Báo cáo đồ án

1. Tổng quan về tổng hợp tiếng nói , các phương pháp tổng hợp tiếng nói

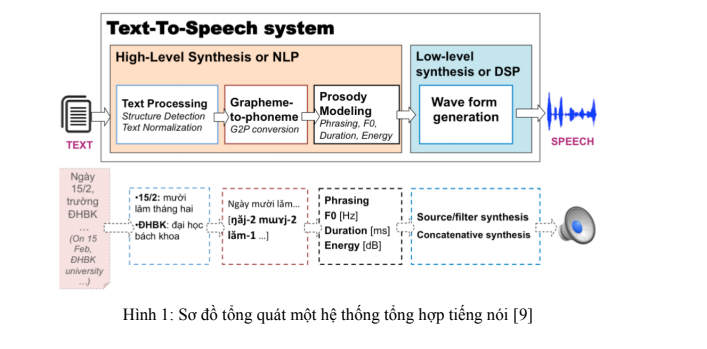
* Tổng hợp tiếng nói là quá trình tạo ra tiếng nói của con người từ văn bản, hệ thống tổng hợp tiếng nói là hệ thống nhận đầu vào là một văn bản và tạo ra tín hiệu tiếng nói tương ứng ở đầu ra.
* Nghiên cứu về tổng hợp tiếng nói đã bắt đầu từ rất lâu, năm 1779 nhà khoa học người đan mạch Christian Kratzenstein đã xây dựng mô phỏng đơn giản hệ thống cấu âm của con người, mô hình này đã có thể phát ra được âm thanh của một số nguyên âm dài
* Sang thế kỷ 20 khi mà có sự lớn mạnh của hệ thống điện, điện tử thì mới thực sự xuất hiện những hệ thống tổng hợp tiếng nói chất lượng

VD : VODER lần đầu được giới thiệu năm 1939 . Hiện nay, có rất nhiều các sản phẩm như sách nói, đồ chơi,.. sử dụng công nghệ tổng hợp tiếng nói. Đặc biệt các mô đun tổng hợp tiếng nói còn được tích hợp trong các trợ lý ảo trên điện thoại và máy tính như Siri hay Cortana

* Qua quá trình phát triển, hiện nay về cơ bản một hệ thống tổng hợp tiếng nói bao gồm hai thành

+) Phần xử lý ngôn ngữ tự nhiên: chuẩn hóa, xử lý các văn bản đầu vào thành các thành phần có thể phát âm được.

+) Phần xử lý tổng hợp tiếng nói: Tạo ra tín hiệu tiếng nói từ các thành phần phát âm được nêutrên.



* Một số phương pháp tổng hợp giọng nói như :

+) Tổng hợp mô phỏng hệ thống phát âm

+) Tổng hợp tần số formant

+) Tổng hợp ghép nối

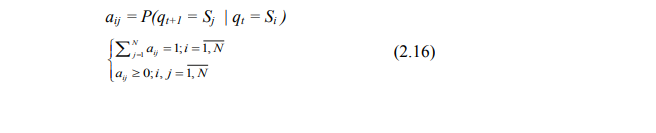
+) Tổng hợp dùng tham số thống kê(HMM)

+) Tổng hợp tiếng nói dựa trên phương pháp học sâu (DNN)

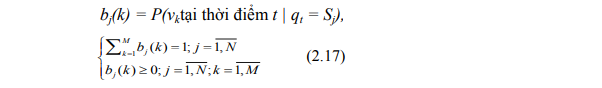
1. Tổng quan về tổng hợp tiếng nói bằng phương pháp (HMM)

2.1- Tổng quan về HMM.

