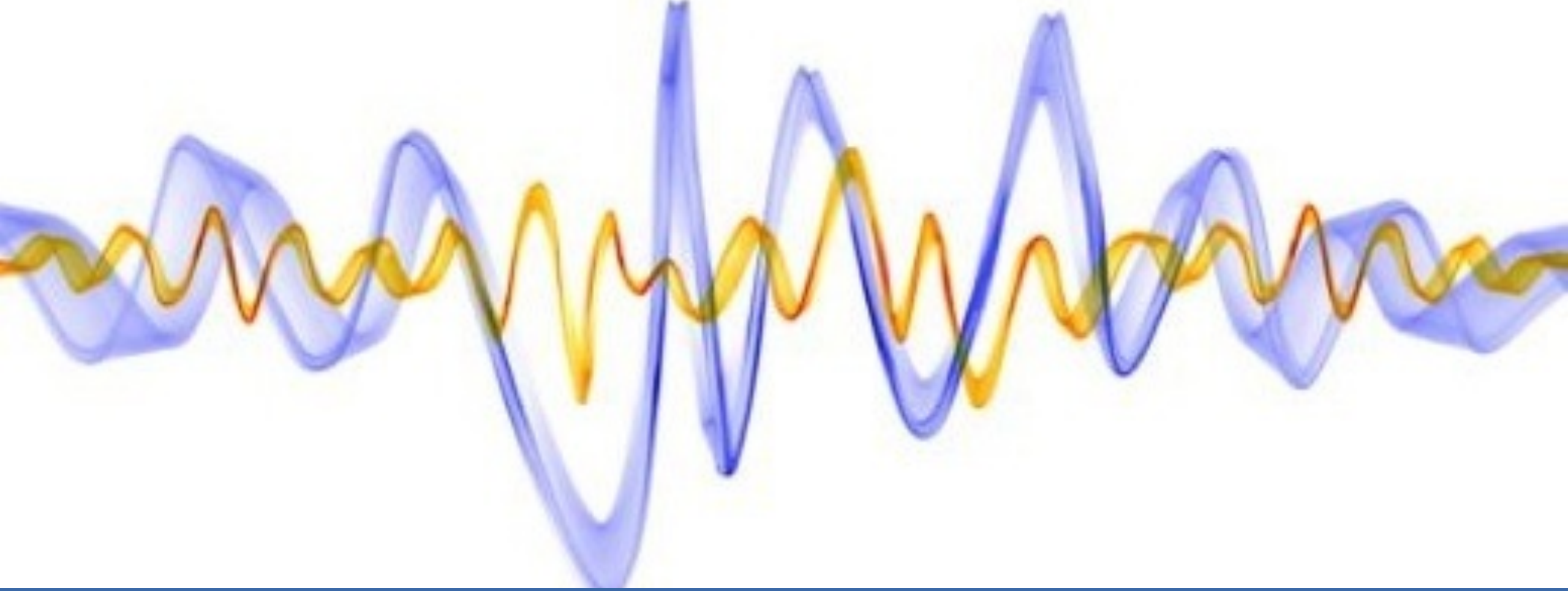


MFCC | Mel-Frequency Cepstral Coefficients

Aluno: Leonardo Manhães Gomes
Profª: Débora Christina Muchaluat Saade
Discipina: Sistemas Multimídia

11/12/2018

Agenda

- Introdução sobre características da voz
- MFCC
- Comparações entre MFCC e outras técnicas
- Exemplos de aplicações em sistemas multimídia

Características da Voz, da Fala e da Linguagem

Voz

- A voz é a ferramenta de comunicação mais primária e mais imediata de que dispomos para interagir na sociedade, pois ela não requer qualquer acessório nem mecanismo especial para ser utilizada.
- Captamos pela voz de uma pessoa: emoções, sensações, intenções e se as pessoas estão alegres, tristes, apressadas, ou seguras.
- A voz interfere em nossa comunicação social ou profissional e determina a própria personalidade e o estado de espírito de quem fala.
- A frequência da voz pode variar entre 50 e 3.400 Hz.
- As cordas vocais vibram rapidamente. Nos homens, que possuem cordas com mais massa e menos esticadas que as das mulheres, o ciclo vibratório fica em torno de 125 vezes por segundo. Nas mulheres, que possuem voz mais aguda, o número aumenta 250 vezes por segundo. Essa característica vibratória é conhecida como frequência.
- O mecanismo para gerar a voz humana pode ser subdividido em três partes: os pulmões, as pregas vocais dentro da laringe e os articuladores - lábios, língua, dentes, palato duro, véu palatar e mandíbula.

Características da Voz, da Fala e da Linguagem

Fala

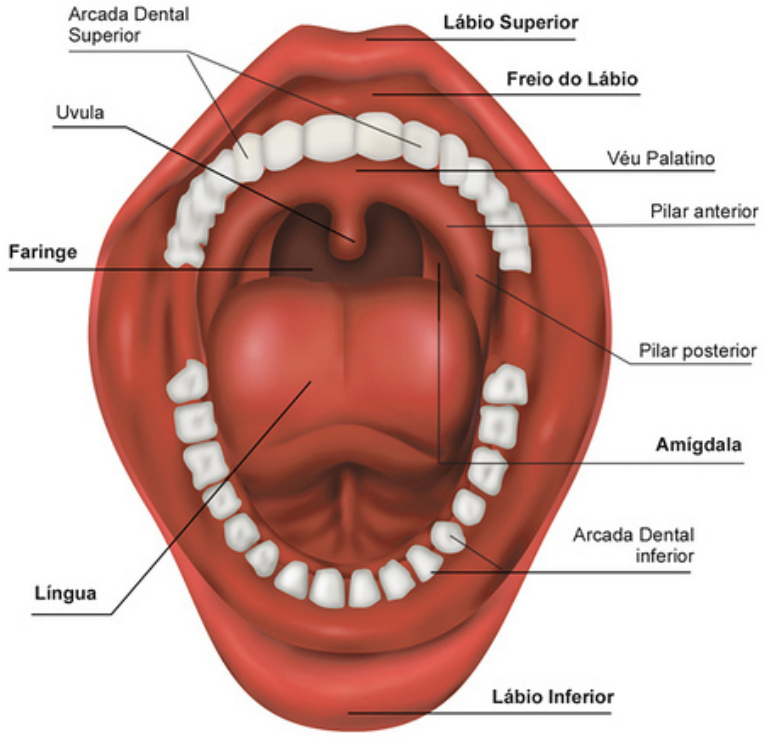
É a capacidade mecânica de emitirmos sons.

