UNIVERSIDADE DE CAXIAS DO SUL
Centro de Computação e Tecnologia da Informação
Curso de Bacharelado em Ciência da Computação

Rodrigo Bergamaschi de Azevedo

DIARIZAÇÃO AUTOMÁTICA DE LOCUTOR UTILIZANDO DISTÂNCIAS PROBABILÍSTICAS ENTRE MODELOS

Caxias do Sul

2010

Rodrigo Bergamaschi de Azevedo

DIARIZAÇÃO AUTOMÁTICA DE LOCUTOR UTILIZANDO DISTÂNCIAS PROBABILÍSTICAS ENTRE MODELOS

Trabalho de Conclusão de Curso para obtenção do Grau de Bacharel em Ciência da Computação da Universidade de Caxias do Sul.

André Gustavo Adami
Orientador

Caxias do Sul

2010

If more of us valued food and cheer and song above hoarded gold, it would be a merrier world.

— J. R. R. Tolkien

DEDICATÓRIA

Aos meus pais, Antonio e Flavia

AGRADECIMENTOS

Muitas pessoas caminharam ao meu lado nestes sete anos de faculdade. Sou extremamente grato a todos que, de uma forma ou de outra, colaboraram no meu crescimento, no meu amadurecimento e na minha formação.

Em primeiro lugar, agradeço aos meus pais, Antonio e Flavia. Agradeço pelo incentivo, pela compreensão, pela paciência, pelo amor e, acima de tudo, por acreditarem em mim e me apoiarem em todos os momentos.

Agradeço à minha namorada Amanda pelo carinho, pela dedicação, pelas inúmeras ofertas de ajuda, pelas palavras de conforto e pelo abraço amigo com que me recebeu sempre que a nuvem da dificuldade pairou sobre mim.

Sou grato aos meus amigos, especialmente Maurício – que esteve ao meu lado e me apoiou durante esta jornada – e Jorge – por todas as conversas, conselhos e palavras amigas –, pelos tantos momentos bons e por tudo o que me ensinaram. Longe ou perto, sei que torcem por mim e que comemoram minhas conquistas.

Agradeço também aos colegas com os quais me reuni para estudar e fazer trabalhos – aqueles fins de semana discutindo códigos e coroados com “salchipão” deixarão saudades.

Também sou grato a André Adami, meu orientador, que me auxiliou e aconselhou na feitura deste trabalho, e aos professores que souberam transmitir seu conhecimento e seus valores nestes anos de faculdade.

RESUMO

Este trabalho analisa a diarização de locutor, importante processo prévio a tarefas como reconhecimento de locutor e de voz e tarefas de indexação. O objetivo da diarização de locutor é obter segmentos de fala de apenas um locutor. Esses segmentos são então agrupados em conjuntos, de forma que cada conjunto contenha fala de somente um locutor.

Essa implementação aborda a diarização de dois locutores, em que existem somente dois locutores no áudio. Para a realização da tarefa, assume-se que não há conhecimento prévio dos locutores e que esses locutores não falam simultaneamente. Para a etapa de detecção de mudança de locutor é utilizado o método DISTBIC. O agrupamento dos segmentos é feito com base na distância Kullback Leibler.

Os resultados obtidos são avaliados por um programa disponibilizado pelo NIST, o Instituto Nacional de Padrões e Tecnologia dos Estados Unidos, para a tarefa de diarização de locutor.

Palavras-chave: Diarização de Locutor, Detecção de Mudança de Locutor, Segmentação de Locutor, Agrupamento, DISTBIC

ABSTRACT

This work analyses speaker diarization, important process prior to tasks such as speaker and speech recognition and indexing tasks. Speaker diarization aims to obtain segments of speech of a single speaker from audio data. These segments are then clustered so that each cluster will contain speech from only one speaker.

This implementation addresses the two-speaker diarization, where an audio contains only two speakers. To achieve this task it is assumed there is no previous knowledge about the speakers and the speakers do not speak simultaneously. DISTBIC method is used in speaker change detection step. The distance measure employed in the clustering step is Kullback Leibler.

The results are evaluated by a program provided by NIST – the National Institute of Standards and Technology – to the speaker change detection task.

Keywords: Speaker Diarization, Speaker Change Detection, Speaker Segmentation, Clustering, DISTBIC

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 Introdução

As bases de dados de áudio vêm crescendo e seu acesso vem se tornando cada vez mais fácil. Esse crescimento implica na necessidade de métodos eficientes e precisos para classificar, buscar e recuperar informações desses arquivos (JOHNSON; WOODLAND, 2000). O armazenamento de grandes quantidades de áudio é facilitado pela diminuição do custo e aumento do acesso ao poder de processamento, capacidade de armazenamento e largura de banda de rede (TRANTER; REYNOLDS, 2006).

Apesar da facilidade que se tem hoje em dia para armazenar áudio, seu acesso é mais difícil que o acesso a documentos de texto, pois a audição dos arquivos de áudio é mais lenta que a leitura dos documentos de texto. Além disso, o áudio não proporciona a vantagem da procura por palavras-chave. É muito mais fácil acessar determinada parte de um texto do que encontrar um trecho específico em uma gravação de áudio sem ter de percorrer o arquivo do início ao fim (OLSEN, 1995).