* HMM là một mô hình thống kê, được sử dụng để mô hình hóa các tham số tiếng nói của một đơn vị ngữ âm, trong một ngữ cảnh cụ thể.
* Mô hình Markov ẩn (HMM) cho phép chúng ta xem xét đến hai thành phần là các sự kiện quan sát được (như các đặc trưng của tiếng nói) và các sự kiện ẩn (ví dụ như các từ).
* Mỗi mô hình Markov ẩn được sử dụng để mô hình hóa một âm vị, và các mô hình markov ẩn được móc nối với nhau để mô hình hóa chuỗi âm vị.
* Mô hình Markov ẩn là một mô hình học máy dựa trên thống kê, do đó hệ thống tổng hợp tiếng nói dựa trên mô hình markov ẩn hoạt động bao gồm hai quá trình là quá trình huấn luyện và quá trình tổng hợp.
* Quá trình tổng hợp dựa trên mô hình markov ẩn sẽ nhận đầu vào là một đoạn văn bản, chuyển hóa đoạn văn bản này thành chuỗi âm vị, sau đó dựa vào các mô hình Markov ẩn mô hình hóa chuỗi các âm vị tương ứng ta sẽ tìm ra được các tham số mel và tần số cơ bản f0.
* Từ các tham số mel xây dựng nên chuỗi các bộ lọc MLSA (Mel Log Spectral Approximation) và kết hợp với tín hiệu kích thích được tạo từ f0 sẽ tạo ra được tín hiệu tiếng nói
* Quá trình huấn luyện dựa trên mô hình markov ẩn bao gồm các bước: Trích chọn đặc trưng tiếng nói và huấn luyện mô hình dựa trên các véc tơ đặc trưng trích được. Các đặc trưng tiếng nói được trích trong quá trình huấn luyện là các véc tơ như véc tơ hệ số mel và véc tơ mô tả f0.
* Sử dụng Multi-Space Probability Distribution Hidden để giải quyết vấn đề tần số cơ bản f0 chỉ tồn tại ở âm hữu thanh còn các âm vô thanh lại là nhiễu
* Các đặc trưng ngôn ngữ của văn bản được mô tả bằng cách sử dụng một bộ phân cụm (thường là cây quyết định) để gom các cụm trạng thái của mô hình markov ẩn có đặc tính ngôn ngữ gần nhau nhất và bầu chọn ra một trạng thái tiêu biểu để thay thế cho các trạng thái còn lại trong cụm.
* Hệ thống tổng hợp tiếng nói dựa trên mô hình markov ẩn là một hệ thống có khả năng tạo tiếng nói mang phong cách nói khác nhau, với đặc trưng của nhiều người nói khác nhau, thậm chí là mang cảm xúc của người nói.
* Ưu điểm của phương pháp này là cần ít bộ nhớ lưu trữ và tài nguyên hệ thống hơn so với tổng hợp ghép nối, và có thể điều chỉnh tham số để thay đổi ngữ điệu
* Nhược điểm của hệ thống này đó là độ tự nhiên trong tiếng nói tổng hợp của hệ thống bị suy giảm hơn so với tổng hợp ghép nối, tiếng nói tổng hợp nghe có vẻ đều đều mịn và đôi khi trở thành bị “nghẹt mũi”.
* Một mô hình Markov ẩn được đặc trưng bởi các thành phần cơ bản sau : N, số trạng thái (state) trong mô hình Markov. Các trạng thái thường được ký hiệu bằng S= {S1, S2, S3, ...} và trạng thái của mô hình tại thời điểm t được kí hiệu là qt , M là số ký hiệu quan sát (observation symbol), đây là kích thước của bảng từ vựng của mô hình. Các ký hiệu quan sát được biểu diễn bằng V= {v1, v2, ...}, A = {aij} là xác suất chuyển trạng thái (state transition probability distribution). Trong đó aij là xác suất để trạng thái j xuất hiện tại thời điểm t+1 khi trạng thái i đã xuất hiện tại thời điểm t .



B={bj(k)} xác suất phát xạ quan sát trong mỗi trạng thái (observation symbol probability distribution in state), bj(k) là xác suất của quan sát vk tại trạng thái j tại thời điểm t



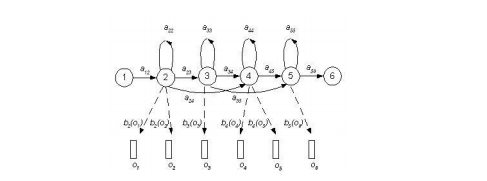
π = {π1, π2, ..., πN} xác suất trạng thái khởi đầu (initial state distribution), πi là xác suất để trạng thái i được chọn tại thời điểm khởi đầu t=1:



Với các giá trị thích hợp A, B, π, M, N, một mô hình Markov ẩn được dùng để sinh ra một dãy các quan sát: O= {O1 , O2 , O3 , ...} Trong đó Oi lấy một trong các giá trị trong V. Hoạt động của HMM được mô tả như sau: Chọn một trạng thái khởi đầu q1 tương ứng với xác suất trạng thái khởi đầu π. Gán t=1. Chọn Oi = vk tương ứng với xác suất quan sát tại trạng thái Si : bi(k). Chuyển sang trạng thái mới qt+1 = Sj tương ứng với xác suất chuyển trạng thái aij. Gán t=t+1 và quay lại lại bước 3) nếu t<T, nếu không kết thúc.

