UNIVERSIDADE FEDERAL DE PERNAMBUCO

GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO

CENTRO DE INFORMÁTICA

2011.2

UM SISTEMA DE BUSCA DE ATOR EM FILME ATRAVÉS DO RECONHECIMENTO DE LOCUTOR

TRABALHO DE GRADUAÇÃO

Aluno: Leonardo Valeriano Neri (lvn@cin.ufpe.br)
Orientador: Tsang Ing Ren (tir@cin.ufpe.br)

Recife, dezembro de 2011

UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

UM SISTEMA DE BUSCA DE ATOR EM FILME ATRAVÉS DO RECONHECIMENTO DE LOCUTOR

Leonardo Valeriano Neri

Trabalho de Graduação apresentado no Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para a obtenção do título de Engenheiro da Computação.

Orientador: *Prof. Dr. Tsang Ing Ren*

Recife, dezembro de 2011

Agradecimentos

Agradeço primeiramente a Deus, por ter me dado força, perseverança, por me ensinar a nunca desistir daquilo em que se acredita e por me guiar em momentos tão difíceis que passei até chegar neste momento tão grandioso em minha vida.

Agradeço aos meus pais, Luciano Neri do Nascimento e Jandira Valeriano Neri por sempre me apoiarem nos meus objetivos, por me ensinarem a seguir um caminho de honestidade para alcançar meus objetivos e por me ensinarem a ter calma e paciência para alcançar bons resultados.

Agradeço aos meus irmãos Leandro Valeriano Neri e Luana Valeriano Neri por terem me ajudado na minha longa caminhada, vindo me buscar em muitas noites na universidade. Agradeço muito a vocês dois por isso.

Agradeço a todo o restante da minha família, avó, tios, tias, primos e primas por terem me amado sempre.

Agradeço aos amigos que criei durante o curso, por serem verdadeiros exemplos de esforço, dedicação e sucesso em tudo o que fizeram ao longo do curso.

Agradeço à Paula Cristiane Gomes, por me amar como sou e por toda a felicidade que ela trouxe em minha vida. Te amo muito!

A todos vocês, muito obrigado!

“O seu tempo é limitado, então não o gaste vivendo a vida de alguém. Não fique preso pelos dogmas, que é viver com o resultado do pensamento de outras pessoas. Não deixe que o barulho da opinião dos outros cale a sua própria voz interior. E o mais importante: tenha coragem de seguir o seu próprio coração e a sua intuição. Eles de alguma maneira já sabem o que você realmente quer se tornar. Todo o resto é secundário.”

Steve Jobs

Resumo

A fala é a forma de comunicação mais simples, ágil e eficiente entre os seres humanos. Através da voz são expressas diversas informações: as palavras que compõe a mensagem, o estado emocional do transmissor da mensagem e as características fisiológicas do mesmo, as quais permitem ao receptor da mensagem identificar se o transmissor é alguém conhecido.

Todas essas informações são estudadas por muitas áreas de conhecimento, e consequentemente, dão origem às diversas aplicações com interface via voz. Uma dessas áreas é a área de reconhecimento de locutor, com o objetivo de tentar identificar unicamente um indivíduo através da sua voz e possibilitando a construção de um sistema computacional para realizar o reconhecimento automático de locutor.

Dentre as aplicações para sistemas de reconhecimento de locutor destacam-se o controle de acesso de pessoas a locais restritos, confirmação de identidade em transações bancárias e telefônicas e, entre outras, a que será abordada neste trabalho, uma aplicação em desenvolvimento conhecida como *Speaker Diarization*, que tem como objetivo extrair partes específicas de arquivos de áudio e/ou vídeo onde um determinado indivíduo está falando.

Este trabalho visa o desenvolvimento de um sistema de busca de atores em filmes, utilizando o reconhecimento de locutor, capaz de identificar em quais cenas do filme um determinado ator está presente com fala ativa. O sistema utilizará o método de extração de características *Mel Frequency Cepstral Coefficients* (MFCC) e o método de modelagem e classificação *Gaussian Mixture Models* (GMM). Será explicado todo o processo de desenvolvimento do sistema, desde a obtenção da base de dados até os experimentos realizados.

Palavras-chave: Reconhecimento de locutor, Speaker Diarization, Modelos de misturas Gaussianas, Processamento de sinais.

Abstract

The speech is the simplest, faster and most efficient means of communication among human beings. Through the speech, several information are expressed: the words that compose the message, the emotional state of the speaker and his physiological features, which allow to the listener to identify the speaker.

All of these information are studied by many areas of knowledge, and consequently, originating several applications with voice interface. One of these areas is the speaker recognition area, in order to try to identify the speaker by his voice, enabling the construction of a computer system to perform automatic speaker recognition.

Among the applications of the speaker recognition system we have people access control, identity confirmation in bank transactions, phone transactions and, the application that will be approached in this study, *Speaker Diarization*, in order to extract specifics parts of audio or video files where the speaker is present.

The aim of this work is to develop a search system of actors in movies, using speaker recognition, able to identify which scenes in the film a particular actor is speaking. The system will utilize the feature extraction method *Mel Frequency Cepstral Coefficients* (MFCC) and the modeling and classification method *Gaussian Mixture Models* (GMM). Will be explained all of the system development process, since the database building until the experiments performed.

Keywords: Speaker recognition, Speaker Diarization, Gaussian mixture models, Signal Processing.

SUMÁRIO

LISTA DE FIGURAS

Lista de Tabelas

1. INTRODUÇÃO

A fala é a forma de comunicação mais simples, ágil e eficiente entre os seres humanos. Através da voz são expressas diversas informações: as palavras que compõe a mensagem, o estado emocional do transmissor da mensagem e as características fisiológicas do mesmo, que permitem ao receptor da mensagem identificar se o transmissor é do sexo masculino ou feminino e se é alguém conhecido.

Todas essas informações são estudadas por muitas áreas do conhecimento e dão origem às diversas aplicações com interface via voz. Uma das áreas que estudam essas informações é a biometria, com o objetivo de identificar unicamente o indivíduo através das características intrínsecas da voz e possibilitando a construção de um sistema computacional para realizar o reconhecimento automático de locutor, definido como sistema de reconhecimento de locutor [1].

Os sistemas de reconhecimento de locutor analisam o sinal de voz com a finalidade de extrair as características fisiológicas da voz do indivíduo, que representam a forma do trato vocal, tamanho da laringe e entre outros órgãos responsáveis pela geração da voz, o jeito de falar do indivíduo, seu vocabulário, entonação, ritmo. A partir dessa análise, as características da voz são extraídas do sinal de voz é realizada uma modelagem dos sinais e feito o reconhecimento de padrões, identificando-se o indivíduo.

São muitas as aplicações para sistemas de reconhecimento de locutor, destacando-se o controle de acesso de pessoas a locais restritos e confirmação de identidade em transações bancárias e telefônicas, dando um nível a mais de segurança.

Existe também outro tipo de aplicação, em desenvolvimento, conhecida como *Speaker Diarization* [2], tendo como objetivo extrair partes específicas de arquivos de áudio e/ou vídeo onde um determinado indivíduo está falando. Essa aplicação pode determinar quando e quem está falando, sendo bastante útil para a área forense, para realizar buscas por bandidos em gravações telefônicas e também na área cinematográfica, buscando as cenas onde a voz de um ator específico está presente.

1.1. OBJETIVOS

1.1.1. Objetivos Principais

Com base no contexto apresentado, o trabalho visa o desenvolvimento de um sistema de busca de atores em filmes, utilizando o reconhecimento de locutor, capaz de identificar em quais cenas do filme um determinado ator está presente com fala

ativa. Após a identificação do ator, sua fala será extraída da cena, com o objetivo de sintetizar uma base de dados específica para aquele ator.

1.1.2. Objetivos Específicos

O sistema utilizará o método de extração de características *Mel Frequency Cepstral Coefficients* (MFCC) e o método de modelagem e classificação *Gaussian Mixture Models* (GMM). Serão usadas combinações dos dois métodos, no que se diz respeito ao número de coeficientes do MFCC e quantidade de misturas da GMM, para medir o desempenho de cada uma e escolher a melhor combinação para construção do sistema de reconhecimento de locutor, que por sua vez será utilizado no sistema de busca.

1.2. ORGANIZAÇÃO DO TRABALHO

O trabalho apresenta a seguinte organização:

- O Capítulo 2 fará uma apresentação dos sistemas de reconhecimento de locutor. Será apresentado um breve histórico, citando as principais técnicas do estado da arte, e os conceitos dos sistemas de reconhecimento. Também será esboçada uma arquitetura básica para a construção dos sistemas de identificação de locutor;
- O Capítulo 3 fará uma apresentação da aplicação *Speaker Diarization*. Serão apresentados os conceitos dessa aplicação e citadas as principais técnicas do estado da arte. Também será esboçada a arquitetura usada na construção do sistema de busca proposto;
- No Capítulo 4 será feita a descrição do sistema de busca, que utiliza um sistema de reconhecimento de locutor;
- No Capítulo 5 são realizados os experimentos com o sistema. Este capítulo explica como foi feita a confecção da base de dados, os tipos de experimentos e métricas utilizadas;
- O Capítulo 6 apresenta a conclusão do trabalho e apresenta propostas para trabalhos futuros.

2. SISTEMAS DE RECONHECIMENTO DE LOCUTOR

Este capítulo irá apresentar uma descrição dos sistemas de reconhecimento de locutor. Inicialmente é mostrado um breve histórico, mostrando como ocorreu a evolução desses sistemas e citando as principais técnicas do estado da arte, e em seguida serão mostrados os seus conceitos e arquitetura básica para o desenvolvimento desses sistemas.

2.1. BREVE HISTÓRICO

Durante muitos anos vêm sendo realizadas pesquisas em reconhecimento de pessoas através da voz. Esses estudos, pelo que se tem registrado, foram iniciados na década de 1930, através de um estudo realizado pelo Dr. Francis McGehee, professor de Psicologia na Universidade Johns Hopkins, onde ele queria determinar o quão bem uma pessoa pode identificar outras pessoas escutando apenas suas vozes [3].