Além da mensagem falada, o áudio pode conter diversas informações importantes. Além do conteúdo da fala, a análise do áudio permite a obtenção de informações sobre o locutor e sobre a dinâmica da gravação. Quanto ao locutor, pode-se descobrir, por exemplo, seu sexo, sua escolaridade, devido à correção e diversidade do vocabulário empregado, seu estado emocional, a língua que está falando e seu sotaque. Já no que diz respeito à dinâmica da gravação, pode-se saber os períodos em que há fala, a que locutor a fala pertence, o assunto que está em pauta e a relação entre os locutores. Naturalmente, o ser humano é capaz de identificar tais características no áudio com certo grau de facilidade. Entretanto, sua capacidade de processamento é bastante limitada.

Atualmente é possível, através de computadores, processar muito mais informações em muito menos tempo que o cérebro humano. Dessa forma pode-se, via de programas de computador, extrair informações do áudio com relativa precisão em um curto espaço de tempo.

O objeto de interesse deste trabalho é a diarização automática de locutor, cujo objetivo é estimar os períodos de tempo em que cada locutor está falando em determinada conversa (ADAMI et al, 2002). Assume-se geralmente que não há conhecimento prévio do número de locutores presentes na conversa, de forma que não é possível construir um modelo de locutor antes da realização da diarização. Informações não relevantes, tais como ruído e música, dificultam o processo, pois interferem no processo de caracterização dos locutores.

A diarização de locutor é de extrema importância nas tarefas de reconhecimento de voz e reconhecimento de locutor, pois os segmentos obtidos são utilizados para estimar os modelos de locutor, de forma que a diarização incorreta resulta na obtenção de modelos também incorretos (ADAMI et al, 2002). Os segmentos podem ser usados para treinar modelos que permitam localizar a fala de determinado locutor em um arquivo de áudio (MORARU et al, 2003), bem como para indexar arquivos contendo dados de áudio, facilitando a busca dessa informação (DELACOURT; WELLEKENS, 2000). Outras aplicações que usam os segmentos obtidos a partir da diarização de locutor são a legenda automática de filmes e auxílio ativo para pessoas com deficiência auditiva. O rastreamento de locutor, por exemplo, permite a mudança de cor dinâmica das legendas com base na identidade do locutor (PIETQUIN et al, 2001). As mudanças de locutor detectadas pela diarização em arquivos de áudio são frequentemente utilizadas para estruturar o documento para navegação pelos ouvintes. Um exemplo são os telejornais, nos quais as mudanças de locutor geralmente coincidem com mudanças ou transições entre matérias. Outros exemplos são painéis, apresentações e gravações de encontros, em que a organização dos segmentos de áudio de acordo com a identidade dos locutores pode facilitar a navegação para os ouvintes. Outra motivação para a diarização de locutor é a transcrição automática da fala. A execução do reconhecimento automático de fala pode se beneficiar da adaptação de locutor em muitos casos, independente de ser inspecionada. Embora não seja um pré-requisito estrito para a adaptação de locutor, a detecção de mudança de locutor á importante para executar a adaptação sobre dados com múltiplos locutores, uma vez que pode prover o reconhecedor com dados homogêneos destes (PARK; GLASS, 2006).

1.1 Objetivo

O objetivo deste trabalho é desenvolver e avaliar as etapas de detecção de mudança de locutor (baseado no método DISTBIC) e agrupamento para um sistema de diarização de locutor. O sistema utilizará áudio contendo somente dois locutores e será avaliado utilizando o paradigma de avaliação do Instituto Nacional de Padrões e Tecnologia (NIST – *National Institute of Standards and Technology*) para as competições de detecção de mudança de locutor (NIST, 2002). Os resultados serão refinados através do agrupamento dos segmentos obtidos.

1.2 Estrutura do Trabalho

O trabalho está estruturado da seguinte forma: o Capítulo 2 aborda a diarização de locutor, analisando as fases de sua estrutura. O Capítulo 3 trata da configuração experimental: as avaliações do Instituto Nacional de Padrões e Tecnologia, as bases de dados utilizadas, a avaliação do desempenho do algoritmo implementado e o sistema de referência para tal avaliação. O Capítulo 4 apresenta a detecção de mudança de locutor baseada no Critério de Informação Bayesiano, o método DISTBIC e seus resultados. O Capítulo 5 é dedicado ao estudo do agrupamento dos segmentos. A conclusão do trabalho é apresentada no Capítulo 6.

2 Diarização de Locutor

Este capítulo é dedicado à estrutura da diarização de locutor, processo no qual a tarefa de detecção automática de mudança de locutor está inserida. O processo é explicado e as suas partes são enumeradas na seção 2.1. A Seção 2.2 trata da extração das características do sinal de áudio. A detecção de fala é analisada na Seção 2.3. Na Seção 2.4 é apresentada a detecção de mudança de locutor. O agrupamento dos segmentos, que é a etapa final de uma parte dos sistemas, é descrito na Seção 2.5. A Seção 2.6 aborda a re-segmentação, etapa posterior ao agrupamento dos segmentos em alguns sistemas.