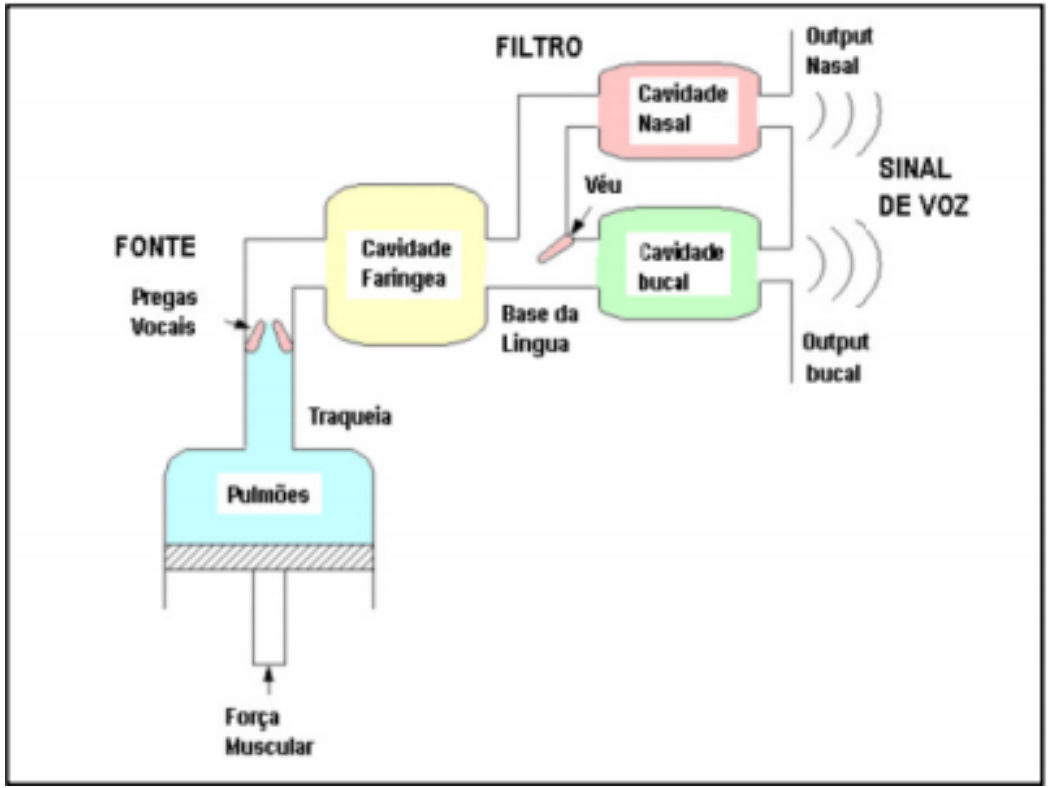
Linguagem

É um sistema de comunicação natural, artificial, humano ou não humano. Constitui a base de todas as nossas relações sociais, políticas, afetivas, culturais e históricas.

Filtros Naturais da Fala Humana

- Os sons gerados por um humano são filtrados pela forma do trato vocal, incluindo a língua, os dentes, etc.
- A forma do trato vocal se manifesta no envelope do curto espectro de tempo e o trabalho das técnicas de extração é representar com precisão este envelope.





Técnicas de Extração de Características da Fala

- Extraem características da fala de um orador permitindo trabalhar tais características e apresentá-las a um Sistema.

- Encaminham os dados da fala como *input* dos sitemas que trabalham com reconhecimento da fala, de palavras, idiomas, sentimentos e oradores.

- Extraem e identificam os componentes do sinal de áudio que são adequados para processamento do conteúdo linguístico e descartam todas as outras partes que transportam informações desnecessárias. Ex. de descarte: ruído de fundo.

MFCC

Escala Mel

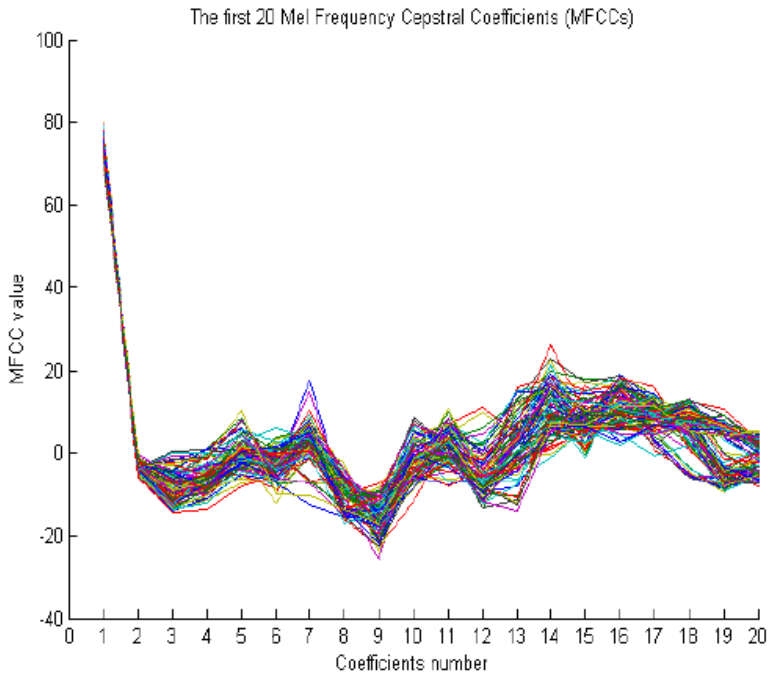
- Mel é uma unidade de medida da altura do som proposta por Stevens, Volkmann and Newmann em 1937.
- O objetivo da sua criação foi construir uma escala que refletisse exatamente como as pessoas ouvem os tons musicais.
- Foram realizados experimentos com ouvintes e aplicado um método utilizando critérios perceptivos, conhecido na psicofísica como diferença apenas perceptível (*just-noticeable difference* - JND) ou limiar diferencial (*differential threshold*).
- “mel” faz referência à palavra “melodia”.
- O ponto de referência entre mel e a frequência normal é definido atribuindo uma altura sonora perceptível de 1.000 mels a um tom de 1.000 Hz, 40 dB acima do limiar do ouvinte.
- Em 1976, Makhoul e Cosell publicaram a atual fórmula popular para conversão da frequência da escala em Hertz para a escala em Mels:

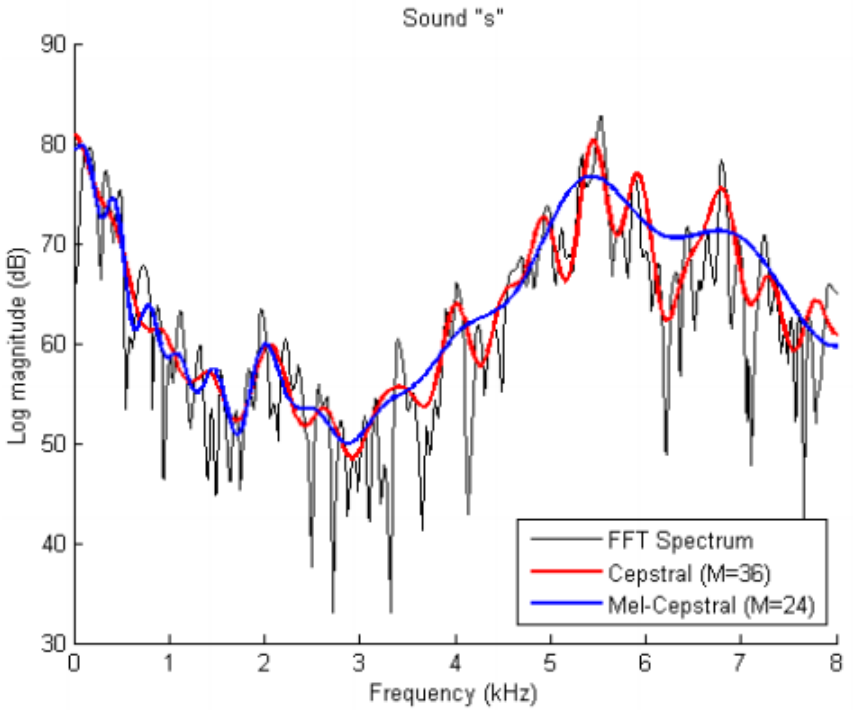
$$Mel = 2595\ log_{10}(1 + f/700) = 1127\ ln(1 + f/700)$$

