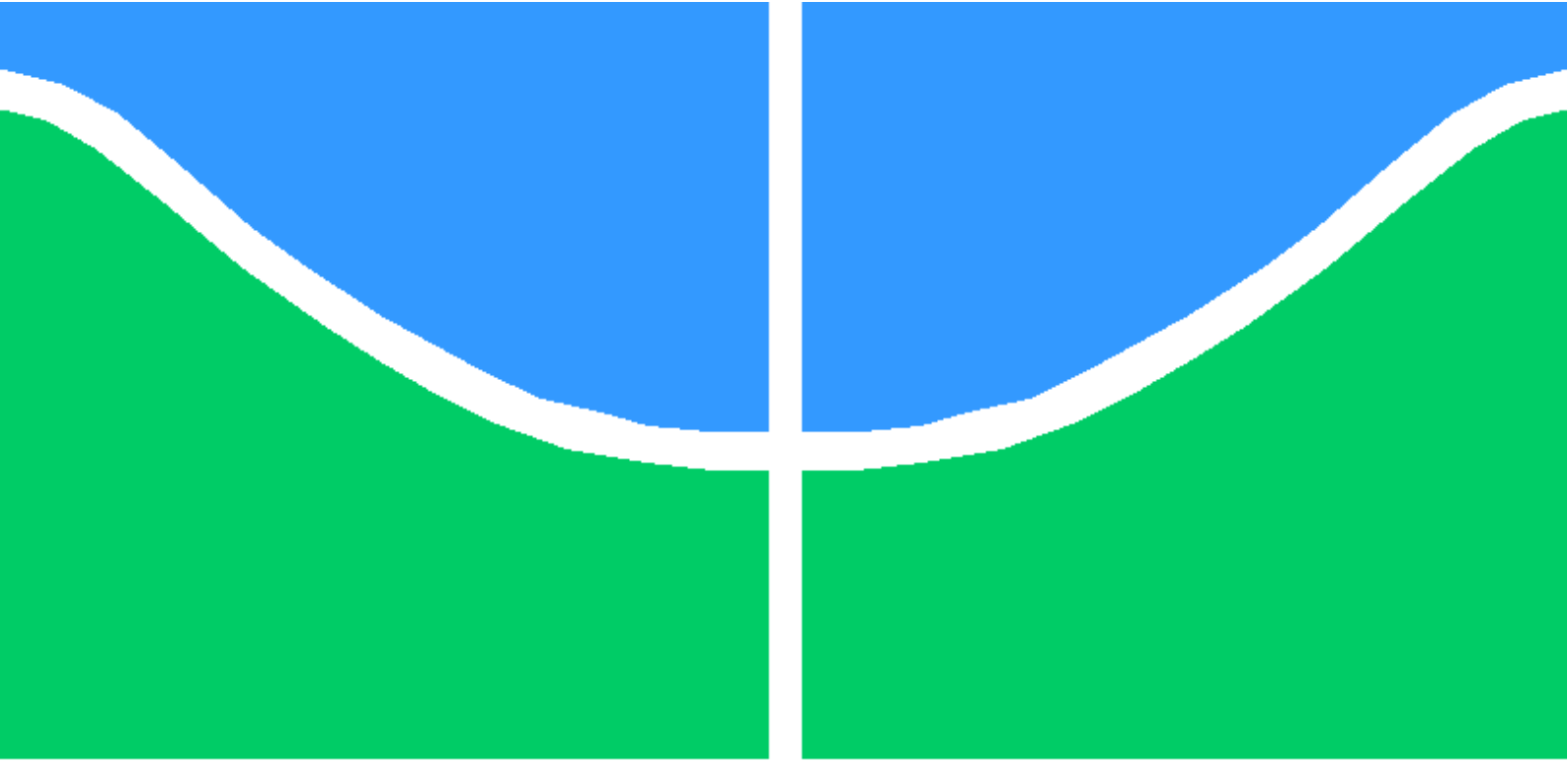

Trabalho de Graduação

Identificação Forense de Múltiplos Locutores baseada em ICA convolutiva usando coeficientes cepstrais de frequência Mel e modelo de mistura gaussiana

Matheus Almeida Silveira

Brasília, dezembro de 2013

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASILIA
Faculdade de Tecnologia

Trabalho de Graduação

Identificação Forense de Múltiplos Locutores baseada em ICA convolutiva usando coeficientes cepstrais de frequência Mel e modelo de mistura gaussiana

Matheus Almeida Silveira

Relatório submetido ao Departamento de Engenharia
Elétrica como requisito parcial para obtenção
do grau de Graduaçao em Engenharia de redes de comunicaçao

Banca Examinadora

Prof. João Paulo Carvalho Lustosa da Costa, ENE/UnB _____
Orientador

Prof. Ricardo Zelenovsky, ENE/UnB _____
Examinador interno

M. Sc. Marco A. M. Marinho, ENE/UnB _____
Examinador interno

Agradecimentos

Agradeço primordialmente meu orientador Dr. João Paulo que sempre acreditou em mim e me motivou. Também agradeço o estudante Florian Denk, da alemanha, com quem trabalhei junto um dos capítulo deste documento, e o Marco Marinho que me ajudou na confecção deste trabalho. Agradeço também a todos meus amigos que me apoioram e em especial minha companheira Caroline Cordova.

Matheus Almeida Silveira

RESUMO

Técnicas de identificação automática de locutores são amplamente utilizadas em aplicações forenses. Porém a sua precisão cai severamente quando a voz do locutor de interesse está imersa em uma gravação que contenha mais de uma voz. Esta é uma situação comum de investigações onde a voz dos alvos são obtidos por meio de gravações de escuta ambiente. Em aplicações forenses onde temos microfones escondidos, ruídos interferentes em gravações são comuns e degradam severamente o desempenho das técnicas de identificação de locutores. Nesse trabalho, é proposto um método para atenuar esse problema separando espacialmente a voz de cada pessoa usando uma técnia de Separação Cega de Sinais chamada Análise de Componente Independente , e então aplicando nos sinais de voz separados um sistema de identificação de voz baseado em Coeficientes Cepstrais de Frequência Mel e Modelos de Mistura Gaussianas. Para identificar mais que um locutor, o método proposto tem mais acurácia que os presentes na literatura.

ABSTRACT

Automatic speaker identification techniques are widely used nowadays in forensic applications, but its accuracy harshly degrades when the voice of the speaker of interest is immersed in a recording containing more than one voices, common situation of investigations where the targets voice are obtained through ambient recordings. In forensic applications where microphones are hidden, such interferent sound sources in recordings are common and they severely degrade the performance of speaker identification techniques. In this paper, we propose a method to mitigate this problem by spatially separating the voice of each speaker using a Blind Source Separation technique called Convolutive Independent Component Analysis, and then applying the separated speech signals to a speaker identification system based on Mel Frequency Cepstral Coefficients and Gaussian Mixture Models. For identifying more than one speaker, the proposed system has a better accuracy than the state-of-the-art solutions.

SUMARIO

LISTA DE FIGURAS

LISTA DE TABELAS

Capítulo 1

Introdução

1.1 Contextualização

Identificação de locutor lida com o problema de identificar a pessoa que está falando ao se analisar um sinal de voz. Contudo, verificação de locutor nos dá a aceitação ou rejeição da identidade de uma pessoa dado um sinal de voz em análise, Figura 1.1. Este trabalho lida com o problema de identificação de locutor com texto independente, ou seja, para fim da comparação não é necessário que as palavras sejam as mesmas nos dois sinais de voz, em cenários onde há mais de uma pessoa falando, o que gera uma adversidade onde uma, ou mais, pessoa(s) não pode(m) ser identificada(s), Figura 1.2. Um exemplo dessa aplicação é analisar áudios gravados de escuta ambiente forense, que é uma técnica de investigação muito importante e utilizada, especialmente quando o suspeito é bastante cuidadoso e não passa informações importante, utilizada por telefone ou e-mail.

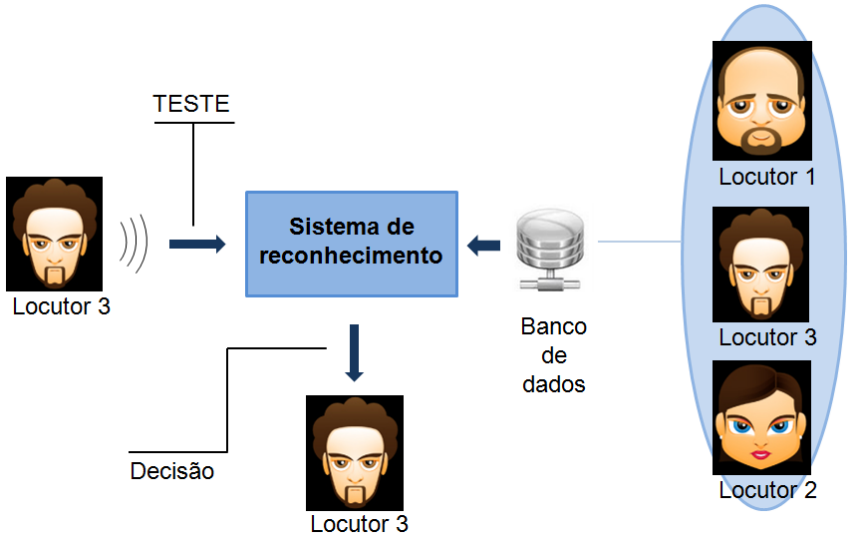

Figura 1.1: Exemplo de realização de identificação de locutor. Um locutor profere algumas palavras, o sistema extrai suas características a partir desse trecho de fala, compara com características modeladas de outros locutores em um banco de dados, estima a compatibilidade mais provável e retorna sua decisão.

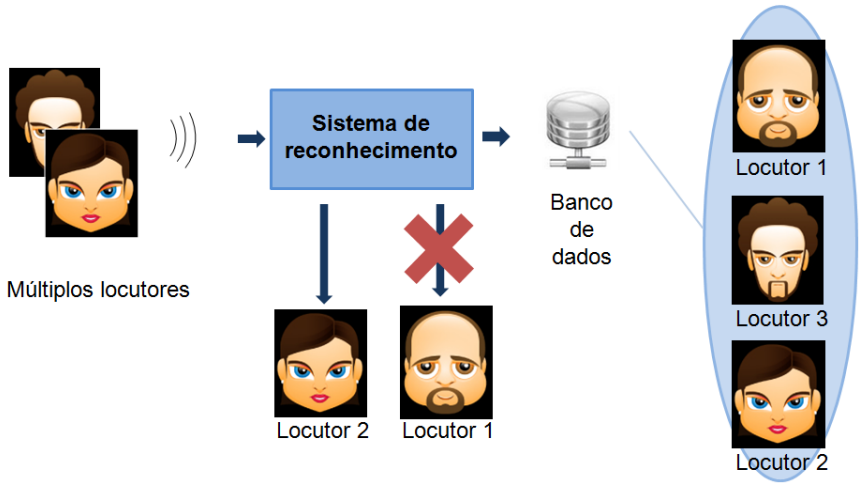

Figura 1.2: Exemplo do problema que é abordado neste trabalho. Quando se tem um áudio contendo a voz de mais de uma pessoa, o sistema erra sua estimativa pelo menos para um dos locutores.

1.2 Definição do problema

A identificação pode ser feita extraindo a informação do áudio gravado e comparando com a informação do locutor previamente armazenado em um banco de dados. Há várias abordagens sobre o problema de identificação de locutor [1, 2, 3]. Como em vários sistemas de reconhecimento por biometria, a identificação de locutor pode ser dividida em duas etapas: uma etapa de treinamento e outra de teste. Geralmente a etapa de teste é dividida em outras duas etapas: a extração de características e a comparação de características. A primeira lida com a extração das características do sinal de fala do locutor. A segunda lida com o problema de comparar essas características extraídas com as contidas em um banco de dados para determinar a entrada que melhor se combina com as características extraídas da gravação, assim, estimando a identidade do locutor.

O problema de extração de características pode ser trabalhado por várias abordagens, tais como LPC (*Linear Predictive Coding*), LDB (*Local Discriminant Bases*) [10], e MFCC (*Mel Frequency Cepstral Coefficient*) [10, 6]. O MFCC, que é utilizado neste trabalho, atualmente é a caracterização mais usada para falas e identificação de locutor.

Para lidar com o problema de comparação de características há várias abordagens disponíveis [3], tais como: VQ (*Vector Quantization*), SVM (*Support Vector Machine*), HMM (*Hidden Markov Model*) e GMM (*Gaussian Mixture Model*) [6, 2, 4, 14]. Dentre essas abordagens o GMM se tornou o padrão utilizado para sistemas de identificação de locutor, logo esta abordagem também será explorada neste trabalho.

1.3 Objetivos do projeto

Neste trabalho será apresentada uma técnica que recebe sinais de voz de duas ou mais pessoas falando ao mesmo tempo. Estes áudios serão gravados utilizando um arranjo de microfones. Esta técnica separa as fontes de áudio (locutores) usando um método de separação cega de fontes chamada ICA (*Independent Component Analysis*)convolutivo [7, 12], e submete os arquivos contendo as fontes de áudio separadas a um sistema de identificaçao de locutor baseado em MFCC e o GMM, a Figura 1.3 ilustra esta abordagem. Como será mostrado neste trabalho, a técnica de ICA combinada com o tradicional MFCC-GMM provê melhores resultados que as abordagens típicas de MFCC-GMM. De acordo com [14], o aprimoramento da gravação de áudio para melhorar a

inteligibilidade e audibilidade de baixos níveis de RSR (Razão sinal ruído) é uma das principais preocupações quando se trabalha com áudios forenses. RSR's baixas são muito comuns em investigações criminais porque os microfones ficam escondidos. Tendo isso em mente, foi proposta uma segunda abordagem onde uma técnica cancelamento de ruído é aplicada para se obter bons resultados mesmo em regimes de baixo RSR, Figura 1.4.

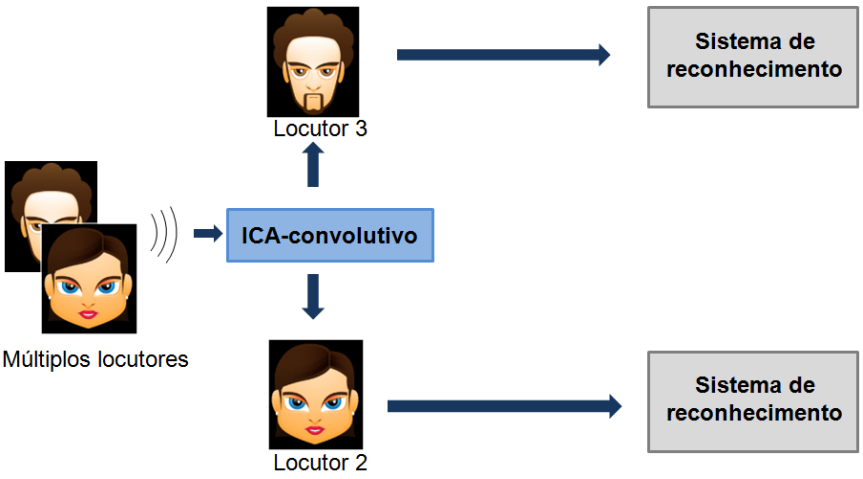

Figura 1.3: Abordagem proposta para aprimorar a acurácia do sistema de reconhecimento automático de locutor separando os sinais de cada fonte utilizando o ICA-concolutivo

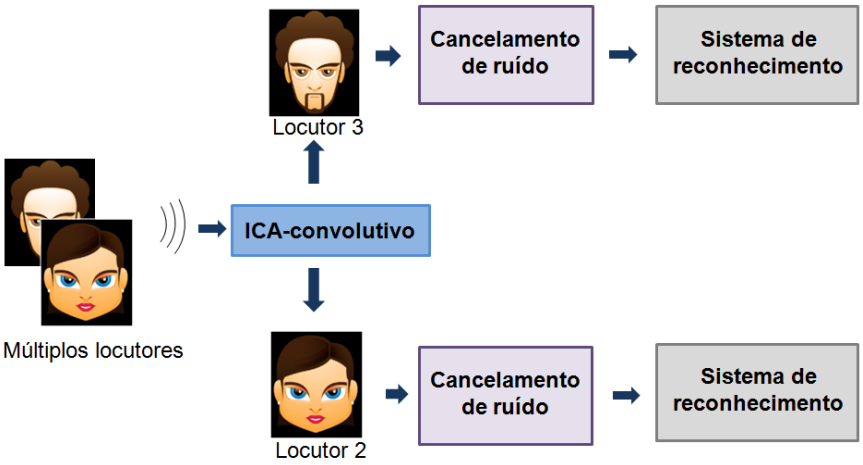

Figura 1.4: Abordagem proposta para aprimorar a acurácia do sistema de reconhecimento automático de locutor separando os sinais de cada fonte utilizando o ICA-concolutivo e uma técnica de cancelamento de ruído

1.4 Apresentação do manuscrito

O trabalho apresentado está organizado como segue: no Capítulo 2 será apresentado alguns conceitos básicos de estatística que serão abordados ao decorrer deste trabalho, no Capítulo 3 é apresentado a tradicional identificação de locutor via MFCC-GMM, explicando como é implementada as etapas de extração e comparação de características do sistema. Em seguida, no Capítulo 4 é acoplado a abordagem de ICA convolutivo ao conteúdo apresentado no Capítulo 3 para alcançar a solução do problema. Capítulo 5 apresenta os resultados das simulações com áudios sem ruído e uma análise envolvendo ruído branco. No Capítulo 6 será apresentado um esquema para cancelamento de ruído, e no Capítulo 7 haverá simulações comparando diferentes abordagens para o sistema de reconhecimento de locutor. E por fim, no capítulo 8 temos uma conclusão do estudo.

Capítulo 2

Conceitos Estatísticos

2.1 Variável Aleatória

Antes de apresentar qualquer processo que levou ao desenvolvimento deste trabalho é necessário antes definir alguns conceitos estatísticos que servirão como base para sustentar os métodos que serão apresentados posteriormente. Como o principal objeto de manipulação deste trabalho é o sinal de voz, os conceitos a seguir serão tratados considerando-se apenas informações coletadas com uma dimensão, o tempo. Um sinal de voz digitalizado pode ser representado como uma variável aleatória $X = [x_1, x_2, x_3, ..., x_n]$ de tamanho n.

Uma das principais informações para inferência estatística que podemos estimar de um sinal de voz descrito acima é a sua média, denotada por $\hat{\mu}$, que nos diz em torno de qual valor os dados coletados estarão situados.

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.1}$$

Enquanto a média é uma tendência central dos valores da variável aleatória, a variância mede o nível de dispersão desses valores da média. Intuitivamente estamos pensando na distância quadrada média.

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} |x_i - \mu|^2 \tag{2.2}$$

Note que a variância é representada por σ^2, onde σ é o desvio padrão, uma representação estatística que remete à distância entre dois pontos. Algumas pessoas estimam a variância usando $n-1$ como divisor, mas para grandes quantidades de amostras, como é o caso de um sinal de voz digitalizado com mais de 30 segundos de duração, esse fator não possui relevância.