2.1 Processo de Diarização de Locutor

O processo de diarização de locutor pode ser dividido em cinco etapas: extração de características, detecção de fala, detecção de mudança de locutor, agrupamento e re-segmentação (como ilustrado na Figura 1). A meta da extração de características é coletar os atributos do sinal (i. e. as características fisiológicas do locutor, calculadas sobre a amplitude e a freqüência do sinal) em um vetor. O objetivo da detecção de fala é encontrar as regiões do sinal de áudio que contêm fala. Na detecção de mudança de locutor são encontrados pontos no sinal de áudio em que é provável que tenha havido mudança de locutor. A etapa de agrupamento tem por meta agrupar os segmentos obtidos na etapa anterior que pertencem a um mesmo locutor em um conjunto. A finalidade da re-segmentação é refinar as fronteiras originais dos segmentos (TRANTER; REYNOLDS, 2006).

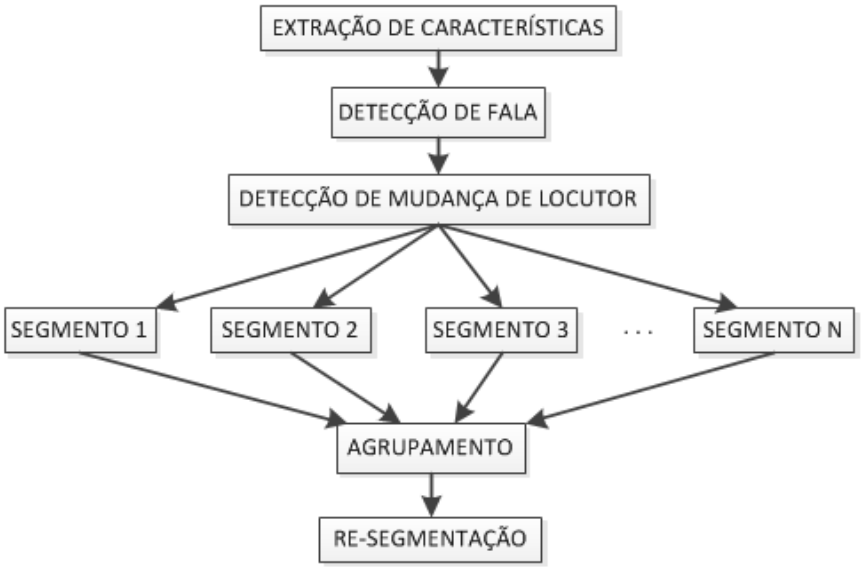

Figura 1. Estrutura de um sistema de Diarização de Locutor.

2.2 Extração de Características

A extração de características é o processo através do qual os atributos do sinal são calculados e coletados no formato de um compacto vetor, denominado vetor de características. Esse vetor de características pode ter mais de uma dimensão, sendo então uma matriz de características, em que cada ocorrência do sinal é um vetor cujo tamanho é o número de dimensões da matriz. Esse processo pode ser considerado uma compressão de dados que remove informações irrelevantes do sinal ao mesmo tempo que preserva suas informações relevantes (KIL; SHIN, 1996).

Para realizar a extração das características é necessário capturar dos espaços de projeção os atributos do sinal que são mais favoráveis à discriminação de classes e à compactação da energia que o espaço original. O ideal é que se utilize o menor número possível de *bits* para transportar a mensagem completa do áudio original com a menor taxa de erro possível (KIL; SHIN, 1996).

A idéia fundamental da transformação é que, uma vez que o sinal desejado será confinado a um pequeno número de funções de base, interferências e ruídos serão distribuídos uniformemente em diversas funções de base ou ocuparão diversas funções de base. Essa propriedade leva à filtragem de subespaço e/ou compactação da energia, que, por sua vez, leva a uma melhoria na razão do sinal para o ruído. A filtragem do subespaço é possível porque o sinal desejado e a interferência residem em diferentes subespaços no espaço de projeção (KIL; SHIN, 1996).

A Figura 2 mostra o deslocamento da janela ao longo do sinal de fala (unidimensional). As setas representam os vetores (uni ou multidimensionais) que são extraídos a cada deslocamento. O conteúdo dos vetores é a representação das características fisiológicas do locutor (extraídas com base na freqüência e na amplitude do sinal).