MFCC

Mel-Frequency Cepstral Coefficients

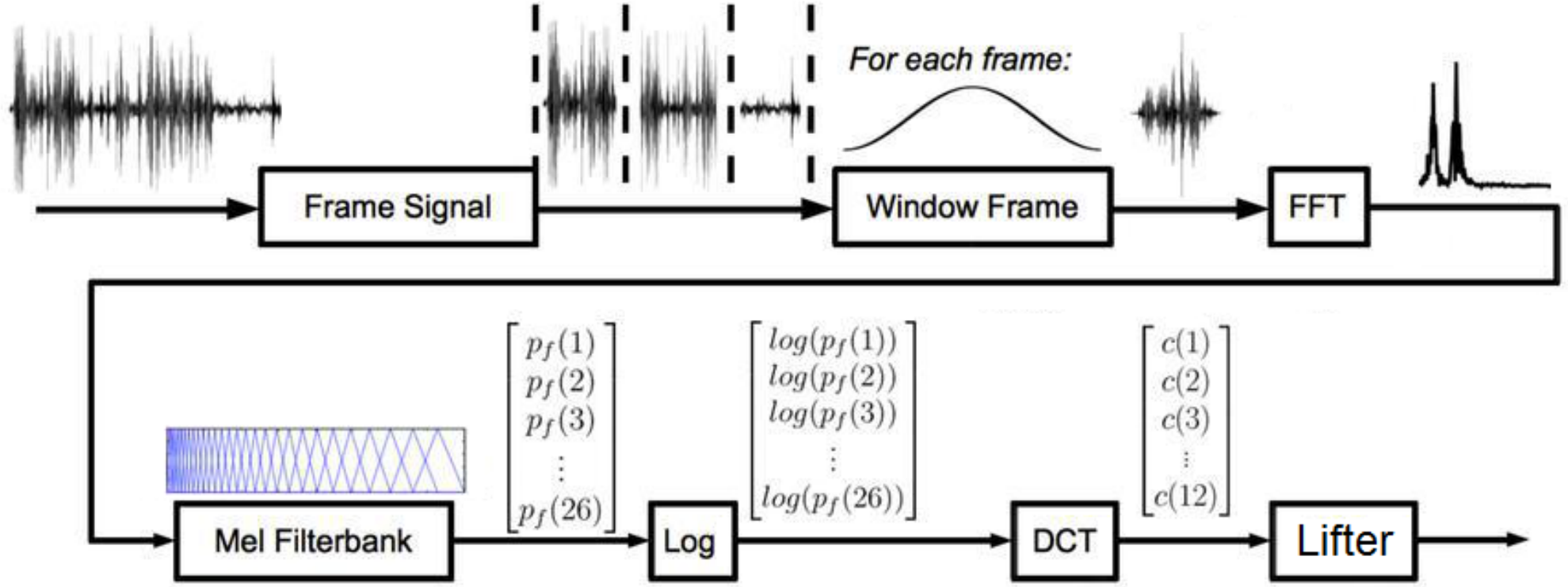
- **Cepstrum** → é uma representação de curto prazo do espectro de potência de um som.
- Os coeficientes cepstrais de frequência mel (MFCCs) coletivamente formam uma MFC.
- O uso de cerca de 20 coeficientes MFCC é comum na ASR (*Automatic Speech Recognition*), embora entre 10 a 12 sejam frequentemente considerados suficientes para codificar a fala.





MFCC

Mel-Frequency Cepstral Coefficients – Processamento para extração das *features*



MFCC

Mel-Frequency Cepstral Coefficients – Processamento para extração das *features*

1. Pré-ênfase

Este passo processa a passagem do sinal através de um filtro que enfatiza frequências mais altas aumentando a sua energia de sinal.

2. Framing

Processo de segmentação das amostras de fala obtidas da conversão analógico-digital em um quadro pequeno com o comprimento dentro da faixa de 20 a 40 mseg. O sinal de voz é dividido em quadros de N amostras.

MFCC

Mel-Frequency Cepstral Coefficients – Processamento para extração das *features*

3. Janela de Hamming

Quando o sinal medido é periódico e um número inteiro de períodos preenche o intervalo de tempo de aquisição, o FFT fica bem, pois corresponde a essa suposição.

Quando o número de períodos na aquisição não é um número inteiro, os pontos finais são descontínuos. Estas descontinuidades artificiais aparecem na FFT como componentes de alta frequência não presentes no sinal original.

O espectro obtido da FFT, portanto, é uma versão borrada. Parece que a energia de uma frequência vaza para outras freqüências. Esse fenômeno é conhecido como vazamento espectral.

O janelamento reduz a amplitude das descontinuidades nos limites de cada seqüência finita adquirida pelo digitalizador.