Considerando agora duas variáveis aleatórias, $X = [x_1, x_2, x_3, ..., x_n]$ e $Y = [y_1, y_2, y_3, ..., y_n]$, podemos, a partir do conceito de variância, estimar a covariância entre os sinais, que nos diz o quanto que a ocorrência de um evento influencia no outro

$$\operatorname{cov}(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)^*}{n}, \tag{2.3}$$

onde μ_x representa a média de X, e μ_y representa a média de Y.

Pelo fato da covariância não ser normalizada, normalmente calcula-se o valor da correlação entre os sinais, também chamado de coeficiente de correlação, que visa normalizar o problema de escala da covariância. Seja o coeficiente de correlação ρ_{xy}, onde $-1 < \rho_{xy} < 1$, então:

$$\rho_{xy} = \frac{\operatorname{cov}(X, Y)}{\sigma_x \sigma_y} \tag{2.4}$$

onde σ_x representa o desvio padrão de X, e σ_y representa o desvio padrão de Y. Valores de $|\rho_{xy}|$ próximos a 1 indicam um forte relação entre os sinais, enquanto um valor próximo a zero indicam uma fraca relação entre os sinais. Um dos tópicos abordados neste trabalho é a independência entre sinais, o que, neste caso. é indicado por uma covariância, ou correlação, igual a zero.

2.2 Momento

Os Momentos de uma distribuição de probabilidade são importantes medidas para caracterizar sua dispersão e assimetrias. O momento natural de ordem t de uma variável aleatória $X = [x_1, x_2, \ldots, x_n]$ é definido por

$$m_t = \frac{(\Sigma\,(x_i))^t}{n},$$

onde Σ representa o somatório de $i = 1$ até n. Para $t = 1$, fica evidente a relação do momento de primeira ordem m_1 com a estimativa da média $\hat{\mu}$ em (2.1). Uma forma mais interessante de se observar a dispersão dos valores em uma distribuição é através da definição de Momento Central, em que o valor de média desta distribuição é usado como referência para a dispersão dos dados, definido por

$$m_t = \frac{(\Sigma\,(x_i - \mu))^t}{n},$$

onde μ é o valor de média da distribuição. Para $t = 2$, fica evidente a relação entre o Momento Central de ordem 2, m_2, com a estimativa de variância visto em (2.2).
Na seção seguinte será mostrado a importância do Momento de ordem 4.

2.3 Função Massa de Probabilidade

A função massa de probabilidade (FMP), de uma variável aleatória X é uma função que nos diz a probabilidade de um determinado evento ocorrer. A notação típica de uma FMP é $P(X = x)$.

Para ficar mais clara a ideia considere um experimento onde deseja-se encontrar a FMP do número de vezes que se rola um dado honesto até se obter o número seis.

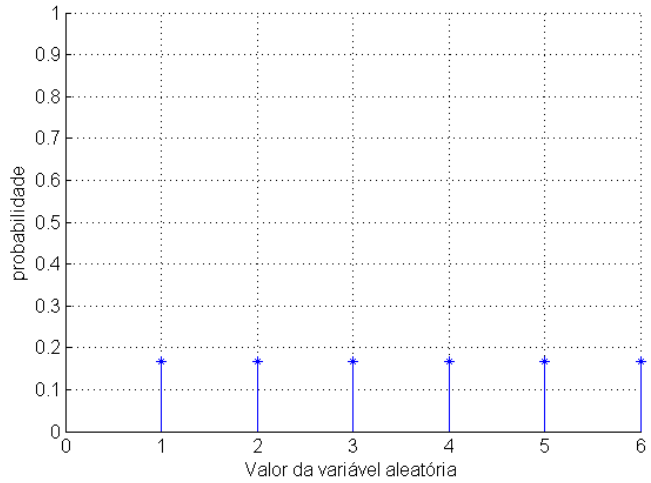

Figura 2.1: Massa de probabilidade da variável aleatória dado

Sendo X a variável aleatória, vide Figura 2.1, teremos $P(X = 1) = 1/6$ que corresponde a probabilidade de ocorrer seis na primeira jogada, por seguinte $P(X = 2) = \frac{5}{36}$ em que a primeira jogada necessariamente não foi um seis e a segunda jogada sim, e etc.
Com isso podemos definir a FMP do experimento como sendo $P(X = x) = \frac{1}{6}\left(\frac{5}{6}\right)^{x-1}$, representado pelo gráfico abaixo. Na Figura 2.2 é mostrado a probabilidade para até 7 jogadas, mas de fato, esse valor pode se estender ao infinito.

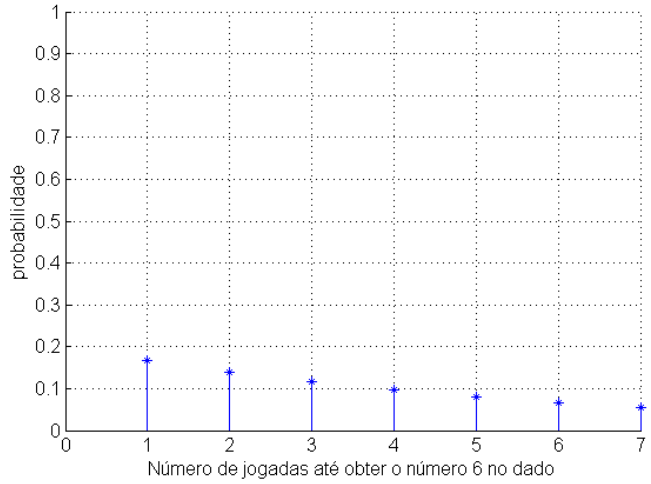

Figura 2.2: Massa de probabilidade de uma variável aleatória que determina número de jogadas até aparecer o número seis

2.3.1 Curva Gaussiana

Em vários experimentos probabilísticos há variáveis aleatórias que costumam apresentar o mesmo comportamento, ainda que pertencendo a experimentos distintos. Estas variáveis aleatórias costumam ter seu comportamento catalogado.

Neste trabalho será bastante explorada a função densidade de probabilidade que descreve o comportamento de variáveis aleatórias com distribuição Normal, também chamada de curva gaussiana:

$$P(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \tag{2.5}$$

Esta é bastante caracterizada pelos seus dois primeiros Momentos, média e variância, e além disso este trabalho irá explorar seu quarto Momento Central para a obtenção da curtose. O coeficiente de curtose é representada por κ e pode ser obtido expandindo-se a expressão:

$$\kappa = -3 + \frac{m_4}{(m_2)^2} \tag{2.6}$$

Como foi visto anteriormente, o Momento Central de ordem 2 equivale à variância da distribuição. Substituindo m_2 por σ^2 e aplicando a definição de Momento Central em m_4, obtém-se

$$\kappa = -3 + \frac{1}{\sigma^4 n} \sum_{i=1}^{n} |x_i - \mu|^4, \tag{2.7}$$

onde μ é a média da distribuição .O termo -3 da equação, que as vezes é omitido, é utilizado para dar um resultado $\kappa = 0$ em caso de curvas gaussianas. Se o resultado for uma curtose positiva, então a curva é supergaussiana e possui formato estreito e pontudo, se o resultado for uma curtose negativa, então a curva é subgaussiana e apresenta uma forma achatada e mais larga, como mostra a Figura 2.3.

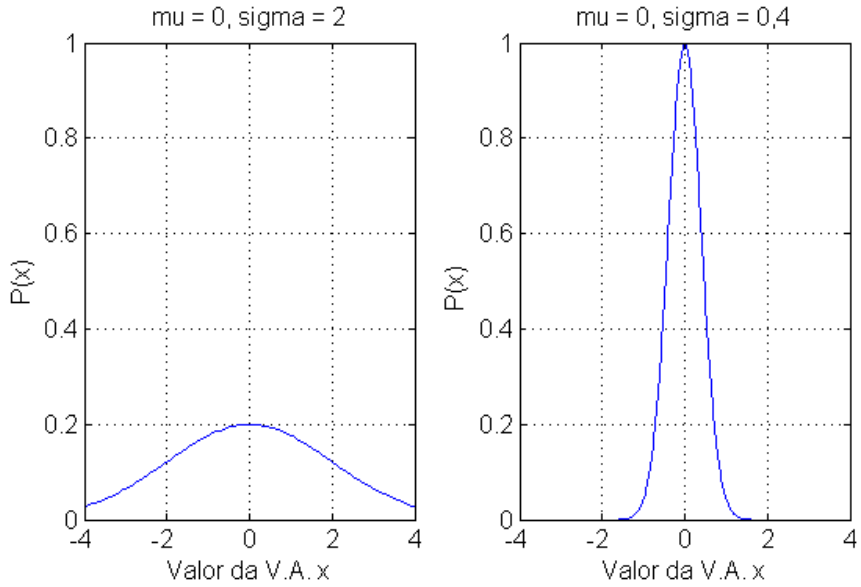

Figura 2.3: Lado esquerdo curva subgaussiana. Lado direito curva supergaussiana

Capítulo 3

Sistema de identificação de locutor

3.1 MFCC e GMM

Como foi dito, um sistema de identificação de locutor pode ser divido em duas etapas: uma etapa de treinamento e outra de teste. Na primeira etapa, como mostra a Figura 3.1 os coeficientes cepstrais de frequência Mel do sinal de fala de treinamento são computados, então a técnica de GMM é usada para determinar a soma de gaussianas que melhor descrevem o histograma formado pela distribuição de cada coeficiente.

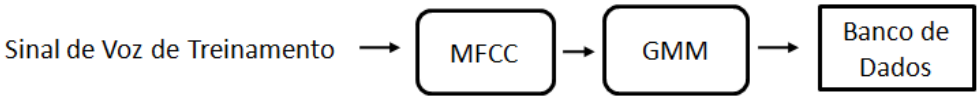

Figura 3.1: Diagrama de blocos da etapa de treinamento

Os parâmetros obtidos dessa distribuição são únicos para cada indivíduo, e é essa informação que é armazenada no banco de dados. No segundo passo da Figura 3.2 os coeficientes cepstrais extraídos do sinal de fala de teste são comparados com os parâmetros da soma de gaussianas de cada indivíduo contido no banco de dados. Essa comparação é feita analisando o log-likelihood entre dois sinais, que representa o quão verossimel estes são. No bloco de lógica de decisão, o indivíduo com os parâmetros que melhor representa os coeficientes cepstrais é indicado como o mais provável locutor daquele sinal de fala de teste.

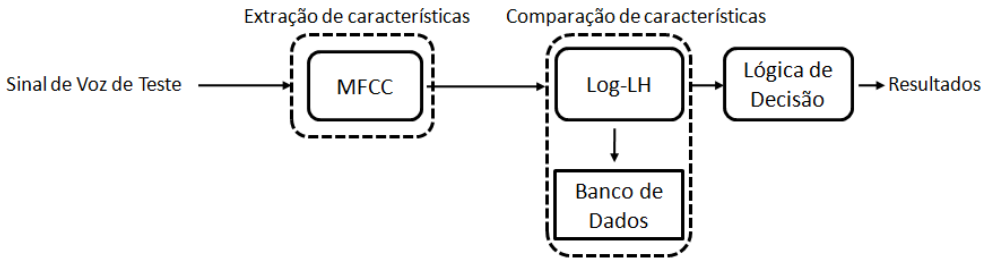

Figura 3.2: Diagrama de blocos da etapa de teste

Como mencionado no Capítulo 1, a abordagem atual do problema de identificação de locutor usa o MFCC para extração de características e o GMM para modelagem dessas características. As metodologias usuais do MFCC e GMM são descritas nas subseções seguintes.

3.1.1 Extração de características via MFCC

A extração de características é um método não reversível, ou seja, não é possível reconstruir o sinal de voz devido a perda de informação durante o processo.

Coeficientes Cepstrais de frequência Mel de um sinal de voz $s(t)$ representam a variação do trato vocal e são o resultado da transformada discreta do cosseno do logaritmo da energia do espectro de um curto-tempo de fala expressado na escala de frequência mel [4], matematicamente expresso como

$$C_n = dct(log(\mathcal{F}\{s(t)\} \otimes \mathcal{F}\{s(t)\}^*)),$$

onde $\mathcal{F}\{.\}$ representa uma mudança do domínio do tempo para frequência, e $dct(.)$ a transformada discreta do cosseno. Todos esses processos são necessários para isolar o melhor possível o som produzido no trato vocal, pois o sinal de áudio que ouvimos quando alguém fala é uma convolução entre o som produzido no trato vocal com o som produzido na boca [19], e pode ser expresso como $s(t) = e(t) * g(t)$, onde $s(t)$ representa o sinal de voz que escutamos, $e(t)$ o som chamado de excitação, ou fricativo, que é produzido na boca, e $g(t)$ o som produzido na glótis, ou trato vocal. Estes processos são aplicados pelo fato do som do trato vocal apresentar uma lenta variação de magnitude no domínio da frequência em relação ao som produzido na boca. Para ilustrar, na Figura 3.3 $E(f)$, representa um sinal fricativo com variações bruscas de energia no domínio da frequência e $G(f)$ um sinal produzido pela glótis com variação lenta de energia entre as frequências.

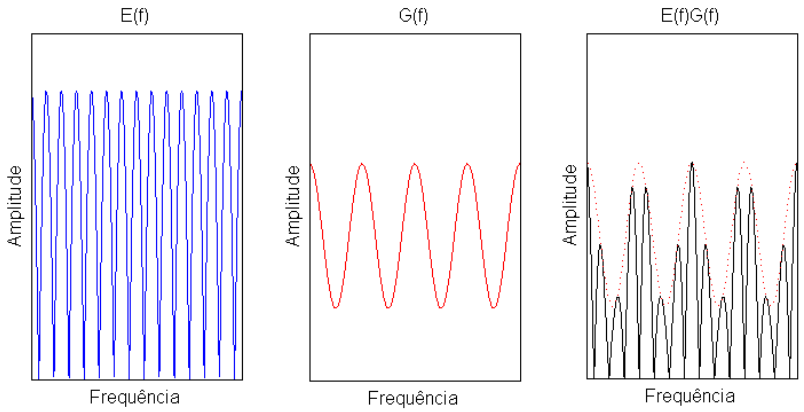

Figura 3.3: Complexidade do sinal de voz no domínio da frequência

Como os sinais são convoluídos no domínio do tempo, tem-se uma simples multiplicação dos sinais no domínio da frequência. Aplicando uma função logarítmica o problema é simplificado mais ainda para apenas uma soma de sinais, Figura 3.4. Com isto aplica-se a transformada discreta do cosseno, que será mostrada mais adiante, devido à sua característica de concentrar as energias dos sinais de baixa frequência nos primeiros índices, assim, a informação da glótis estará primordialmente no início do sinal resultante podendo-se descartar o restante da informação que representa apenas sons fricativos, estes não carregam informação para biometria, Figura 3.5

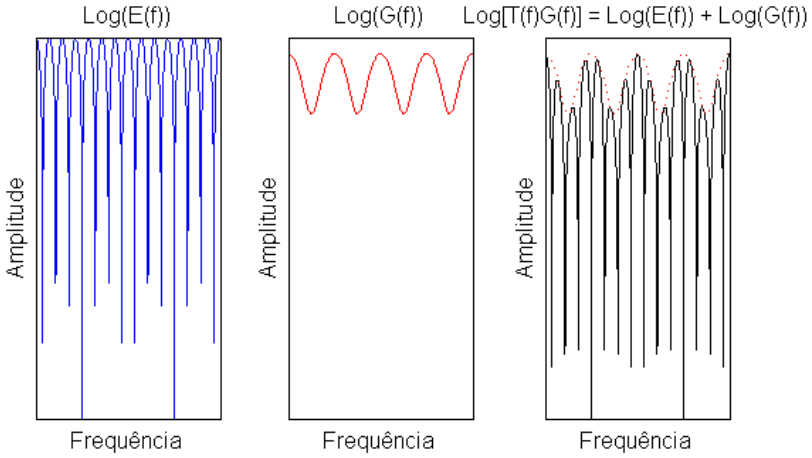

Figura 3.4: Complexidade do sinal de voz simplificado com logaritmo

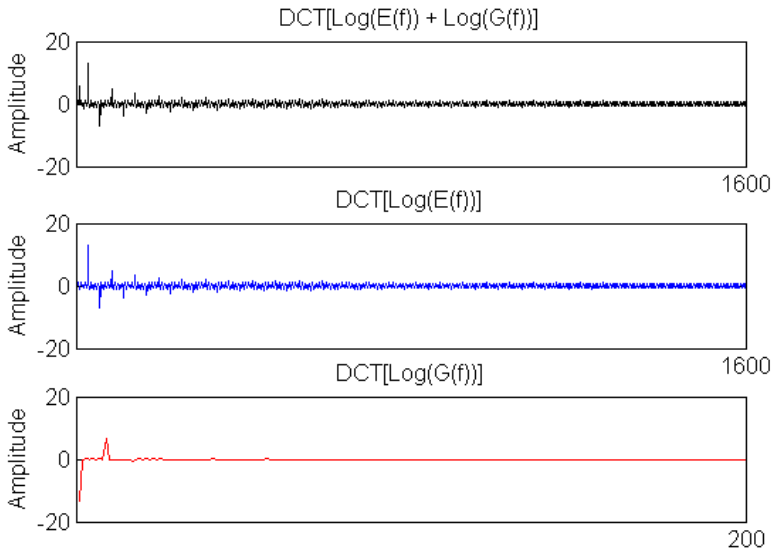

Figura 3.5: Cepstrum do sinal de voz, fricativo, e glótis

O gráfico dos coeficientes ceptrais da glótis, na Fgura 3.5, mostram apenas as 200 primeiras amostras, isto foi feito para uma melhor visualização da variação de amplitude no início do sinal, de fato suas amostras continuam nulas até o índice 1600. Os sinais acima foram foram simulados com o intuito de validar os processos de extração de características,na realidade os sinais produzidos são mais detalhados mas a característica de variação de frequência continua a mesma.

A escala Mel é utilizada para criar o banco de filtros triangulares que simula a percepção auditiva dos humanos. Assim é possível reduzir a informação processada mantendo as características da voz da pessoa. Figura 3.6 Mostra um diagrama de blocos da técnica do MFCC.

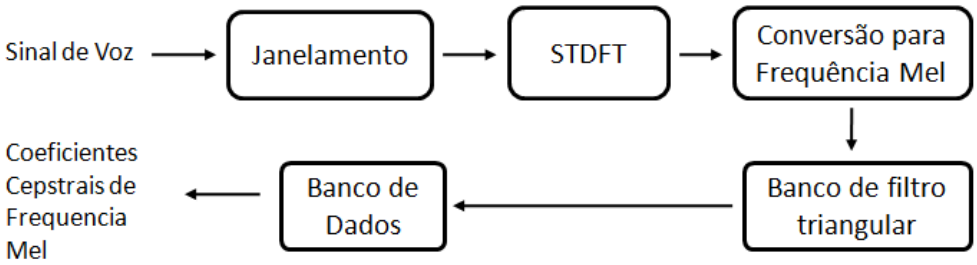

Figura 3.6: Diagrama de blocos da técnica do MFCC

Para obter os coeficientes cepstrais, primeiramente converte-se o sinal sonoro em sinal elétrico por meio de um microfone. Utilizando um conversor analógico/digital digitaliza-se o sinal elétrico. Uma vez que o sinal está digitalizado fica fácil implementar os processos para a extração das características. O sinal de voz está constantemente mudando, mas em um curto período de tempo

ele não muda muito e podemos considerá-lo estatisticamente estacionário [2], portanto, antes de computar seu espectro é necessário segmentá-lo em várias janelas com tamanho N amostras de duração aproximada de 30 milissegundos, se a janela é muito curta não há amostras suficientes para obtermos uma boa resolução do espectro e se a janela é muito grande o sinal muda muito ao longo da janela. Além disso, há sobreposição de 50% entre as janelas, Figura 3.7, que será explicado logo adiante.

