- **التعقيد (التكلفة) complexity**: هناك عاملان مهمان في تحديد درجة تعقيد النظام وهما التكلفة واستهلاك القدرة power. بالطبع فإن التكلفة تكون عاملاً مهماً في اختيار أي مشفر للصوت كذلك استهلاك القدرة وبالذات في الأجهزة التي تعمل لاسلكيا وتعمل من خلال بطاريات حيث هنا يكون استهلاك القدرة عاملاً مهماً في اختيار النظام. معظم مشفرات الصوت تكون مبنية باستخدام شرائح معالجات أو شرائح معالجة إشارة DSP أو حتى على شرائح خاصة بذلك، وفي كل هذه الأحوال تكون سرعة الشريحة ونبضات التزامن التي تعمل عندها عاملاً مهماً في تحديد تكلفة النظام واختياره. بعض هذه الشرائح تحتوي على كميات من الذاكرة RAM أو ROM وبعضها يكون ١٦ بتاً أو ٣٢ بتاً أو حتى ٦٤ بت وكل ذلك يكون عاملاً مهماً في تحديد سرعة النظام. بعض هذه الأنظمة أيضاً تتعامل مع الأرقام الصحيحة integers والبعض يتعامل مع الأرقام الحقيقية floating point، وبالطبع فإن التعامل مع الأرقام الصحيحة يكون أسهل من وجهة نظر البرمجة والحسابات.

### (١٠,٣,١) نموذج شفرة التنبؤ الخطي لإنتاج الصوت LPC speech production

في نموذج التنبؤ الخطي يتم التنبؤ بالعينة القادمة أو التالية كمركبة خطية من العينات السابقة، وعلى ذلك يمكن كتابة العينة x[n] عند اللحظة n كما يلي:

$$x[n] = \sum_{i=0}^{I} a_i x_{n-i} \qquad (١٠,١)$$

حيث $a_i$ هي معاملات التنبؤ، وهناك أكثر من طريقة لحساب هذه المعاملات بحيث تجعل الفرق بين العينة المتوقعة والعينة الحقيقية أقل ما يمكن. بأخذ تحويل z لطرفي المعادلة رقم (١٠,٤) يمكن كتابة المعادلة التالية:

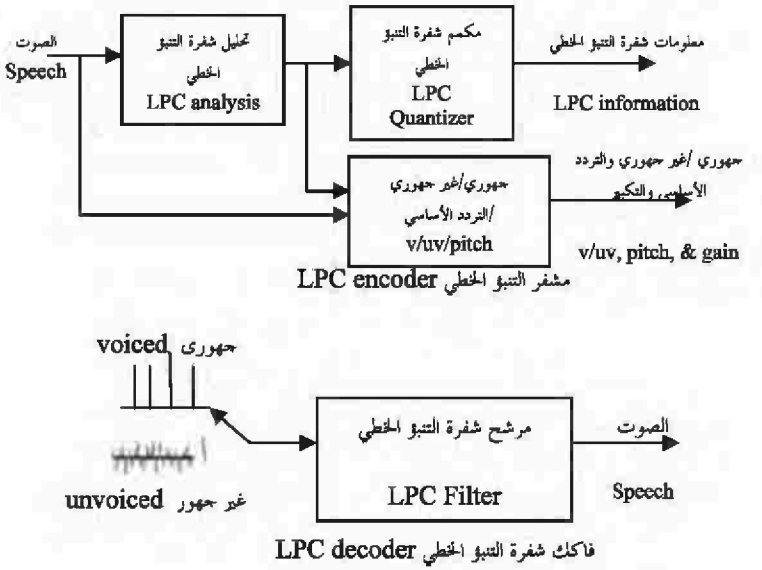
$$A(z) = 1 - \sum_{i=1}^{I} a_i z^{-i} \qquad (١٠,٢)$$

حيث 1/A(z) يسمى مرشح التنبؤ الخطي. أقطاب هذا المرشح تكون قريبة جداً من دائرة الوحدة بحيث إنه أي تعديل أو خطأ يطرأ على هذه الأقطاب يمكن أن يجعل هذا المرشح غير مستقر. لذلك فإن التكميم quantization للمعاملات $a_i$ يجب أن يكون بدرجة عالية جداً من الدقة (التمثيل بعدد كبير من البتات) حتى نتجنب دخول هذا المرشح في حالة عدم استقرار.