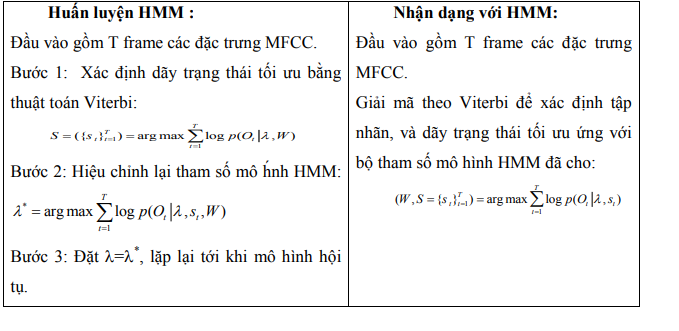
Người ta thường dùng bộ ba λ=(A, B, π) được coi là bộ ký pháp gọn để biểu diễn một mô hình Markov ẩn. A, B và π được gọi là các tham số (parameters) của mô h́nh λ.

Vd:



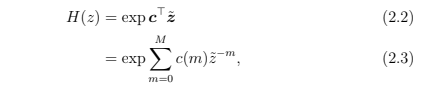
2.2- Nhận dạng tiếng nói dựa trên HMM với đặc trưng MFCC

* Các hệ thống nhận dạng dựa trên HMM thường sử dụng đặc trưng chuẩn MFCC và quy trình huấn luyện và nhận dạng theo thuật toán sau:



2.3- Mel-Cepstrum analysis

* Trong phân tích mel-cepstral, hàm truyền âm thanh H (z) được điều chỉnh bởi các hệ số mel-cepstral thứ tự c = [c (0), c (1), ..., c (M) ] T (siêu ký hiệu biểu thị chuyển vị ma trận) như sau:



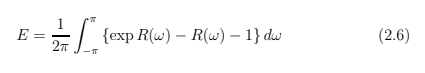
trong đó z ̃ = [1, z ̃ 1, ..., z ̃ M] T và z ̃ 1 qua hàm



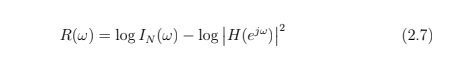
tần số bị biến dạng β (ω)



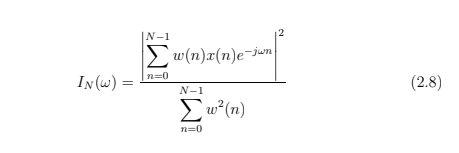
ước lượng công suất phổ | H (ejω) | 2 không biến thiên theo nghĩa công suất tương đối . Nên ta giảm thiểu tiêu chí E :



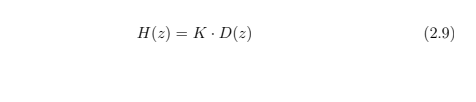
Với



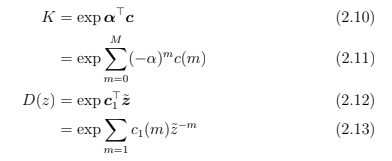
và In(ω) là biểu đồ sửa đổi của quá trình đứng yên yếu x (n) được xác định bởi :



Trong đó w (n) là cửa sổ có độ dài là N. Hệ số khuếch đại K bên ngoài từ H (z) trong biểu thức :



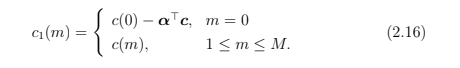
Tại



Với :



Mối quan hệ giữa các hệ số c và c1 được xác định :



Nếu hệ thống H (z) được coi là bộ lọc tổng hợp của lời nói, D (z)

phải ổn định . Do đó, giả sử rằng D (z) là hệ pha tối thiểu



Sử dụng phương trình trên, tiêu chí phổ E trở thành

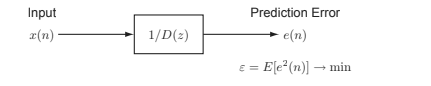


Với



Do đó, việc bỏ qua các số hạng không đổi, tối thiểu hóa E với phổ c dẫn đến giảm thiểu đối với c1 và việc nhân đôi E đối với K. bằng cách lấy đạo hàm của E đối với K và đặt kết quả về 0, K thu được như sau:



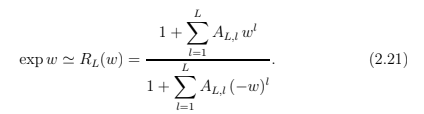


Trong đó εmin là giá trị tối thiểu của ε. Nó đã được chỉ ra rằng tối thiểu hóa phương trình. ε dẫn đến việc giảm thiểu năng lượng dư ε -> min như trong

Chỉ tồn tại một điểm tối thiểu vì tiêu chí E là lồi đối với c. Do đó, vấn đề tối thiểu hóa của E có thể được giải quyết bằng thuật toán lặp hiệu quả dựa trên FFT và các công thức đệ quy. Ngoài ra, độ ổn định của giải pháp mô hình H (z) luôn được đảm bảo .

2.4- Bộ lọc tổng hợp(Filter)

* Để tổng hợp lời nói từ các hệ số mel-cepstral, cần phải nhận ra hàm truyền hàm mũ D (z). Mặc dù hàm truyền D (z) không phải là hàm hợp lý, bộ lọc MLSA (Mel Log Spect xấp xỉ) có thể xấp xỉ D (z) với độ chính xác đủ. Hàm hàm mũ phức tạp exp w được xấp xỉ hợp lý



Ví dụ: nếu AL, l (l = 1, 2, ..., L) được chọn thì



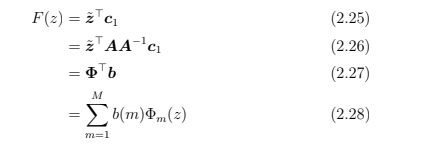
[L/L] xấp xỉ exp (w ) tại w = 0. Do đó D(z) xấp xỉ



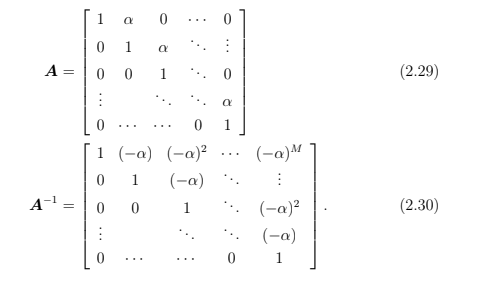
Với F(z) :



Do AL, l (l = 1, 2, ..., L) có các giá trị cố định trong khi c1 (m) là biến. Để xóa vòng lặp không có độ trễ khỏi F (z) đã được sửa đổi



Với :



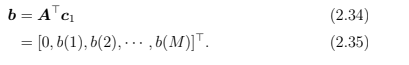
Vector Φ xác định bởi :



Với :



Các hệ số b có thể được lấy từ c1 bằng cách sử dụng phép biến đổi



Hoạt động ma trận trong phương trình có thể được thay thế bằng công thức đệ quy:



Vì phần tử đầu tiên của b bằng 0 do ràng buộc



giá trị của đáp ứng xung của F (z) là 0 tại thời điểm 0, nghĩa là F (z) không có đường dẫn không trễ.

Vì hàm truyền F (z) không có đường dẫn không có độ trễ, RL (F (z)) không có vòng lặp không có độ trễ, nghĩa là RL (F (z)) có thể thực hiện được.

Nếu b (1), b (2), ..., b (M) bị chặn, | F (ejω) | cũng bị ràng buộc, và ở đó

tồn tại một giá trị hữu hạn dương r sao cho:

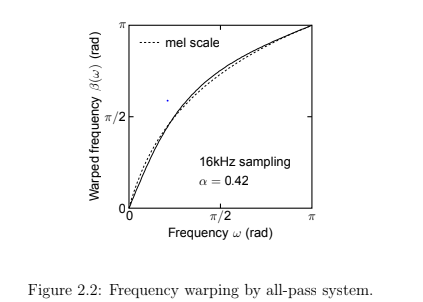


Các hệ số AL, l có thể được tối ưu hóa để giảm thiểu tối đa sai số tuyệt đối với | w | = r , max| EL (w) | sử dụng một kỹ thuật xấp xỉ Chebyshev phức tạp trong đó



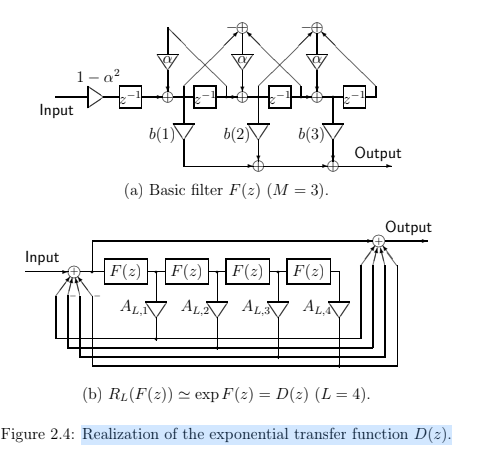
Các hệ số thu được với L = 5, r = 6.0 được thể hiện trong hình dưới đây