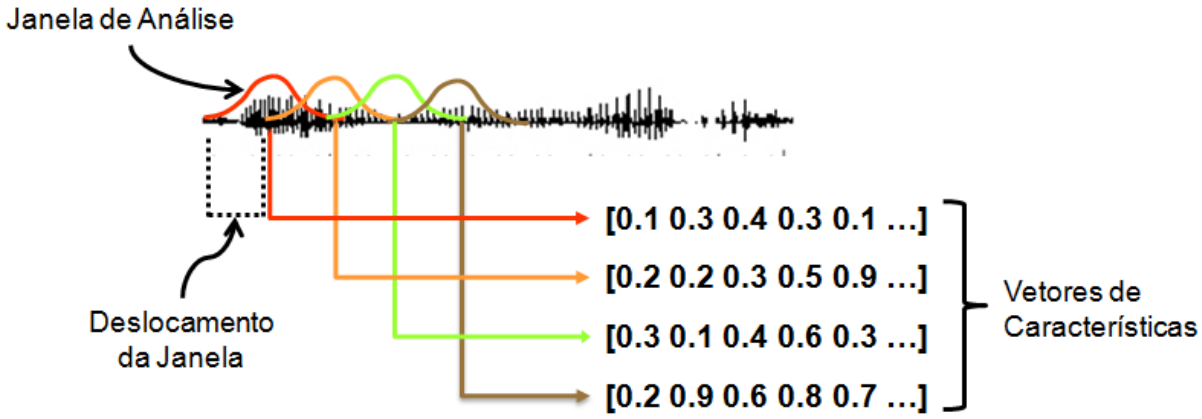

Figura 2. Extração de Características.

2.3 Detecção de Fala

A etapa de detecção de fala (também chamada detecção de atividade de fala e detecção de atividade de voz) tem como objetivo encontrar as regiões do sinal de áudio que contêm fala. Dependendo do tipo de dados que é utilizado, as regiões sem fala que devem ser descartadas podem incluir diversos fenômenos acústicos tais quais silêncio, ruído de fundo, ruído ambiente, música ou fala cruzada (TRANTER; REYNOLDS, 2006).

Para realizar a detecção de fala geralmente é utilizada uma abordagem baseada na detecção da energia presente no sinal. Essa técnica foi desenvolvida na telefonia para aproveitar o tempo ocioso das chamadas. Através da detecção de atividade de fala, foi possível aproximadamente dobrar a usabilidade dos longos e caros canais (como o sistema de cabos no fundo do mar) (BULLINGTON; FRASER, 1958).

A idéia por trás da técnica é que, em uma conversa entre duas pessoas, cada uma fala por apenas metade do tempo. O tempo ocioso das chamadas poderia então ser utilizado para interpolar outras ligações (BULLINGTON; FRASER, 1958).

Na detecção de atividade de voz baseada em energia a energia do sinal é monitorada e comparada com um limiar pré-definido. Assume-se que a energia de todo o sinal na presença de fala é suficientemente maior que a energia do ruído de fundo, de forma que as regiões em que de fato há atividade de fala possam ser detectadas. O valor pré-definido do limiar é recalculado para cada segmento analisado caso o nível de ruído seja variável (TANYER; ÖZER, 2000).

A sensibilidade do detector é um ponto crucial, pois o que determina se está havendo fala ou não na chamada é a energia estar acima ou abaixo de um limiar, respectivamente. A sensibilidade do detector é aumentada com a diminuição do limiar. Entretanto, a diminuição

do limiar implica no aumento da atividade de fala, de forma que um limiar muito baixo reduz a utilidade do detector de atividade de voz (BULLINGTON; FRASER, 1958).

Quando a fase de detecção de fala é primeiramente executada em um sistema ou a saída é requerida para outro procedimento (e. g. para transcrição), é mais importante minimizar a taxa de perda de detecção (i. e. não detectar mudanças de locutor que de fato ocorreram) do que a taxa de falso alarme (i. e. criar hipóteses de mudanças de locutor onde tais mudanças não ocorreram), uma vez que a primeira constitui erro irreparável na maioria dos sistemas. Todavia, a taxa de erro da diarização, que é usada para avaliar o desempenho da diarização de locutor, trata igualmente das duas formas de erro (TRANTER; REYNOLDS, 2006).

2.4 Detecção de Mudança de Locutor

A etapa seguinte à detecção de atividade de fala é a detecção de mudança de locutor, que tem como meta encontrar pontos no sinal de áudio em que é provável que tenha havido mudança de locutor. Se a entrada é um sinal de áudio não segmentado, então a etapa é destinada a detectar tanto mudanças de locutor quanto a presença ou não de fala. Se um detector de fala ou um classificador de gênero/largura de banda já tiver sido executado, então o detector de mudança de locutor busca apenas pelos pontos candidato a mudança de locutor dentro de cada segmento de fala.

Dentre as abordagens que utilizam o cálculo da distância entre segmentos adjacentes de dados para estimar se estes pertencem a um mesmo locutor ou a locutores diferentes, duas merecem destaque. As diferenças entre elas estão na escolha da distância utilizada e em decisões de limiar (TRANTER; REYNOLDS, 2006).