$$w(n) = 0.54 - 0.46 \cos(2\pi nM-1) \qquad 0 \le n \le M-1$$

M → número de pontos na janela de saída

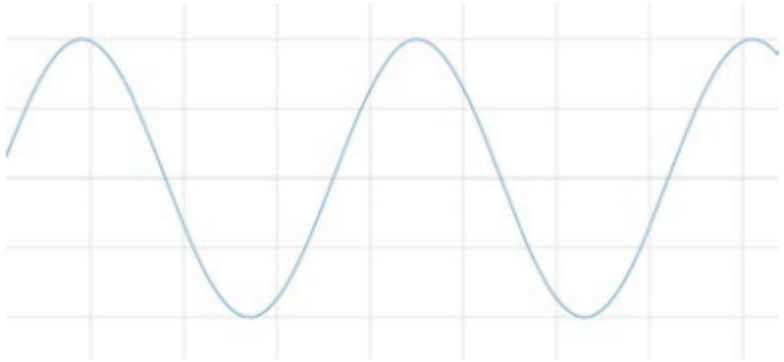
MFCC

Mel-Frequency Cepstral Coefficients – Processamento para extração das *features*

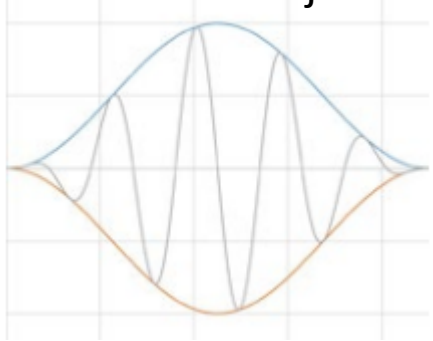
3. Janela de Hamming

1 - Medindo número de períodos não inteiros

A

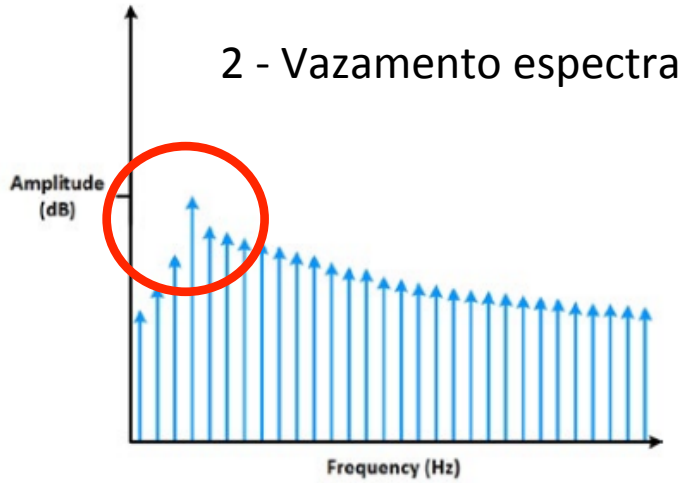


3 - Processando o janelamento



2 - Vazamento espectral adicionado na FFT

B



4 – Efeitos do vazamento espectral minimizados

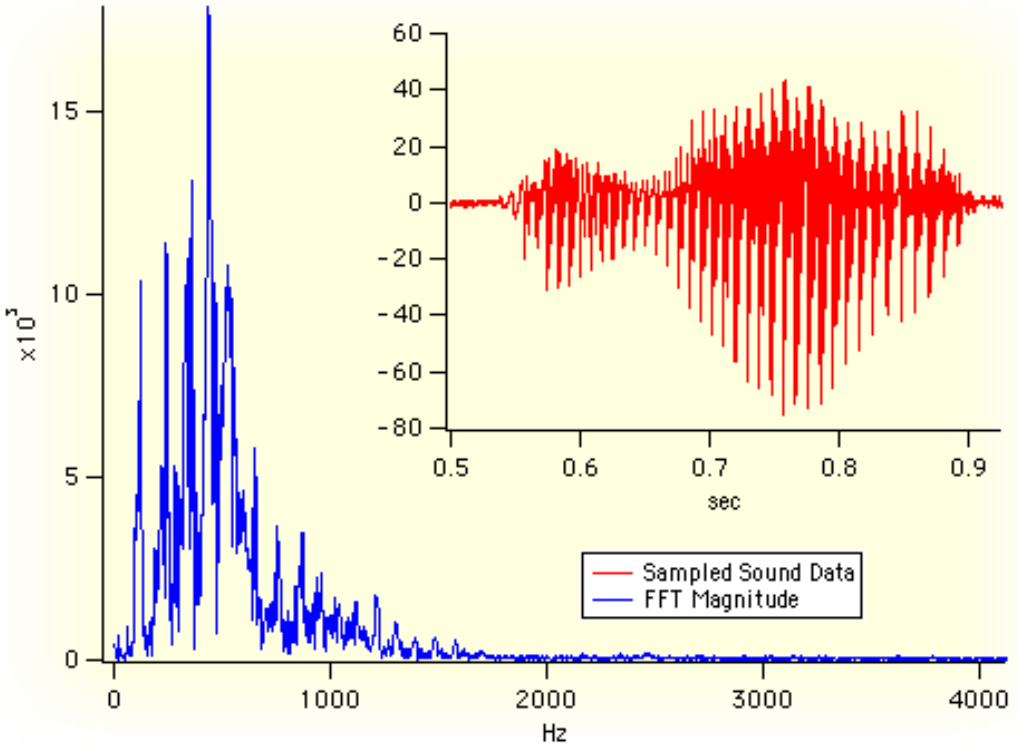
MFCC

Mel-Frequency Cepstral Coefficients – Processamento para extração das *features*

4. *Fast Fourier Transform* (FFT)

Para converter cada quadro de N amostras do domínio de tempo em domínio de frequência. Um sinal complicado pode ser dividido em ondas mais simples.

$$S_n = \sum_{k=0}^{N-1} s_k e^{-2\pi jkn/N} \quad , n = 0,1,2,\ldots, N-1$$



MFCC

Mel-Frequency Cepstral Coefficients – Processamento para extração das *features*

5. Processamento de Banco de Filtros Mel

O banco de filtros Mel é importante devido aos seguintes motivos:

- Aplica o escalonamento Mel-frequency, que é uma escala de percepção que ajuda a simular o funcionamento do ouvido humano. Corresponde a melhor resolução em baixas frequências e menos em alta.
- O uso do banco de filtros triangular ajuda a capturar a energia em cada banda crítica e fornece uma onda aproximada da forma do espectro, além de suavizar a estrutura harmônica.

Para calcular um Mel para uma determinada frequência, usamos a seguinte equação aproxi

$$Mel(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right)$$

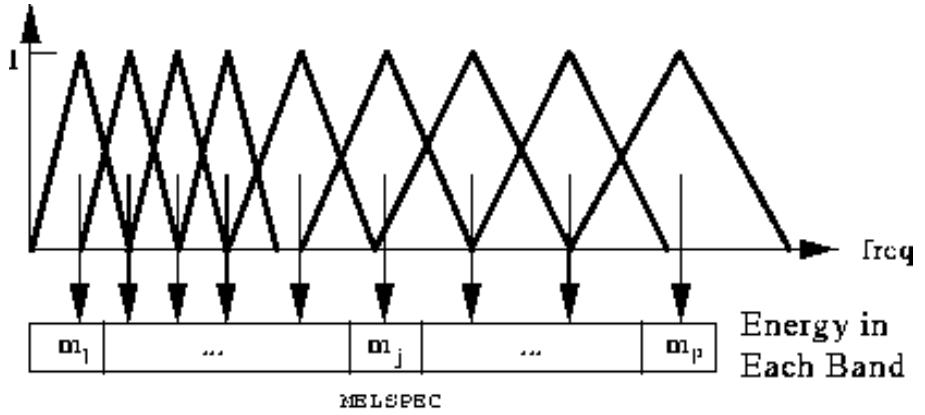


Fig. 5.3 Mel-Scale Filter Bank

MFCC

Mel-Frequency Cepstral Coefficients – Processamento para extração das *features*

6. Log

Em seguida, o módulo de log é aplicado como função de suavização para antecipar as informações de perda quando o processo de filtragem usando o banco de filtros mel é aplicado.

7. *Discrete Cosine Transform* (DCT)

Este é o processo para converter o espectro log Mel no domínio do tempo usando Transformada Cosseno Discreta (DCT). O resultado da conversão é chamado Mel Frequency Cepstrum Coefficient. O conjunto de coeficiente é chamado de vetores acústi[illegible] de entrada é transformado em uma seqüência de vetore

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cdot \cos\left(\frac{\pi i}{N}(j - 0{,}5)\right)$$

MFCC

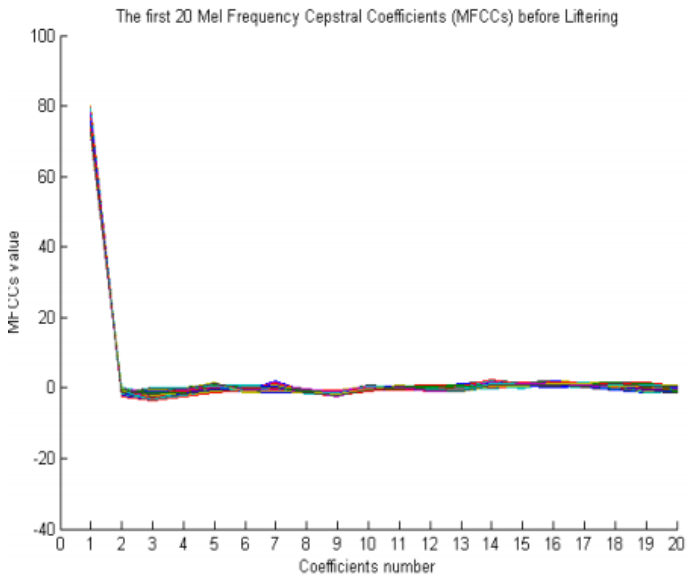
Mel-Frequency Cepstral Coefficients – Processamento para extração das *features*