Durante a Segunda Guerra Mundial houve um interesse maior em identificar pessoas pela voz. Com isso, em 1941, o Bells Laboratories inventou uma máquina que fazia o espectrograma da voz, com a esperança que essa informação pudesse ajudar a identificar as vozes de alemães que fossem interceptadas por rádio [4]. Os resultados obtidos para a identificação do indivíduo utilizando espectrograma não foram satisfatórios na época, e os estudos nessa área foram abandonados por um bom tempo.

Já em 1962, Lawrence Kersta, um dos que inventaram o espectrógrafo de som, publicou na revista *Nature* o trabalho "Voiceprint Identification" [5], no qual ele defendia a infalibilidade de seu método e relatava taxas de identificação corretas de 99%. Contudo, seus métodos e resultados foram controvertidos e sua aceitação na comunidade científica em geral foi restrita [6]. Apesar das controvérsias, seus estudos foram muito importantes para o início da área de reconhecimento de locutor.

Como foi visto até agora, o processo de reconhecimento de locutor era um processo executado manualmente, até meados de 1960, sendo preciso ter um profissional especializado, treinado para poder comparar visualmente os espectrogramas. A primeira técnica de reconhecimento automático de locutor surgiu em 1963, com os estudos de Pruzansky no Bell Laboratories [7]. Nesse sistema foi usado banco de filtros e dois espectrogramas digitais para medir a similaridade dos sinais de voz. Ainda durante o final da década de 1960 e toda a década 1970, surgiram vários outros sistemas baseados em analisar a evolução temporal de certos parâmetros da voz, especialmente da frequência fundamental, formantes, intensidade, e coeficientes de predição linear ([8] [9] [10] [11] [12] [13]).

Com o aumento do poder computacional, na década de 1980, as técnicas de reconhecimento de locutor ficaram progressivamente mais complexas, proporcionando melhorias de desempenho dos sistemas. Em 1986 Soong et. al. [14], após verificar o sucesso da utilização de técnicas de Quantização Vetorial (*Vector Quantization – VQ*) no reconhecimento de padrões, propuseram um sistema baseado nesse tipo de técnica para reconhecimento de locutor. Ainda que os experimentos com VQ demonstrassem bons resultados, em geral, havia uma restrição quanto ao tamanho do vocabulário, devido à própria característica da modelagem. Visando suprir isto surgiram as modelagens probabilísticas.

A partir da década de 1990, houve uma grande popularização dos sistemas baseados em modelagens probabilísticas. Entre eles se destacam os modelos ocultos de Markov (*Hidden Markov Models – HMM*) e os modelos de misturas gaussianas (*Gaussian Mixture Models – GMM*). O HMM é um modelo estatístico baseado em cadeias de Markov que incorporam informações sobre a evolução temporal dos parâmetros, sendo bastante utilizado tanto para o reconhecimento de voz como para o reconhecimento de locutor em modo texto dependente. Sua utilização em reconhecimento de locutor pode ser vistos em Rosemberg et. al. 1990 [15], Webb et. al. 1993 [16], Che and Lin 1995 [17] e Colombi et. al. 1995 [18]. Apesar de o HMM obter bons resultados para o reconhecimento em modo texto dependente, a informação temporal incorporada nesses modelos não mostrou vantagem nos sistemas de reconhecimento automático de locutor independente do texto.

O GMM também é um modelo estatístico, mas diferente do HMM, ele não leva em consideração a relação temporal, basicamente os GMM são HMM desprovidos da informação temporal. Assim sendo, o GMM está sendo amplamente utilizado em sistemas de reconhecimento automático de locutor independente do texto. A utilização do GMM em sistemas de reconhecimento automático de locutor foi introduzida por Reynolds, em 1992 [19], na sua tese de doutorado e posteriormente publicado em 1995 [20].

2.2. CONCEITOS

Os sistemas de reconhecimento de locutor executam o processo de reconhecimento automático de locutor, com o objetivo de extrair, categorizar e reconhecer a identidade do locutor vinda como informação do sinal de voz [21].

O processo de reconhecimento automático de locutor é subdividido em duas tarefas: *Verificação de Locutor* e *Identificação de Locutor*. Na verificação, o indivíduo afirma ter uma identidade (utilizando um cartão de identificação, senha, etc.) e o sistema verifica se a afirmação é verdadeira através da voz do indivíduo. Enquanto que, na identificação, não existe nenhuma informação *a priori* sobre a identidade do indivíduo, e o sistema precisa identificar através da voz, quem é o indivíduo, a qual

grupo ele pertence, ou então, caso o sistema seja utilizado por qualquer pessoa, se é um desconhecido [22].

Neste trabalho é abordada a problemática relacionada aos sistemas de identificação de locutor.

Dependendo do nível de cooperação do usuário, os sistemas de reconhecimento de locutor podem ser caracterizados como *dependentes de texto* ou *independentes de texto*. Nos sistemas dependentes de texto, existe o conhecimento *a priori* da fala que será pronunciada pelo usuário. Enquanto que nos sistemas independentes de texto o usuário poderá usar uma fala espontânea para a tarefa executada.

Em Campbell, 1997 [22], é visto que, em geral, os sistemas dependentes de texto tendem a apresentar um desempenho melhor em virtude do conhecimento prévio do que será dito, baseando-se no reconhecimento do texto e partindo da divergência entre a fala gerada e a fala de teste de um modelo selecionado, identificar o locutor. Os sistemas independentes de texto não possuem essa vantagem, pois não possuem o conhecimento prévio das características fonéticas da fala utilizada.

2.3. ARQUITETURA DE UM SISTEMA DE IDENTIFICAÇÃO DE LOCUTOR

Os sistemas de identificação de locutor compartilham uma arquitetura básica em comum. A arquitetura básica é apresentada pela Figura 1 e será utilizada neste trabalho. Todo o sinal de voz a ser utilizado pelo sistema é capturado por um microfone e depois convertido para o formato digital.

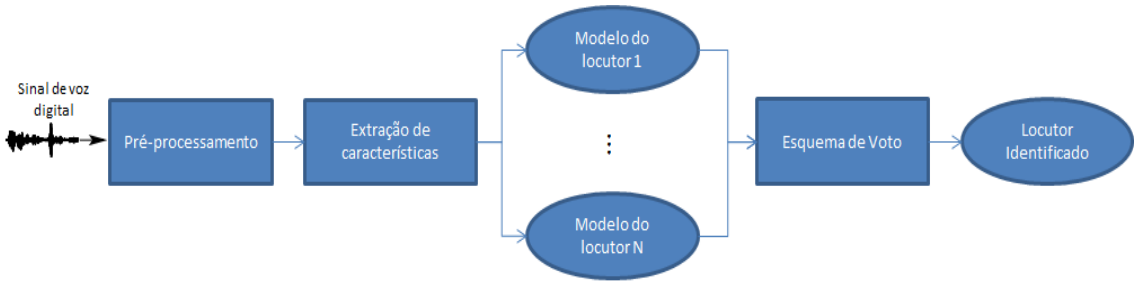

Figura 1: Representação de uma arquitetura básica de um sistema de Identificação de Locutor

A arquitetura será detalhada no capítulo 4, onde explicaremos como o sistema de Identificação de Locutor utilizado neste trabalho foi concebido.

3. SPEAKER DIARIZATION

Este capítulo irá apresentar uma descrição da tarefa *Speaker Diarization*, mostrando seus conceitos. Apresentará também uma arquitetura básica para desenvolvimento dessa aplicação.

3.1. CONCEITOS

Em geral, arquivos de áudio podem conter diversas informações, podem conter trechos com pessoas diferentes falando, trechos de música, silêncio, ruído, etc. A indexação do áudio seria o processo de atribuir a um determinado trecho do áudio uma informação, no sentido de categorizar esse trecho como música, fala, ou ruído. *Speaker Diarization* trata-se de uma tarefa de indexação do áudio, em que são adicionadas informações para marcar a mudança entre a fala dos locutores, segmentar o áudio separando as partes em que cada locutor fala para criar partições com esses trechos. É geralmente definida pela expressão *"quem fala quando"* [2].

O objetivo desta tarefa é particionar o áudio em regiões homogêneas de acordo com a identidade de cada locutor presente [23]. Ter o conhecimento de quando cada locutor está falando pode ser bastante útil como uma etapa de pré-processamento para sistemas que convertem fala em texto (speech-to-text systems – STT systems), melhorando a qualidade do texto produzido [24].

3.2. ARQUITETURA PARA REALIZAR A TAREFA SPEAKER DIARIZATION

Segundo Tranter et. al., em 2006 [2], para a realização da tarefa *Speaker Diarization*, são necessárias as seguintes etapas:

- Detecção de fala: Tem o objetivo de identificar as regiões de fala, as regiões que contiverem ruído, silêncio ou música.
- Detecção de mudança de locutor: Após a etapa de detecção de fala ter sido realizada, esta etapa irá procurar os instantes onde houver a mudança de locutores nas regiões de fala identificadas.
- Segmentação e criação de clusters: Esta etapa irá separar e juntar os trechos das regiões de fala, de acordo com a identidade de cada locutor, observando as mudanças na fala indexadas na etapa anterior. No final, cada locutor possuirá um cluster, contendo os segmentos com sua fala.

De acordo com as etapas citadas acima, a Figura 2 mostra uma arquitetura básica que foi montada para a construção de sistemas que realizam essa tarefa e a que foi utilizada na construção do sistema proposto.