### (١٠,٣,٢) أنواع مشفرات الصوت

١- **مشفرات التنبؤ الخطي LPC coders**: الشكل رقم (١٠,٢) يبين هذه الطريقة حيث من إشارة الصوت الحقيقية يتم حساب معاملات مشفر التنبؤ الخطي LPC coefficients أو $a_i$ كما ذكرنا سابقاً، وهذه المعاملات يتم تكميمها وإرسالها إلى المستقبل. مع هذه المعاملات يتم إرسال التردد الأساسي pitch وهل هذا البلوك أو الإطار جهوري أم لا ومعامل التكبير حيث يتم إرسال هذه المعاملات مع كل إطار من إشارة الصوت كما في الشكل. عند المستقبل إذا كان الإطار الذي تم استقباله جهوريا

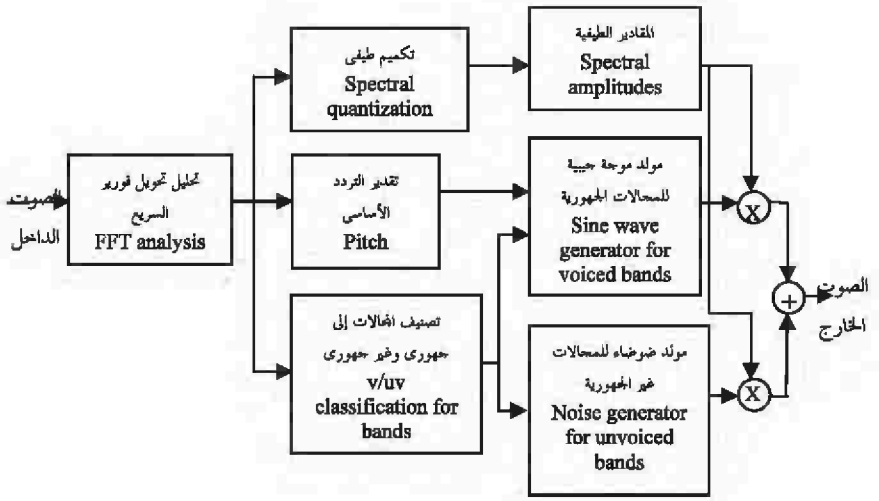
فإنه بمعرفة التردد الأساسي يتم توليد إشارة دورية وبمساعدة معاملات التنبؤ يتم إنتاج الصوت. أما إذا كان الإطار غير جهوري فإن إشارة الإثارة عند المستقبل تكون ضوضاء white noise وباستخدام معاملات التنبؤ يمكن أيضاً إنتاج الصوت. هناك الكثير من المراجع الموجودة في الفصل ٤٥ من المرجع [Vijay K. Madisetti and Douglas 1999] والتي يمكن الاستزادة منها. هذه الأنظمة عند المرسل والمستقبل يطلق عليها عادة فوكودر vocoders أو المشفرات الصوتية.