A primeira dessas abordagens é uma variação do Critério de Informação Bayesiano, introduzido por Chen e Gopalakrishnan (1998), e foi usada por Reynolds e Torres-Carrasquillo (2004). Essa técnica busca por pontos de mudança dentro de um segmento usando um teste da razão da verossimilhança penalizada, verificando se os dados no segmento são mais bem modelados por uma única distribuição (i. e. não houve mudança de locutor) ou por duas distribuições diferentes (i. e. houve mudança de locutor). Se um ponto candidato a mudança de locutor é encontrado, então a janela que representa o segmento é reajustada para aquele ponto e a busca é reiniciada. Se nenhum ponto candidato a mudança é encontrado, a janela é deslocada e a busca é refeita. A detecção de mudança de locutor através do Critério de Informação Bayesiano tem alta taxa de perda de detecções sobre segmentos pequenos

(menores que 2-5 segundos), o que pode se revelar um problema em áudio que contenha muitas mudanças de locutor, como conversações. Outro problema é o custo computacional da busca completa (que tem ordem N^2), o que leva a algumas implementações com formas de reduções computacionais (TRANTER; REYNOLDS, 2006).

A segunda abordagem utiliza janelas de tamanho fixo e representa cada janela por uma Gaussiana e a distância entre elas pela Divergência Gaussiana (a distância Kullback Leibler simétrica – KL2) (SIEGLER et al, 1997). Foi proposta também uma implementação passo-a-passo similar, mas que usa a taxa de verossimilhança logarítmica generalizada como distância (MEIGNIER et al, 2006). Os picos na função de distância são então encontrados e definem os pontos candidatos a mudança de locutor se seu valor absoluto excede um limiar pré-definido. É possível prevenir o sistema de gerar mais mudanças do que realmente há nos limites verdadeiros suavizando a distribuição das distâncias ou eliminando os menores dentre um conjunto de picos em determinada duração mínima. O uso de Gaussianas simples é preferido aos Modelos de Misturas Gaussianas devido à simplificação nos cálculos de distância. Os segmentos geralmente têm tamanho de 1-2 segundos quando uma Gaussiana de covariância diagonal é usada, ou 2-5 segundos quando uma Gaussiana de covariância completa é utilizada. Da mesma forma que com o Critério de Informação Bayesiano, o tamanho da janela restringe a detecção de mudanças de locutor rápidas.

Uma vez que a detecção de pontos candidatos a mudança de locutor frequentemente provê apenas uma segmentação inicial básica para sistemas de diarização de locutor (cujos segmentos serão agrupados e, em muitos casos, haverá re-segmentação), geralmente é mais importante estar habilitado a executar a detecção de pontos de mudança de forma bem rápida do que a perda de desempenho (TRANTER; REYNOLDS, 2006). De fato, existem sistemas que não têm perda de desempenho significativa quando é feita uma segmentação uniforme inicial (WOOTERS et al, 2004; MEIGNIER et al, 2006).

As duas técnicas de detecção requerem um limiar de detecção para serem ajustadas empiricamente para mudanças de tipo de áudio e de características. O ajuste do detector de mudanças é um balanceamento entre o desejo de se obter segmentos longos e puros para auxiliar na inicialização da fase de agrupamento e a minimização das perdas de pontos de mudanças que produzem contaminações no agrupamento.

Em adição (ou alternativamente), um passo de decodificação de palavras ou de fonemas com regras heurísticas pode ser usado para auxiliar no encontro de supostos pontos em que exista mudança de locutor. Entretanto, essa aproximação segmenta demais os dados

de fala e requer combinação e agrupamento adicional para formar segmentos de fala viáveis, e limites em rápidas intercalações de locutor em que não há silêncio ou mudança de gênero entre os locutores podem ser perdidos (TRANTER; REYNOLDS, 2006).

2.5 Agrupamento

Após a detecção de mudança de locutor é realizado o agrupamento, cujo objetivo é agrupar os segmentos que pertencem a um mesmo locutor em um conjunto. O ideal é que se obtenha um conjunto para cada locutor presente na conversa. A aproximação mais utilizada para realizar o agrupamento é o agrupamento hierárquico aglomerativo com critério de parada baseado no Critério de Informação Bayesiano (TRANTER; REYNOLDS, 2006). Agrupamento hierárquico é aquele em que os segmentos de fala são iterativamente separados ou combinados até que o número desejado de conjuntos seja obtido. O termo "aglomerativo" se refere à forma de agrupamento em que o número de segmentos de fala é maior que o número de conjuntos desejado, de forma que os segmentos vão sendo combinados até que o número de conjuntos desejado seja alcançado (MIRÓ, 2006).

Os passos do processo de agrupamento são os seguintes:

1. Os segmentos-folha da árvore são inicializados com segmentos de fala.
2. As distâncias entre cada par de segmentos são calculadas.
3. Os dois segmentos mais próximos são combinados.
4. As distâncias dos segmentos restantes são atualizadas em relação ao novo agrupamento.

Os passos 2-4 são repetidos até que o critério de parada seja alcançado (TRANTER; REYNOLDS, 2006).

O critério de parada é um ponto crítico para o desempenho da aplicação, independente da técnica de agrupamento utilizada, e depende de qual será o uso do resultado. Um sub-agrupamento pode produzir um número elevado de conjuntos, ao passo que um agrupamento excessivo pode produzir conjuntos contaminados (isto é, que contenham fala de mais de um locutor). Nenhum dos casos é bom para a indexação de informação por locutor. No caso do uso dos resultados para treinamento de modelos para reconhecimento de fala, entretanto, o sub-agrupamento pode ser satisfatório quando um locutor aparece em múltiplos ambientes acústicos, e o agrupamento excessivo pode ser vantajoso para agregar a fala de locutores similares ou ambientes acústicos (TRANTER; REYNOLDS, 2006).