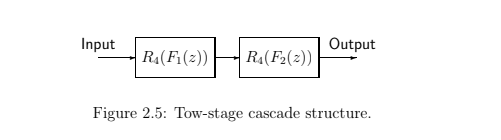
Khi | F (ejω) | <r = 6.0



Lỗi xấp xỉ không vượt quá 0,2735 dB



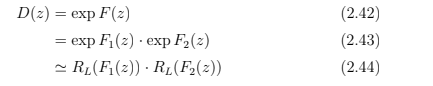




Khi F (z) được biểu thị là :



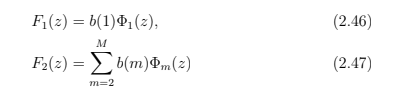
hàm truyền theo hàm mũ được tính gần đúng theo dạng :



như thể hiện trong hình 2.5 Nếu như :



Ước vọng RL (F1 (e^jω)) · RL (F2 (e^jω)) xấp xỉ D (e^jω) nhiều hơn so với RL (F (e^jω)). Với các chức năng sau đây đã được công nhận.



2.5- Đặc trưng âm thanh MFCC

* MFCC (Mel Frequency Cepstral Coefficients) là các hệ số biểu diễn phổ của phổ (spectrum-of-a-spectrum) của đoạn âm thanh. Các hệ số ceptral c(k) là cách thuận tiện cho việc mô hình hóa phân bổ năng lượng phổ
* 
* Các hệ số ceptral được tính toán cho mỗi khoảng thời gian ngắn của tín hiệu âm thanh. Hệ số ceptral được sử dụng trong MMDBMS (tìm kiếm âm thanh) và trong nhận dạng tiếng nói:
* Một cặp âm thanh cảm nhận có độ cao nhƣ nhau nếu giá trị mel của chúng như nhau. Mel-scale: Xấp xỉ tuyến tính dưới 1 kHz, và loga trên 1 kHz.
  + 1. **- Phương pháp phân tích MFCC**

1. **Quá trình lọc theo thang Mel Cepstral**

- MFCC gồm các bƣớc chính sau:

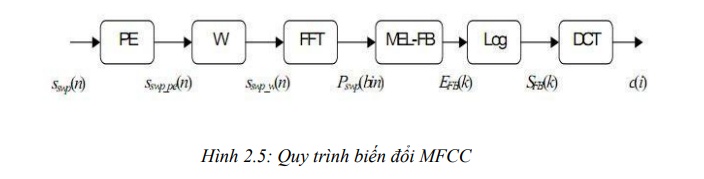
1. Phân khung tín hiệu

2. Cửa sổ hóa

3. Chuyển đổi sang miền tần số

4. Chuyển đổi sang thang Mel

5. Thực hiện biến đổi Cosine rời rạc



Quan sát quá trình trên ta thấy, âm thanh đƣợc chia thành những khung có độ dài cố định. Mục đích là để lấy mẫu những đoạn tín hiệu nhỏ (theo lý thuyết là ổn định). Hàm cửa sổ bỏ đi những hiệu ứng phụ và vector đặc trưng cepstral được thực hiện trên mỗi khung cửa sổ. Biến đổi Fourier rời rạc của mỗi khung được tính toán và lấy logarithm biên độ phổ. Thông tin về pha bị bỏ qua do biên độ phổ là quan trọng hơn pha. Thực hiện lấy logarithm biên độ phổ do âm lượng của tín hiệu là xấp xỉ logarith. Tiếp theo biến đổi phổ theo thang Mel. Từ kết quả này, trong vector

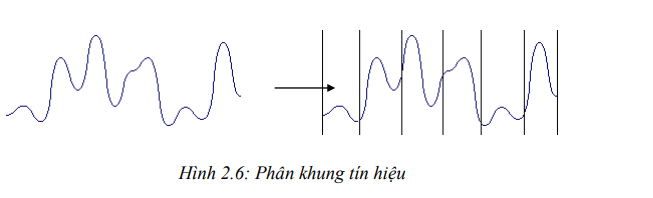
Mel – spectral của các thành phần tương quan cao, bước cuối cùng là thực hiện biến đổi cosine rời rạc để tổng hợp vector phổ Mel để tương quan lại các thành phần này