**الشكل رقم (١٠,٥). تشفير الصوت وإعادته vocoder.**

٢- **المشفرات متعددة النطاق Multiband excitation coders**: يبين الشكل رقم (١٠,٦) رسماً تخطيطياً لهذه الطريقة، حيث هنا يتم حساب معاملات محول فورير (المكونات الطيفية لإشارة الصوت)، كما يتم تحديد التردد الأساسي وهل هذا الإطار جهوري أم لا. كل هذه المعلومات يتم حسابها وإرسالها إلى المستقبل حيث بمعرفة

التردد الأساسي ومعلومات الطيف spectral amplitude يتم توليد الأصوات الجهورية باستخدام مولد نبضات جيبي كما في الشكل. أما الأصوات غير الجهورية فباستخدام مولد ضوضاء مع معلومات الطيف والتردد الأساسي يمكن توليد الأصوات غير الجهورية. بعد ذلك يمكن جمع الإشارة الجهورية وغير الجهورية للحصول على إشارة الصوت كاملة. هذه الطريقة والطريقة السابقة تسمى بطرق النمذجة حيث يتم استخدام نموذج لتخليق الصوت ولا يستخدم الشكل الأساسي لموجة الصوت كما في بعض الطرق الأخرى التي سنراها.




**الشكل رقم (١٠,٢). استخدام المكونات الطيفية لإشارة الصوت فى تشفيره.**

٣- هناك طرق عديدة لتشفير الصوت تعمل على شكل الإشارة في النطاق الزمني وهذه الطرق معروفة ولمن يريد معلومات أكثر عن هذه الطرق عليه اللجوء إلى أي كتاب في معالجة الصوت. من هذه الطرق: التشفير بالتعديل النبضي Pulse code

Differential pulse code modulation PCM والتشفير بالتعديل النبضي الفرقي Adaptive differential pulse code modulation DPCM والتشفير بالتعديل النبضي الفرقي المتكيف modulation ADPCM.

## (١٠.٤) تحويل النصوص إلى كلام
## Text to Speech Conversion

إن عملية تحويل النصوص إلى كلام تحاكي تماماً القراءة من نص. غير أن الإنسان حينما يقرأ نصاً فإنه يضيف بعض التعبيرات مثل الوقفات عند بعض الكلمات أو رفع الصوت وخفضه مع كلمات معينة لتأكيد معنى هذه الكلمات ولكن عند استخدام آلة أو حاسب لمحاكاة هذه العملية فإن إضافة هذه التعبيرات تكون صعبة جداً وهي أقصى ما يمكن الوصول إليه.

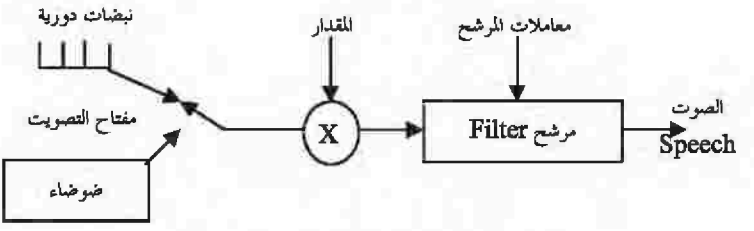
المرحلة الأولى من مراحل تحويل النص إلى كلام هي تهيئة النص، ويشمل ذلك تحديد نهاية كل كلمة وكل جملة والتخلص من الاختصارات إن وجدت مثل Mr. حيث يحول هذا الاختصار إلى كلمة مستر مثلاً، كما أن النقطة هنا لا تعني نهاية جملة لذلك يجب التفريق بينها وبين النقطة في نهاية الجملة.

المرحلة الثانية بعد عملية عزل كل كلمة على حده هي عملية النطق بهذه الكلمة. أبسط الطرق لذلك هي تكوين قاموس يحتوي الكلمات المكتوبة وما يقابلها نطقاً وعند التعرف على الكلمة المكتوبة يتم إخراج منطوقها من القاموس. هذه الطريقة تعاني من عدة عيوب منها أن حجم القاموس سيكون بالطبع كبيراً جداً حتى يحتوي على كل كلمات اللغة التي نتعامل بها، كما أن القاموس لن يحتوي على كل مشتقات الكلمات مثل يكتب وكاتب ومكتوب وكاتبة ومشتقات أخرى كثيرة، وهذه مشكلة خاصة باللغة العربية ونعتقد أنها لا توجد في اللغات الأخرى. كما أنه لا يمكن

للقاموس أن يضم كل الأسماء الموجودة في اللغة، لذلك فإن فكرة القاموس نادراً ما تستخدم لهذا الغرض.