2.6 Re-segmentação

Muitos sistemas apresentam uma última etapa, que é a re-segmentação do áudio através da decodificação de Viterbi (com ou sem iterações) usando os modelos dos conjuntos finais e modelos sem fala. Essa fase tem como propósito refinar as fronteiras originais dos segmentos e/ou agregar pequenos segmentos que possam ter sido removidos para um processamento mais robusto na fase de agrupamento. Um reconhecedor de palavras ou de fonemas pode ser usado para filtrar os limites dos segmentos, ajudando a reduzir o falso-alarme da taxa de erro (TRANTER; REYNOLDS, 2006).

3 Configuração Experimental

Neste capítulo é apresentada a configuração experimental do trabalho. A Seção 3.1 aborda as avaliações do NIST. Dados e estatísticas sobre as bases de dados utilizadas são apresentados na Seção 3.2. A Seção 3.3 demonstra como o NIST avalia os resultados obtidos pelos sistemas. A Seção 3.4 descreve o sistema de referência utilizado.

3.1 Avaliações do NIST

As avaliações do NIST são um importante meio de direcionar pesquisas na área de processamento de áudio voltada para tecnologias da fala e auxiliam na calibração das capacidades técnicas. Entre as diversas campanhas, o público-alvo dessas avaliações são os pesquisadores da área de reconhecimento de locutor independente de texto. As avaliações em reconhecimento de locutor foram criadas para serem simples, apoiadas e acessíveis. O objetivo dessas avaliações é encorajar o processo de pesquisa, e suas metas são (1) a exploração de novas idéias promissoras em reconhecimento de locutor, (2) o desenvolvimento de tecnologia avançada que incorpore essas idéias e (3) medir o desempenho dessa tecnologia (NIST, 2000). Elas ocorreram anualmente desde 1997 e passaram a ocorrer de dois em dois anos desde 2006 (NIST, 2000).

A avaliação utilizada como base nesse trabalho é a de 2000, que tem foco nas tarefas de detecção, rastreamento e detecção de mudança de locutor. Os dados são conversas telefônicas, e há dados limitados para treinamento. A avaliação de 2000 incluiu quatro tarefas: detecção de um locutor, detecção de dois locutores, rastreamento de locutor e detecção de mudança de locutor. No caso do reconhecimento de locutor, o objetivo é identificar os intervalos de tempo nos quais locutores desconhecidos estão falando em um segmento de conversa. A avaliação desdobra a tarefa de detecção de mudança de locutor em duas sub-tarefas. A primeira consiste em segmentar conversações que possuem somente dois locutores (conhecida por segmentação de 2 locutores – *2-Speaker Segmentation*). A segunda consiste da soma de segmentos de dois canais em múltiplos idiomas, geralmente de longa duração (até um máximo de dez minutos) e podem conter até dez locutores por segmento (conhecida por segmentação de *n* locutores – *N-Speaker Segmentation*). Os sistemas não são informados da quantidade de locutores por segmento (NIST, 2000).

3.2 Bases de Dados

Os dados utilizados na avaliação de 2000 foram extraídos da base *SwitchBoard-2 Corpus, Phases I e II*[1]. São 1000 conversas telefônicas entre duas pessoas com duração média de um minuto. Os arquivos totalizam aproximadamente 997 minutos de conversas, em que 90,1% dos quais correspondem a fala, em que 54% do tempo que corresponde a fala é utilizado para avaliação (o restante da fala corresponde a trechos em que os dois locutores falam ao mesmo tempo e a segmentos com menos de meio segundo de duração). Do total de arquivos, 269 correspondem a conversas entre duas pessoas do sexo masculino, 323 a conversas entre duas pessoas do sexo feminino, e 408 são conversas entre um homem e uma mulher.

O áudio das conversas é fornecido no formato *sphere* (formato criado pelo NIST para arquivos *wave* de fala). Os arquivos são convertidos para o formato *wave*, e então suas características são extraídas (o resultado é o vetor de características da conversa).

3.3 Avaliação de Desempenho

Para avaliar o desempenho dos resultados recebidos dos competidores, o NIST determina os segmentos em que cada locutor fala através de um detector de fala baseado em energia. Este detector é usado nos canais individuais (os "lados" da conversa), antes que estes sejam somados. Em alguns casos os pontos em que ocorre a mudança de locutor foram determinados por transcritores humanos. A pontuação é dada de acordo com trechos em que apenas um locutor está falando, e exclui variações de até 250 milissegundos das extremidades de cada intervalo. Dessa forma, garante-se que apenas segmentos de fala de meio segundo ou mais sejam utilizados para determinar a pontuação (NIST, 2000).

O NIST avalia separadamente os segmentos em que o número de locutores presentes na conversa é conhecido com antecedência e aqueles em que esse número não é conhecido. Todavia, o mesmo processo de pontuação é usado nos dois casos (NIST, 2000).