1. **Lấy mẫu**

- Trong việc lấy mẫu dữ liệu, chúng ta xem xét đến tín hiệu âm thanh đã được số hóa bằng việc rời rạc hóa các giá trị trên những khoảng đều nhau vì vậy cần phải chắc chắn rằng tốc độ lấy mẫu là đủ lớn để mô tả tín hiệu dạng sóng. Tấn số lấy mẫu nên ít nhất gấp đôi tần số dạng sóng như trong định lý của Nyquist. Tốc độ lấy mẫu phổ biến là 8000, 11025, 22050, 44000. Thông thường sử dụng tần số trên 10kHz.

1. **Phân khung tín hiệu**

- Phân khung là quá trình chia mẫu tín hiệu thành một số khung chồng lấp lên nhau hoặc không. Mục đích của phân khung là để lấy mẫu các đoạn tính hiệu nhỏ (theo lý thuyết là ổn định). Vấn đề là bản chất của âm thanh là không ổn định. Vì vậy, biến đổi Fourier sẽ thể hiện tần số xảy ra trên toàn miền thời gian thay vì thời gian cụ thể. Bởi thế khi tín hiệu là không ổn định, tín hiệu đó nên đƣợc chia nhỏ thành các cửa sổ rời rạc nhờ đó mỗi tín hiệu trong một cửa sổ trở nên tĩnh và phép biến đổi Fourier có thể thực hiện trên mỗi khung. Quá trình phân khung đƣợc thể hiện trong hình sau:



Trong khối này tín hiệu hiệu chỉnh ~ s(n) đƣợc phân thành các khung, mỗi khung có N mẫu, hai khung kề lệch nhau M mẫu. Khung đầu tiên chứa N mẫu, khung thứ hai bắt đầu chậm hơn khung thứ nhất M mẫu và chồng lên khung thứ nhất N-M mẫu. Tƣơng tự, khung thứ ba chậm hơn khung thứ nhất 2M mẫu (chậm hơn khung thứ hai M mẫu) và chờm lên khung thứ nhất N-2M mẫu. Quá trình này tiếp tục cho đến khi tất cả các mẫu tiếng nói cần phân tích 37 thuộc về một hoặc nhiều khung

1. **Lấy cửa sổ tín hiệu**

- Bước tiếp theo là lấy cửa sổ cho mỗi khung riêng rẽ nhằm giảm sự gián đoạn của tín hiệu tiếng nói tại đầu và cuối mỗi khung. Nếu w(n)

0<= n <=N-1

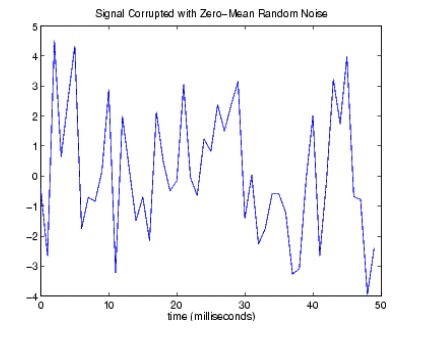
Thông thường, cửa sổ Hamming được sử dụng, cửa sổ này có dạng:

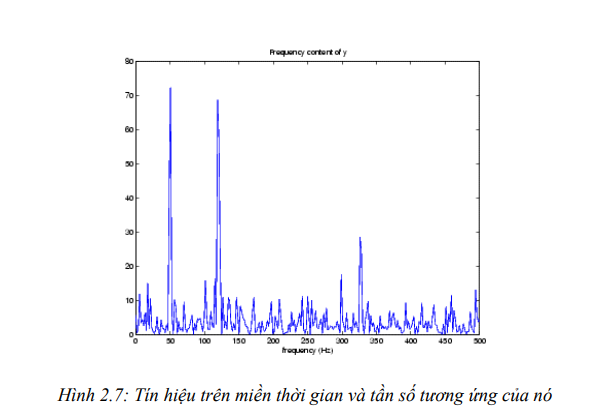


Ý tưởng ở đây là giảm bớt sự méo phổ bằng việc sử dụng các cửa sổ để giảm tín hiệu về không tại điểm bắt đầu và kết thúc mỗi khung. Sự chồng lấp các khung để làm nhẵn khung đến khung kế tiếp. Lấy cửa sổ tín hiệu đảm bảo tất cả các phần của tín hiệu được khôi phục và loại trừ được khe hở giữa các khung. Việc này được thực hiện để khử tính không liên tục tại đường viền khung cho biến đổi Fourier thực hiện sau đó. Các khung có độ dài lớn hơn có cao độ pitch chính xác hơn và nén dữ liệu tốt hơn nhƣng chất lượng giảm