الطريقة المثلى لتخليق الكلام هي عن طريق تقسيم كل كلمة إلى وحدات صوتية تسمى فونيمات ، والفونيم هو حرف تقريباً، لذلك فإننا نتوقع أن قاموس فونيمات أي لغة سيساوى تقريباً عدد أحرف هذه اللغة أو ربما يزيد قليلاً. لذلك فإن عملية التعرف على الفونيم واستخراج منطوقه ستكون أسرع بكثير في هذه الحالة عن طريقة استخدام قاموس للكلمات الكاملة، كما أن مشكلة مشتقات الكلمات لن يكون لها تأثير الآن.

بعد تحديد الفونيمات (أو الوحدات الصوتية) تأتي مرحلة النطق أو تخليق هذه الوحدات. الشكل رقم (١٠,٧) يبين رسماً صندوقياً لأحد أنظمة تخليق الصوت الشائعة الاستخدام. نلاحظ أن الجزء الأيمن في هذا الشكل هو الاختيار بين الإشارة النبضية الدورية والتي تتحدد دورتها من التردد الأساسي formant لهذا الفونيم، أو الضوضاء كما في الشكل. الاختيار بين هذين المصدرين يتم عن طريق هل الصوت المطلوب جهوري أو غير جهوري. بعد ذلك يتم التحكم في مقدار الإشارة أو جهارتها loudness، ثم يأتي في النهاية دور المرشح النهائي الذي يضمن تنعيم الإشارة وجعلها طبيعية بقدر الإمكان عن طريق التحكم في معاملات هذا المرشح.




**الشكل رقم (١٠,٧) . نموذج لتخليق الصوت.**

## (١٠,٥) التعرف على الكلام
## Speech Recognition

منذ أن اضطر الإنسان للتعامل مع الآلة وهو يبحث عن وسائل أكثر راحة وأكثر ملاءمة لظروفه للتعامل مع الآلة بدأ من المفاتيح البسيطة (فتح أو غلق) إلى لوحة المفاتيح التي يستطيع من خلالها إعطاء أوامر أكثر تعقيداً، إلى الفأرة أو الماوس إلى التعامل مع الشاشات باللمس وهكذا نرى أن هناك طرقاً عديدة من خلالها يستطيع الإنسان أن يتعامل مع الآلة. من الطرق الحديثة للتعامل مع الآلة، التعامل معها من خلال الصوت والذي كثرت تطبيقاته هذه الأيام ومن أهمها الإملاء الآلي، بأن يقوم الشخص بإملاء الحاسب مثلاً ويقوم الحاسب بالتعرف على الكلمات التي ينطقها المستخدم ثم يخزنها في ملف للاستفادة منها. منها أيضاً إعطاء الأوامر الصوتية للآلة وغير ذلك الكثير. نلاحظ أن أنظمة أو خوارزمات التعرف على الكلام يكون الدخل لها كلام منطوق أما خرجها فأحياناً يكون نصاً (وفي هذه الحالة فهي تسلك المسلك العكسي لخوارزمات تحويل النصوص إلى كلام كما رأينا) أو فعلاً معيناً يتم اتخاذه بناءً على الكلمة أو النص الذي تم التعرف عليه.

هناك العديد من العوامل أو المتغيرات التي تحكم عملية التعرف على الكلام ومنها ما يلي:

١- الطريقة التي يتكلم بها الإنسان إلى الماكينة أو الحاسب وهناك ثلاث طرق لذلك:

أ) التكلم بكلمات أو عبارات محددة مفصولة عن بعضها isolated words، وهذه هي أسهل طرق التعرف على الكلام.