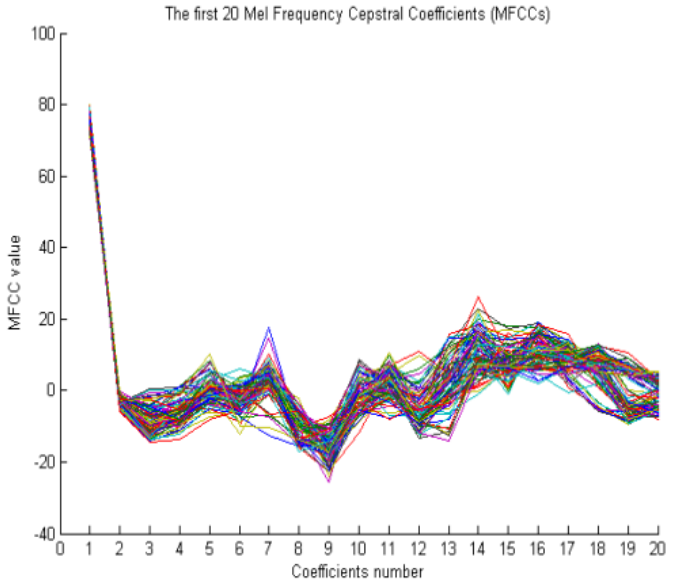
8. Lifter

A principal vantagem dos coeficientes cepstrais é que eles não são correlacionados. No entanto, o problema com eles é que os coeficientes cepstrais de ordem superior são relativamente pequenos. Para isso, é essencial redimensionar esses coeficientes cepstrais para magnitudes bastante semelhantes. Isto foi realizado por Liftering os coeficientes cepstrais de acordo com a seguinte equação:

$$c'_n = \left(1 + \frac{L}{2} \cdot \sin\left(\frac{\pi \cdot n}{L}\right)\right) \cdot c_n$$

onde L é o parâmetro do ascensor do seno Cepstral. Neste caso foi usaso L = 22.





Comparativo entre MFCC e outras técnicas

Caso 1 - Comparativo entre as técnicas MFCC e LPC para o reconhecimento de palavras isoladas da língua Marathi (língua indo-ariana, falada na Índia ocidental e central)

- O banco de dados de fala Marathi é gravado em ambiente Ruidoso.
- O BC consiste em palavras marathi simples que começam com vogais e consoantes.
- Cada palavra foi repetida 10 vezes por um orador masculino e um feminino.
- Para a identificação dos oradores foi utilizado um método de quantização vetorial baseado na distância euclidiana.

Comparativo entre MFCC e outras técnicas

Caso 1 - Comparativo entre as técnicas MFCC e LPC para o reconhecimento de palavras isoladas da língua Marathi (língua indo-ariana, falada na Índia ocidental e central)

Tabela 1: Acurácia no reconhecimento de *features* usando LPC

WORD	SPEAKER 1	SPEAKER 2
AAI	75%	73%
ANANAS	78%	74%
BAL	80%	78%
KSHATRIYA	81%	80%
AVERAGE	78.5%	76.25%

Tabela 2: Acurácia no reconhecimento para *features* usando MFCC

WORD	SPEAKER 1	SPEAKER 2
AAI	98%	99%
ANANAS	100%	100%
BAL	100%	100%
KSHATRIYA	100%	100%
AVERAGE	99.5%	99.75%

Comparativo entre MFCC e outras técnicas

Caso 2 - Comparativo entre as técnicas MFCC, LPCC e BFCC para o reconhecimento de palavras Hindi (língua indo-ariana falada principalmente na Índia central e norte) utilizando RNA.

Gráfico 1: Reconhecimento de palavras isoladas

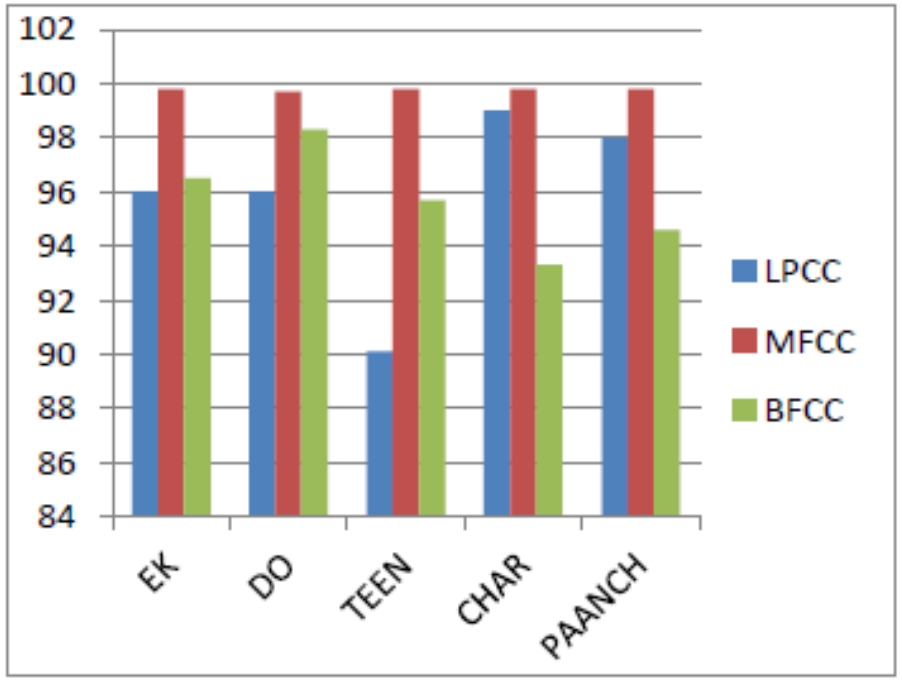
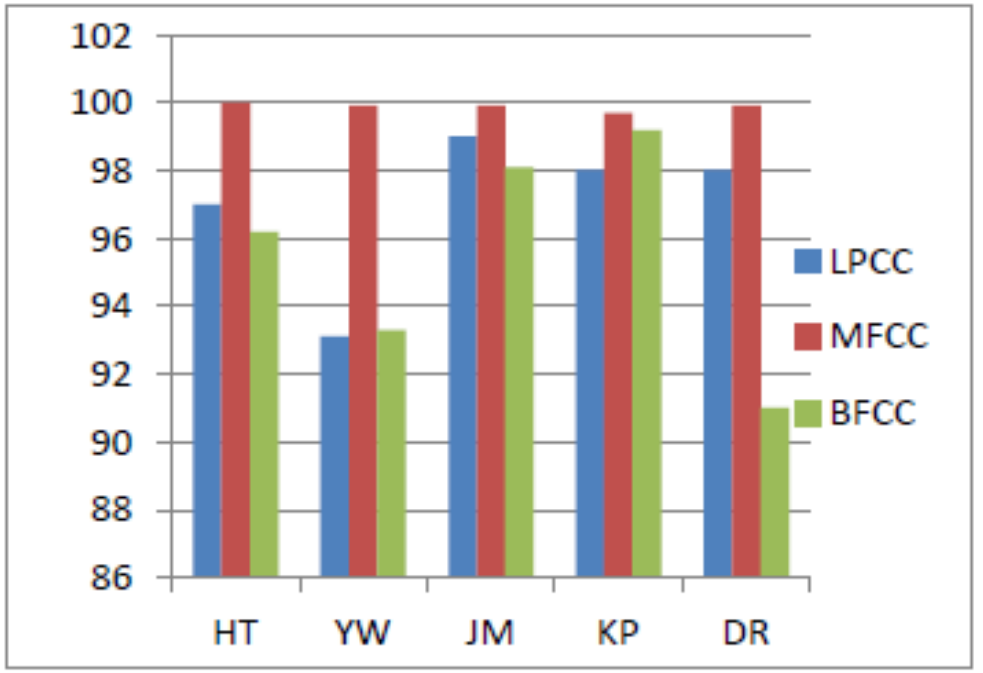


Gráfico 2: reconhecimento de pares de palavras (representadas no gráfico apenas pela primeira letra)



Comparativo entre MFCC e outras técnicas

Caso 2 - Comparativo entre as técnicas MFCC, LPCC e BFCC para o reconhecimento de palavras Hindi (língua indo-ariana falada principalmente na Índia central e norte) utilizando RNA.