1. **biến đổi Fourier nhanh**

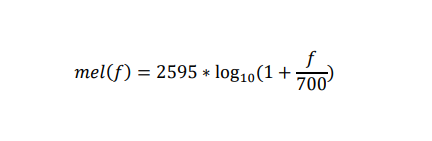
- Biến đổi Fourier rời rạc (DFT) hoặc biến đổi Fourier nhanh (FFT) được thực hiện để chuyển đổi mỗi khung với N mẫu từ miền thời gian sang miền tần số. Tín hiệu gốc cần được thực hiện biến đổi Fourier qua bộ lọc thông dải để xử lý độ lệch tần số Mel. Biến đổi Fourier chuẩn không được sử dụng do tín hiệu âm thanh không xác định trên toàn miền thời gian. Thông thường hay sử dụng biến đổi DFT. Hình sau thể hiện tín hiệu trên miền thời gian và mô tả tần số tƣơng ứng của nó



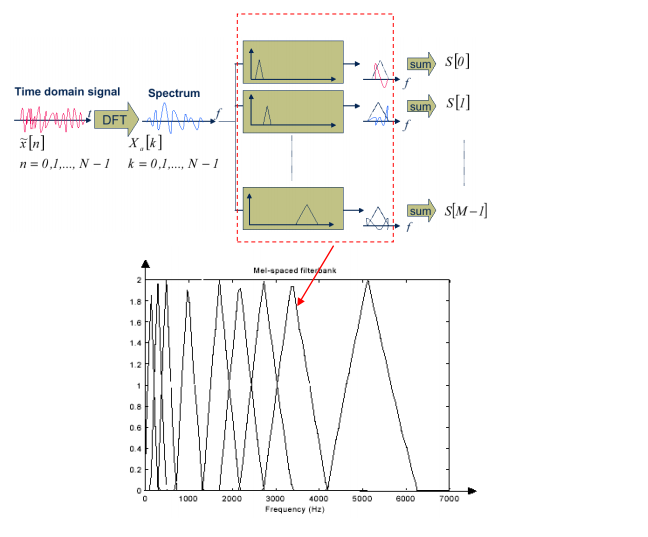


1. **Chuyển đổi sang thang tần số Mel**

- Để mô tả chính xác sự tiếp nhận tần số của hệ thống thính giác, người ta xây dựng một thang khác – thang Mel. Việc chuyển đổi sang miền tần số Mel làm nhẵn phổ và làm nổi lên các tần số cảm thụ có nghĩa. Biến đổi Fourier lên tín hiệu qua bộ lọc thông dải để làm đơn giản phổ mà không làm mất dữ liệu. Điều này được thực hiện bằng cách tập hợp các thành phần phổ thành một dải tần số. Phổ được làm đơn giản hóa do sử dụng một dàn bộ lọc để tách phổ thành các kênh. Các bộ lọc được đặt cách đều nhau trên thang Mel và lấy logarit trên thang tần số, các kênh có tần số thấp là không gian tuyến tính trong khi các kênh có tần số cao là không gian logarit. Tai người không cảm nhận sự thay đổi tần số của tiếng nói tuyến tính mà theo thang mel. Thang tần số Mel tuyến tính ở tần số dưới 1kHz và logarit ở tần số cao hơn 1kHz. Ta chọn tần số 1kHz, 40 dB trên ngưỡng nghe 1000 Mel. Do đó công thức gần đúng biểu diễn quan hệ tần số ở thang mel và thang tuyến tính như sau:

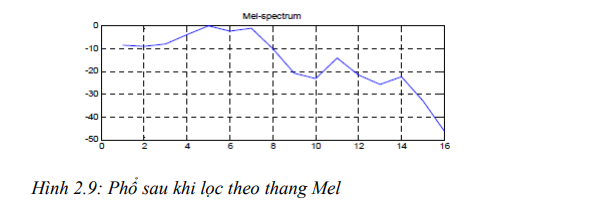


Một phương pháp để chuyển đổi sang thang Mel là sử dụng băng lọc. Khoảng cách của băng lọc được định nghĩa bởi một hằng số tần số mel theo thời gian. Biến đổi phổ S(w) bao gồm đầu ra của bộ lọc khi S(w) là đầu vào. Băng lọc này được áp dụng trong miền tần số, nó có thể xem như các điểm thu được của bộ lọc chính. (Hình 2.8). Với các khung nhỏ tốt nhất là sử dụng các bộ lọc dạng tam giác hoặc thậm chí hình chữ nhật vì độ phân giải là quá thấp trong miền tần số thấp