ب) كلمات متصلة continuous words أو حديث مسترسل أو طليق ولكن كل الكلمات تكون مأخوذة من قاعدة بيانات معروفة ومحددة.

ج ) كلمات متصلة أو حديث مسترسل غير محدد الكلمات، وهذه هي أصعب طرق التعرف على الكلام.

٢- حجم قاموس الكلمات المستخدمة في عملية التعرف:

أ ) قاموس كلمات صغير يحتوى على أقل من ١٠٠ كلمة.

ب ) قاموس كلمات متوسط يحتوي من ١٠٠ حتى ١٠٠٠ كلمة.

ج ) قاموس كلمات كبير يحتوي على أكثر من ١٠٠٠ كلمة.

٣- عوامل تعتمد على المتكلم:

أ ) هل النظام يعتمد على شخصية المتكلم، أو بمعنى آخر هل النظام مصمم لشخص معين وإذا تغير هذا الشخص فإن النظام يفشل في التعرف speaker dependent system.

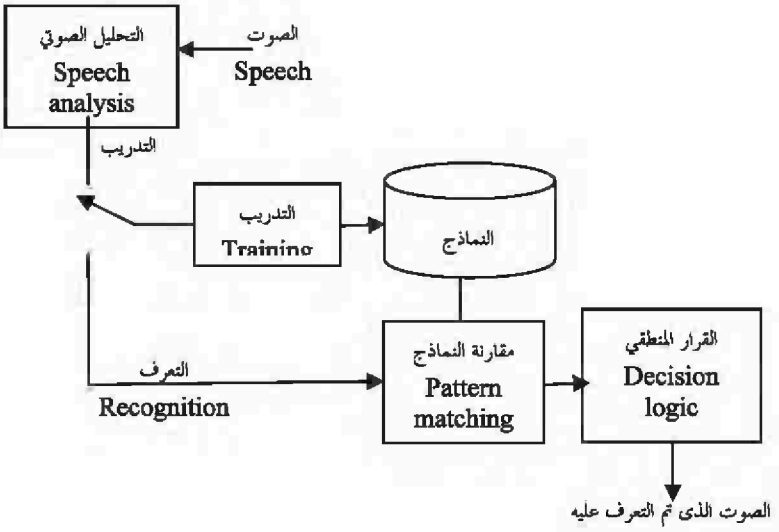
ب ) هل النظام لا يعتمد على شخصية المتكلم speaker independent بمعنى أن النظام يعطي نفس نسبة التعرف الصحيحة بصرف النظر عن الشخص المتكلم.

ج ) هل النظام متكيف adaptive بمعنى أنه مع زيادة مستخدمي النظام فإن النظام يكيف نفسه بحيث يستوعب المستخدمين الجدد.

إن عملية التعرف على الكلام عملية صعبة نتيجة متغيرات عديدة في إشارة الصوت نفسها، فمثلاً هل يضمن المستخدم أن ينطق الكلمات المطلوب التعرف عليها دائماً بنفس الطريقة ودون أي تغيير ونحن نعلم أنه حتى الحالة المزاجية للشخص يمكن أن تغير من طريقة نطقه للكلمات، ناهيك عن إذا كان الشخص عنده برد أو أي مرض يمكن أن يغير من طبيعة جهاز النطق عنده. أيضاً لهجة المتكلم يمكن أن تنطق نفس الكلمة بأكثر من لهجة وهذا قد يؤدي إلى التعرف الخاطئ على الكلمة. هناك أيضاً

عوامل تتوقف على طبيعة الميكروفون المستخدم ومدى جودته في إنتاج إشارة الكلمة وحتى المسافة بين الميكروفون وفم المتكلم تعد ذات تاثير كبير في نسبة التعرف الصحيح على الكلمة. الوسط الذي يتكلم فيه الشخص وهل هو وسط هادئ صوتياً أم أنه وسط به ضوضاء.