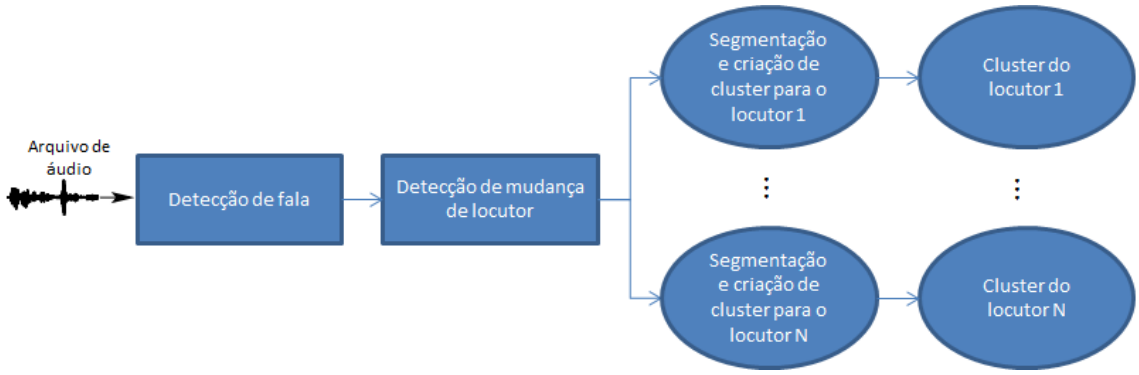

Figura 2: Representação de uma arquitetura básica para um sistema que realiza a tarefa *Speaker Diarization*

As técnicas utilizadas pelo sistema proposto para a implementação das etapas serão detalhadas no capítulo 4.

Outros tipos de arquitetura foram esboçados para a realização da tarefa *Speaker Diarization* nos trabalhos de Tranter et. al., em 2006 [2], Zhu et. al., em 2005 [23], Wooters et. al., em 2008 [24] e Gauvian et. al., em 1998 [25].

4. SISTEMA DE BUSCA DE ATORES EM FILMES

Este capítulo apresenta a descrição do sistema de busca de atores em filmes proposto. Esse sistema realiza a tarefa *Speaker Diarization* com o auxílio de um sistema de identificação de locutor. A escolha de utilizar um sistema de identificação de locutor para esses fins foi feita por ser mais simples de ser construída, se comparada com os trabalhos realizados por Tranter et. al., em 2006 [2], Zhu et. al., em 2005 [23], Wooters et. al., em 2008 [24] e Gauvian et. al., em 1998 [25] para a realização da mesma tarefa.

O sistema de identificação de locutor é peça chave para a construção do sistema de busca de atores em filmes, portanto, será mostrado primeiro como esse sistema foi desenvolvido neste trabalho, de acordo com a arquitetura apresentada na Figura 1, descrevendo as técnicas utilizadas para sua construção. Em seguida será mostrado como o sistema de busca foi desenvolvido, de acordo com a arquitetura apresentada na Figura 2, descrevendo o algoritmo feito para a realização da tarefa *Speaker Diarization*.

4.1. SISTEMA DE IDENTIFICAÇÃO DE LOCUTOR

4.1.1. Pré-processamento

A primeira etapa do sistema é o pré-processamento do sinal. Nesta etapa são aplicados técnicas e algoritmos que visam melhorar a qualidade do sinal ou visam acentuar determinadas características. O principal objetivo com isso é aperfeiçoar o processo posterior, que é o de extração de características, e como consequência melhorar o sistema como um todo. Nesta etapa, temos duas atividades em destaque:

- Filtros de ruído: Servem para eliminar os ruídos do sinal, dependendo dos tipos de ruídos que o sinal contém.
- Remoção de silêncio: Serão aplicados detectores de atividade de voz, com o objetivo de identificar a presença de voz no trecho analisado e descartar os trechos que não têm atividade de voz.

4.1.1.1. Detecção de Atividade de Voz

Os *frames* do sinal de voz que não possuírem informação útil ao sistema devem ser excluídos da análise para que não interfiram no desempenho final. Isso pode ocorrer quando os *frames* são sequencias de silêncio ou ruído. Para resolver este problema são aplicadas técnicas para detectar atividade de voz (*Voice Activity Detection – VAD*), verificando a presença de voz ou não em um determinado *frame*. Essas técnicas podem ser a detecção de energia ao longo do sinal, verificação do número de cruzamentos por zero, entre outras.

Os algoritmos VAD que utilizam técnicas de medição de energia são bem simples, porém, bastante sensíveis à presença de ruído [26]. A técnica baseada na energia utilizada neste trabalho foi desenvolvida por Rabiner, em 1974 [26].

A energia do sinal é calculada em uma janela do sinal, de duração de 10ms, a partir de medições centradas. É a soma das magnitudes de cada amostra neste intervalo, ou seja:

$$E = \sum_{i=n_0}^{i=-n_0} |s(n+i)|. \quad (1)$$

Para a utilização desta técnica é preciso definir um limiar para que seja possível classificar a energia medida como silêncio ou atividade de voz. A determinação desse limiar foi feita de maneira experimental neste trabalho. Em grande maioria das amostras colhidas, o limiar da energia entre o silêncio e uma atividade diferente de silêncio ficou no mínimo 10 decibéis, como mostra a Figura 3 e a Figura 4.

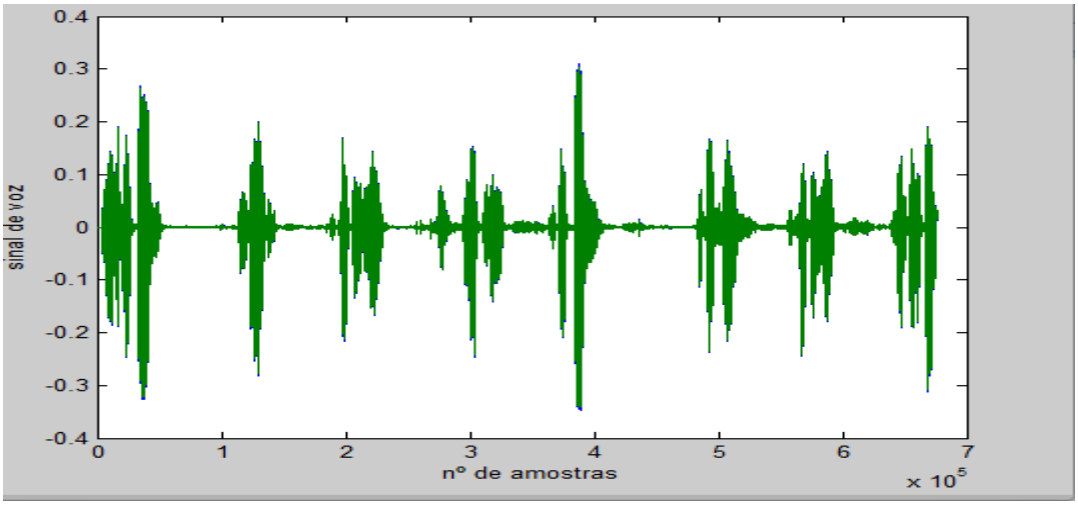

Figura 3: Sinal de voz.

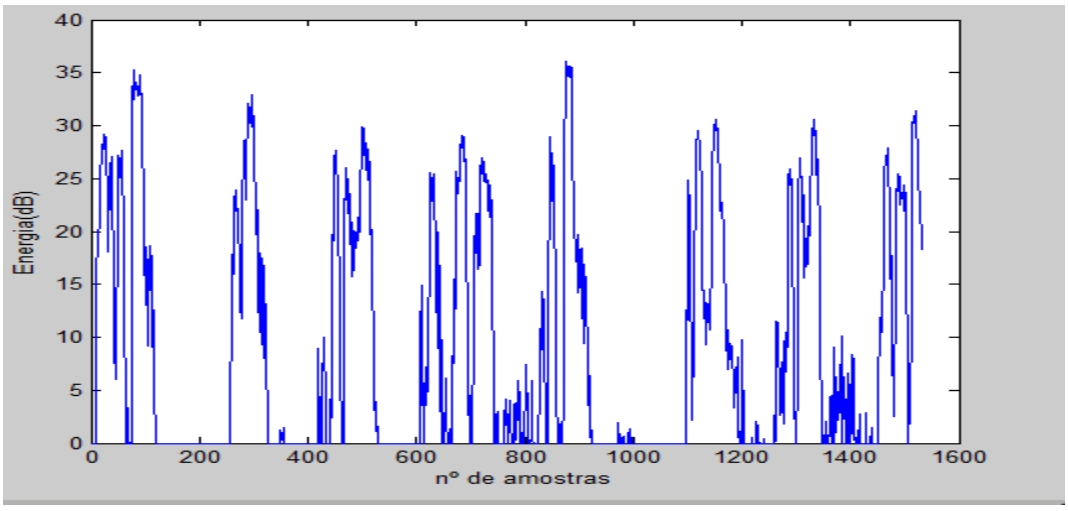

Figura 4: Distribuição de Energia do sinal de voz da Figura 3.

As partes do sinal que estão abaixo do limiar são eliminadas, considerando-se que não existe nenhuma atividade de voz.

Os algoritmos VAD baseados somente na medição de energia podem trazer bastantes problemas, caso a razão entre sinal e ruído (SNR) for muito baixa. A técnica de cruzamento pelo zero consiste em verificar a frequência de alternância entre os sinais positivos e negativos dentro da janela do sinal. O número de cruzamentos pelo zero encontra-se em um intervalo fixo de valores. Para cada janela de 10ms, o número de cruzamentos de um sinal puro de voz fica entre 10 e 20 vezes. Já o número de cruzamentos do sinal com ruído é imprevisível e aleatório, permitindo criar uma regra de decisão independente da energia.

A Figura 5 mostra a aplicação da técnica de cruzamento pelo zero no sinal de voz mostrado pela Figura 3.