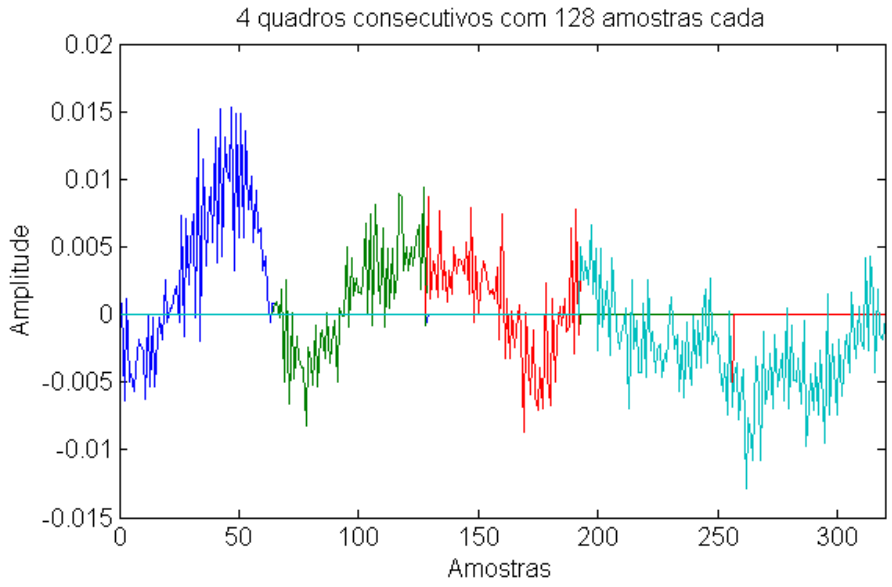

Figura 3.7: Segmentação do sinal de voz

Em seguida cada janela passa pela função (3.1) para amenizar o problema de vazamento espectral e ter uma representação mais precisa no domínio da frequência. O vazamento espectral ocorre devido à descontinuidade entre o valor inicial e o valor final em cada janela. Na Figura 3.8 pode ser visto o efeito causado por esta descontinuidade, no gráfico **a** temos uma senoide simples de 50 Hertz e em **b** seu respectivo espectro, em **c** temos gráfico do sinal **a** truncado de forma que o valor inicial não coincida com o final, o espetro resultante é mostrado em **d** que comparado a **b** houve um alargamento do pulso, computou-se energia em componentes de frequência que não possuem energia no sinal original. Este problema é parcialmente contornado aplicando uma função Hamming (3.1) na no sinal janelado, assim forçando os valores extremos a um valor comum, gráfico **e** da Figura 3.8, no gráfico **f** tem-se o espectro de **e**, note que apesar da perda de magnitude o alargamento do pulso foi suprimido.

$$U(n) = 0.54 - 0.46 \cos\left(2\frac{\pi n}{N-1}\right), n = [0, 2, \ldots, N-1] \tag{3.1}$$

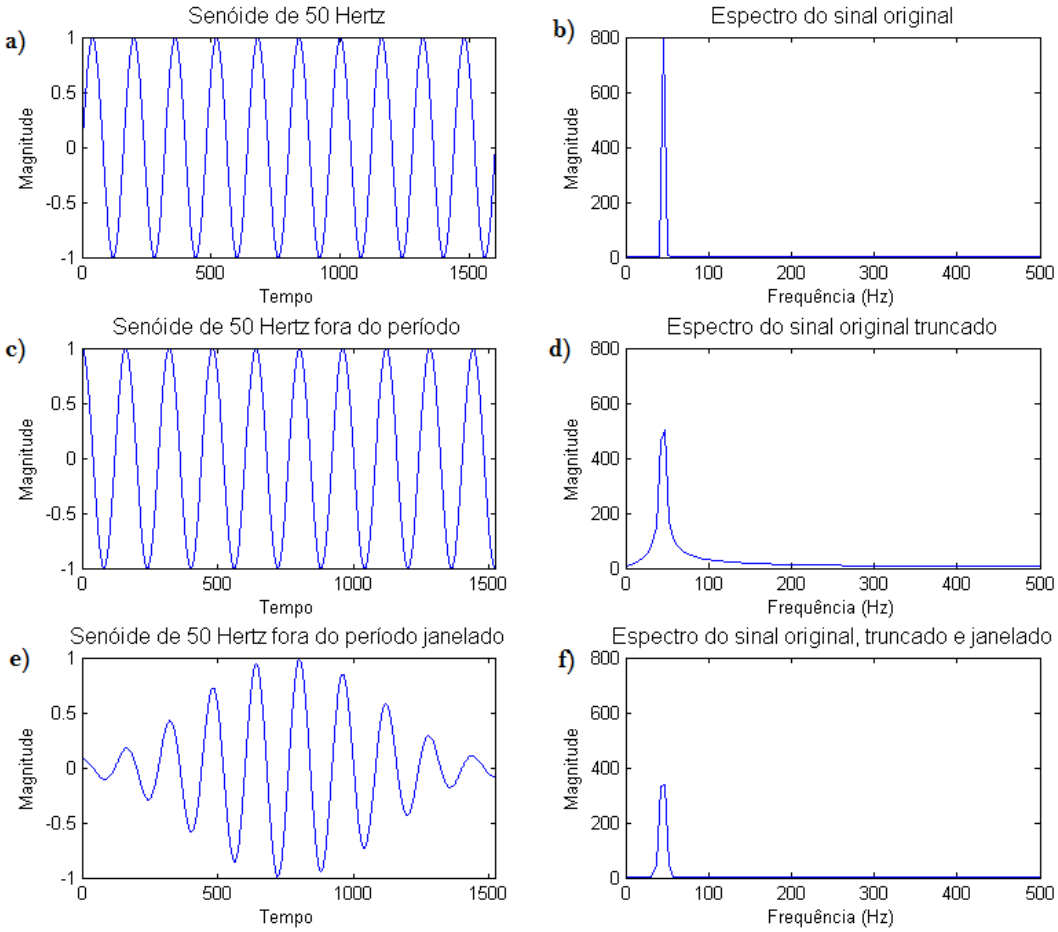

Figura 3.8: Demonstração da amenização do problema de vazamento espectral utilizando funçao de Hamming

A sobreposição de 50% das janelas minimiza a perda de informação causada pelo janelamento, que pode ser visto significativamente nos extremos do sinal do gráfico **e** na Figura 3.8. Após a aplicação da função (3.1) as janelas dos sinal segmentado devem apresentar um aspecto como o da Figura 3.9.

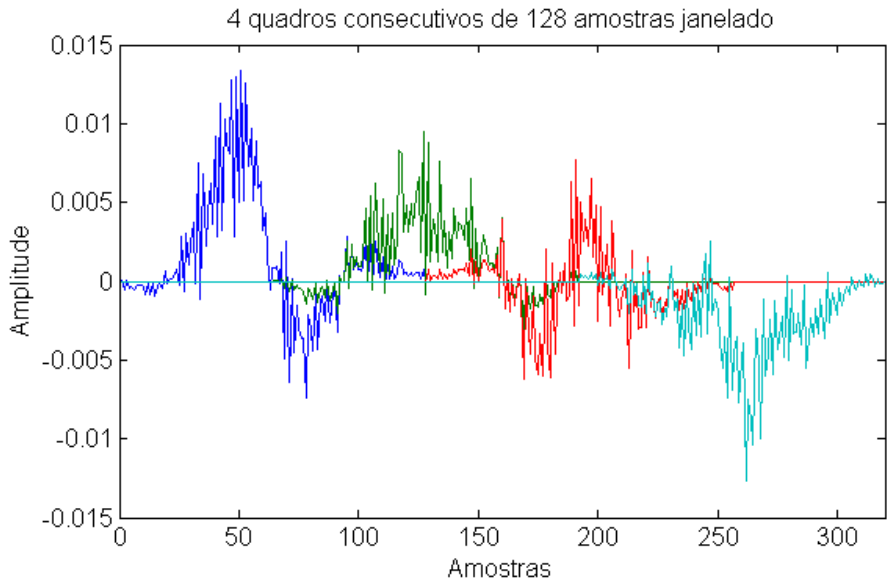

Figura 3.9: Janelamento dos quadros

Após esse processo, as janelas são organizadas em uma matriz $J_{nj,tj}$ com número de linhas nj igual a quantidade de janelas extraídas do trecho de voz, e número de colunas tj igual ao tamanho de uma janela. Sendo assim, cada linha desta matriz contém um quadro que através da STDFT, sigla em inglês para Transformada Discreta de Fourier de Curto-Tempo, terá seu espectro computado usando a expressão da DFT 3.2

$$S_{nj}(k) = \sum_{n=1}^{N} s_{nj}(m)\, e^{-2\frac{\pi jm}{N}}, \tag{3.2}$$

onde N é tamanho da janela e m a m-ésima amostra de s_{nj} . Estimado o espectro de potência de cada janela, computa-se o seu periodograma que é dado por:

$$P_{nj}(k) = \frac{(|S_{nj}(k)|)^2}{N} \tag{3.3}$$

Deve-se obter algo semelhante à Figura 3.10, onde cada cor é um trecho do sinal de voz janelado.

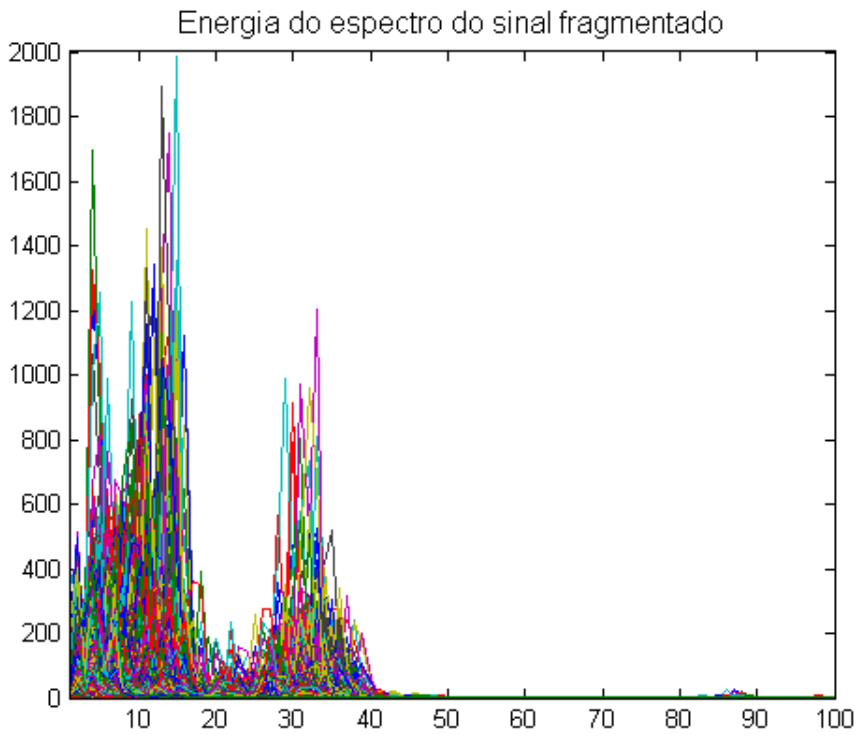

Figura 3.10: Periodograma do sinal de voz segmentado

O próximo passo é criar um banco de filtros na escala hertz que tem espaçamento igual entre cada filtro na escala Mel [5], e filtrar a maior parte da informação útil para se obter as características do sinal de voz. A Figura 3.11 mostra a curva de conversão entre Hertz e Mel baseado na equação (3.4).

$$Mel\,(f) = 2595\,\log\left(1 + \frac{1}{700}\,f\right) \tag{3.4}$$

Para um sistema automático de reconhecimento de locutor use-se um banco de filtros com aproximadamente 26 filtros, mas para simplificar o entendimento será mostrado a seguir como é construído um banco de filtros com 10 filtros, projetado para um trecho de voz de 25 ms amostrado a 16000 Hz. O primeiro passo é definir a frequência do primeiro e do último ponto, como o sinal foi a amostrado a 16000 Hz usamos 8000 kHz como a frequência do último ponto, para a frequência do primeiro ponto podemos escolher 300 Hz, que é aproximadamente quando o sinal de voz começa a apresentar valores significantes de magnitude. Definido a frequência dos pontos extremos, converte-se os valores para a escala *mel* usando-se a (3.4), então obtém-se 401,25 Mel de 300 Hz, e 2834,99 Mel de 8000 Hz, gráfico **a** da Figura 3.12. Como serão usados 10 filtros, precisa-se marcar 10 pontos

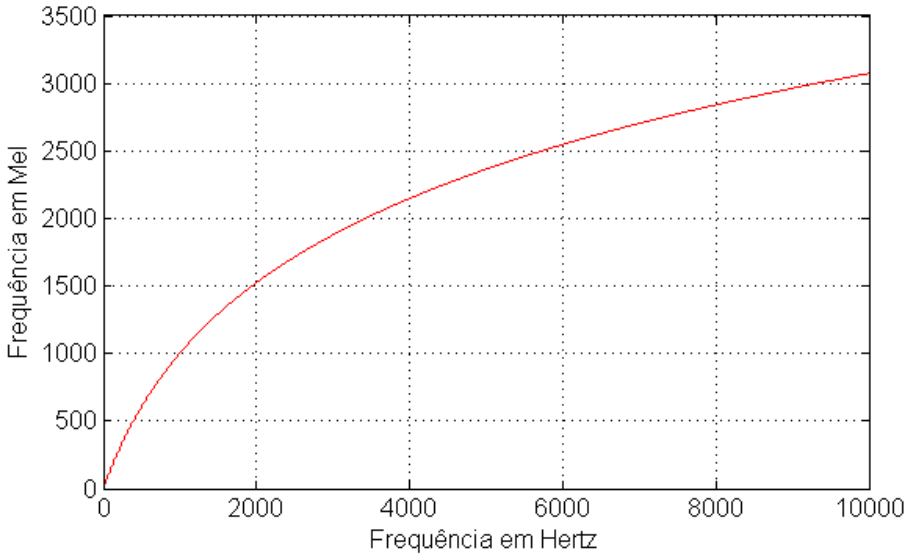

Figura 3.11: Curva de conversão de escala de frequência entre Hertz e Mel

igualmente espaçados entre os valores extremos totalizando 12 pontos, o que nos dá as seguintes frequências:

$mel(f)$ = [401,25 622,50 843,75 1065,00 1286,25 1507,50 2839,74 1949,99 2171,24 2392,49 2613,74 2834,99], gráfico **b** da Figura 3.12.

Os 10 pontos marcados tornam-se o centro de cada filtro triangular, gráfico **a** da figura 3.13. Agora converte-se os valores para a escala Hertz obtendo-se:

$hertz(f)$ = [300 517,33 781,90 1103,97 1496,04 1973,32 2554,33 3261,62 4122,63 5170,76 6446,70 8000] como as frequências de cada ponto. O banco de filtros construído pode ser visto no gráfico **b** da Figura 3.13.

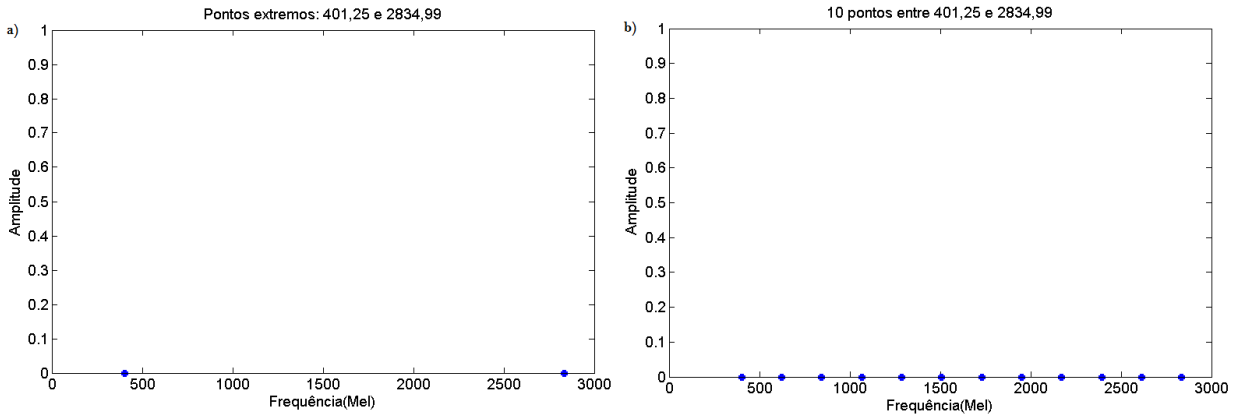

Figura 3.12: Marcação de pontos para criar banco de filtro triangulares

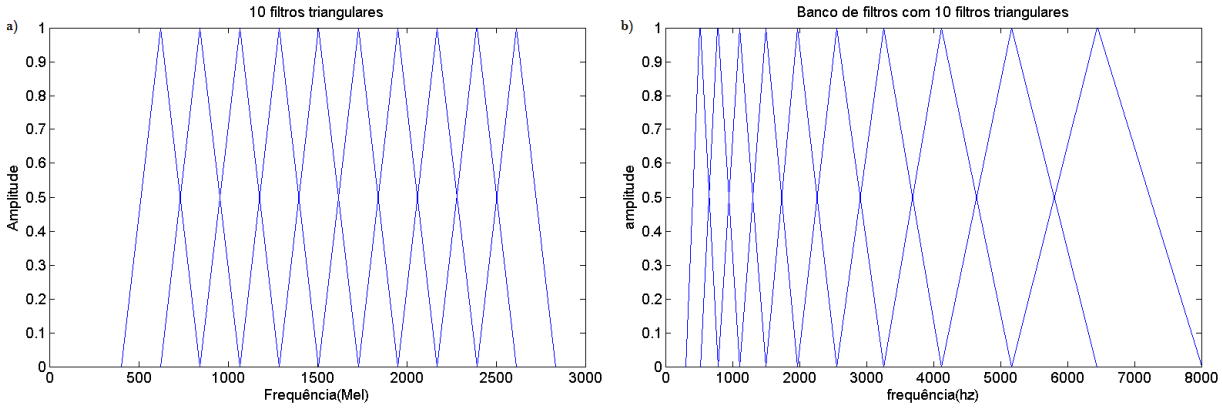

Figura 3.13: Banco de filtros triangulares

Aplica-se então, o banco de filtros em cada linha da matriz $J_{m,n}$, e para cada filtro computa-se o logaritmo da energia neste contida. E como último passo, para se obter os coeficientes cepstrais, o logaritmo da energia do periodograma é transformado de volta para o domínio do tempo utilizando a transformada discreta do cosseno (3.5). A vantagem de se usar esta transformada é que os coeficientes resultantes sempre serão valores reais e a maior parte da energia se concentra nos primeiros coeficientes, o que torna seu manuseio e armazenamento mais simples.

$$C_n = \sum_{t=0}^{tj} log(P_{nj}(k)) cos\left[(2t+1)\frac{\pi n}{2tj}\right] \quad (3.5)$$

O resultado é uma série de coeficietnes cepstrais que caracterizam a fonte de cada som, no caso da voz o trato vocal, como mostra a Figura 3.14. De fato, o histograma gerado por cada índice do cepstrum é caracterizado como um coeficiente cepstral, Figura 3.15. Na seção seguinte veremos que o GMM é utilizado para modelar esses histogramas.