**التعرف على النماذج عن طريق مقارنة النماذج pattern matching:** الشكل رقم (١٠,٨) عبارة عن رسم صندوقي يبين هذه الطريقة. بعد الحصول على إشارة الكلام من الميكروفون وتهيئتها إلكترونياً يتم تقسيمها إلى مقاطع زمنية قصيرة تتراوح من ١٠ – ٣٠ مللي ثانية نتيجة طبيعة إشارة الصوت المتغيرة. بعد ذلك يتم تحويل كل مقطع من هذه المقاطع إلى مجموعة من المعاملات. هذه المعاملات قد تكون في النطاق




**الشكل رقم (١٠,٨) . التعرف على الكلام عن طريق مقارنة النماذج.**

الزمني مثل عدد مرات عبور الإشارة لمستوى الصفر أو أي مستوى آخر، وقد تكون في النطاق الترددي مثل معاملات تحويل فورير المعينن DFT، أو معاملات التنبؤ الخطي المشفر LPC، أو حتى خرج مجموعة من المرشحات للإشارة في نطاقات ترددية مختلفة. ولقد وجد عمليا أن معاملات طيف القدرة power spectrum و السيبسترم cepstrum (تحويل فورير للطيف اللوغاريتمى (log spectrum) تعطي أفضل نتائج في التعرف على الكلام. كل ذلك يقوم به الصندوق الخاص بتحليل الإشارة Recognition في الشكل رقم (١٠,٨).

بعد عملية التحليل السابقة تأتي عملية التدريب، حيث يتم إدخال نماذج من الكلمات المعروفة ويطلب من النظام التعرف على هذه النماذج فإذا تم التعرف عليها خطأ يتم تغيير معاملات النظام بحيث يتم تعديل نتيجة التعرف على هذه النماذج، أو حتى يتم ضم هذه النماذج إلى قاعدة بيانات النظام. تستمر هذه العملية إلى أن يتم التعرف الصحيح على كل محتويات قاعدة بيانات النظام. تأتي بعد ذلك عملية الاختبار testing أو التعرف على كلمات حقيقية حسب متغيرات النظام، حيث يتم إدخال كلمات بها بعض الضوضاء مثلاً، أو منطوقة بمتحدث آخر، ويطلب من النظام التعرف على هذه النماذج وحساب دقة النظام. آخر بلوك في الشكل رقم (١٠,٨) هو القرار الذي يتم أخذه بناءً على المقارنة في البلوك السابق وقد يكون هذا القرار صحيحاً أو خطئاً أي أنه حصل على الكلمة أو النموذج الصحيح أم لا. هناك أكثر من خوارزم يتم استخدامها في هذا الشأن ومنها نماذج ماركوف الخفية Hidden Markoff Models HMM وهي الأكثر استخداماً، والشبكات العصبية والخوارزمات الجينية والكثير من الخوارزمات الأخرى.

## (١٠,٦) التعرف على الأشخاص والتحقق منهم
## Person Recognition and Verification

هناك كتب كثيرة تم تأليفها وأبحاث مازالت مستمرة في مجال القياسات الحيوية biometrics وكلها تصب في التعرف على الشخص من خلال بعض الخواص الحيوية والتي منها صورة وجهه، أو بصمة يده، أو شكل يده، أو شكل أذنه، أو بصمة دمه DNA، أو بصمة صوته عن طريق كلمة أو عبارة ينطقها الشخص ونتعرف عليه منها.

هناك فرق بين التعرف على الأشخاص من خلال أصواتهم والتحقق منهم. التعرف على الأشخاص هو الأعم حيث من كلمة أو عبارة منطوقة تبحث في قاعدة بيانات لتقرر هل هذا الصوت يخص أي شخص في قاعدة بيانات النظام. أما التحقق من الشخص فإن شخص ما يأتي ويقول مثلاً أنا محمد وفي هذه الحالة فإن نظام التعرف عليه أن يتحقق هل هذا الشخص هو في الحقيقة محمد أم لا بناءً على كلمة أو عبارة ينطقها. تطبيقات التعرف على الأشخاص والتحقق منهم من خلال الصوت له تطبيقات كثيرة جداً وأهمها في مجال مكافحة الجريمة والتعرف على المجرمين.