Como saída da detecção de mudança de locutor o sistema deve produzir mudanças de locutor hipotéticas (ou segmentos de fala produzidos por um único locutor) e um rótulo genérico do locutor para cada mudança (e. g. locutor 0, locutor 1, ...). O que se espera é que todas as mudanças para um mesmo locutor tenham o mesmo rótulo genérico de locutor (NIST, 2000).

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99S79

Diferente das outras tarefas propostas pela Avaliação de Reconhecimento de Locutor do NIST, a detecção de mudança de locutor é uma tarefa de detecção, e pode ser caracterizada por uma única taxa de erro, já que toda a fala deve ser levada em consideração e uma detecção perdida para um rótulo de locutor hipotético gerará um falso alarme correspondente para outro rótulo de locutor hipotético, de forma que os dois erros não são independentes (NIST, 2000).

Para calcular o erro na detecção é efetuada uma busca pelo melhor mapeamento (i. e. erro mínimo) de rótulos de locutor hipotético para locutores verdadeiros em cada conversa, e então os erros são acumulados ao longo do conjunto de conversações de teste. A taxa de erro é calculada da seguinte forma: assumindo que um sistema produza M rótulos de locutor hipotético para uma conversa que contenha na realidade N locutores verdadeiros (quando o número de locutores é conhecido de antemão tem-se $M = N$), quando M for menor que N são gerados automaticamente $N - M$ rótulos de locutor não associados a segmentos de fala. É produzido então um suposto mapeamento de um para um entre os locutores verdadeiros $\{i\}$ e os rótulos de locutor hipotético $\{j\}$

$$map(i) = j; i = 1, \dots, N, 1 \leq j \leq M.$$

Considerando $duração\_verdadeira(i, conv)$ = a duração total da fala do locutor verdadeiro i em uma conversa denotada $conv$, e $duração\_hipotética(i, j, conv)$ = a duração total da fala comum ao locutor verdadeiro i e ao rótulo de locutor hipotético j nessa conversa, a taxa de erro para esse mapeamento é calculada da seguinte forma:

$$erro(map, conv) = 1 - \frac{S_i\ duração\_hipotética(i, map(i), conv)}{S_i\ duração\_verdadeira(i, conv)}.$$

O melhor mapeamento de um para um, $map^*(i)$, usado para essa conversa é aquele que produz o erro mínimo para $erro(map, conv)$. Para a conversa são então registrados os valores

$$acerto(conv) = S_i\ duração\_hipotética(i, map^*(i), conv)$$

$$total(conv) = S_i\ duração\_verdadeira(i, conv).$$

Segundo NIST (2000), o erro (ponderado) total sobre um conjunto de conversas é finalmente calculado como

$$erro\_total = 1 - \frac{S_{conv}\ acerto(conv)}{S_{conv}\ total(conv)}.$$

3.4 Sistema de Referência

O sistema utilizado como referência para a avaliação dos resultados obtidos é uma implementação de um detector de mudanças de locutor baseado em energia desenvolvido por André Gustavo Adami. Sistemas para detecção de mudança de locutor baseados em energia assumem que existe grande probabilidade de ter ocorrido mudança de locutor em torno das regiões com silêncio (ADAMI et al, 2002). Nessa abordagem, o sistema de detecção de mudança de locutor recebe a entrada de áudio e detecta os períodos em que há silêncio no sinal. O silêncio é detectado através de um decodificador ou por meio de um limiar imposto à energia do áudio (KEMP et al, 2000). No caso de uso de decodificador, é executado um sistema de reconhecimento completo, obtendo os pontos em que ocorreu mudança de locutor a partir dos locais em que se detectou silêncio (MIRÓ, 2006). Já no segundo caso, os segmentos que se encontram abaixo do limiar (i. e. sua energia possui um valor baixo) são considerados silêncio, ao passo que os segmentos acima do limiar (i. e. os segmentos cuja energia apresenta um valor mais alto) são considerados fala (KEMP et al, 2000). Uma alternativa ao uso do limiar é o uso do desvio absoluto médio, que mede a variabilidade da energia nos segmentos (MIRÓ, 2006).

Restrições adicionais podem ser impostas à detecção de mudança de locutor baseada em energia, como a duração mínima de um período de silêncio, que é utilizada para reduzir o número de alarmes falsos. Satisfeitas as restrições, limiares hipotéticos dos segmentos de fala são determinados (KEMP et al, 2000).

O sistema realiza a detecção de mudança de locutor e, em seguida, o agrupamento dos segmentos obtidos. A distância utilizada na etapa de agrupamento dos segmentos é a Razão da Verossimilhança Generalizada. A Tabela 1 mostra os resultados obtidos com este sistema. Pares de Linhas Espectrais (PLE) e Coeficientes Cepstrais de Predição Linear (CCPL) foram utilizados nestes experimentos.