Gráfico 3: Formação aleatória de pares de palavras, representadas apenas pela primeira letra

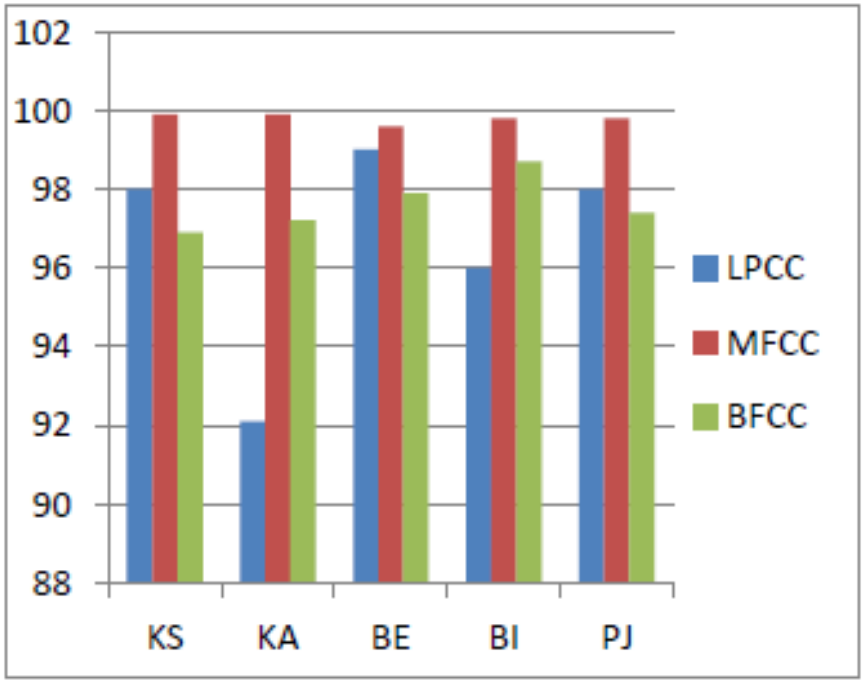
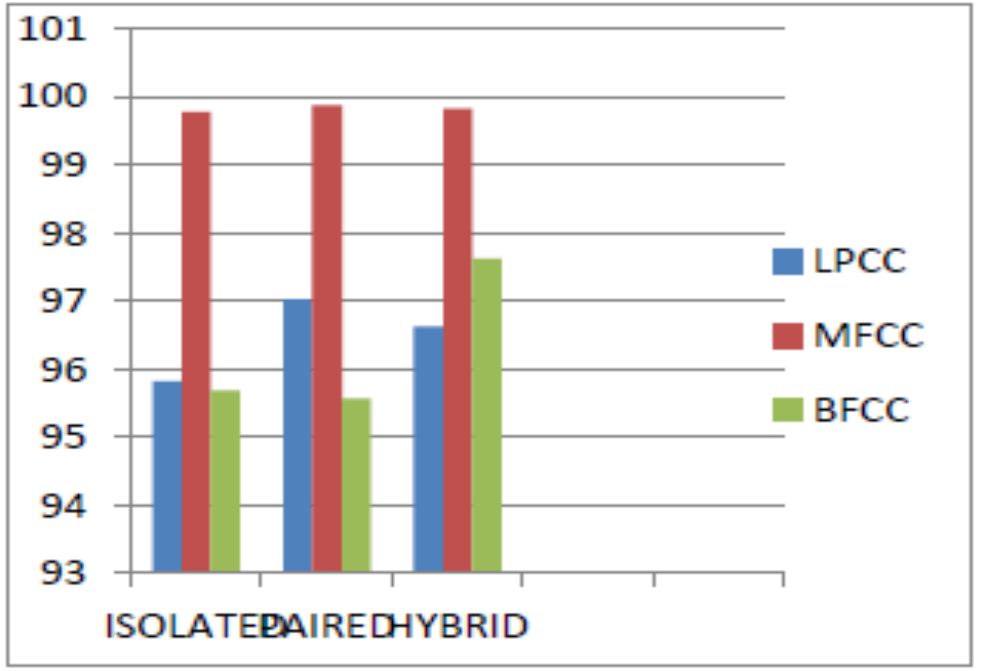


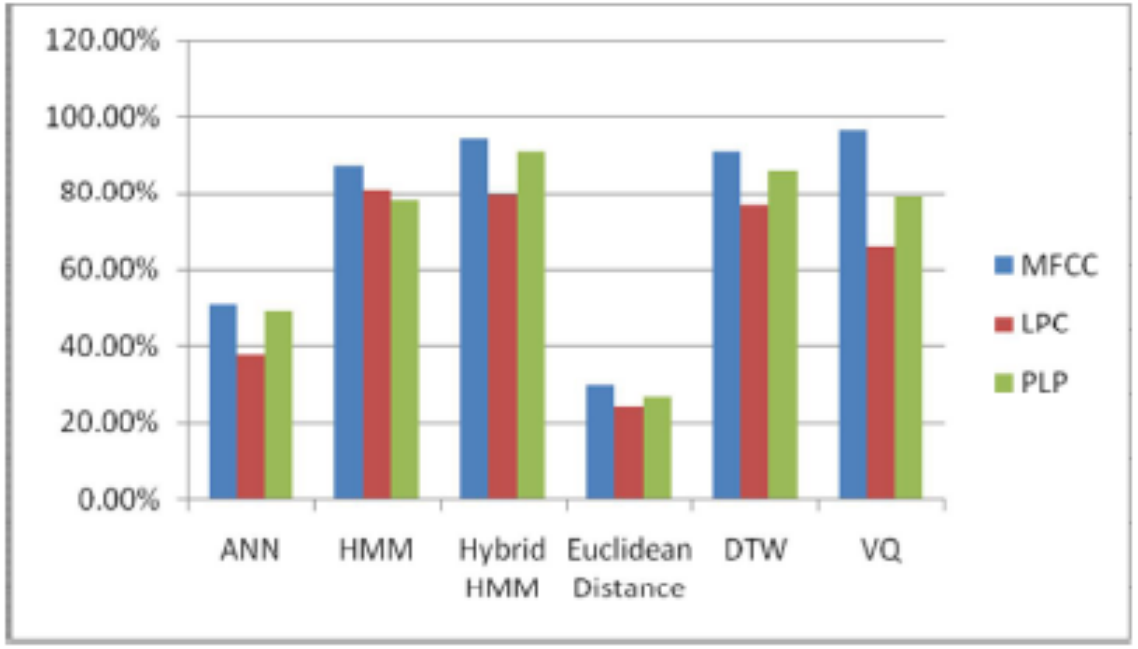
Gráfico 4: Média dos 3 gráficos anteriores



Comparativo entre MFCC e outras técnicas

Caso 3 - Análise comparativa utilizando as técnicas de extração MFCC, LPC e PLP e diferentes métodos classificadores para identificação de oradores.