من أهم الخواص التي يتم استخلاصها من إشارة الصوت لاستخدامها في التعرف على الشخص هي الخواص الطيفية spectral features مثل التردد الأساسي formants، أو طاقة الطيف spectral energy، ويتم حساب ذلك على مقاطع صوتية تبلغ ١٠ – ٣٠ ميلليثانية باستخدام نافذة هامنج كما أشرنا مسبقاً. من المعاملات الكثيرة الاستخدام هي خرج مجموعة من المرشحات التى يبلغ عددها أحياناً ١٦ مرشحاً من النوع BPF والتي يتم توزيع مراكز ترددتها بحيث تكون المسافة الترددية بين كل منها والآخر حوالى ٥٠٠ هرتز فى مدى ترددات الصوت وهو أربعة كيلوهرتزات، وأحياناً يتم توزيع هذه المرشحات توزيعاً خطياً أو غير خطي.

معاملات التنبؤ الخطي المشفر يتم استخدامها بكثرة في التعرف على المتكلمين،

وكما ذكرنا فإن كل عينة من عينات إشارة الصوت يمكن كتابتها في صورة كثيرة الحدود من العينات السابقة كما في المعادلة التالية:

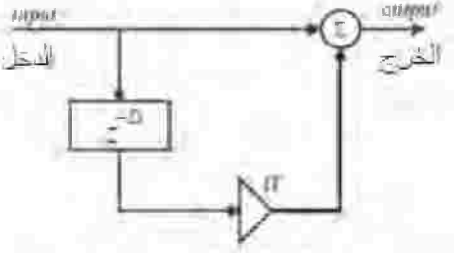
$$S(t)=a_1s(t-1)+a_2s(t-2)+a_3s(t-3)+\dots+a_ps(t-p)+Gu(t) \qquad (١٠.٨)$$

هناك أكثر من طريقة لحساب المعاملات $a_p,\dots,a_2,a_1$ وهي التي يتم استخدامها لتحديد الشخص المتكلم. ولقد وجد أيضاً أن الطاقة الصوتية على مقاطع معينة من إشارة الصوت أو حتى على عبارات كاملة يتم نطقها وهي في النطاق الزمني تعطي نتائج جيدة في التفريق بين الأشخاص مثل المعاملات الطيفية.

## (١٠.٧) مرشحات لبعض التأثيرات الصوتية الخاصة
## Digital Filters for Special Sound Effects

### مرشح الصدى الأحادي Single Echo Filter

إضافة صدى للصوت يعد أحد التأثيرات التي يرغبها البعض في الكثير من التطبيقات ويتم ذلك باستخدام مرشح يضيف للإشارة نسخة منها نفسها متأخرة عنها بزمن معين يمكن التحكم فيه، وكذلك يتم التحكم في معامل a لإحباط أو اضمحلال attenuation هذه الإشارة المضافة كما في الشكل رقم (١٠.٩).




الشكل رقم (١٠.٩) ـ مرشح الصدى الأحادي.

يمكن التعبير عن ذلك بالمعادلة التالية :

$$y(n)=x(n)+ax(n-D) \qquad (١٠,٤)$$

حيث D هى زمن التأخير و $|a|\leq 1$ هى معامل الاضمحلال attenuation. بإجراء تحويل z على طرفي المعادلة رقم (١٠,٤) نحصل على دالة العبور كما يلي :