Apesar de ter sido utilizado um banco de filtros triangulares, há vários bancos de filtros diferentes. Particularmente o banco de filtro gammatone, filtros com curvas de distribuição gamma, tem mostrado uma assemelhação ao sistema auditivo humano [10, 11].

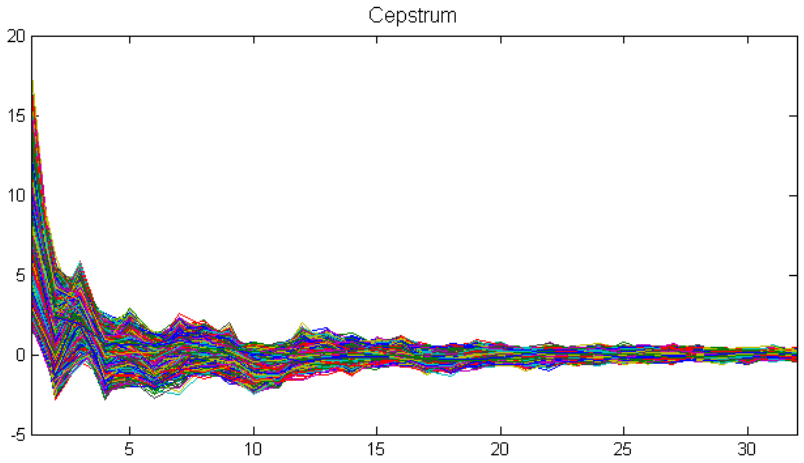

Figura 3.14: Cepstrum de um fragmento de fala

3.1.2 GMM e comparação de características

O GMM é uma função densidade de probabilidade paramétrica representada por uma soma ponderada de M componentes gaussianas [8], Figura 3.16. O GMM geralmente é usado como um modelo paramétrico de distribuições de probabilidade arbitrárias. Os parâmetros do GMM dos dados de treinamento podem ser obtidos com o algoritmo EM (Maximizaçao da Esperança)

O modelo de misturas gaussianas (GMM) é uma soma ponderada de M gaussianas dada por

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g\left(x|\mu_i, \Sigma_i\right), \tag{3.6}$$

onde x é um vetor de valores contínuo D-dimensional que representa o histograma dos coeficientes cepstrais, $w_i, i = 1, ..., M$ são os pesos de cada componente gaussiana e $g\left(x|\mu_i, \Sigma_i\right), i = 1, ..., M$, são as componentes gaussianas. Cada componente é uma função gaussiana D-variável da forma

$$g(x|\mu_i, \Sigma_i) = \frac{1}{|\Sigma_i|^{1/2}(2\pi)^{D/2}} e^{-\frac{1}{2}(x-\mu_i)'(x-\mu_i)\,\Sigma_i^{-1}}, \tag{3.7}$$

com vetor média μ_i e matriz de covariância Σ_i. A soma dos pesos das componentes gaussianas satisfaz a expressão $\sum_{i=1}^{M} w\left(i\right) = 1$. O GMM é parametrizado em sua totalidade utilizando os pesos das componentes, os vetores de média e as matrizes de covariância de todas componentes gaussianas. Esse conjunto de parâmetros pode ser representado com a seguinte notação, $\lambda = (w_i, \mu_i, \Sigma_i)$.

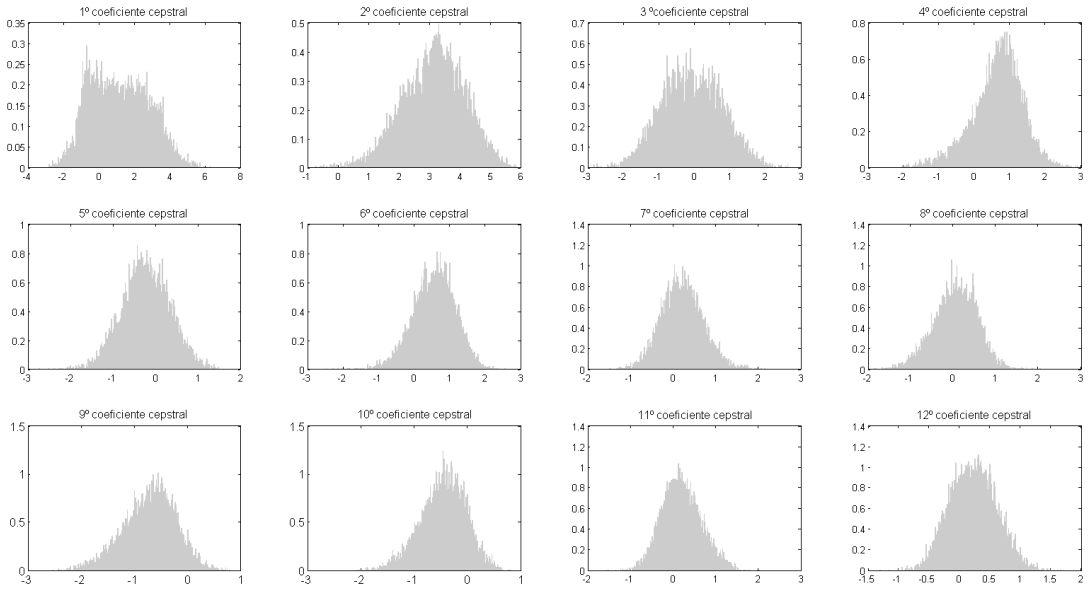

Figura 3.15: 12 primeiros coeficientes cepstrais de um cepstrum

3.1.3 Estimação de parâmetros

Existem várias técnicas para estimar os parâmetros do GMM [8]. O método de estimação mais popular é o de ML, sigla em inglês para máxima-verossimilhança. Esse método visa encontrar o modelo de parâmetros que maximizam a verossimilhança do GMM dado um vetor de dados de treino. A dificuldade dessa tarefa está na característica não linear dessa operação. Para encontrar os parâmetros de ML é utilizado o algorítimo EM, do inglês Expectation-Maximization [8]. A função do EM é partir de um modelo inicial de variadas componentes gaussianas para estimar novos parâmetros que resultarão em componentes mais acuradas, tal que $p(x|\lambda) \leq p(x|\lambda')$. O novo modelo então se torna o modelo inicial para a próxima iteração, e esse processo continua até que a diferença entre o valor gerado pela última iteração e o valor da penúltima iteração forem menor que que um determinado valor, ou atinja uma número máximo de iterações, a Figura 3.18 mostra um diagrama de blocos do algoritmo EM implementado nos experimentos deste trabalho. O algorítimo EM é aplicado em cada parâmetro de cada conjunto de gaussianas segundo as fórmulas,

$$\mu_k{}^{(novo)} = \frac{\sum_{n=1}^{T} x_n P\left(q_k|x_n,\Theta\right)}{\sum_{n=1}^{T} P\left(q_k|x_n,\Theta\right)} \tag{3.8}$$

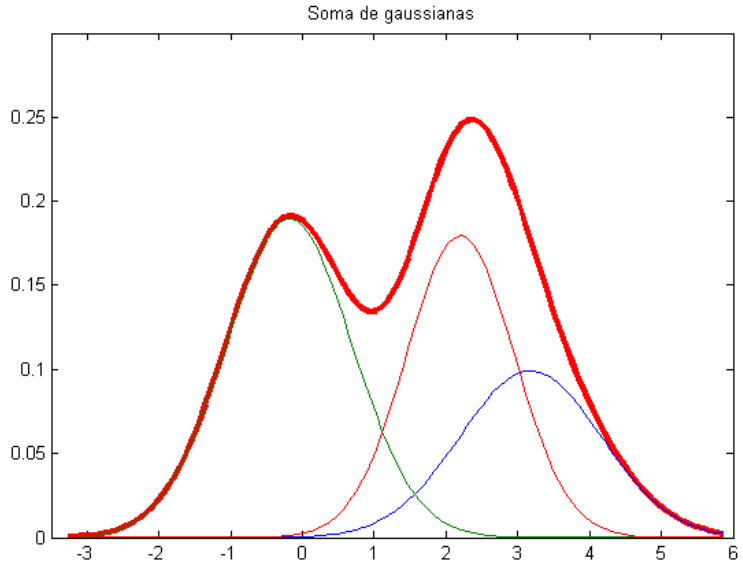

Figura 3.16: Modelagem de coeficiente cepstral

$$\Sigma_k{}^{(novo)} = \frac{\sum_{n=1}^{T} P\left(q_k|x_n,\Theta\right)\left(x_n-\mu_k{}^{(novo)}\right)\left(x_n-\mu_k{}^{(novo)}\right)^T}{\sum_{n=1}^{T} P\left(q_k|x_n,\Theta\right)} \tag{3.9}$$

$$P\left(q_k{}^{novo}|\Theta^{novo}\right) = \frac{\sum_{n=1}^{T} P\left(q_k|x_n,\Theta\right)}{T} \tag{3.10}$$

onde μ_k representa a média da mistura k de tamanho T, e Σ_k a variância da mistura k. $P\left(q_k|x_n,\Theta\right)$ é interpretado por vários autores como a responsabilidade que a componente gaussiana q_k tem sobre a amostra x_n, representa a probabilidade da amostra x_n pertencer à componente q_k dado um parâmetro Θ. Uma vez designada todas as responsabilidades de cada gaussiana em relação às n amostras, novos parâmetros de média, variância e peso podem ser estimados gerando novas componentes. Agora com novas componentes, todo processo designação de responsabilidades é refeito para obter parâmetros mais precisos. Este processo se repete até que haja uma convergência, seja por número de iterações definidos ou variação significativa entre iterações subsequentes.

Quando o algoritmo converge, o modelo λ final é armazenado no banco de dados, como mostra a Figura 3.1. Na Figura 3.17, logo abaixo, é ilustrado um exemplo de modelagem de coeficiente cepstral. Três gaussianas são utilizadas para modelar o histograma do primeiro coeficiente cepstral de um dado trecho de fala.

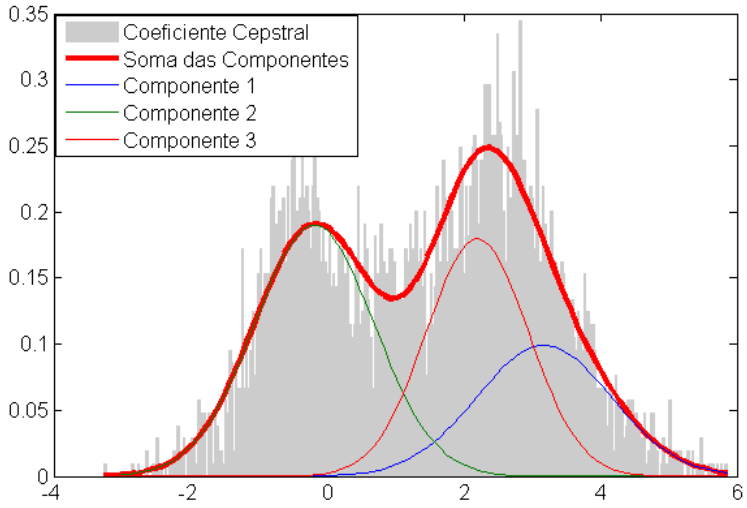

Figura 3.17: Modelagem de coeficiente cepstral

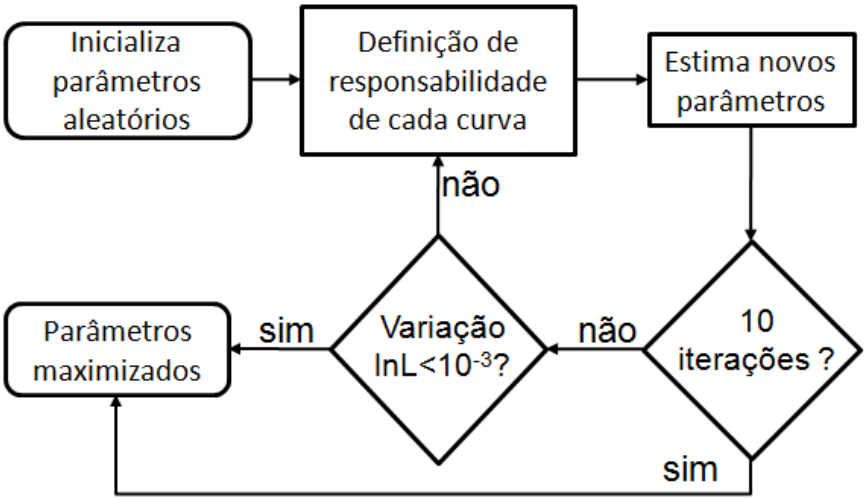

Figura 3.18: Diagrama de blocos do algoritmo EM

3.1.4 Identificação de N locutores

O modelo λ encontrado pelo EM é utilizado na etapa de comparação de características na fase de teste. Nesse ponto, o modelo de cada entrada do banco de dados é comparado com o histograma dos coeficientes cepstrais extraídos da fala em teste. Usando a Figura 3.17 para ilustrar o caso, as gaussianas seriam o modelo extraído na etapa de treino e comparadas ao coeficiente ceptral extraído na etapa de teste. Essa comparação é feita calculando o valor da função de verossimilhança utilizando os parâmetros armazenados no banco de dados com o coeficiente cepstral extraído da fala de teste.

A função de verossimilhança calcula, a partir das amostras observadas de uma variável

aleatória, a probabilidade dos parâmetros média μ_1 e variância σ_1 fixados representarem o comportamento desta variável aleatória, Figura 3.19.

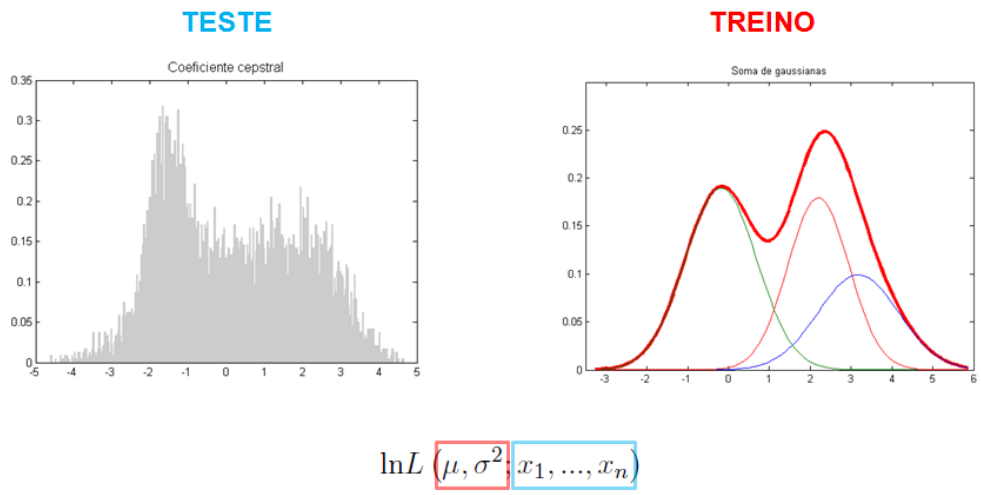

Figura 3.19: Comparação entre o modelo de treinado de um coeficiente cepstral e um coeficiente ceptral de teste para se obter o valor de *log-likelihood*

Dada uma variável aleatória X, retira-se uma amostra aleatória simples de X, de tamanho n, $X_1, \ldots, X_n$, e sejam $x_1, \ldots, x_n$ os valores efetivamente observados. Para um dado parâmetro θ a função de verossimilhança é definida como:

$$L\left(\theta; x_1, ..., x_n\right) = p\left(x_1; \theta\right) p\left(x_2; \theta\right) ... p\left(x_n; \theta\right) = \prod_{i=1}^{n} p\left(x_i; \theta\right) \tag{3.11}$$

O que se procura fazer é encontrar o estimador de máxima verossimilhança de θ que vai maximizar a função $L\left(\theta; x_1, ..., x_n\right)$. Essa função é amplamente usada para encontrar parâmetros que descrevem uma dada curva observada. Tratando-se de curva gaussiana, o parâmetro θ de (3.11) é representado pela média e variância, μ e σ. Logo, a função em questão toma a forma (3.12), mas por simplicidade computacional usa-se a forma logarítmica da função de verossimilhança (3.13), transformando o produtório em um somatório, e como as funções possuem uma relação monótona, maximizar o valor de uma é o mesmo que maximizar o valor da outra.

Dessa comparação é obtido um vetor coluna com valores de verossimilhança M-dimensional de tal forma que o elemento a_i representa o *log-likelihood* entre o modelo gaussiano do i-ésimo indivíduo no banco de dados e o histograma dos coeficientes cepstrais da fala em teste. O valor de log-likelihood de uma curva gaussiana é obtido com a equação abaixo:

$$L\left(\mu,\sigma^2;x_1,...,x_n\right)=\left(2\,\pi\right)^{-\frac{n}{2}}\left(\sigma^2\right)^{-\frac{n}{2}}\mathrm{e}^{-\frac{1}{2}\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{\sigma}} \tag{3.12}$$

$$\ln L\left(\mu,\sigma^2;x_1,...,x_n\right)=-\frac{n}{2}\ln\left(2\,\pi\right)-\frac{n}{2}\ln\left(\sigma^2\right)\frac{1}{2}\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{\sigma^2}. \tag{3.13}$$

Depois que o vetor coluna a é gerado no bloco de lógica de decisão da figura 3.2, o indivíduo com o maior valor de *log-likelihood* é indicado como provável locutor do sinal de voz em teste. Na Figura 3.20 é apresentado um exemplo de resultado quando se testa 3 locutores diferentes em sistema de reconhecimento automático contendo o modelo destes 3 mesmo locutores, e como esperado, o valor de *log-likelihood* é maior se compara o teste de um locutor com seu próprio modelo.

A =

	Teste 1	Teste 2	Teste 3
Modelo 1	-17,1068	-21,399	-26,5288
Modelo 2	-19,3234	-17,728	-20,592
Modelo 3	-22,2834	-22,0115	-16,7265

Figura 3.20: Tabela do resultado do teste de 3 locutores com o modelo de cada um deles. Em verde temos em cada coluna a decisão do sistema por maior valor de *log-likelihood*.