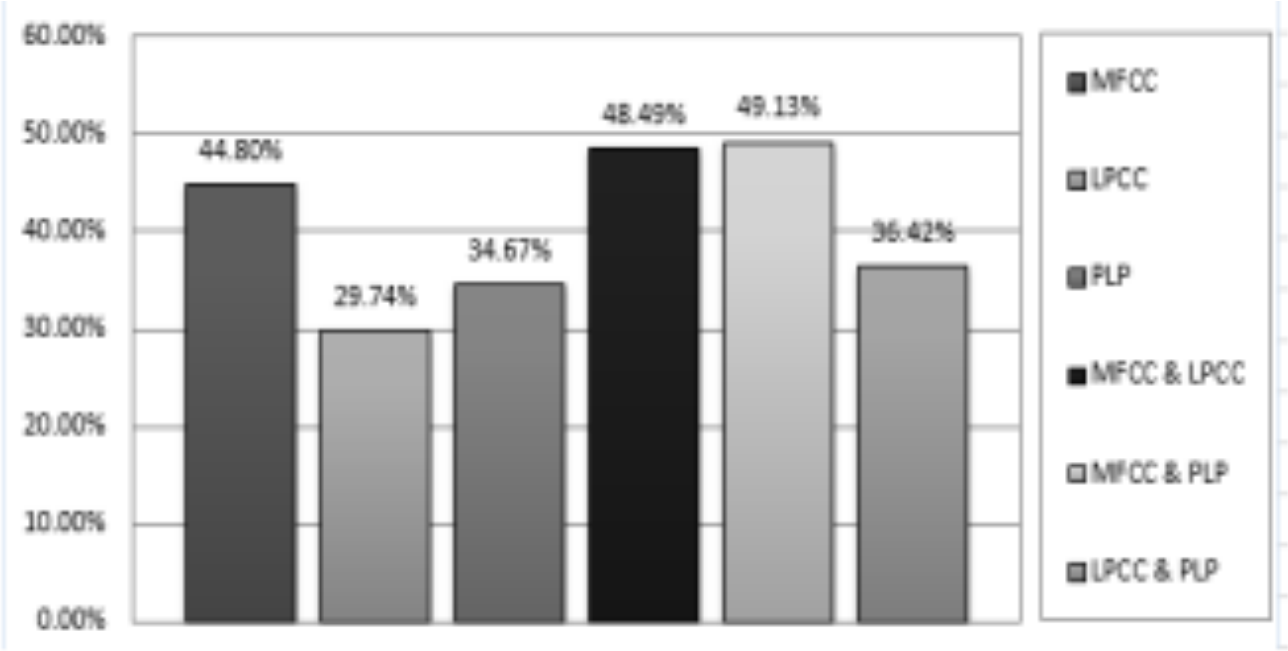
- Os métodos classificadores comparados foram: Redes Neurais Artificiais, Modelo Oculto de Markov, Modelo Oculto de Markov Híbrido, Distância Euclidiana, DTW (*Dynamic Time Warping*) e Quantização de Vetor



Comparativo entre MFCC e outras técnicas

Caso 4 - Análise comparativa da utilização individual e combinada das técnicas de extração MFCC, LPCC e PLP.

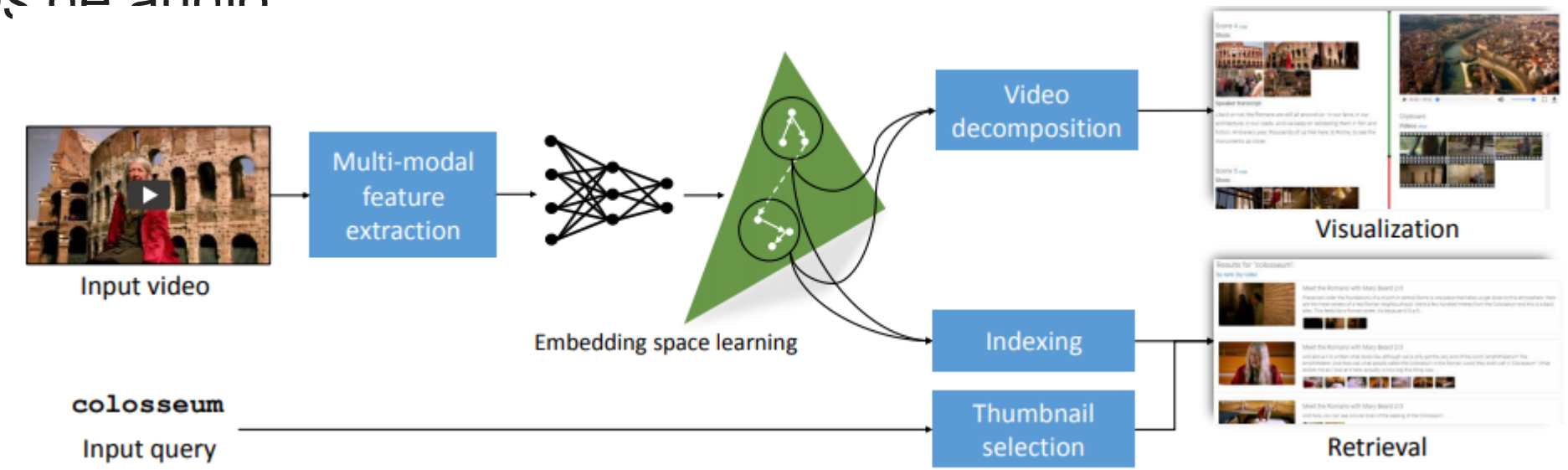
- A base de dados consiste em 2000 falas de quatro palavras árabes isoladas pronunciadas por 50 oradores árabes nativos sendo que cada orador repete a palavra 10 vezes



Aplicações em sistemas multimídia

Caso 1 - Um sistema multimídia interativo para indexação de vídeo e reutilização.

O vídeo de entrada é decomposto em partes coerentes por meio de uma rede *Triplet Deep* treinada em recursos multimodais: essa decomposição é a base da interface de visualização e também permite uma pesquisa detalhada dentro de clipes de vídeo. Uso do MFCC na extração de *features* de áudio



Aplicações em sistemas multimídia

Caso 2 - Classificação e recuperação de áudio baseada em conteúdo inteligente para aplicativos da Web.

- Identificação de várias classes durante a análise de áudio, utilizando os extratores de *features* MFCC e LPC.