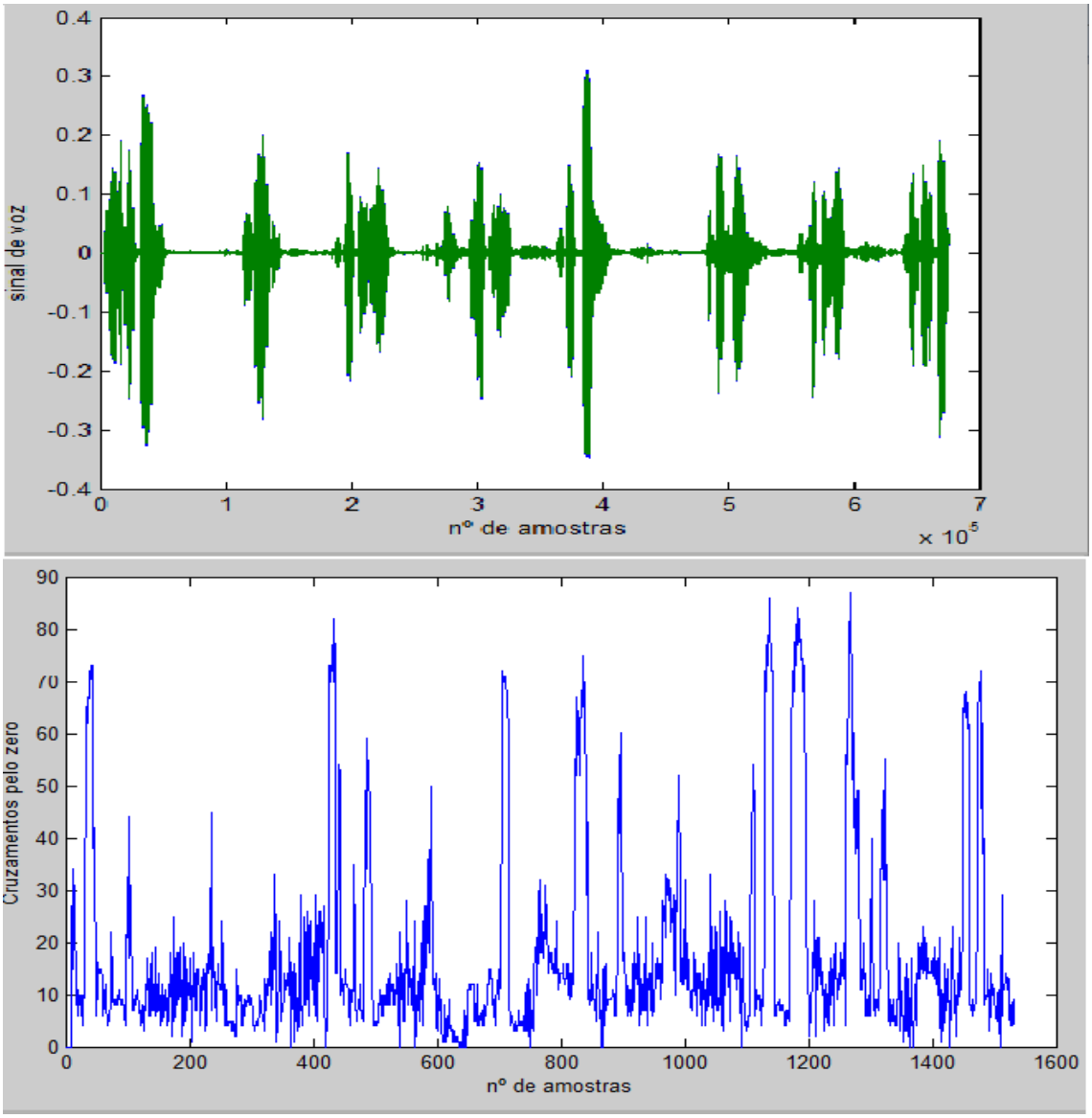

Figura 5: Número de cruzamentos pelo zero (Figura de baixo) do sinal de voz da Figura 3 (Figura de cima).

As partes do sinal que não estão fora do intervalo são eliminadas, considerando-se ruído.

Este trabalho é proposto um VAD que utiliza uma combinação entre as duas técnicas descritas, com o objetivo de aproveitar suas vantagens. Inicialmente o sinal é dividido em janelas. Serão calculados para cada janela a energia e o número de cruzamentos pelo zero. Caso o número de cruzamentos esteja dentro do intervalo [10,20] a janela é classificada como voz, caso esse número esteja fora do intervalo e a energia estiver acima do limiar definido, será considerado ruído, e caso este número esteja fora do intervalo e a energia abaixo do limiar, será considerado silêncio. Esse intervalo [10,20] foi escolhido de acordo com os estudos de Yamamoto et. al., em 2006 [27].

As janelas que forem consideradas ruído serão eliminadas, mas as que forem consideradas silêncio ainda serão analisadas com mais detalhe. O silêncio precisa ser removido, já que não faz parte da voz do locutor, mas, de acordo com Prasad et. al., em 2002 [28], a pausa entre a pronúncia de duas sílabas não precisa ser considerada como silêncio. As janelas consideradas como silêncio e classificadas como pausa não serão descartadas, caso contrário serão descartadas.

A Figura 6 mostra um sinal de voz antes e depois da utilização do VAD proposto.

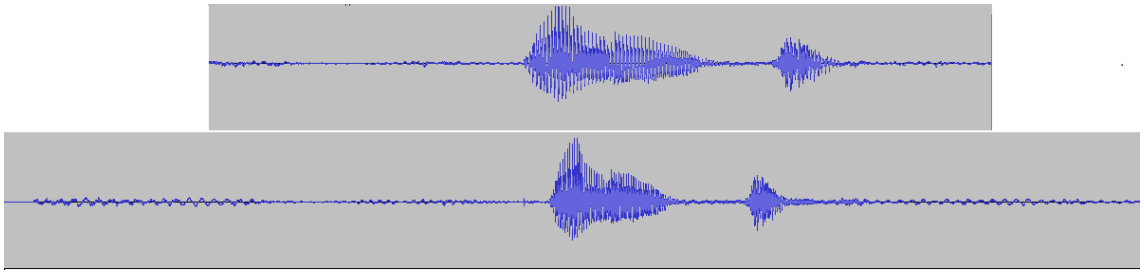

Figura 6: O sinal de voz original (acima) e o mesmo sinal após a utilização do VAD proposto (em baixo).

4.1.2. Extração de características

O processo de extração de características é responsável por transformar o sinal de voz em vetores de características, contendo informações únicas do trato vocal para cada locutor.

A técnica de extração de características utilizada neste trabalho foi a *Mel Frequency Cepstral Coefficients* (MFCC), proposta por Davis et. al., em 1988 [29]. MFCC trata-se de uma análise de características espectrais de tempo curto, analisando o sinal de voz em intervalos de 10ms a 30ms, baseando-se no uso do espectro da voz convertido para a escala *Mel*, uma escala representativa que procura aproximar-se de características únicas presentes no ouvido humano. Os coeficientes Mel-cepstrais surgiram de um estudo em psico-acústica, pois se percebeu que a percepção humana de frequência de tons puros ou de sinais de voz não segue uma escala linear. Essa

escala funciona da seguinte maneira: para cada tom com sua frequência em Hz, é atribuído um valor na escala *Mel*, sendo então uma medida de frequência. Realizaram-se vários experimentos para determinar o mapeamento de uma frequência Hz para Mel. Este mapeamento é definido como:

$$mel(f) = 1127 * \ln\left(1 + {}^{f}/_{700}\right). \quad (2)$$

Para a extração dos vetores de características MFCC, são necessárias as seguintes etapas: pré-ênfase, segmentação do sinal, janelamento, transformada de *Fourier*, aplicação de filtros triangulares, transformada discreta do cosseno.

Além das características espectrais de tempo curto (MFCC), também foram extraídas as características temporais do espectro, representando informações como energia, velocidade e a aceleração da fala. Para obter as informações de velocidade e aceleração é necessário estimar a primeira e segunda derivadas do vetor de características, conhecidas como coeficientes *delta* (Δ) e *delta* ao quadrado (Δ^2), respectivamente. Essas informações tornaram-se úteis para o processo de discriminação, de acordo com o trabalho de Soong et. al., em 1986 [14].

4.1.2.1. Pré-ênfase

A pré-ênfase é a primeira etapa para extrair os vetores MFCC. Ela é necessária para compensar a atenuação das componentes de alta frequência causada pelo mecanismo de produção da voz, mais especificamente os lábios e a glote. Isso faz com que as informações presentes nas altas frequências tenham uma menor energia do que nas frequências baixas.

Trata-se então de um filtro passa alta, criando um processo reverso ao da produção da voz. O filtro é do tipo

$$s_2(n) = s(n) - a * s(n-1), \quad (3)$$

onde $s_2(n)$ é o sinal de saída do filtro, $s(n)$ é o sinal original e a é um valor entre [0.9 e 1.0].

4.1.2.2. Segmentação do sinal

O sinal de voz possui uma natureza não estacionária, o que significa dizer que as componentes de frequência variam ao longo do tempo. Esse tipo de natureza não é bom para a determinação de padrões do sinal. Para resolver o problema dessa natureza utiliza-se a segmentação do sinal em *frames* menores, o suficiente para que características estacionárias comecem a aparecer. Quando essas características são observadas, é possível determinar padrões para o sinal analisado. Neste trabalho os

frames foram segmentados de acordo com Rabiner et. al., em 1993 [30], com duração de 20ms a 30ms com uma sobreposição entre *frames* cerca de ½ a 1/3 de sua duração.

4.1.2.3. Janela de Hamming

Após a etapa de segmentação, é necessário atenuar as descontinuidades causadas no início e no final do sinal após cada segmentação. Para isso aplica-se a técnica de janelamento, que consiste em multiplicar cada segmento do sinal por uma função janela, que nesse caso é a janela de Hamming:

$$w(n) = 0.54 - 0.46 * \left(\frac{2n\pi}{N-1}\right), \quad (4)$$

sendo N a quantidade de amostras do segmento.

4.1.2.4. Transformada de Fourier

Após a etapa de janelamento, é necessário converter cada *frame* para o domínio da frequência, para que seja possível observar as características espectrais. Então se utiliza a transformada de *Fourier* para realizar a conversão. Existe uma implementação bastante eficiente do algoritmo que faz a transformada, chamado de *Fast Fourier Transform* (FFT). Este algoritmo realiza a transformada Discreta de Fourier (DFT), sendo mais rápido e viável que a transformada comum.