Capítulo 4

Separação Cega de Fontes

4.1 Introdução

A ideia de separação cega de fontes surgiu da necessidade de isolar as fontes de sinais presentes em uma mistura de sinais sem informações, a priori, das fontes e da estatística de seus sinais. Imagine que você esta em uma sala com duas pessoas conversando simultaneamente. Há dois microfones em locais diferentes, e cada microfone grava um sinal de áudio, podemos caracterizar cada microfone como $x_1(t)$ e $x_2(t)$, onde t é um índice temporal. Cada sinal de áudio de mistura é uma soma ponderada do sinal de voz emitido por cada pessoa, e podemos representar por $s_1(t)$ e $s_2(t)$. Essa situação pode ser expressa como uma equação linear do tipo:

$$x_1(t) = a_{11}s_1 + a_{12}s_2 \tag{4.1}$$

$$x_2(t) = a_{21}s_1 + a_{22}s_2 \tag{4.2}$$

onde a_{ij} depende da distância entre os locutores e os microfones. Seria ótimo se pudéssemos a partir dessas equações estimar os sinais $s_1(t)$ e $x_2(t)$. Esse problema é conhecido como *cocktail-party problem*. Essa é uma forma simplificada de olhar para o problema. Várias situações poderiam ainda ser consideradas como atrasos gerados por reflexões, obstáculos, ambiente ruidoso, dentre outras. Considerando a situação do problema mais simples para um melhor entendimento, uma forma de estimar os valores a_{ij} seria fazer uso das propriedades estatísticas dos sinais.

O método utilizado para a realização do experimento neste trabalho, ICA, parte da suposição de que em um determinado instante de tempo os sinais $s_1(t)$ e $2_2(t)$ são estatisticamente independentes e não-gaussianos, e com isso estima os valores a_{ij} para poder separar os sinais das misturas $x_1(t)$ e $x_2(t)$. É importante notar que o número de microfones deve ser igual ou maior que o número de fontes presentes na mistura, caso contrário haverá mais incógnitas do que equações no sistema linear.

4.2 ICA instantâneo

O ICA instantâneo modela a situação acima desconsiderando atrasos nos sinais emitidos pelas fontes, ou seja, os sons enviados pelas fontes no instante $t + 1$ são detectados pelos sensores no mesmo instante t_2, Figura 4.1.

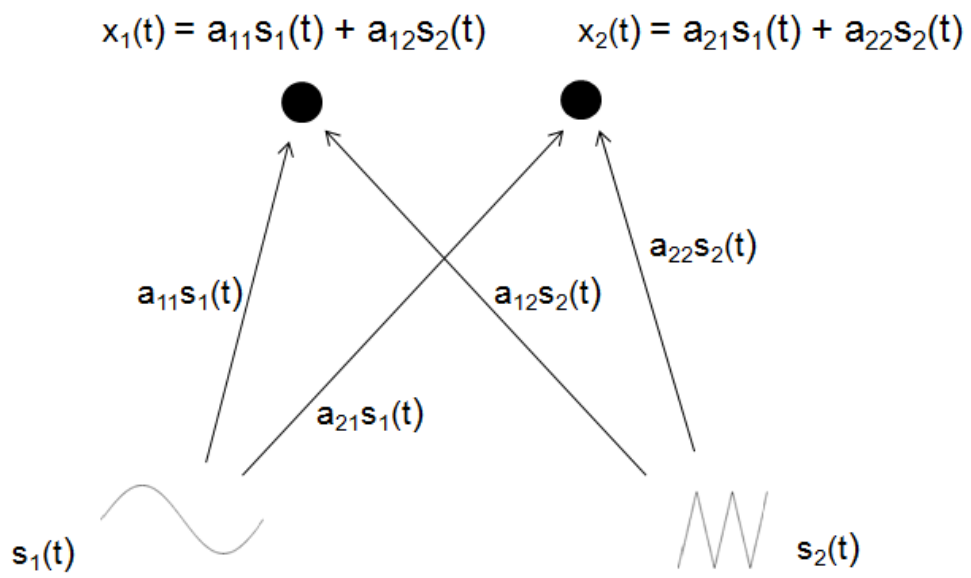

Figura 4.1: Ilustração de captação do sinal de áudio em uma sala fechada por dois microfones (círculos preto)

Podemos representar os sinais de mistura como um vetor coluna **X**:

$$\mathbf{X} = \begin{bmatrix} x_1(t) \\ \\ x_2(t) \end{bmatrix}$$

Definir a matriz de mistura A:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} \\ \\ a_{2,1} & a_{2,2} \end{bmatrix}$$

O vetor coluna de sinais s como

$$\mathbf{s} = \begin{bmatrix} s_1(t) \\ \\ s_2(t) \end{bmatrix}$$

O problema pode ser matricialmente expresso como:

$$\begin{bmatrix} x_1(t) \\ \\ x_2(t) \end{bmatrix} = \begin{bmatrix} a_{1,1}(t)\, a_{1,2}(t) \\ \\ a_{2,1}(t)\, a_{2,2}(t) \end{bmatrix} \begin{bmatrix} s_1(t) \\ \\ s_2(t) \end{bmatrix} \tag{4.3}$$

Uma vez estimada a matriz de mistura $\mathbf{A}$, pode-se computar a sua inversa $\mathbf{W} = \mathbf{A}^{-1}$ e obter os sinais das fontes resolvendo:

$$\begin{bmatrix} s_1(t) \\ \\ s_2(t) \end{bmatrix} = \begin{bmatrix} w_{1,1}(t)\, w_{1,2}(t) \\ \\ w_{2,1}(t)\, w_{2,2}(t) \end{bmatrix} \begin{bmatrix} x_1(t) \\ \\ x_2(t) \end{bmatrix} \tag{4.4}$$

Essa é uma forma simplificada para mostrar como o ICA trabalha, sem considerar situações adversas como presença de ruído e distorção de canal.

4.2.1 Ambiguidade do ICA

Não há como determinar a ordem da saída do algoritmo. Como $\mathbf{A}$ e $\mathbf{s}$ são desconhecidos, os termos de cada mistura podem estar livremente permutados entre si, pode-se então nomear cada componente independente como sendo a primeira. Formalmente estima-se uma matriz de permutação $\mathbf{P}$ e sua inversa $\mathbf{P}^{-1}$ e as substituem em $\mathbf{X} = \mathbf{A}\mathbf{P}^{-1}\mathbf{P}\mathbf{s}$, onde $\mathbf{P}\mathbf{s}$ são os valores orginais da variáveis independentes s_i, mas em uma outra ordem, e $\mathbf{A}\mathbf{P}^{-1}$ é apenas uma matriz de mistura para ser estimada.

4.2.2 Aplicação do ICA

O ICA faz uso de dados estatísticos para realizar a separação cega de fontes, e para isso ele parte de duas suposições importante:

- Os sinais a serem separados são independentes entre si;
- Os sinais a serem separados são não-gaussianos

Se a ideia do método é separar as fontes, é totalmente coerente pensar que os sinais são independentes. Se os sinais a serem separados contém informação útil, como é o caso do sinal de voz, é intuitivo pensar que não serão separados dados com variações aleatórias como é o caso de ruído AWGN, singla em inglês para *Addctive White Gaussian Noise*, ruído branco.
Será mostrado a seguir conceitualmente os principais processos para realizar a separação de fontes. Em [17] é mostrado detalhadamente como é feita a separação de fontes.

4.2.3 Braqueamento

Antes de partir diretamente para a separação das fontes é preciso pré-processar o sinal para garantir que as fontes apresentem independência entre si após a mistura dos sinais, já que os sinais têm sua correlação comprometida quando misturados.

Um conjunto de sinais é dito branco se os sinais possuírem média nula, variância unitária e forem descorrelacionados, ou seja, valor de correlação igual a zero. Uma interpretação geométrica para o branqueamento é a restauração do "formato"dos sinais, deixando para o ICA apenas a tarefa de achar uma matriz $\mathbf{W}$ que rotaciona esse "formato".

Para chegar ao sinal branqueado $\mathbf{Z} = \mathbf{V}\mathbf{X}'$ é preciso obter o sinal $\mathbf{X}$ normalizado com média nula e variância unitária, e a matriz branqueadora $\mathbf{V}$. A versão normalizada do sinal $\mathbf{X}$, denotada por **X'** pode ser encontrada por

$$\mathbf{X}' = \frac{\mathbf{X} - \mu_x}{\sigma_x}. \tag{4.5}$$

Como é desejado uma matriz $\mathbf{Z}$ de sinais descorrelacionados, iguala-se sua auto-covariância $\mathbf{\Sigma}_{zz}$ à matriz identidade de mesma ordem, onde $\mathbf{\Sigma}_{zz}$ é definido como

$$\mathbf{\Sigma}_{zz} = \begin{bmatrix} \operatorname{cov}(\mathbf{X},\mathbf{X}) & \operatorname{cov}(\mathbf{X},\mathbf{V}) \\ \\ \operatorname{cov}(\mathbf{V},\mathbf{X}) & \operatorname{cov}(\mathbf{V},\mathbf{V}) \end{bmatrix} \tag{4.6}$$

Para sinais com média nula, e tamanho K, fica bem simples a expansão

$$\mathbf{\Sigma}_{zz} = \frac{\mathbf{Z}\mathbf{Z}^{\mathrm{H}}}{K} = \mathbf{V}\frac{\mathbf{X}'\mathbf{X}'^{\mathrm{H}}}{K}\mathbf{V}^{\mathrm{H}} = \mathbf{V}\mathbf{\Sigma}_{xx}\mathbf{V}^{\mathrm{H}} = \mathbf{V}\mathbf{\Sigma}_{xx}\mathbf{V}^{\mathrm{H}} = \mathbf{I} \tag{4.7}$$

Haja vista que Σ_{xx} é uma matriz normal, por ser uma matriz de covariância. Fazendo sua decomposição de autovalores de teremos

$$\mathbf{\Sigma}_{xx} = \mathbf{E}\mathbf{\Lambda}\,\mathbf{E}^{\mathrm{H}} \tag{4.8}$$

onde $\mathbf{E}$ é a matriz ortogonal de autovetores e tem a sua forma transposta igual à inversa [18], e Λ uma matriz diagonal de autovalores. Substituindo 4.8 em 4.7,

$$\mathbf{V}\mathbf{E}\mathbf{\Lambda}\,\mathbf{E}^{\mathrm{H}}\mathbf{V}^{\mathrm{H}} = \mathbf{I}, \tag{4.9}$$

o que nos dá uma matriz branqueadora $\mathbf{V} = \mathbf{E}\mathbf{\Lambda}^{-1/2}\,\mathbf{E}^{\mathrm{H}}$. A vantagem do branqueamento é que agora existe uma nova matriz de mistura **A'** que é ortogonal.

$$\mathbf{E}\mathbf{\Lambda}^{-1/2}\mathbf{E}^{\mathrm{H}}\mathbf{A}\mathbf{s} = \mathbf{A'}\mathbf{s} \tag{4.10}$$

Sabendo que a mistura branqueada $\mathbf{Z}$ corresponde a uma rotação da fonte $\mathbf{s}$, fica mais fácil estimar a matriz separadora $\mathbf{W}$. Nota-se que o branqueamento reduz o número de parâmetros a serem estimados. Ao invés de estimar os n^2 parâmetros que são os elementos originais da matriz $\mathbf{A}$, precisa-se apenas estimar na nova matriz de mistura ortogonal **A'**. Uma matriz ortogonal contém $n(n-1)/2$ graus de liberdade.

Para ilustrar a situação considere duas variáveis aleatórias M e N com distribuição uniforme. No gráfico abaixo M é plotado no eixo das abscissas e N no eixo das ordenadas.

Para simplicidade do exemplo, os sinais são misturados linearmente. Plotando as duas misturas obtemos o gráfico abaixo.

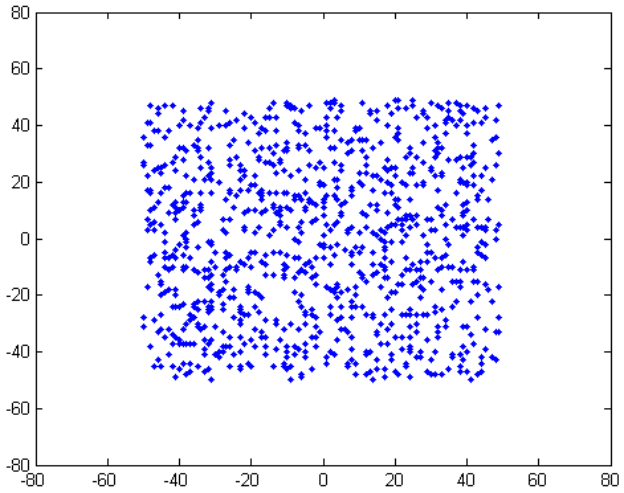

Figura 4.2: Plotagem das variáveis aleatórias independentes M e N

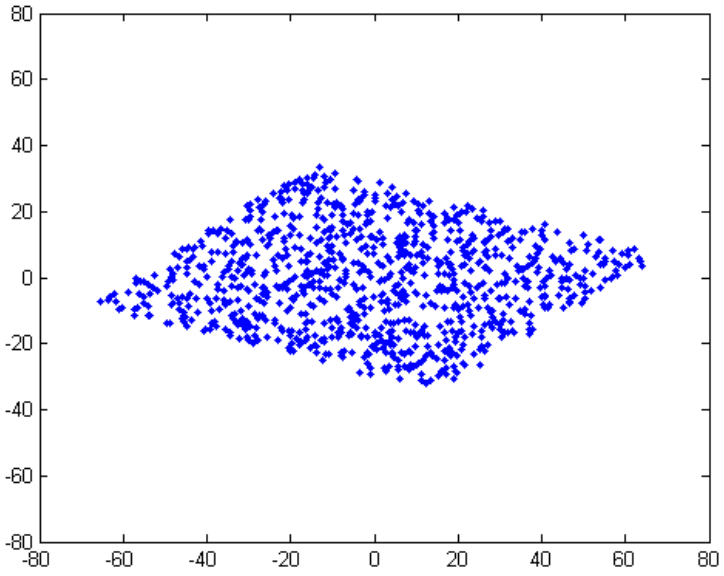

Figura 4.3: Plotagem da mistura linear das variáveis aleatórias M e N

E após o processo de branqueamento das misturas teremos algo como o gráfico abaixo.

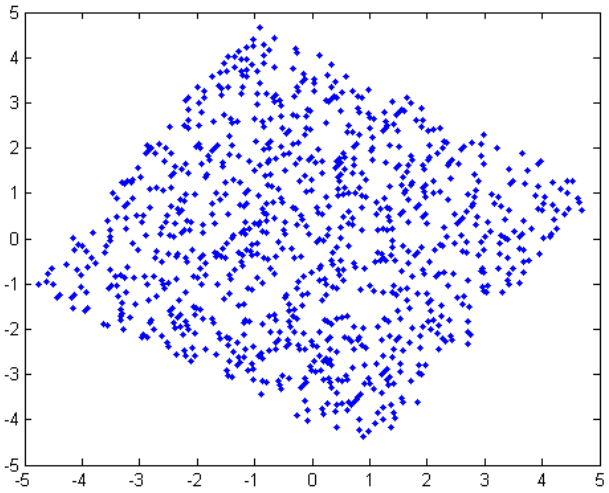

Figura 4.4: Plotagem da mistura linear das variáveis aleatórias M e N normalizadas

4.2.4 Separação por maximização da não-gaussianidade

Partindo do princípio que os sinais das fontes não são gaussianas, pode-se separar os sinais da mistura branqueada maximizando a não-gaussianidade da saída.

No gráfico abaixo a projeção em ambos os eixos é bastante gaussiana, o que faz evidente o Teorema do Limite Central, onde é demonstrado que a sobreposições de duas variáveis aleatórias é mais gaussiana do que as originais.

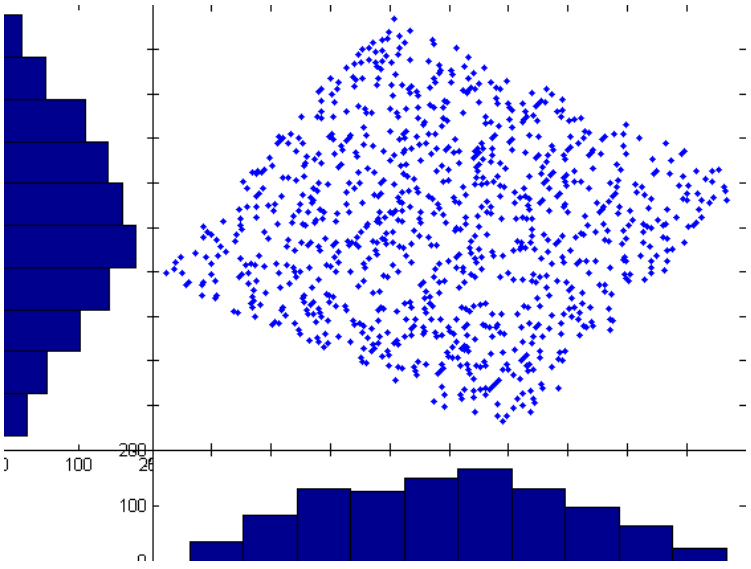

Figura 4.5: Histograma da mistura das variáveis aleatórias M e N

Rotacionando o eixos e minimizando a gaussianidade do gráfico acima, o ICA consegue

recuperar as fontes dos sinais misturados que são estatisticamente independentes.

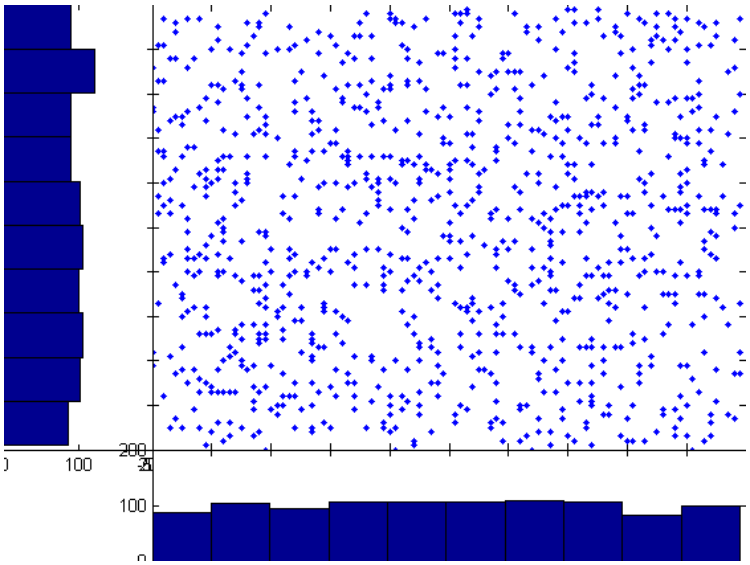

Figura 4.6: Histograma da mistura das variáveis aleatórias M e N rotacionadas

A matriz **W** que rotaciona a mistura branqueada pode ser achada através da maximização da não-gaussianidade. Dada uma mistura separada **Y** a definição de curtose é simplificada para uma média nula e variância unitária.

$$\kappa_Y = -3 + \frac{1}{\sigma^4 n} \sum_{i=1}^{n} |\mathbf{Y}|^4 (4.11)$$

Primeiramente encontra-se o gradiente da curtose de **Y** para acrescentar à matriz **W** a informação necessária para que esta rotacione a matriz de mistura **Z** eficientemente, de modo a maximizar a não-gaussianidade da mistura na saída.

$$\nabla \kappa_y = 2 \begin{bmatrix} \frac{\partial(\kappa_{y_1})}{\partial({w_{11}}^C)} & \frac{\partial(\kappa_{y_1})}{\partial({w_{12}}^C)} \\ \frac{\partial(\kappa_{y_2})}{\partial({w_{21}}^C)} & \frac{\partial(\kappa_{y_2})}{\partial({w_{22}}^C)} \end{bmatrix} \tag{4.12}$$

Resolvendo as derivadas parciais chega-se em 4.13,

$$\nabla (\kappa_Y) = \Big((|\mathbf{Y}|)^3 \otimes \text{sign}(\mathbf{Y}) \Big) (Z) \Big)^{\text{H}} \tag{4.13}$$

Onde $\otimes$ representa uma multiplicação termo a termo. Com 4.13 pode-se acrescentar W iterativamente com um limite de convergência até que ele aponte na direção do gradiente.

$$\mathbf{W} = \mathbf{W} + \gamma sign(\kappa_y) \otimes \Big((|\mathbf{Y}|)^3 \otimes sign(\mathbf{Y}) \Big) (\mathbf{Z})^{\mathrm{H}} \quad (4.14)$$

É importante notar que se a curtose κ_y for positiva deve-se maximiza-la, mas se a curtose κ_y for negativa deve-se minimizá-la para afastar da gaussianidade, isto é, $\kappa_y = 0$. Para maximizar a não gaussianidade basta então multiplicar a matriz $\mathbf{W}$ pela matriz branqueada $\mathbf{Z}$ a fim de se obter a matriz separada $\mathbf{Y}$.