$$H(z)=1+az^{-D} \qquad (١٠,٥)$$

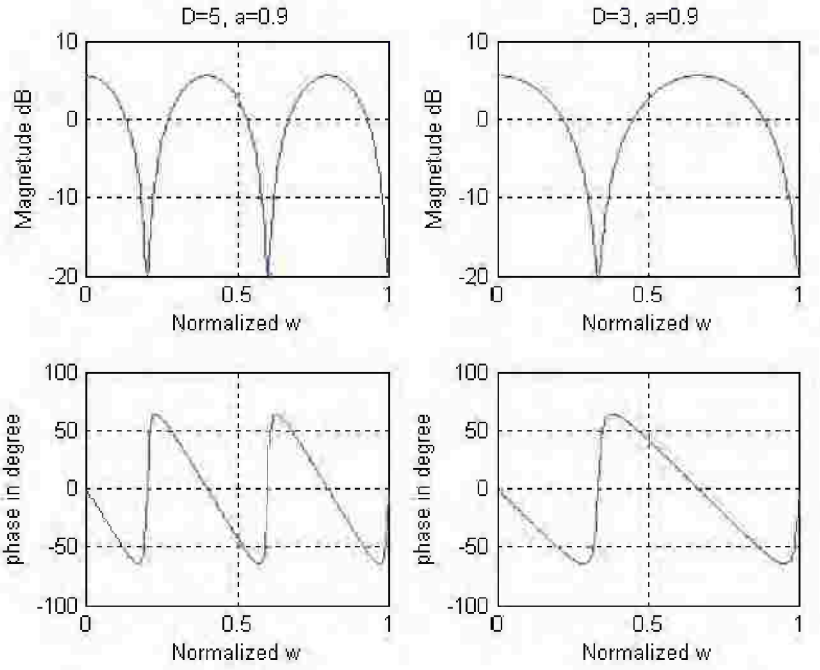
والاستجابة الترددية كما يلي أيضاً :

$$H(e^{jw})=1+ae^{-jwD} \qquad (١٠,٦)$$

الشكل رقم (١٠,١٠) يبين الاستجابة الترددية لهذا المرشح عند قيم مختلفة للتأخير D مع ثبوت معامل الاضمحلال a. يمكن تكرار هذه الرسم لقيم مختلفة لمعامل الاضمحلال مع ثبوت التأخير D. البرنامج التالي هو المستخدم للحصول على الشكل رقم (١٠,١٠).

```
%Frequency response for an echo filter
h1=[1 0 0 0 0  0.9];
h2=[1 0 0  0.9];
w=0:pi/255:pi;
H1=freqz(h1,1,w);
H2=freqz(h2,1,w);
mag1=20*log10(abs(H1));
mag2=20*log10(abs(H2));
set(gcf, 'color', 'white');
subplot(2,2,1);
```

```
plot(w/pi,mag1);grid;
ylabel('Magnetude dB'); xlabel('Normalized w');
title('D=5, a=0.9');
ph1=angle(H1)*180/pi;
ph2=angle(H2)*180/pi;
subplot(2,2,3);
plot(w/pi,ph1);grid;
ylabel('phase in degree'); xlabel('Normalized w');
subplot(2,2,2);
plot(w/pi,mag2);grid;
ylabel('Magnetude dB'); xlabel('Normalized w');
title('D=3, a=0.9');
ph1=angle(H1)*180/pi;
subplot(2,2,4);
plot(w/pi,ph2);grid;
ylabel('phase in degree'); xlabel('Normalized w');
```




الشكل رقم (١٠,١٠). الاستجابة الترددية لمرشح الصدى الأحادي.

كما أشرنا في مقدمة هذا الفصل أننا لن ندخل في التفاصيل الدقيقة لموضوعات معالجة الكلام المختلفة والتطبيقات الخاصة بها. ولكننا فقط مررنا عليها مروراً سريعاً كأحد التطبيقات الهامة للموضوع الأساسي في هذا الكتاب وهو المعالجة الرقمية للإشارات. وعلى القارئ المهتم بموضوع معالجة الكلام اللجوء إلى الكتب والمراجع التي تناولت هذا الموضع بالشرح المفصل.