4.1.2.5. Aplicação de filtros triangulares

Após converter cada *frame* para o domínio da frequência, é preciso convertê-los para a escala *Mel* de frequência. Essa conversão de escala pode ser obtida através da aplicação de uma sequência de filtros triangulares, com um espaçamento de acordo com a escala *Mel*, apresentada pela Equação (2).

Um sinal, quando submetido a um filtro triangular, tem as componentes que estão próximas ao centro deste filtro enfatizadas e as demais, atenuadas. Dessa forma, ao se empregar um banco de filtros o que está se fazendo é enfatizar as frequências *Mel*, assim escalando o espectro na escala *Mel*.

Para Rabiner et. al., em 1993 [30], são dois os principais motivos para a utilização de filtros triangulares para realizar a conversão de escala:

- Redução da quantidade de características envolvidas. O cálculo de conversão de escala envolve agora a energia de cada filtro, e não o sinal como um todo;
- Suavização da magnitude do espectro do sinal. Assim haverá a suavização da frequência fundamental (*pitch*) na extração de características, garantindo que o sistema de reconhecimento comporte-se de maneira parecida se ocorrer variações no timbre da voz ou entonação.

4.1.2.6. Transformada discreta do cosseno

Nessa etapa final, converte-se o logaritmo do espectro *Mel* de volta ao domínio do tempo. Como o logaritmo do espectro *Mel* fornece números reais, eles podem ser convertidos usando a transformada discreta do cosseno (*Discrete Cosine Transform* - DCT). Os valores dessas energias no domínio do tempo estão numa frequência chamada *quefrequency*, ou o domínio do tempo na escala *Mel*. A DCT pode ser calculada da seguinte forma:

$$X_k = \sum_{n=1}^{N} x_N * cos\left[\frac{\pi}{N} * \left(n + \frac{1}{2}\right) * k\right], \quad k = 0, \dots, L \tag{5}$$

onde L é o número de coeficientes MFCC que desejamos extrair e N o número de filtros triangulares utilizados na etapa anterior. Neste trabalho foram utilizados os valores $L = 12$ e $N = 20$.

No final teremos o vetor de características MFCC de tamanho L, que já pode ser usado pelo processo de modelagem.

4.1.2.7. Obtenção da energia do sinal

A energia do *frame* também é uma característica importante para a discriminação dos locutores e que é facilmente obtida. Assim, o log da energia do *frame* é adicionado na última posição do vetor de características MFCC, resultando em MFCC + energia.

4.1.2.8. Obtenção dos coeficientes delta

Através da combinação MFCC + energia, é possível extrair características temporais como velocidade e aceleração, os chamados *Cepstrum Delta*. Adicionando a velocidade ao vetor de características temos um total de 26 coeficientes. Acrescentando a aceleração, o vetor resultante terá 39 coeficientes. Este último vetor é utilizado pela maioria de sistemas de identificação/verificação de locutor, segundo Rabiner et. al., em 1993 [30] e Reynolds et. al., em 2000 [31].

4.1.3. Modelagem Utilizando GMM

Os modelos de misturas gaussianas (*Gaussian Mixtures Models – GMM*) foram primeiramente utilizados para reconhecimento de locutor por Reynolds, em 1992 [19]. Desde então é considerada referência para modelagem de locutores.

GMM representa de forma geral a dependência das características da voz associadas ao locutor, em conjunto com a capacidade de modelar densidades de probabilidades desconhecidas.

Em Reynolds et. al., em 2000 [31], as principais vantagens de se usar GMM como técnica de modelagem de locutores em sistemas de reconhecimento de locutor independentes são o seu baixo custo computacional, pela teoria estatística bem fundamentada e, principalmente, o fato de que não são sensíveis aos aspectos temporais do sinal de voz, representando com precisão os aspectos diretamente ligados ao locutor, refletindo a sua dependência com o seu trato vocal.

GMM é uma função de densidade de probabilidade parametrizada representada pela soma de pesos de M componentes Gaussianas, representada pela equação,

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{6}$$

onde $\mathbf{x}$ é vetor de características com dimensão D, $w_i,\ i = 1, \dots, M$, são os pesos das misturas, e $g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),\ i = 1, \dots, M$, são as densidades das componentes Gaussianas. Cada densidade de componente é uma função Gaussiana D-variada da forma,

$$g(\mathrm{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}\,|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)'^{\,\Sigma_i^{-1}}(\boldsymbol{x} - \boldsymbol{\mu}_i)\right\} \tag{7}$$

com vetor de média $\boldsymbol{\mu}_i$, e a matriz de covariância $\boldsymbol{\Sigma}_i$. Os pesos das misturas devem obedecer à condição: $\sum_{i=1}^{M} w_i = 1$.

O modelo completo da mistura gaussiana é parametrizado pelos vetores de média, as matrizes de covariância e os pesos das misturas de todas as componentes. Estes parâmetros são representados coletivamente pela notação,

$$\lambda = \{w_i,\ \boldsymbol{\mu}_i,\ \boldsymbol{\Sigma}_i\} \qquad i = 1, \dots, M \tag{8}$$

A Equação (8) pode sofrer diversas variações. As matrizes de covariância $\boldsymbol{\Sigma}_i$ podem ser representadas de forma completa, com todos os seus elementos, ou podem ser representadas apenas pela sua diagonal. Em geral é utilizada a representação pela diagonal, principalmente por razões de desempenho computacional, sendo bastante custoso calcular a inversa da matriz de covariância na Equação (7) e pelo desempenho do sistema não ser afetado de forma significativa [31]. Além da maneira de representar as matrizes de covariância, as mesmas podem ser de três tipos:

- Uma única matriz para cada componente gaussiana do modelo;
- Uma única matriz para todas as componentes gaussianas do modelo;
- Uma única matriz para todas as componentes gaussianas de todos os modelos.

Neste trabalho a representação das matrizes de covariância escolhida foi a representação pela diagonal e o tipo escolhido foi uma única matriz para cada componente gaussiana do modelo.

O treinamento de um GMM para sistemas de reconhecimento de locutor consiste em construir modelos que irão representar os locutores no sistema. Essa modelagem será feita através das características extraídas dos sinais de voz de treinamento. No treinamento serão estimados os parâmetros de cada modelo λ mostrados na Equação (8). Isso pode ser realizado através de um algoritmo que maximize a verossimilhança entre os dados de treinamento $X = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$,

$$\hat{\lambda} = arg \max_{\lambda} p(X|\lambda), \tag{9}$$

$$p(X|\lambda) = \prod_{t=1}^{T} p(\mathbf{x}_t\,|\lambda). \tag{10}$$

O algoritmo escolhido para o treinamento foi o *Expectation-Maximization* (EM).

4.1.3.1. Algoritmo EM

Esse é um algoritmo iterativo que se inicia com um modelo λ^0 e a cada iteração estima um novo modelo λ^{n+1} de forma que esteja mais correlacionado com o conjunto de observações X em comparação ao modelo da iteração anterior λ^n, observando-se a propriedade:

$$p(X|\lambda^{n+1}) \geq p(X|\lambda^n). \tag{11}$$

O algoritmo EM possui duas etapas. Na primeira etapa, E (*Expectation*), é calculada a verossimilhança entre o modelo atual e os dados de treinamento. A verossimilhança entre o modelo atual e os dados de treinamento X é calculada, para cada vetor de treinamento $\mathbf{x}_t$, como:

$$\Pr(i|\mathbf{x}_\mathrm{t}, \lambda) = \frac{w_i g(\mathbf{x}_t|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{k=1}^{M} w_k\, g(\mathbf{x}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}, \tag{12}$$

onde $i = 1, \dots, M$, representa cada uma das componentes gaussianas presentes no modelo.

Na segunda etapa, M (*Maximization*), é feita a atualização dos parâmetros do modelo. Essa etapa busca a construção de um modelo com uma maior similaridade com os dados de treinamento do que o modelo da iteração anterior apresentava. As equações que estimam o novo modelo $\bar{\lambda}$ a partir do modelo antigo λ, são as seguintes:

$$\bar{w}_i = \frac{1}{T}\sum_{t=1}^{T} \Pr(i|\mathbf{x}_\mathrm{t}, \lambda)\,; \tag{13}$$

$$\bar{\boldsymbol{\mu}}_i = \frac{\sum_{t=1}^{T} \Pr(i|\mathbf{x}_t, \lambda)\, \mathbf{x}_t}{\sum_{t=1}^{T} \Pr(i|\mathbf{x}_t, \lambda)}; \tag{14}$$

$$\bar{\boldsymbol{\Sigma}}_i = \frac{\sum_{t=1}^{T} \Pr(i|\mathbf{x}_t, \lambda)\, (\mathbf{x}_t - \bar{\boldsymbol{\mu}}_i)(\mathbf{x}_t - \bar{\boldsymbol{\mu}}_i)'}{\sum_{t=1}^{T} \Pr(i|\mathbf{x}_t, \lambda)}. \tag{15}$$

Visando melhorar desempenho computacional, são utilizados apenas os elementos da diagonal principal da matriz de covariância $\boldsymbol{\Sigma}_i$. Assim o cálculo para atualização da matriz de covariância fica:

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} \Pr(i|\mathbf{x}_t, \lambda)\, {x_t}^2}{\sum_{t=1}^{T} \Pr(i|\mathbf{x}_t, \lambda)} - \bar{\mu}_i^2, \tag{16}$$

onde σ_i^2, x_t e μ_i referem-se a elementos arbitrários dos vetores $\boldsymbol{\sigma}_i^2$, $\mathbf{x}_t$ e $\boldsymbol{\mu}_i$, respectivamente.

4.1.3.1.1. Inicialização

O algoritmo EM necessita de um modelo inicial λ^0 para que seja possível estimar um novo modelo. Logo o EM tem que ter uma inicialização (iteração 0) para a construção deste modelo inicial. Existem duas formas de inicialização:

- Aleatoriamente: As médias são inicializadas escolhendo-se vetores de características aleatórios, oriundos da base de treinamento. A inicialização da matriz de covariância é a matriz identidade. Enquanto que os pesos das misturas são inicializados uniformemente.
- Clusterização: As médias são inicializadas escolhendo-se os centros de cada grupo, onde a quantidade de grupos é igual à quantidade de componentes do GMM utilizado. A matriz de covariância é inicializada calculando-se a variância entre os dados e o centro de cada grupo. Enquanto que os pesos das misturas são inicializados uniformemente.