Hình 2.8: Băng lọc khoảng cách theo tần số mel

Mỗi bộ lọc trong băng lọc đƣợc nhân với phổ tín hiệu vì vậy chỉ có một giá trị đơn của cường độ trên bộ lọc được trả lại. Điều này có thể đạt được qua các tính toán của ma trận đơn. Kết quả là tổng của biên độ trong dải lọc và vì vậy làm giảm độ chính xác tới mức tai của con người. Hình 2.9 cho thấy kết quả này. Trục hoành mô tả chỉ số của bộ lọc và theo thang mel :



Quá trình chuyển đổi sang thang tần số mel được thực hiện theo ba bước :

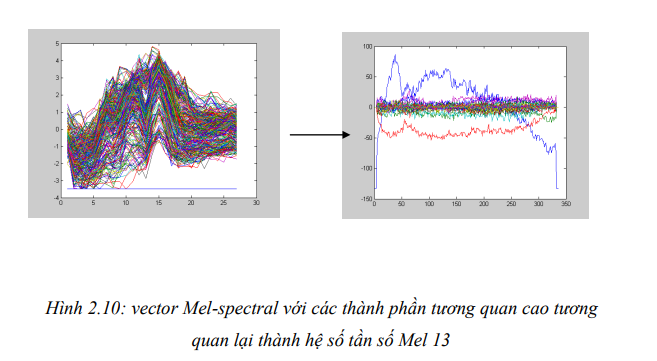
1. Cố định vùng giá trị dưới mỗi bộ lọc và đôi khi đưa thang về 1. Đặt M = số băng lọc yêu cầu

2. Phân bố đều trên thang tần số Mel

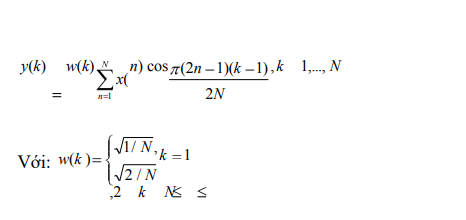
3. Chuyển đổi từ Hz sang w i' s trên thang tuyến tính. Mối quan hệ giữa mel và frq đƣợc cho bởi công thức: m=ln(1+f/700)\*1000/ln(1+1000/700)

1. **Biến đổi Cosin rời rạc**

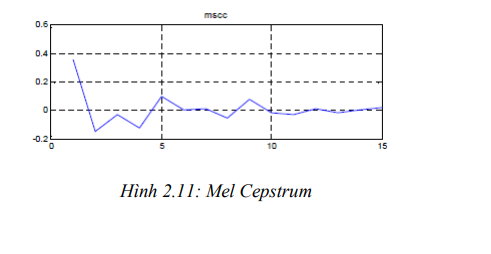
- Ở bước cuối cùng này, sử dụng biến đổi cosin rời rạc để làm tương quan cường độ phổ logarit mel với hệ số tần số mel MFCC. Cepstrum là biến đổi ngược của tín hiệu theo log biên độ. Vì nguồn phổ là cân xứng nên thay biến đổi Fourier ngược thành biến đổi cosin rời rạc (DCT) . Thêm vào đó, DCT có khả năng tương quan cao hơn và hệ số cepstral chặt chẽ hơn. Hình dưới mô tả vector Mel-spectral với các thành phần tương quan cao tương quan lại thành hệ số tần số Mel 13



Sự rời rạc tính hiệu x(n) đƣợc định nghĩa trong biểu thức:



Thực hiện DCT, thu được tần số Mel Cepstrum (Hình 2.11) Có thể xem như hệ số thứ 0: C0 được loại trừ. Lý do là nó đại diện cho các giá trị của tín hiệu vào mang ít thông tin. Beth Logan đã nói rằng hệ số cepstral thứ 0 chỉ chứa thông tin về năng lượng.



Quan sát hình trên chúng ta có thể thấy đƣợc độ suy giảm biên độ ở các tần số cao hơn.