4.3 ICA convolutivo

Para a situação em que se deseja separar sinais de voz captados por microfones, caso de estudo deste trabalho, precisa-se levar em conta no algoritmo que as misturas recebidas pelos sensores são convolutivas. Trazendo o caso para o lado da investigação forense, onde geralmente se realiza uma escuta ambiente, os sensores são instalados em locais estratégicos, o que torna impossível que se obtenha os suspeitos em distâncias exatamente iguais dos sensores, e como o som tem velocidade limitada, os sinais de áudio emitidos pelas fontes no instante t chegam aos sensores em instantes diferentes. Além disso também devemos lidar com o problema de múltiplo percusso, em que a existência de qualquer objeto, ou obstáculo, no ambiente pode gerar, por reflexões, versões do sinal voz que chegarão ligeiramente atrasadas nos sensores, Figura 4.7.

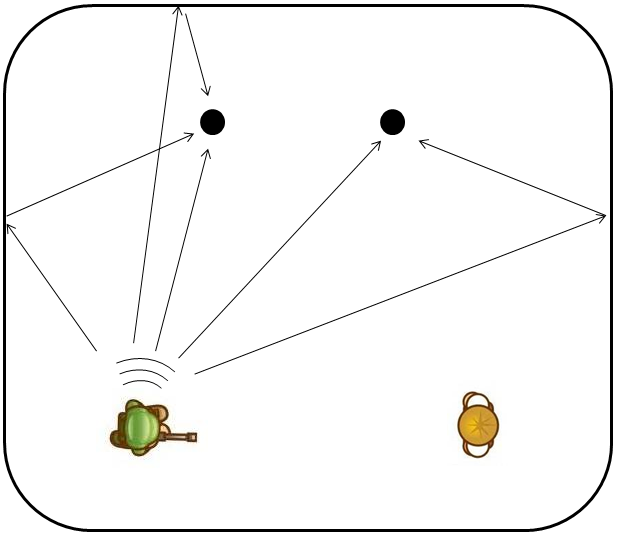

Figura 4.7: Ilustração de captação do sinal de áudio em uma sala fechada por dois microfones (círculos preto)

Ilustrando um cenário análogo ao apresentado no ICA instantâneo, onde se deseja separar os sinais de voz emitidos por duas pessoas em uma sala fechada, captados por dois microfones, os sinais de misturas recebidos pelos sensores serão do tipo $X = Hs$,

$$\begin{bmatrix} x_1(t) \\ \\ x_2(t) \end{bmatrix} = \begin{bmatrix} h_{1,1}(t)\, h_{1,2}(t) \\ \\ h_{2,1}(t)\, h_{2,2}(t) \end{bmatrix} \begin{bmatrix} s_1(t) \\ \\ s_2(t) \end{bmatrix} \tag{4.15}$$

onde $\mathbf{H}$ é a matriz de mistura, e uma vez estimada pode-se computar a sua inversa $\mathbf{W} = \mathbf{H}^{-1}$ e obter os sinais das fontes resolvendo:

$$\begin{bmatrix} s_1(t) \\ \\ s_2(t) \end{bmatrix} = \begin{bmatrix} w_{1,1}(t)\, w_{1,2}(t) \\ \\ w_{2,1}(t)\, w_{2,2}(t) \end{bmatrix} \begin{bmatrix} x_1(t) \\ \\ x_2(t) \end{bmatrix} \tag{4.16}$$

4.3.1 ICA no domínio da frequência

Devido aos ambientes reverberantes que são comuns para aplicações de áudio, o ICA linear apresenta baixo desempenho. Isto porque um atraso em um sinal no domínio do tempo implica em uma defasagem no domínio da frequência, que pode ser facilmente verificada na transformada de Fourier.

Seja $x(n - n_0)$ uma versão atrasada do sinal $x(n)$ e $X(k) = \mathcal{F}\{x(n)\}, k = [1, 2, \ldots, K]$ a representação deste sinal no domínio da frequência, então

$$X(k) = \frac{1}{K} \sum_{n=A}^{K-1} x(n) \mathrm{e}^{-2\frac{\pi jnk}{K}}, \tag{4.17}$$

substituindo $x(n)$ por $x(n - n_0)$

$$\frac{1}{K} \sum_{n=A}^{K-1} x(n - n_0) \mathrm{e}^{-2\frac{\pi j(n-n_0)k}{K}} = \frac{1}{K} \sum_{n=A}^{K-1} x(n) \mathrm{e}^{-2\frac{\pi jnk}{K}} \mathrm{e}^{-2\frac{\pi jn_0k}{K}}, \tag{4.18}$$

logo,

$$\mathcal{F}\{x(n - n_0)\} = X(k) \mathrm{e}^{-2\frac{\pi jn_0k}{K}}. \tag{4.19}$$

Para sinais de banda estreita, supergaussianos, uma defasagem no domínio da frequência pode acarretar em um deslocamento completo do sinal no espectro, ou seja, apenas uma multiplicação por uma constante complexa.

Portanto, para lidar com as amostras atrasadas, utiliza-se o ICA convolutivo [11]. Esse método pode ser usado para separar misturas de sinais de banda estreita, devido a sua capacidade de corrigir o deslocamento de fase causado por múltiplas amostras repetidas no ambiente. No domínio da frequência os sinais de mistura podem ser representados matricialmente como

$$\begin{bmatrix} \mathcal{F}\{x_1(t)\} \\ \\ \mathcal{F}\{x_2(t)\} \end{bmatrix} = \begin{bmatrix} h_{1,1}c_1h_{1,2}c_2 \\ \\ h_{2,1}c_3h_{2,2}c_4 \end{bmatrix} \begin{bmatrix} \mathcal{F}\{s_1(t)\} \\ \\ \mathcal{F}\{s_2(t)\} \end{bmatrix}, \tag{4.20}$$

onde $\mathcal{F}\{.\}$ representa a transformada de Fourier, c_n os coeficientes responsáveis pela defasagem no espectro, S contém dois sinais de banda estreita que serão misturados e H mantém os pesos no domínio da frequência.

Quando se trabalha com sinais no domínio da frequência a aplicação do ICA não é suficiente. Como a variação de fase nas frequências positivas a menos do sinal deve ser igual a variação nas frequências negativas. Quando se computa a transformada de Fourier de um sinal em função do tempo, o resultado é um espectro com frequências positivas, e frequências negativas espelhadas de mesmo módulo. Isso pode ser observado no gráfico abaixo Figura 4.8, onde a primeira metade localiza-se na parte negativa do espectro, e a segunda metade na parte positiva.

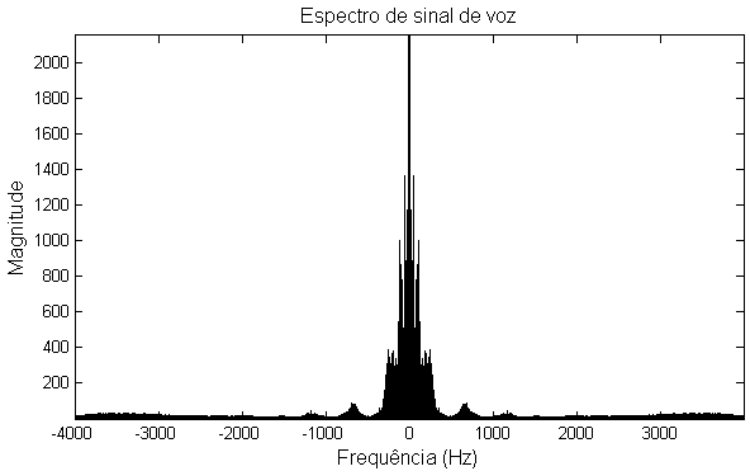

Figura 4.8: Espectro de um trecho de fala

Como as frequências negativas não podem ser ignoradas devemos conjugar estas raias de frequência do sinal de mistura $\mathcal{F}\{X(t)\}$, assim, uma operação de soma nestas frequências conjugadas será o mesmo que uma subtração nas frequências não conjugadas, resultando na variação de fase desejada. Após a aplicação do ICA deve-se conjugar novamente a primeira metade de raias de frequência antes de retornar o sinal para o domínio do tempo.

4.3.2 Ambiguidade de permutação

É importante ainda considerar o problema de ambiguidade de permutação, pois as componentes de frequência no sinal separado provavelmente estarão permutadas. A separação só ocorrerá de fato quando cada componente de frequência estiver no seu respectivo sinal de origem.

Apesar dos sinais estarem no domínio da frequência, os mesmos carregam informação estatísticas assim como na sua forma temporal, e através de análise de covariância e correlação pode-se encontrar uma relação entre dos componentes de frequência permutados pertencentes ao mesmo sinal. O gráfico abaixo, que é um exemplo disponível no Matlab, mostra o espectrograma do som *Chirp*, um canto de pássaro. Nesse tipo de gráfico pode-se analisar a variação de energia das frequências ao longo do tempo, especificamente neste gráfico há uma variação de cores do vermelho ao azul escuro, onde as maiores amplitudes estão na região vermelha. É perceptível a variação sincronizada de amplitude ao longo do tempo, apesar das amplitude se situarem em frequências maiores com o tempo, a sua distribuição permanece praticamente a mesma.

A fim de resolver o problema de permutação com maior eficiência, em [17] o autor sugere envolver as raias de frequência com um envelope e maximizar a correlação de suas formas. O envelope suaviza variações bruscas e não leva em consideração a fase da componente. O envelope de um sinal pode ser razoavelmente fácil de se obter. Seja Ey_k o envelope do sinal separado Y, defina

$$\mathbf{Ey}_k(i,m) = \frac{\sum_{t=1}^{T} |\mathbf{Y}_k(i,t+m)|}{T}, \tag{4.21}$$

em que $\mathbf{Ey}_k$ é o envelope de um sinal separado $\mathbf{Y}$, no domínio da frequência, em uma raia de frequência k, de fonte i, em função de um índice de quadro m. Conceitualmente o envelope é a media entre o módulo de T amostras consecutivas, com o intuito suavizar as variações bruscas.

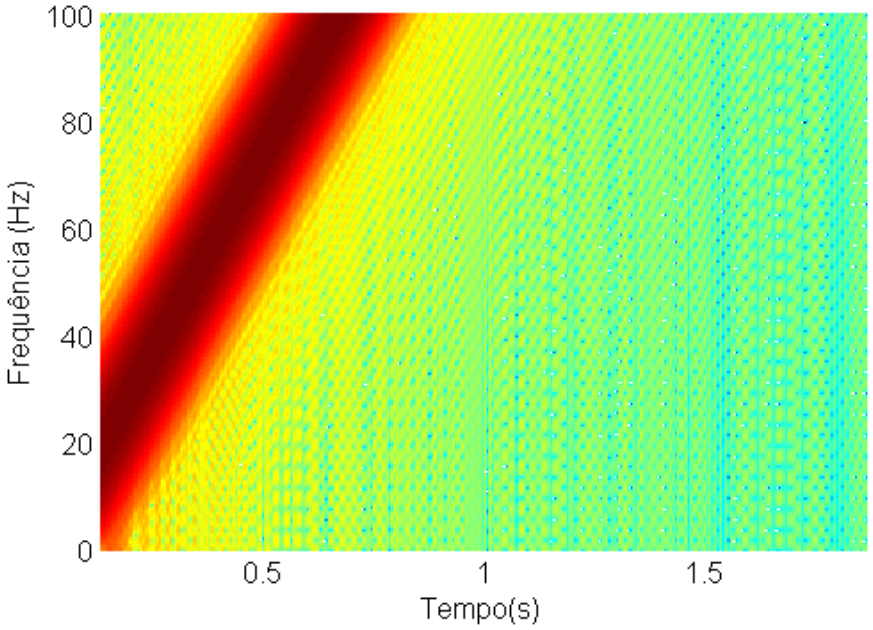

Figura 4.9: Espectrograma de canto de pássaro

Para exemplificar o resultado, na Figura 4.10, (4.21) é aplicado em um sinal de trecho de voz no domínio do tempo.

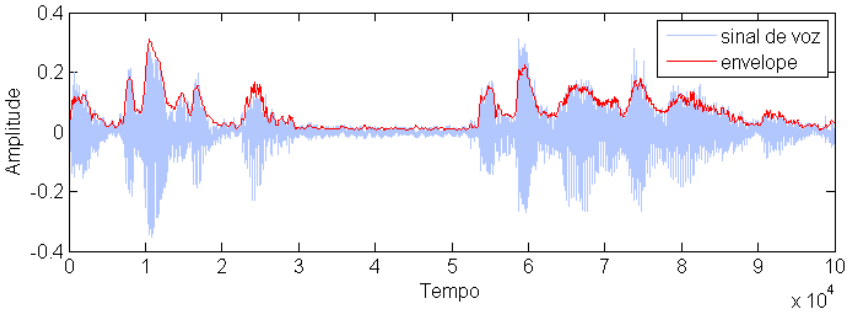

Figura 4.10: Diagrama de blocos da etapa de treinamento

A comparação se dará entre um conjunto de estimativa de envelopes de cada fonte e a estimativa de envelope de cada raia de frequência. Para isso é preciso construir conjunto de envelopes, que pode ser matematicamente expresso como um somatório dos envelopes de todas as raias de frequência k,

$$\mathbf{E'y}\left(i'\right) = \frac{\sum_k \mathbf{Ey}_k\left(i\right)}{\boldsymbol{\sigma}_{E'y(i')}}. \tag{4.22}$$

Em [17] o autor normaliza os envelopes dividindo-os pela sua respectiva variância, com isso

é preciso fazer a comparação estimando a covariância entre os envelopes ao invés da correlação, uma vez que a correlação desconsidera a amplitude dos sinais para computar suas relações. Tendo em vista que pessoas enfatizam frequências para produzir a fala, principalmente quando se deseja separar uma voz feminina de uma voz masculina, é importante considerar a diferença de amplitude além do formato do envelope.

$$\mathbf{CV}_k\left(i', i\right) = \text{cov}\left[\mathbf{E'y}\left(i'\right), \mathbf{Ey}_k\left(i\right)\right] \tag{4.23}$$

A equaçao (4.23) é a matriz de covariância da raia de frequência k que servirá de base para obter a matriz de permutação $\mathbf{P}$. A matriz $\mathbf{CV}_k$ computa a covariância entre os envelopes $\mathbf{Ey}_k$ dos sinais i contidos na raia k, com o conjunto de envelopes $E'y$ do sinal separado i', em que i e i' referem-se às fontes que estão sendo comparadas. Como o método aplica-se a N fontes, desde que haja N sensores, $(i, i') = [2, 3, \ldots, N]$. Com isso temos $textbfCV_k$ uma matriz quadrada de ordem N em que cada envelope de sinal da coluna i é comparado ao envelope do sinal separado $textbfY_i$ com todos os N conjunto de envelopes.

Computado as covariâncias que relacionam a semelhança entre as componentes de frequência de da raia k com seus respectivos sinais separados $\mathbf{Y}_i$, basta criar uma matriz de permutação $\mathbf{P}_k$ da raia k que maximiza as covariâncias na diagonal principal de $\mathbf{CV}_k$,

$$\mathbf{P}_k = argmax[\mathbf{CV}_k],$$

Argmax representa uma função que maximiza a diagonal principal da matriz $\mathbf{CV}_k$ permutando o conjunto de envelopes $\mathbf{Ey}_k$. A seguir será mostrado de forma bem intuitiva como esse função trabalha.

$$\mathbf{CV}_k = \begin{bmatrix} 0.27 & 0.71 & 0.09 \\ 0.38 & 0.12 & 0.68 \\ \mathbf{0.86} & 0.58 & 0.45 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & \mathbf{0.71} & 0.09 \\ 0 & 0.12 & 0.68 \\ 1 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & \mathbf{0.68} \\ 1 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

O primeiro passo é encontrar o elemento de maior valor na matriz CV_k, substituí-lo por 1 e igualar a zero os demais elementos pertencentes à mesma linha e coluna. Repete-se o passo

desconsiderando o valor 1 para a comparação até que haja apenas elementos iguais a 1 e 0. A matriz de permutação P_k vai ser igual à matriz CV_k em seu último estado.
Agora basta multiplicar o envelope Ey_k pela matriz de permutação P_k

$\mathbf{Ey}_k = \mathbf{Ey}_k\mathbf{P}_k$, para que a próxima vez que a matriz $\mathbf{CV}_k$ seja computada, resulte nos elementos

$$\begin{bmatrix} \mathbf{0.71} & 0.09 & 0.27 \\ 0.12 & \mathbf{0.68} & 0.38 \\ 0.58 & 0.45 & \mathbf{0.86} \end{bmatrix}$$

, e a nova matriz $\mathbf{P}_k$ será igual à matriz identidade. É possível que seja preciso estimar a matriz de permutação mais de uma vez antes da mesma se igualar à matriz identidade, logo, faz-se necessário realizar várias iterações até que não haja mais mudanças na matriz $\mathbf{P}_k$.

A Figura 4.11 mostra um diagrama de blocos que nos dá uma visão geral de como o ICA convolutivo funciona.

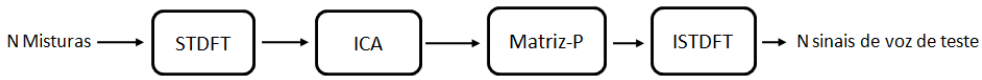

Figura 4.11: Diagrama de blocos da etapa de treinamento

O arranjo de microfones gera N sinais, um para cada microfone. Esses sinais representam N misturas que serão aplicadas ao algorítimo do ICA convolutivo. O STDFT é utilizado para analisar o sinal no domínio da frequência a fim de rastrear as versões original e atrasadas das amostras. Assim, o método ICA valor-complexo é aplicado a cada componente banda estreita dos sinais de mistura. Para resolver o problema de ambiguidades de permutação para cada raia de frequência, aplica-se uma matriz de permutação, que pode ser calculada baseada na amplitude da componente de frequência adjacente, também conhecido como envelope [11]. Por fim, a STDFT inversa é aplicada, resultando em N sinais de teste de fala.