Neste trabalho as duas formas de inicialização foram avaliadas, e a segunda forma de inicialização, utilizando o algoritmo *k-means*, presente na *toolbox* da ferramenta *Matlab*, foi escolhida devido aos seus melhores resultados.

4.1.3.1.2. Condição de Parada

O EM é um algoritmo de maximização que visa o máximo local. Logo, podem existir duas condições de parada: quando o algoritmo atinge um número máximo de iterações, ou quando a diferença relativa do log da verossimilhança entre o modelo atual e o anterior atinge um determinado limiar, representando que a etapa de treinamento não consegue mais melhorar os parâmetros do modelo. A diferença relativa é expressa pela equação:

$$\frac{log\ p\left(X\middle|\bar{\lambda}\right) - log\ p(X|\lambda)}{log\ p(X|\lambda)} < \theta. \tag{17}$$

Escolhendo-se um limiar mais significativo, podemos reescrever a Equação (17) por:

$$\left|\frac{log\ p\left(X\middle|\bar{\lambda}\right) - log\ p(X|\lambda)}{log\ p(X|\lambda)} < \theta\right|. \tag{18}$$

4.1.4. Processo de Identificação de Locutor

Na identificação, deseja-se descobrir o modelo de locutor λ_s dentre o conjunto de locutores $\lambda = \{\lambda_1, \dots, \lambda_T\}$, mais verossímil a uma amostra de teste $Y = \{y_1, \dots, y_T\}$. Para isso é necessário calcular a máxima verossimilhança:

$$\hat{\lambda} = \arg\max_S p(\lambda_s|Y) = \arg\max_S \frac{p(Y|\lambda_s)p(\lambda_s)}{p(Y)}, \tag{19}$$

onde a segunda parcela da equação corresponde a regra de Bayes. Supondo que todos os locutores são igualmente prováveis $p(\lambda_s) = 1/S$ e que $p(Y)$ é a mesma para todos os locutores, temos:

$$\hat{\lambda} = \arg\max_S p(Y|\lambda_s). \tag{20}$$

Admitindo a independência entre os elementos da amostra de testes, usando o logaritmo para evitar problemas numéricos e realizando a normalização para que a duração da amostra de teste não afete a decisão, temos:

$$\hat{\lambda} = \arg\max_S \frac{1}{T}\sum_{t=1}^{T} log\ p(y_t|\lambda_s). \tag{21}$$

Por fim, o locutor que maximiza esta expressão, será o locutor identificado e será considerado correto se de fato o locutor que originou a amostra for o locutor identificado.

4.2. CONSTRUÇÃO DO SISTEMA DE BUSCA DE ATORES EM FILMES

Após a construção do sistema de identificação de locutor, será detalhada a construção do sistema de busca de atores de acordo com a Figura 2 e como o sistema de identificação se encaixa nessa arquitetura.

O sistema de busca de atores irá identificar quais atores estão presentes em uma determinada cena. A partir da identificação, os trechos em que cada ator está falando serão segmentados e armazenados em um cluster próprio para cada ator.

4.2.1. Detecção de fala

O sistema de busca preocupa-se apenas em identificar a fala de cada ator. Os trechos que contiverem ruídos ou intervalos de silêncio irão prejudicar as demais etapas do sistema, portanto, devem ser descartados.

Com isso, a primeira etapa do sistema de busca faz reuso da etapa de pré-processamento do sistema de identificação de locutor, pois para descartar os trechos de silêncio e ruído será utilizado o VAD.

Agora que o VAD será usado como etapa inicial do sistema de busca, a etapa de pré-processamento do sistema de identificação de locutor que utiliza o mesmo VAD não precisa ser executada.

4.2.2. Detecção de mudança de locutor

Após a remoção dos trechos de silêncio e ruído, começa a segunda e mais importante etapa do sistema: Detecção de mudança de locutor. Esta etapa irá realizar a busca pelos atores na cena que será analisada do filme, através da utilização do sistema de identificação do locutor para detectar as mudanças que ocorrem na cena e identificando o locutor presente antes de cada mudança. Para a realização deste feito, este trabalho propõe que a etapa de detecção de mudança de locutor seja feita através de duas etapas: Segmentação do sinal e Identificação de locutor.

Outras formas de detectar a mudança de locutor são exploradas nos trabalhos feitos por Reynolds, em 2002 [21], Zhu et. al. [23] e Gauvian et. al., 1998 [25].

4.2.2.1. Segmentação do sinal

Nesta etapa o sinal de voz será segmentado, preparando o sinal para a próxima etapa. Os segmentos são de média duração, de modo que cada segmento possa ser uma amostra para o sistema de identificação e que evite, na medida do possível, que dois ou mais locutores estejam presentes no mesmo segmento.

A duração de segmento que foi experimentada neste trabalho foi de 0,5 segundos, por ter sido utilizada nos trabalhos [2], [25] e [23].

Após o sinal ser segmentado, estará pronto para a etapa de Identificação de locutor.

4.2.2.2. Identificação de locutor

Nesta etapa será realizada de fato a detecção de mudança. O algoritmo proposto neste trabalho é o seguinte:

1. Para cada segmento criado, faça:
    2. Utilize o sistema de identificação de locutor construído para identificar quem é o locutor presente no segmento;
    3. Marque o segmento com o resultado da Equação (21) fornecido pelo sistema de identificação;
    4. Caso o segmento anterior ao atual esteja marcado com um resultado diferente, houve mudança de locutor. Registre a mudança de locutor como: $mudança = \begin{bmatrix} (ID\ do\ segmento\ anterior, ID\ do\ locutor); \\ (ID\ do\ segmento\ atual, ID\ do\ locutor) \end{bmatrix}$.

De acordo com o passo nº 4 do algoritmo, a detecção de mudança de locutor é caracterizada como dois segmentos consecutivos do sinal classificados com locutores diferentes.

4.2.3. Segmentação e criação de clusters

Nesta etapa, serão criados novos segmentos de acordo com os segmentos que foram marcados pela identificação de locutor e as mudanças que foram registradas. Um novo segmento é a sequência de segmentos marcados com o mesmo locutor até que ocorra uma mudança. Esta etapa também criará clusters para cada ator com a união dos novos segmentos criados.

Neste trabalho, os clusters irão armazenar os novos segmentos criados a cada cena do filme que será analisada.

5. EXPERIMENTOS E RESULTADOS

Neste capítulo será explicado como foi feita a confecção das bases de dados utilizadas para treinamento e testes do sistema de identificação de locutor e testes do sistema de busca de atores proposto, os experimentos que foram realizados, as métricas utilizadas para análise de desempenho, os resultados e uma discussão sobre os mesmos.

Os filmes que foram escolhidos para a realização deste trabalho foram: *Tropa de Elite* e *Tropa de Elite 2 – O Inimigo agora é outro*. Foram escolhidos três atores: Wagner Moura (presente em ambos os filmes), realizando o papel do *Capitão Roberto Nascimento*, Irandhir Santos (presente no 2º filme), realizando o papel do *Deputado Diogo Fraga* e o ator André Ramiro (presente em ambos os filmes), realizando o papel do *André Matias*.

Primeiramente serão mostradas as explicações deste capítulo acerca do sistema de identificação do locutor. Por último, serão mostradas as explicações para o sistema de busca de atores proposto.

5.1. SISTEMA DE IDENTIFICAÇÃO DE LOCUTOR

5.1.1. Base de dados

A base de dados foi dividida em uma base para treinamento e uma base para testes. Foram escolhidas cenas diversas dos filmes em que os atores estivessem presentes para colher amostras da voz de cada um deles. As amostras foram colhidas através de um microfone e depois convertidas para o formato digital *stereo* de 16bits e frequência de amostragem de 44,1 kHz.

As bases foram montadas em função da duração das amostras, de forma que, para treinamento cada ator possui cerca de 1 minuto e 30 segundos de amostras e para teste cada amostra tem duração em torno de 20 segundos, duas amostras para cada ator, em um total de seis amostras de teste. Mais especificamente, como mostram a Tabela 1 e a Tabela 2, temos:

Amostras/Atores	Wagner	Irandhir	André
Duração Total	1min 37s	1min 23s	1min 36s
Quantidade	6	9	14

Tabela 1: Características da base de dados para treinamento

Atores/Amostras	1ª	2ª
Wagner	20s	18s
Irandhir	12s	20s
André	21s	26s

Tabela 2: Características da base de dados para testes

Houve também uma preocupação, durante o processo de confecção das bases, a respeito das emoções expressas pelos atores durante as cenas do filme, visto que estados emocionais extremos (raiva, estresse, etc.) são fontes de queda de desempenho, segundo Campbell, em 1997 [22]. Com isso, foram escolhidas propositalmente algumas amostras para treinamento e testes, onde eles expressavam emoções de raiva e calma.

5.1.2. Experimentos

Foi necessária a realização de experimentos acerca desse sistema para escolher a configuração que obtivesse o maior desempenho, bem como confirmar se o sistema foi construído de maneira bem sucedida. A configuração do sistema é caracterizada como: a quantidade de coeficientes MFCC utilizada na extração de características, bem como sua natureza (espectrais, ou espectro-temporais) e a quantidade de componentes GMM utilizada na modelagem das vozes dos atores. Além das configurações, foi realizado um experimento sem a utilização do VAD, para poder observar a diferença de desempenho do sistema com e sem o VAD. Os mesmos experimentos realizados com o VAD são também realizados no sistema sem o VAD.

Para a etapa de extração de características, foram estudadas as seguintes configurações:

- MFCC com 12 coeficientes;
- MFCC com 12 coeficientes mais a Energia (representada por E), resultando em 13 coeficientes;
- MFCC com 12 coeficientes mais a Energia, mais as derivadas de 1ª e 2ª ordem (representadas por Δ e Δ^2, respectivamente), totalizando 39 coeficientes.