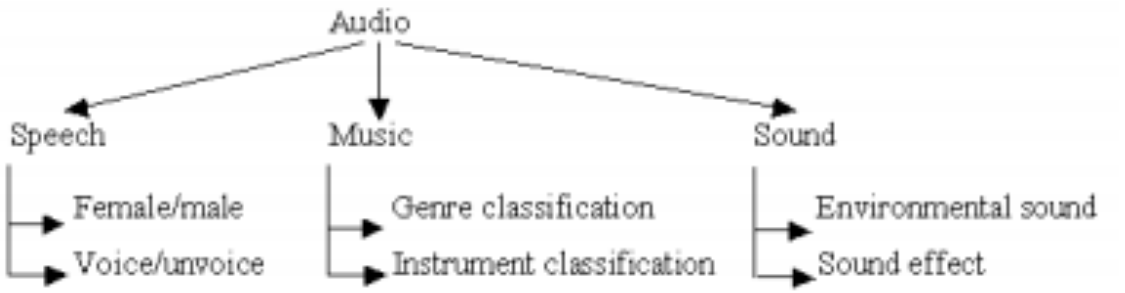


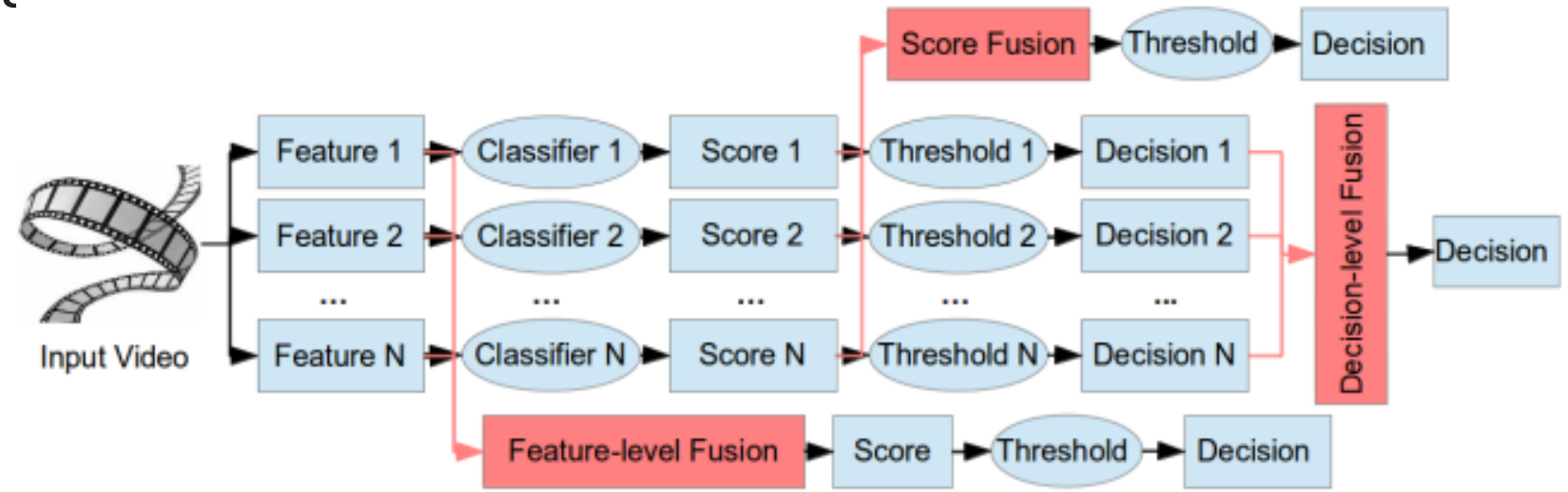
Table 13.1 The audio database structure.

Class name	No of files	Class name	No of files
1.Speech	53	Violin-pizzicato(9)	40
Female(1)	36	3.Sound	62
Male(2)	17	Animal(10)	9
2.Music	299	Bell(11)	7
Trombone(3)	13	Crowds(12)	4
Cello(4)	47	Laughter(13)	7
Oboe(5)	32	Machines(14)	11
Percussion(6)	102	Telephone(15)	17
Tubular-bell(7)	20	Water(16)	7
Violin-bowed(8)	45	Total	414

Aplicações em sistemas multimídia

Caso 3 - Fusão de pontuação ponderada por metadados para detecção de eventos multimídia.

- Detecção de eventos multimídia a partir de vídeos capturados, em especial a fusão de sugestões de vários aspectos do conteúdo do vídeo: objetos detectados, movimentos observados, assinaturas de áudio etc. Empregamos a pontuação de fusão, também conhecida como fusão tardia, e propomos um método que aprende as ponderações locais das várias pontuações do classificador base que respeitam as diferenças de desempenho decorrentes da qualida



Referências

- [1] J. S. Bridle and M. D. Brown (1974), "An Experimental Automatic Word-Recognition System", JSRU Report No. 1003, Joint Speech Research Unit, Ruislip, England.
- [2] Stevens, Stanley Smith; Volkmann; John & Newman, Edwin B. (1937). "A scale for the measurement of the psychological magnitude pitch". Journal of the Acoustical Society of America. 8 (3): 185190.
- [3] Hasan R., Jamil M., Rabbani G., Rahman S. (2004), “SPEAKER IDENTIFICATION USING MEL FREQUENCY CEPSTRAL COEFFICIENTS”, 3rd International Conference on Electrical & Computer Engineering 2004, 28-30 December 2004, Dhaka, Bangladesh.
- [4] S.B. Davis, and P. Mermelstein (1980), "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," in IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), pp. 357–366.
- [5] Hagen A., Connors D.A. & Pellm B.L.: The Analysis and Design of Architecture Systems for Speech Recognition on Modern Handheld-Computing Devices. Proceedings of the 1st IEEE/ACM/IFIP international conference on hardware/software design and system synthesis, pp. 65-70, 2003
- [6] Kumar J., Prabhakar O., Sahu N. (2014), “Comparative Analysis of Different Feature Extraction and Classifier Techniques for Speaker Identification Systems: A Review”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 1, January 2014.
- [7] Kakade M., Salunke D. (2018), “Real Time Speaker Independent Speech Recognition System”, International Journal of Innovations & Advancement in Computer Science (IJIACS) - Volume 7, Issue 3 - March 2018.
- [8] Dhonde S., Jagade S., “Feature Extraction Techniques in Speaker Recognition: A Review”, International Journal on Recent Technologies in Mechanical and Electrical Engineering (IJRMEE) ISSN: 2349-7947 Volume: 2 Issue: 5 – pg. 104 a 106.
- [9] Mehta L., Mahajan S., Dabhade A. (2013), “COMPARATIVE STUDY OF MFCC AND LPC FOR MARATHI ISOLATED WORD RECOGNITION SYSTEM”, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering - Vol. 2, Issue 6, June 2013.
- [10] Gulzar T., Singh A., Sharma S. (2014), “Comparative Analysis of LPCC, MFCC and BFCC for the Recognition of Hindi Words using Artificial Neural Networks”, International Journal of Computer Applications (0975 – 8887) Volume 101– No.12, September 2014.

Referências

- [11] Hasan R., Hussein H., Lazaridis P. et al. (2017), “Improvement of Speech Recognition Results by a Combination of Systems”, Proceedings of the 23rd International Conference on Automation & Computing, University of Huddersfield, Huddersfield, UK, 7-8 September 2017.
- [12] Roopalakshmi R., Reddy G. (2011), “A Novel Approach to Video Copy Detection Using Audio Fingerprints and PCA”, The 2nd International Conference on Ambient Systems, Networks and Technologies (ANT-2011), Procedia Computer Science 5 (2011) 149–156.
- [13] Thiruvengatanadhan R. (2018), “Music Classification using MFCC and SVM”, International Research Journal of Engineering and Technology (IRJET) - Volume: 05 - Issue: 09 - September 2018.
- [14] Jamal N., Shanta S., Mahmud F., Sha’abani M. (2017), “Automatic Speech Recognition (ASR) based Approach for Speech Therapy of Aphasic Patients: A Review”, AIP Conference Proceedings - Published by the American Institute of Physics.
- [15] Wang K., An N., Li B., Zhang Y., Li L. (2015), “Speech Emotion Recognition Using Fourier Parameters”, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, 6(1):69–75, Jan.
- [16] Dave N. (2013), “Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition”, INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY, Volume 1, Issue VI, July 2013.
- [17] Singh V., Jain V., Tripathi N. (2014), “A Comparative Study on Feature Extraction Techniques for Language Identification”, International Journal of Engineering Research and General Science Volume 2, Issue 3, April-May 2014.
- [18] Baraldi L., Grana C., Cucchiara R.(2017), “Neural Story: an Interactive Multimedia System for Video Indexing and Re-use”, 15th International Workshop on Content-Based Multimedia, Italy, 19-21 June 2017.
- [19] McCloskey S., Liu J. (2014), “Metadata-weighted Score Fusion for Multimedia Event Detection”, 2014 Canadian Conference on Computer and Robot Vision, Montreal, QC, Canada, 6-9 May 2014.
- [20] Liu M., Wan C., Wang L. (2004), “Intelligent Content-Based Audio Classification and Retrieval for Web Applications”, Computational Web Intelligence, pág. 257-281 (2004)