4.4 Sistema de identificação de locutor com ICA convolutivo

Sinais de áudio, mais especificamente a fala humana, costumam ter distribuição bi-exponencial que também podem ser vistas como supergaussiana, e no domínio da frequência são ainda mais supergaussiana por ter a energia distribuída majoritariamente nas raias de frequências principais que caracterizam o som. Considerando um ambiente bastante reverberante, onde várias versões atrasadas serão geradas por reflexão, haverá uma enfatização nas frequências principais do som por interferência construtiva, e uma diminuição na energia das demais frequências por interferência destrutiva, o que torna o sinal de áudio ainda mais supergaussiano. Tendo em vista que quanto mais não-gaussiano um sinal, mais eficiente é o processo de separação de fontes, a melhor forma de separar sinais de áudio é aplicar o ICA após passar os sinais para o domínio da frequência. Podemos dividir o sistema de identificação de locutor proposto em duas etapas: de treinamento e de teste. No sistema proposto a etapa de treinamento é exatamente o mesmo mostrado na figura 3.1. O aprimoramento do sistema é feito na etapa de teste. A figura 4.12 mostra um digrama de blocos do método proposto.

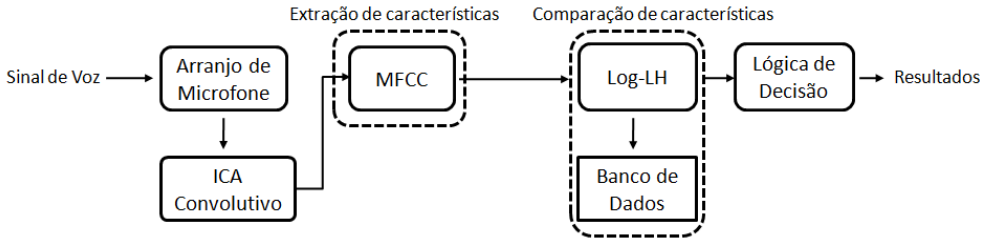

Figura 4.12: Diagrama de blocos da etapa de treinamento

Pode-se notar que o processo se mantém basicamente o mesmo da figura 3.2, exceto pela presença do ICA antes da extração de características. O sistema recebe N áudios obtidos de N locutores por gravação do arranjo de microfones. Usa-se então a técnica de ICA convolutivo para separar a voz dos locutores, resultando em N sinais de áudio de teste, contendo majoritariamente a voz de cada indivíduo. Em seguida os N sinais são aplicados à mesma etapa de teste mostrada no diagrama de blocos da figura 3.2. Como será mostrado adianta há pequenas diferenças entre essas duas etapas de teste.

Capítulo 5

Aplicação em identificação de múltiplos locutores

5.1 Introdução

Neste capitulo é avaliada o desempenho do sistema de identificação de locutor baseada em ICA com o método típico que utiliza apenas o MFCC e o GMM. É avaliada o desempenho do sistema baseado na taxa de acertos ao tentar identificar um determinado locutor, no caso típico onde há apenas uma mistura convolutiva de duas pessoas falando, e no sistema proposto onde as fontes são separadas via o algorítimo do ICA convolutivo.

Na subseção 5.1.1 é caracterizado o ambiente experimental. Na subseção 5.1.2, é apresentado resultado para casos sem ruído. Na subseção 5.1.3, a performance da abordagem proposta é avaliada na presença de ruído aditivo gaussiano branco. Na subseção 5.1.4, é mostrado uma comparação entre os coeficientes cepstrais do sinal de voz de teste e de treinamento do mesmo indivíduo nas duas abordagens de sistema de identificação de locutor.

5.1.1 Ambiente Experimental

O áudio de mistura pode ser obtido utilizando um arranjo de microfones, mas para este experimento a mistura convolutiva foi simulada. Aplicando o algorítimo de ICA convolutivo em um áudio contendo a mistura da voz de duas pessoas, obtemos como saída dois arquivos de áudio, cada

um contendo majoritariamente a voz de cada individuo. Usa-se então esses sinais como entrada no sistema típico de identificação de locutor MFCC-GMM para avaliar o sistema proposto.

A parte do MFCC-GMM foi configurada para trabalhar com janelas de 256 amostras com 50% de sobreposição entre cada janela, um banco de filtros triangulares com 20 filtros igualmente espaçados na escala mel, e um conjunto de 5 gaussianas para modelar cada coeficiente cepstral. Todos os áudios utilizados para montar o banco de dados e fazer os testes foram obtidos com UCLA Phonetics Lab Archive. De cada áudio foi extraído de 2 a 3 minutos de fala ininterrupta para montar o banco de dados, e trechos de 30 a 40 segundos para ser utilizado como entradas de teste. O banco de dados contém modelos de sete pessoas diferentes .

Supondo um arranjo de microfone com dois microfones, o canal é simulado incluindo atrasos de amostras e diferença de amplitudes. Para uma fonte foi considerado o atraso de duas amostras e para outra fonte um atraso de 3 amostras. Para uma das fontes a amplitude percebida por um dos microfones é multiplicada por 1,2 e para o outro microfone por 0,8.

5.1.2 Desempenho com experimentos sem ruído

Em cada execução, duzentas simulações foram efetuadas para se obter a taxa de acertos (TA), definido como,

$$TA\,(\%) = 100\,\frac{N.A.}{N.S.} \tag{5.1}$$

Onde $N.A.$ é o número de acertos em uma execução, e $N.S.$ é o número de simulações em uma execução.

No primeiro cenário a voz de dois homens são misturadas. Como mostra a tabela 5.1, é um aumento da taxa de acertos para o locutor 7 quando o algorítimo ICA convolutivo é aplicado.

No segundo cenário, as vozes de um homem e de uma mulher são misturados. Como pode-se ver na tabela 5.2 há um aumento na taxa de acerto para a locutora 6 possibilitando a identificação da mesma com uma probabilidade maior que 50% quando o algorítimo de ICA convolutivo é aplicado.

Tabela 5.1: Avaliação da taxa de sucesso aplicando a técnica de identificação de locutor MFCC-GMM, e a taxa de sucesso com a abordagem do ICA, utilizando a voz de dois homens adultos

Áudio de entrada	**Taxa de sucesso**	
	Vozes misturadas	**Vozes separadas**
Locutor 7	0%	85,5%
Locutor 3	100%	99,5%

Tabela 5.2: Avaliação da taxa de sucesso aplicando a técnica de identificação de locutor MFCC-GMM, e a taxa de sucesso com a abordagem do ICA, utilizando a voz de dois homens adultos

Áudio de entrada	**Taxa de sucesso**	
	Vozes misturadas	**Vozes separadas**
Locutor 1	100%	97,5%
Locutor 6	6,5%	65,5%

5.1.3 Desempenho na presença de ruído

Nesta subseção o ruído aditivo gaussiano branco é considerado durante o processo identificação de locutor. O ruído utilizado para se obter a RSR desejada foi aplicado nos canais, ou seja, no mmomento em que os sinais das fontes são misturados. Analisando o áudio com a mistura o locutor 3 pode ser identificado, mas o locutor 7 permanece indetectável. A fim de identificar o locutor 7 utilizando o algorítimo de ICA convolutivo com uma taxa de acertos maior que 50% é preciso que o canal deste esteja 10dB menos ruidoso que o do outro locutor. Isto pode ser observado na figura 5.1

Em [16], um esquema de remoção de ruído colorido em um sistema de identificação de Locutor é proposto, onde o autor utiliza a método apresentado neste trabalho para avaliar a sua performance.

5.1.4 Análise de coeficiente cepstral

Nesta subseção será mostra a influência do ICA nos coeficientes cepstrais.

Para fim de uma identificação certeira, os coeficientes cepstrais do sinal de voz de teste têm que ser iguais aos coeficientes cepstrais do sinal de voz de treinamento.

Na figura 5.2 temos o primeiro coeficiente do locutor 7. No lado esquerdo da figura temos o

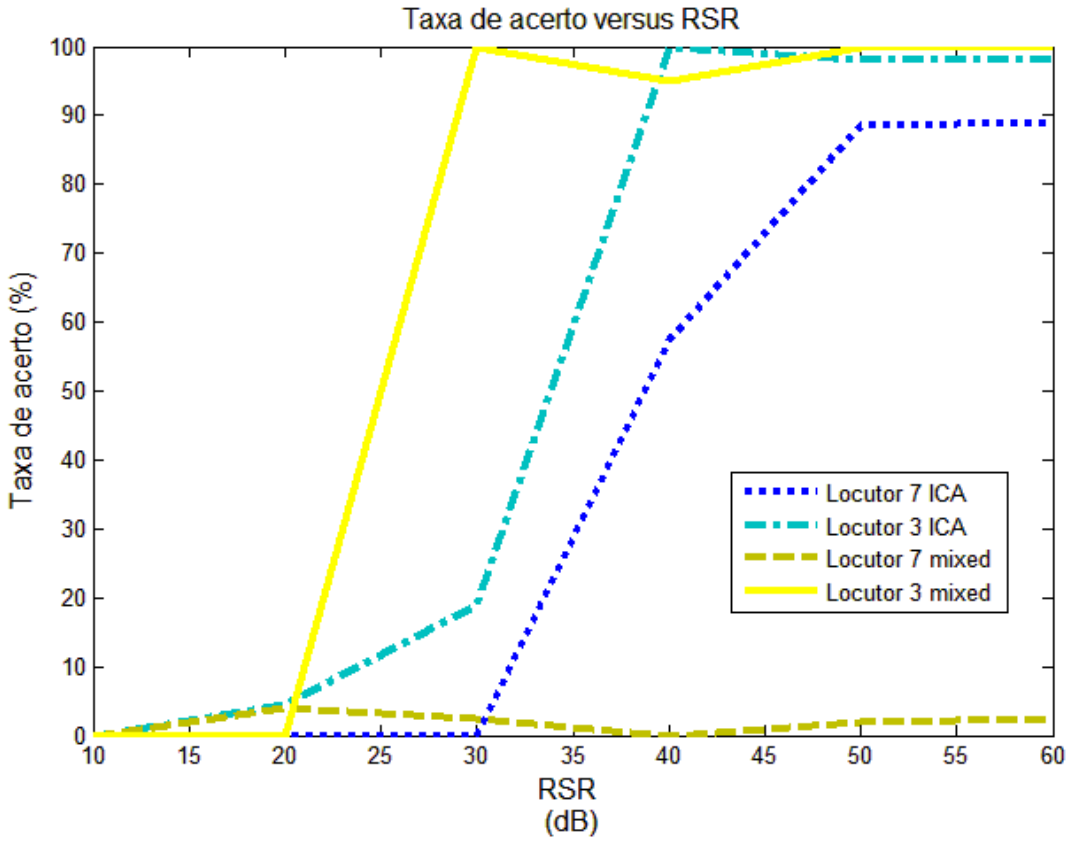

Figura 5.1: Comparação entre as abordagem com ICA, e abordagem típica MFCC-GMM para diferentes níveis de RSR

primeiro coeficiente cepstral do áudio de teste, e no lado direito o primeiro coeficiente cepstral do áudio de treinamento. É notável que os histogramas são bastante parecidos.

A figura 5.3 mostra a influência da mistura de dois locutores captados ao mesmo tempo. No lado esquerdo da figura temos o primeiro coeficiente cepstral do áudio de mistura contendo a voz dos locutores 3 e 7. No lado direito da figura temos o primeiro coeficiente ceptral do áudio de teste do locutor 7 após passar pelo algorítimo de separação de fonte de sinal do ICA convolutivo. É notável a sua curva ficou bastante similar ao primeiro coeficient cepstral contido no banco de dados, o que gera um valor mais alto de *log-likilihood*.

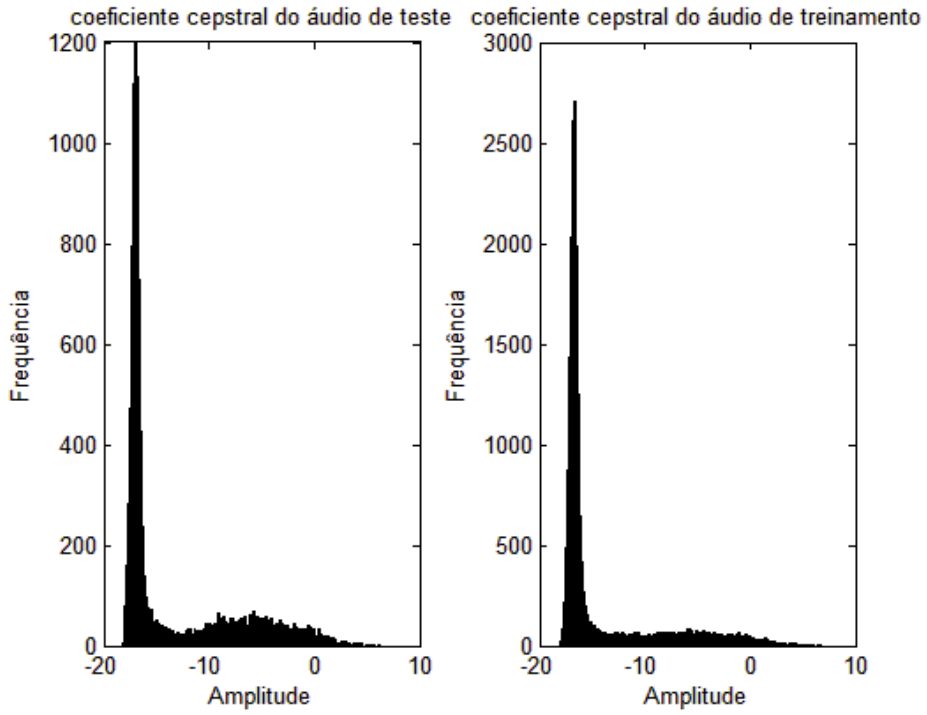

Figura 5.2: Comparação entre o primeiro coeficiente cepstral do áudio de teste na esquerda e do áudio de treinamento na direita

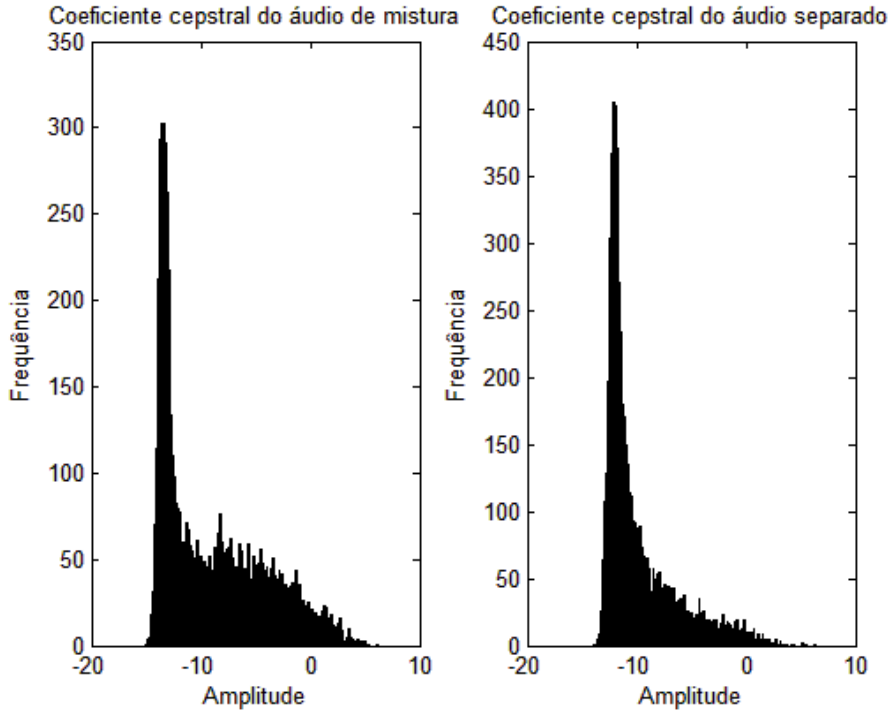

Figura 5.3: Comparação entre o primeiro coeficiente cepstral do áudio de mistura na esquerda e do áudio após separação via ICA

Capítulo 6

Cancelamento de ruído colorido

6.1 Introdução

Como foi visto no Capítulo anterior, o desempenho do sistema de reconhecimento automático de locutor cai drasticamente na presença de ruído. A adição de ruídos ao sinais de voz provoca um tendenciamento do coeficiente cepstral, assim reduzindo a acurácia do sistema. Já foi mencionado anteriormente também que em casos de investigação criminal o microfones geralmente estão escondido, o que resulta em uma fraca captação de voz com bastante ruído.

6.2 Detecção de atividade de voz

Para o esquema de cancelamento de ruído, é preciso saber quando há presença de voz e quando não há para um dado sinal observado. O método de detecção VAD, sigla em inglês para Voice Activity Detection (Detecção de atividade de voz) é utilizada para identificar onde há presença de voz para que se possa fazer um balanço entre a energia dos trechos com, e sem voz. Há diversas medidas relacionadas ao VAD, como abordagens baseadas em limiares, medição de periodicidade, informação espectral, informação dos primeiros quadros de ruído, e até fusão dos métodos.

Como a abordagem por limiar é bastante vulnerável a regimes de baixo RSR, Razão-Sinal-Ruído, e os outros métodos não são muito confiáveis por termos outras vozes agindo como ruído em plano de fundo, é proposto um esquema de VAD iterativo.

Segmentando um sinal de áudio, em função do tempo, em vários quadros de tamanho M, calcula-se o valor RMS de um quadro b por

$$\mathrm{RMS}_b = \sqrt{\frac{\sum_{m=1}^{M} {b_m}^2}{M}}. \tag{6.1}$$

Considera-se a presença de voz em quadros que tenham um valor RMS acima de um limiar estabelecido, caso contrário caracterizados como ruído. Com isso é possível fazer uma primeira estimativa da RSR do sinal.

$$\mathrm{RSR} = 20 \ln\left(\frac{\mathrm{RMS}_s - a_{RMS}\mathrm{RMS}_n}{\mathrm{RMS}_n}\right), \tag{6.2}$$

onde RMS_s é o valor médio do RMS_b de todos os quadros do sinal segmentado com presença de voz, RMS_n o valor respectivo para quadros caracterizados como ruído, e a_{RMS} um fator de equilíbrio levando em conta a presença de ruído nos quadros de voz. A partir dessa primeira estimativa de RSR um novo limiar é estimado por determinação empírica. Esse novo limiar é utilizado para reavaliar a classificação dos quadros quanto à presença ou não de voz. O processo é repetido iterativamente até que não haja mudança significativa entre estimativas consecutivas de novo RSR. A Figura 6.1 mostra um fluxograma do algoritmo. Geralmente o método de VAD proposto não leva mais de 10 iterações para convergir e trás resultados confiáveis para RSR's de até 0 dB. Esse método será aplicado logo após o ICA computar os sinais separados de volta para o domínio do tempo, onde geralmente possuem RSR's maiores que 0 dB, o que torna o esquema em questão válido.