Para a modelagem dos atores utilizando GMM, foram estudadas as seguintes configurações:

- GMM com 8 componentes (ordem 8);
- GMM com 16 componentes (ordem 16);
- GMM com 32 componentes (ordem 32);
- GMM com 64 componentes (ordem 64).

5.1.3. Métricas

A métrica de avaliação utilizada foi a taxa de acerto, uma métrica simples que representada como:

$$taxa\ de\ acerto = \left(\frac{nº\ de\ amostras\ \ de\ teste\ identificadas\ corretamente}{nº\ total\ de\ amostras\ de\ teste}\right) * 100\% \quad (22)$$

5.1.4. Resultados

Após a realização dos experimentos, foram obtidos os seguintes resultados para o sistema de identificação de locutor sem o VAD, como mostra a Figura 7:

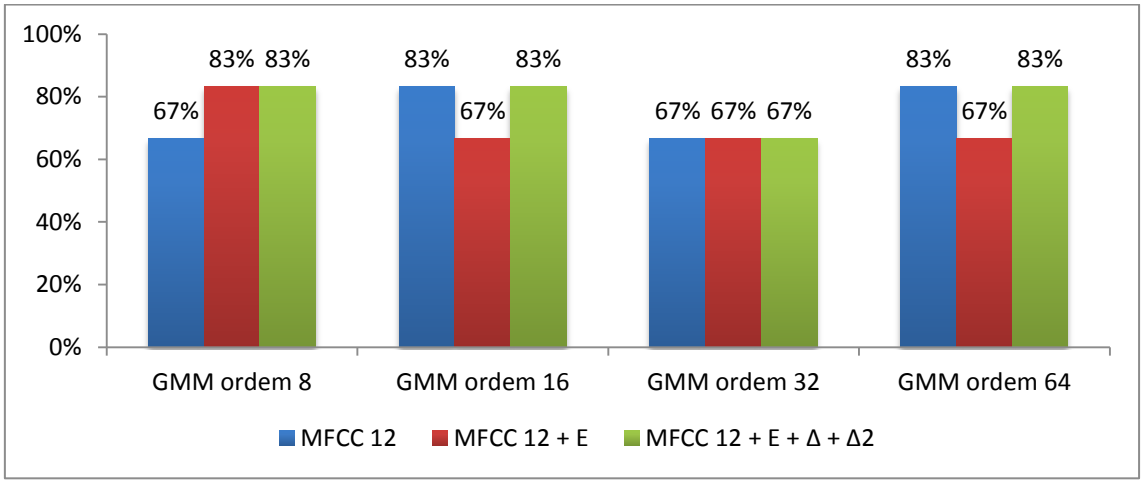

Figura 7: Resultados do sistema de identificação de locutor sem utilizar o VAD.

Após a realização dos mesmos experimentos, agora utilizando o VAD, foram obtidos os seguintes resultados, como mostra a Figura 8:

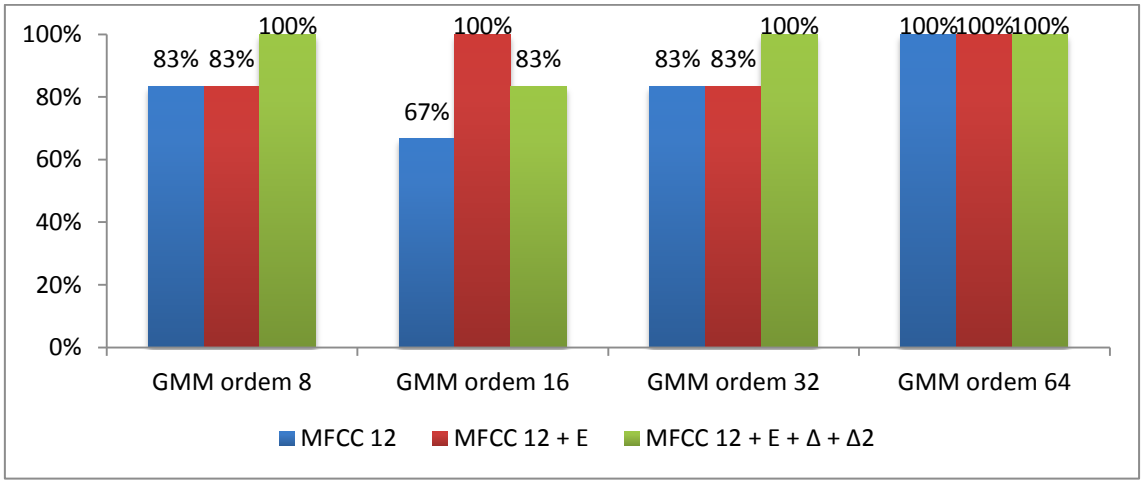

Figura 8: Resultados do sistema de identificação de locutor utilizando o VAD.

5.1.5. Discussão

Fica claro que através dos resultados que utilizar o VAD como uma etapa de pré-processamento aumenta significativamente o desempenho do sistema de identificação. Visto que, sem o seu uso, os GMM iriam modelar trechos de silêncio e ruído, o que não condiz com as vozes dos atores.

De acordo com os resultados, a configuração do sistema de identificação que apresenta o melhor desempenho é a utilização do VAD como etapa de pré-processamento, GMM com 64 componentes e quaisquer das três configurações de extração de características.

A configuração escolhida para a construção do sistema de identificação que será utilizado no sistema de busca de atores foi aquela que obteve maior desempenho e com a configuração para extração de características de MFCC com 39 coeficientes, por apresentar um conjunto mais rico de características do locutor que as demais configurações de extração. Portanto o sistema a ser construído terá as configurações:

- Utilização do VAD;
- GMM com 64 componentes;
- MFCC com 39 coeficientes.

5.2. SISTEMA DE BUSCA DE ATORES EM FILMES

5.2.1. Base de dados

Foram escolhidas algumas cenas dos filmes em que dois dos três atores estivessem presentes para colher amostras contendo as duas vozes. As amostras foram colhidas através de um microfone e depois convertidas para o formato digital *stereo* de 16bits e frequência de amostragem de 44,1 kHz.

As amostras possuem durações de acordo com as durações das cenas, o critério de durações semelhantes entre as amostras usadas na seção 5.1.1 não foi utilizado para confecção da base de dados do sistema de busca de atores. Em todas as amostras havia apenas dois dos atores escolhidos presentes, pois não existia nenhuma cena em que os três escolhidos estivessem presentes no 2º filme (lembrando que o ator Irandhir só está presente no 2º filme). Também não existiam cenas entre o ator Irandhir e o ator André, então, escolheram-se cenas em que o ator Wagner fala com os outros dois. O total de amostras colhidas foi quatro e as características mais específicas da base de dados são mostradas pela Tabela 3 e Tabela 4:

Atores (Presença na amostra)	Amostra 1	Amostra 2	Amostra 3	Amostra 4
Wagner	Sim	Sim	Sim	Sim
Irandhir	Não	Sim	Sim	Não
André	Sim	Não	Não	Sim

Tabela 3: Características da base de dados do sistema de busca de atores quanto à presença dos atores nas amostras.

Amostras	Amostra 1	Amostra 2	Amostra 3	Amostra 4
Duração	43s	30s	15s	39s

Tabela 4: Característica da base de dados do sistema de busca de atores quanto à duração das amostras.

5.2.2. Experimentos

Foi necessária a realização de experimentos acerca desse sistema de busca para análise de seu desempenho. Os experimentos feitos foram: Utilizar o sistema de busca com o sistema de identificação descrito na seção 5.2.5 sobre a base de dados para observar o grau de pureza dos clusters (métrica descrita na próxima seção) e a criação de um novo sistema de identificação de locutor utilizando como base de dados de treinamento os clusters e reutilizando a base de testes da seção 5.1.1. A respeito desse último experimento, ele seguirá o mesmo roteiro dos experimentos da seção 5.1.2, apenas utilizando o VAD, visto que o VAD já foi utilizado na construção dos clusters e, portanto, não haveria sentido construir um novo sistema de identificação sem ele.

5.2.3. Métricas

A métrica utilizada para análise de desempenho do sistema de busca segue a definição dos trabalhos de Gauvian et. al., 1998 [25], sendo como a percentagem de *frames* em um dado cluster vinda do representante do cluster. Segue a equação:

$$pureza = \left(\frac{nº\ frames\ identificados\ com\ o\ ID\ do\ cluster}{nº\ de\ frames\ do\ cluster}\right) * 100\% \qquad (23)$$

Para o experimento de criação de um novo sistema de identificação, será utilizada a métrica descrita na seção 5.1.3.

5.2.4. Resultados

Com a realização dos experimentos, foram obtidos os resultados descritos pela Figura 9 e Figura 10:

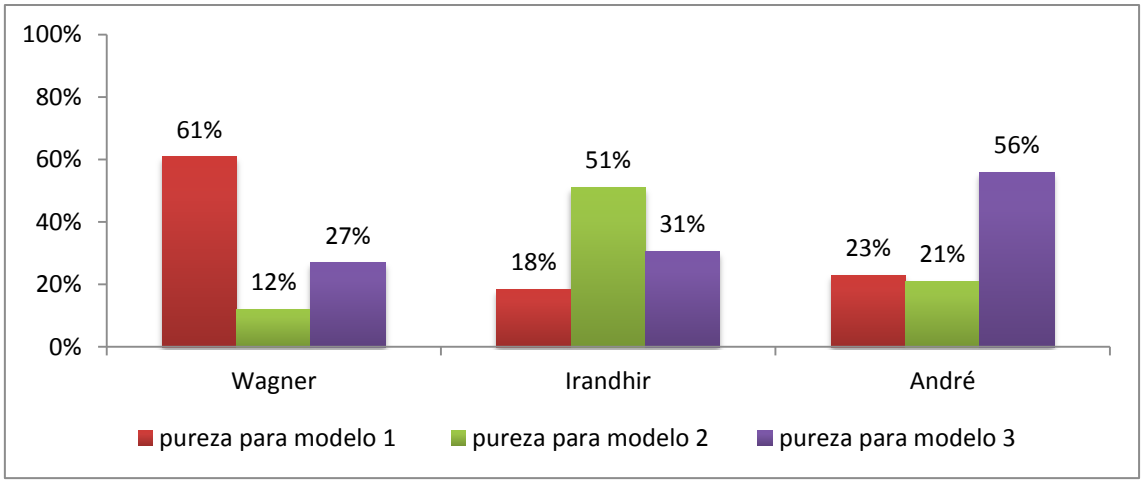

Figura 9: Resultados do sistema de busca em filmes utilizado sobre a base de dados.

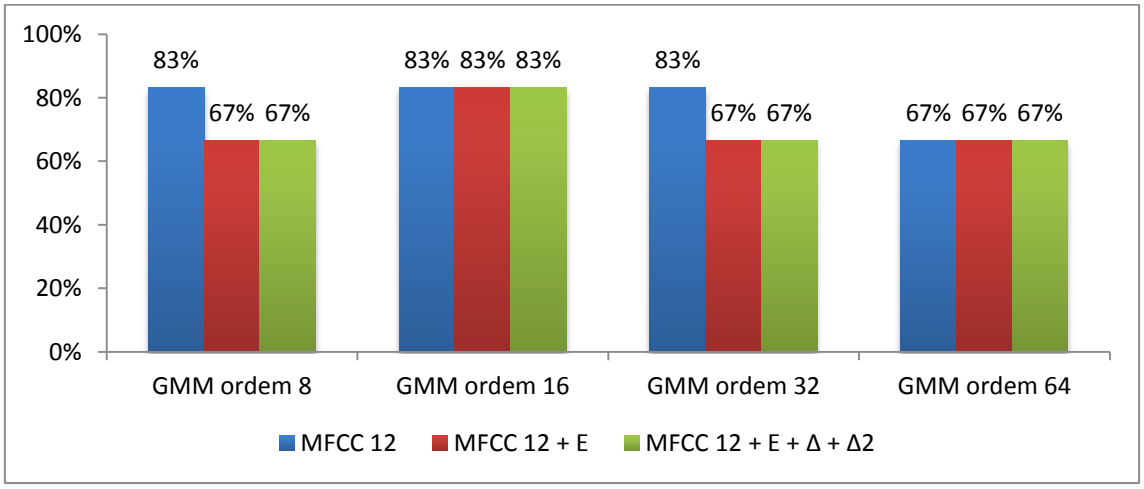

Figura 10: Resultados do novo sistema de identificação treinado com os clusters gerados pelo sistema de busca de atores.

5.2.5. Discussão

Como visto na Figura 9 o grau de pureza dos clusters foi consideravelmente ruim, pela tarefa *Speaker Diarization* ter sido realizada de uma maneira pouco robusta e bastante simples. Esse resultado ocorre justamente na etapa mais importante citada na seção 4.2.2: a detecção de mudança de locutor. Apesar da duração do segmento ser de 0,5 segundos, ainda é possível que haja dois locutores no mesmo segmento. Nesse caso o algoritmo de identificação proposto na seção 4.2.2.2 falha na classificação do segmento. Outro fato, observado na base de dados, é que existem alguns trechos nas amostras que os dois atores falam simultaneamente, prejudicando também na identificação do segmento, visto que não houve um pré-processamento para esse caso em especial.

6. CONCLUSÃO E TRABALHOS FUTUROS

Este capítulo encerra o trabalho desenvolvido ao longo desse período, que foi a proposta e construção de um sistema que busca atores em filmes. Este feito foi realizado através de muitas pesquisas sobre a área de Reconhecimento de Locutor e suas aplicações.

Foi apresentada uma descrição a cerca da arquitetura dos sistemas de identificação de locutor, utilizando técnicas como MFCC para extração de características, VAD para pré-processamento do sinal e GMM para modelagem e reconhecimento de padrões.

Foi apresentada uma descrição sobre a arquitetura de sistemas que realizam a tarefa *Speaker Diarization*, avaliada e discutida uma proposta de algoritmo para realizar esta tarefa.

A partir dos resultados obtidos, podemos concluir que o GMM é uma técnica bastante eficiente para modelagem de características de locutores e que apesar de os resultados apresentados pelo sistema de busca ter sido ruins, é uma proposta simples que tenta resolver uma tarefa com grau considerável de complexidade.

Como trabalhos futuros, fica a missão de melhorar a proposta de forma que:

- Melhore a etapa de detecção de mudança de locutor combinando o uso do sistema de identificação com alguns métodos propostos nos trabalhos [23] e [2];
- Acrescentar na etapa de pré-processamento a técnica de análise de componentes independentes (*Independent Component Analysis* - *ICA*) para separar falas simultâneas de dois ou mais atores, como proposta por Hyvärinen et. al., em 2000 [32].

REFERÊNCIAS

[1] D. A. Reynolds and R. C. Rose, "Speaker identification and verification using Gaussian mixture speaker models," *ELSEVIER - Speech Communication,* vol. 17, pp. 91-108, 1995.

[2] S. Tranter and D. Reynolds, "An Overview of Automatic Speaker Diarisation," *IEEE TRANS. ON SAP,* pp. 1-8, 2006.

[3] H. F. Hollien, “Forensic Voice Identification,” *Academic Press,* p. 204, 2001.

[4] L. Yount, "Forensic science: from fibers to fingerprints," *Infobase Publishing,* p. 206, 2006.

[5] L. G. Kersta, "Voiceprint Identification," *Nature,* pp. 1253-1257, 1996.

[6] R. Vanderslice and P. Landefoged, "The Voiceprinte Mystique," *UCLA Working Papers in Phonetics,* vol. 7, pp. 126-142.

[7] S. Furui, "50 Years of Progress in Speech and Speaker Recognition Research," in *SPECOM conference*, Patras, Greece, 2005.

[8] G. R. Dodington, J. L. Flanagan and R. C. Lummis, "Automatic Speaker Verification by Nonlinear Alignment of Acoustic Parameters". U.S. Patent 700.815.

[9] B. S. Atal, "Automatic Speaker Recognition Based on Pitch Contours," *Journal Acoustic Society of America,* vol. 52, pp. 1687-1697, 1972.

[10] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," *Journal Acoustic Society of America,* vol. 55, no. 6, pp. 1304-1312, 1974.

[11] B. S. Atal, "Automatic Recognition of Speaker From Their Voices," *Proceedings IEEE,* vol. 64, no. 460-475, p. 4, 1976.

[12] R. C. Lummis, "Speaker Verification by Computer Using Speech Intesity for Temporal Registration," *IEEE Trans. Audio Electroacoust.,* pp. 80-89, 1973.

[13] A. E. Rosemberg, "Automatic speaker verification: a review," *Proceedings of the IEEE,* vol. 64, no. 4, pp. 475-486.

[14] F. K. Soong, A. E. Rosemberg, L. R. Rabiner and B. H. Juang, "A vector quantization approach to speaker recognition," *Proc. IEEE Int. Conf. Acoust.,* pp. 387-390, 1986.

[15] A. E. Rosemberg, C. H. Lee and F. K. Soong, "Sub-word unit talker verification using hidden markov models," *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing,* vol. 1, pp. 269-272, 1990.

[16] J. J. Webb and E. L. Rissanen, "Speaker identification experiments using HMMs," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* vol. 2, pp. 387-390, 1993.

[17] C. Che and Q. Lin, "Speaker recognition using HMM with experiments on the YOHO database," in *Proc. EUROSPEECH*, Madrid, Italy, 1995.

[18] J. Colombi, D. Ruck, S. Rogers, M. Oxley and T. Anderson, "Cohort selection and word grammer effects for speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Atlanta, GA, 1996.

[19] D. A. Reynolds, Artist, *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification.* [Art]. Ph. D. Thesis, Georgia Institute of Technology, Department of Eletrical Engineering, 1992.

[20] D. A. Reynolds, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing,* vol. 3, no. 1, 1995.

[21] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology," *IEEE Transactions,* pp. 4072-4075, 2002.

[22] J. P. Campbell, "Speaker Recognition: A Tutorial," *Proceedings of the IEEE,* vol. 85, no. 9, pp. 1437-1462, 1997.

[23] X. Zhu, C. Barras, S. Meignier and J.-L. Gauvain, "Combining Speaker Identification and BIC for Speaker Diarization," in *INTERSPEECH*, Lisboa, Portugal, 2005.

[24] C. Wooters and M. Huijbregts, "The ICSI RT07s Speaker Diarization System," *Springer-Verlag,* pp. 509-519, 2008.

[25] J.-L. Gauvian, L. Lamel and G. Adda, "Partitioning and Transcription of Broadcast News Data," in *Proc. ICSLP*, Sydney, Dec 1998.

[26] L. R. Rabiner, "Algorithm for determining the endpoints of isolated utterances," *Journal of the Acoustical Society of America,* vol. 56, no. S1, 1974.

[27] K. J. Yamamoto, K. F. Reinhard and A. Kawamura, "Robust endpoint detection for speech recognition based on discriminative feature extraction," *Proc ICASSP,* vol. 1, pp. 805-808, 2006.

[28] V. Prasad, R. Sangwan and H. S. Jamadagni, "Comparison of voice activity detection algorithms for VoIP," in *Proceedings ISCC 2002 Seventh International Symposium on Computers and Communications*, 2002.

[29] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuosly Spoken Sentences," *IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING,* Vols. ASSP-28, no. 4, pp. 357-366, 1980.

[30] L. R. Rabiner and B. H. Juang, Fundamentals of speech recognition, PTR Prentice Hall, 1993.

[31] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing,* vol. 10, no. 1-3, pp. 19-41, 2000.

[32] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *PERGAMOM Neural Networks,* vol. 13, pp. 411-430, 2000.

[33] D. Gabriel, Artist, *Verificação de Locutor Utilizando Modelos de Misturas Gaussianas em Combinação com Lógica Difusa Tipo 2.* [Art]. UFPE, 2010.