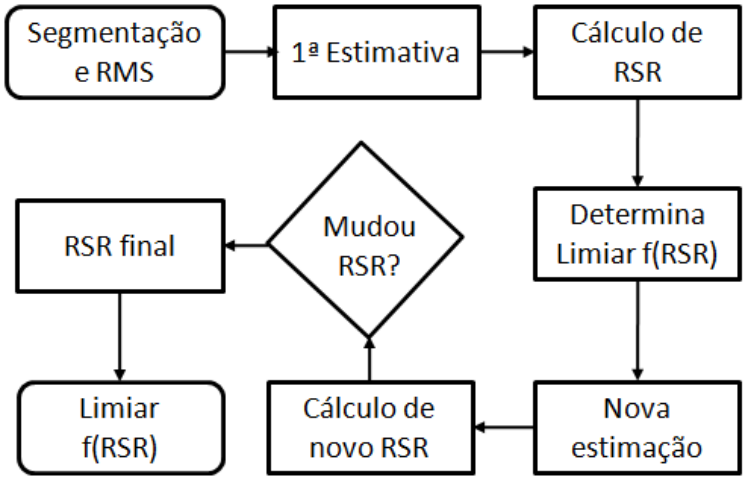

Figura 6.1: Fluxograma para determinação do limiar para detecção de ruído

6.3 Cancelamento de ruído incorporado ao ICA

Como ferramenta extra para o cancelamento do ruído o método do ICA pode ser aplicado considerando-se um sensor a mais que a quantidade de fontes nas misturas, assim o sistema interpretará a N-ésima+1 fonte sendo a causa do ruído e terá suas componentes separadas dos outros sinais. Seguindo fluxograma da Figura 6.2, o VAD recebe as saídas do ICA no domínio do tempo, mas como o cancelamento de ruído é feito por subtração espectral, o VAD gera um vetor de índices apontando se trechos dos sinais de saída do ICA, no domínio da frequência, possuem presença de voz ou ruído. O trechos caracterizados com presença de voz sofrem uma subtração de estimativa de ruído, dos trechos sem presença de voz, e são encaminhados ao sistema de reconhecimento de locutor. Ademais, trechos do sinal separado pelo ICA caracterizados com ruído sofrem uma atenuação e também são encaminhados ao sistema de reconhecimento de locutor.

Com o cancelamento do ruído espera-se um grande aprimoramento no desempenho do sistema automático de reconhecimento de locutor, fato que será discutido no capítulo seguinte.

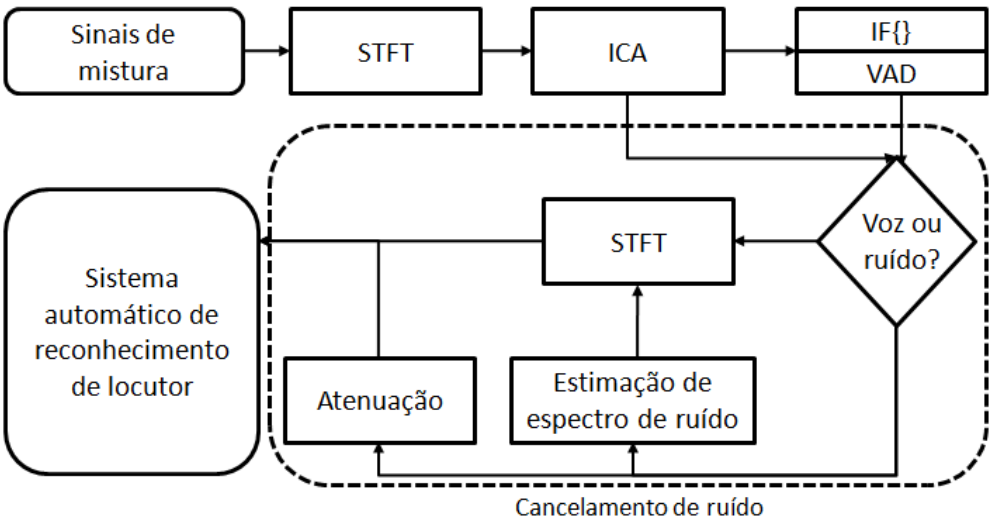

Figura 6.2: Fluxograma explicativo da incorporação do esquema VAD para aprimorar sistemas de reconhecimento automático de locutor

Capítulo 7

Aplicação em identificação de múltiplos locutores em ambiente com ruído colorido

7.1 Ambiente experimental

O método proposto foi testado para dois e três locutores com sinal de voz contaminado por aleatórios ruídos estacionários com diferentes níveis de RSR's e níveis de correlação de ruído ρ. Foi simulado $N+1$ microfones para capturar os sinais com diferentes atrasos. Para simular o ambiente desejado, foi adicionado ruídos coloridos n_c aleatórios em todos canais de mistura. Os testes foram executados com diferentes níveis de ρ criados a partir de um ruído gaussiano n_w.

$$n_c\,(t+1) = \sqrt{1-\rho^2}n_w\,(t+1) + \rho\, n_c\,(t) \tag{7.1}$$

Dentro do esquema de cancelamento de ruído visto na Figura 6.2, para a estimativa da RSR no VAD a subtração pelo RMS_n do ruído foi ponderado por uma fator $a_{RMS} = 0,7$. Os valores numéricos foram encontrados empiricamente para tornar o experimento funcional.

Como medidor de desempenho no sistema de reconhecimento de locutor novamente será usado TA definido em (5.1). Ademais, será usado uma nova grandeza TAC, Taxa de Acerto Cruzado, para medir a capacidade de identificar todos locutores de uma vez, definida por:

$$\mathrm{TAC}\,(\%) = \left(\prod_{i=1}^{N} \mathrm{TA}_i\right)^{N-1} \tag{7.2}$$

Todos os sinais da simulação foram amostrados a uma taxa de 44100 Hz. Para cada configuração de ruído e RSR foram executadas 50 simulações de separação de sinais, cancelamento de ruído e reconhecimento de locutor para determinar a acurácia do sistema. O banco de dados contém características de 9 locutores, sendo 3 homens, 5 mulheres e 1 criança.

Com o intuito de provar a eficácia do método final proposto, o desempenho do mesmo será comparado com o de outras abordagens demonstradas na Figura 7.1.

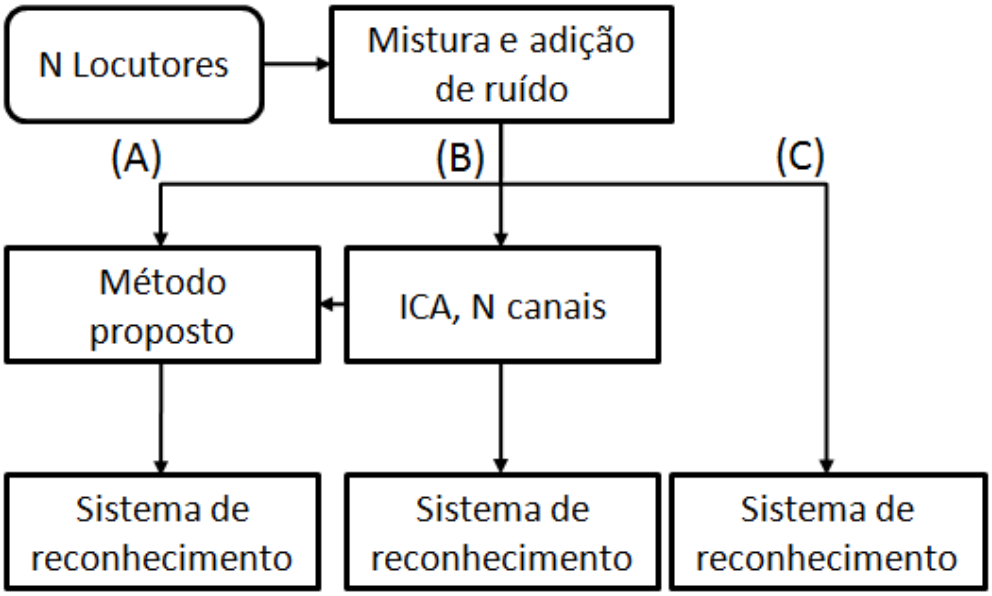

Figura 7.1: Três abordagens diferentes a serem simuladas

7.2 Resultados

Os simulações são executadas considerando dois e três locutores, e as taxas de acertos TA entre as três abordagens são comparadas. Nas Figuras 7.2 e 7.3 é mostrado a taxa de acertos TA em relação a RSR para todas as abordagens. Utilizando o método proposto final a taxa de acertos aumenta constantemente a partir de uma RSR de 10dB enquanto as outras abordagens falham para esses valores de RSR em um regime com coeficiente de correlação $\rho = 0,1$. Em uma situação com alto coeficiente de correlação, $\rho = 0,7$, um dos locutores pode ser reconhecido quase que sem erro em um regime de RSR = 25dB. A abordagem que inclui o método ICA falhou completamente

para todos valores de RSR.

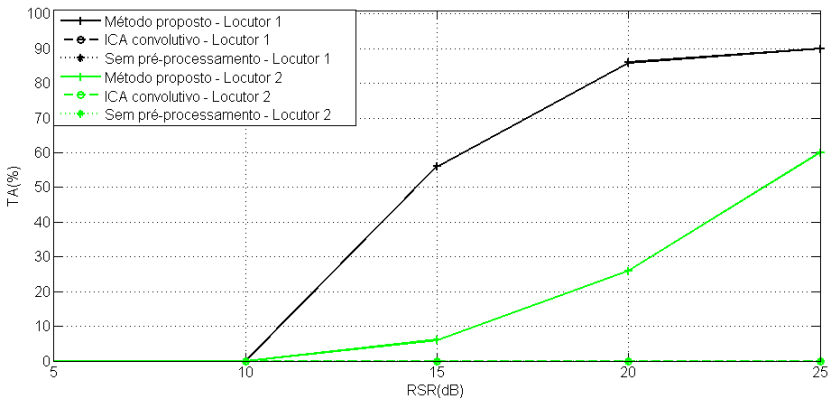

Figura 7.2: Comparação entre o ICA convolutivo com cancelamento de ruído, ICA convolutivo e aplicação de sistema de reconhecimento de locutor sem pré-processamento para diferentes níveis de RSR e coeficiente de correlação de ruído fixado em 0,1

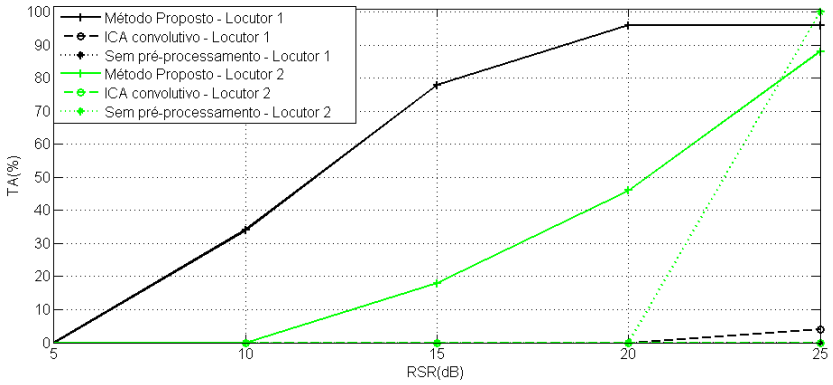

Figura 7.3: Comparação entre o ICA convolutivo com cancelamento de ruído, ICA convolutivo e aplicação de sistema de reconhecimento de locutor sem pré-processamento para diferentes níveis de RSR e coeficiente de correlação de ruído fixado em 0,7

Figura 7.4 mostra as taxas de acertos em relação a diferentes níveis de coeficiente de correlação de ruído ρ com RSR fixada em 20dB, nas três abordagens mostradas na Figura 7.1.Para o método proposto a variação do coeficiente de correlação quase não causa influência , pois temos a TA do Locutor 1 variando próximo a 90% e do Locutor 2 variando em torno de 40%. Apenas a abordagem via ICA não apresentou resultado positivo para qualquer variação do coeficiente de correlação de ruído.

Para comparar o desempenho com diferente número de locutores presentes, na Figura 7.5 a TAC é relacionada com RSR para as diferentes configurações de ambiente com 2 e 3 locutores presentes. Em ambos os casos a TAC cresce junto com RSR. As outras abordagens não apresentaram

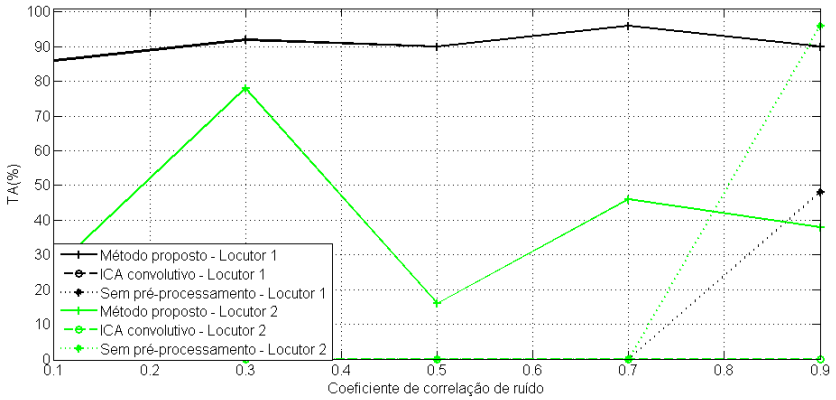

Figura 7.4: Comparação entre o ICA convolutivo com cancelamento de ruído, ICA convolutivo e aplicação de sistema de reconhecimento de locutor sem pré-processamento para diferentes níveis de coeficiente de correlação de ruído e RSR fixado em 20dB

resultados positivos, o que indica que só o método proposto é capaz de realizar o reconhecimento de vários Locutores simultaneamente.

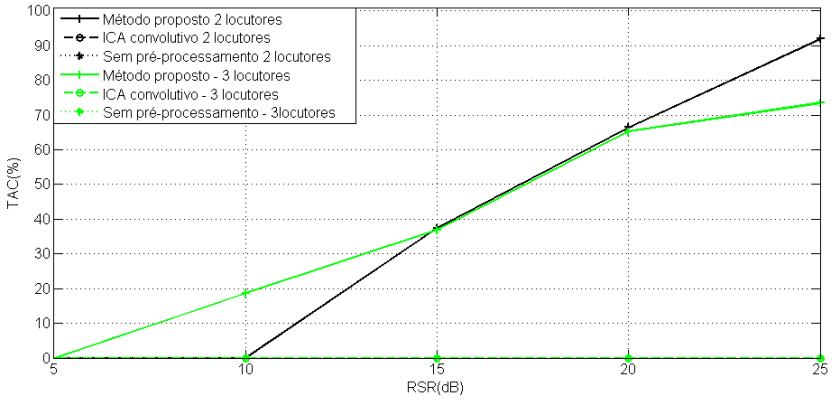

Figura 7.5: Comparação entre o ICA convolutivo com cancelamento de ruído, ICA convolutivo e aplicação de sistema de reconhecimento de locutor sem pré-processamento para reconhecimento simultâneo de locutores e diferentes níveis de RSR

Pode ser atribuído ao comportamento do método proposto uma grande robustez a ruído colorido, capaz de identificar vários locutores simultaneamente em regimes de baixa RSR. As outras abordagens testadas na simulação só mostram resultado satisfatório com níveis muito baixos de ruído. Em quase todos os ambientes o crescimento da taxa de acerto do método proposto é praticamente linear com o aumento da RSR.

Capítulo 8

Conclusões

Neste trabalho foi apresentada uma solução para aprimorar a técnica de identificação de locutor em ambientes com ruído colorido e vozes interferentes para aplicaçao forense, que tipicamente é o caso de gravações em escuta ambiente. É confirmado que usando apenas abordagem atual, MFCC-GMM, apenas um dos locutores pode ser identificado, enquanto que com apenas a aplicação do algorítimo do ICA convolutivo é possível identificar N locutores presentes na gravação, com uma taxa de acertos consideravelmente alta na ausência de ruído. Dando um passo além, aplicando um método de VAD foi possível fazer o sistema ficar tolerante a ambientes bastante ruidosos, que era o objetivo deste trabalho, haja vista que cenários de escuta ambientes geralmente o os microfones capturam bastante ruído.

A abordagem baseada no ICA convolutivo e cancelamento de ruído via VAD permite a identificação de todos os locutores em uma gravação forense, enquanto que a solução padrão permite a identificação apenas de um locutor. Contudo, o esquema proposto é bastante promissor no campo forense de identificação de locutor.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] JR, J. P. C. Speaker recognition: A tutorial. *Proceedings of the IEEE*, v. 85, n. 9, p. 1437–1462, 1997.

[2] REYNOLDS, D. A. An overview of automatic speaker recognition technology. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, v. 4, 2002.

[3] KINNUNEN; TOMI; LI, H. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication 52.1*, p. 12–40, 2010.

[4] TIWARI, V. MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies 1.1*, v. 1, p. 19–22, 2010.

[5] REYNOLDS, D. A.; ROSE, R. C. Robust text independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, p. 72–83, 1995.

[6] REYNOLDS, D. A.; QUATIERI, T. F.; DUNN, R. B. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, v. 10, p. 19–41, 2000.

[7] REYNOLDS, D. A. Gaussian mixture models. *Encyclopedia of Biometric Recognition*, p. 14–68, 2008.

[8] DEMPSTER, A. P.; LAIRD, N. M.; RUBIN., D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, p. 1–38, 1977.

[9] SCHLUTER, R.; BEZRUKOV, L.; WAGNER, H.; NEY, H. Gammatone features and feature combination for large vocabulary speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[10] KUMAR, K.; SINGH, R.; RAJ, B.; STERN, R. Gammatone sub-band magnitude-domain dereverberation for ASR. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011.

[11] DOUGLAS, S. C.; GUPTA, M.; SAWADAAND, H.; MAKINO, S. Spatio temporal FastICA algorithms for the blind separation of convolutive mixtures. *IEEE Transactions on Audio, Speech, and Language Processing*, p. 1511–1520, 2007.

[12] ALMEIDA, F. D.; NASCIMENTO, F.; BERGER, P.; SILVA, L. Noise robust speaker recognition using reduced multiconditional gaussian mixture models,. *The International Journal of Forensic Computer Science IJoFCS*, v. 3, p. 60–69, 2006.

[13] ALMEIDA, F. D.; NASCIMENTO, F.; BERGER, P.; SILVA, L. Automatic speaker recognition with multi-resolution gaussian mixture models (MR-GMM). *The International Journal of Forensic Computer Science IJoFCS*, v. 4, p. 9–21, 2009.

[14] MAHER, R. C. Audio forensic examination: authenticity, enhancement and interpretation. *IEEE Signal Processing Magazine*, v. 26, n. 2, p. 84–94, 2009.

[15] SILVEIRA, M. A.; SCHROEDER, C. P.; DA COSTA, J. P. C. L. Convolutive ica-based forensic speaker identification using mel frequency cepstral coefficients and gaussian mixture models. *The International Journal of Forensic Computer Science âĂŞ IJoFCS*, v. 4, p. 27–34, 2013.

[16] DENK, F.; DA COSTA, J. P. C. L.; SILVEIRA, M. A. Enhanced forensic multiple speaker recognition in the presence of coloured noise. *submitted to Computer, Systems and Signal Processing*, 2013.

[17] OLIVEIRA, C. G. de. Análise de componentes independentes na separação de misturas convolutivas: Bachelor thesis. *Universidade de Brasília - UnB*, 2012.

[18] HORN, R. A.; JOHN, C. R. Matrix analysis. *Cambridge University*, v. 2, 2012.

[19] DUBUISSON, T. Glottal source estimation and automatic detection of dysphonic speakers: Phd thesis. *University of Mons*, 2013.