Tabela 1. Resultados da Diarização de Locutor baseada em Energia

Característica	Detecções Perdidas	Falso Alarme	Erro de Rotulação de Locutor
PLE	2,4%	0,4%	18,4%
CCPL	2,1%	0,5%	17,8%

Com base nos resultados, pode-se observar que as taxas de detecções perdidas, de falso alarme e de erro de rotulação de locutor são praticamente iguais com os dois tipos de características. A diarização com Coeficientes Cepstrais de Predição Linear apresenta taxas de

detecções perdidas e de erro de rotulação de locutor um pouco mais baixas, ao passo que a diarização com Pares de Linhas Espectrais apresenta taxa de falso alarme levemente mais baixa (porém, como os valores são muito próximos, essa diferença não é significante).

4 Detecção de Mudança de Locutor baseada no Critério de Informação Bayesiano

Este capítulo é dedicado ao estudo do Critério de Informação Bayesiano em segmentação de locutor. A Seção 4.1 aborda a adaptação do Critério de Informação Bayesiano para detecção de mudança de locutor. A Seção 4.2 descreve o método DISTBIC, proposto por Delacourt e Wellekens. Os resultados obtidos são discutidos na Seção 4.3.

4.1 Critério de Informação Bayesiano na Detecção de Mudança de Locutor

O Critério de Informação Bayesiano foi proposto por Gideon Schwarz em 1978. É um critério de verossimilhança máxima penalizado pela complexidade do modelo (i. e. o número de parâmetros). Esse critério é utilizado em Estatística para seleção de modelos em meio a uma classe de modelos paramétricos com diferentes números de parâmetros (SCHWARZ, 1978; CHEN; GOPALAKRISHNAN, 1998). O Critério de Informação Bayesiano pode também ser considerado um algoritmo de detecção de mudanças genérico, uma vez que, em detecção de mudança de locutor, não leva em consideração qualquer informação a respeito dos locutores (DELACOURT; WELLEKENS, 2000).

Em 1998, Scott Chen e P. S. Gopalakrishnan, do Centro de Pesquisa da IBM, publicaram um artigo no qual utilizam o Critério de Informação Bayesiano para detectar mudanças de locutor e também no agrupamento hierárquico dos segmentos obtidos. Assumindo que $X = \{x_i : i = 1, \dots, N\}$ é o conjunto de dados que está sendo modelado e $M = \{M_i : i = 1, \dots, K\}$ são os candidatos a modelo paramétrico desejado, a verossimilhança $L(X, M)$ é maximizada para este modelo. Assumindo que m representa o número de parâmetros, o Critério de Informação Bayesiano para M é definido como

$$CIB(M) = log\, L(X, M) - \lambda \frac{m}{2} log\, N,$$

em que o primeiro termo diz respeito à qualidade da combinação entre o modelo e os dados, enquanto o segundo é uma penalidade pela complexidade do modelo. A variável λ, cujo valor teórico para o Critério de Informação Bayesiano é 1, permite o ajuste do balanço entre os dois termos (CHEN; GOPALAKRISHNAN, 1998).

Assumindo que há, no máximo, uma mudança de um locutor no processo Gaussiano, leva-se em consideração o seguinte teste hipotético para mudança de locutor no tempo i:

- H_0: $x_1 \dots x_N \sim N(\mu, \Sigma)$: um único locutor proferiu a sequência; assume-se, portanto, que é representada por um único processo Gaussiano multidimensional,
- H_1: $x_1 \dots x_i \sim N(\mu_1, \Sigma_1)$; $x_{i+1} \dots x_N \sim N(\mu_2, \Sigma_2)$: dois ou mais locutores diferentes proferiram a sequência. μ e Σ, respectivamente a média e a matriz de covariância completa, são os parâmetros dos modelos.

Assim, a razão máxima da verossimilhança entre as hipóteses H_0 (não houve mudança de locutor) e H_1 (houve mudança de locutor no tempo i) é definida por

$$R(i) = N\, log|\Sigma| - N_1\, log\, |\Sigma_1| - N_2\, log\, |\Sigma_2|, \quad (1)$$

em que Σ, Σ_1 e Σ_2 são, respectivamente, as matrizes de covariância de todos os dados, do subconjunto $\{x_1, \dots, x_i\}$, e do subconjunto $\{x_{i+1}, \dots, x_N\}$, e N, N_1 e N_2 são, respectivamente, o número de vetores acústicos na sequência completa, no subconjunto $\{x_1, \dots, x_i\}$, e do subconjunto $\{x_{i+1}, \dots, x_N\}$. A estimativa da verossimilhança máxima do ponto em que ocorre a mudança de locutor é dada por

$$\hat{t} = arg\, \max_i R(i).$$

O teste das hipóteses pode ser visto como um problema de seleção de modelo, uma vez que uma modela os dados como uma Gaussiana e a outra modela os dados como duas Gaussianas diferentes. A diferença entre os valores do Critério de Informação Bayesiano desses modelos é representada por

$$\Delta CIB(i) = -R(i) + \lambda P,$$

em que a razão da verossimilhança $R(i)$ é definida por (1), $P = ½\ (p + ½p(p+1))\ logN$ é a penalidade, sendo p a dimensão do espaço acústico e o peso da penalidade λ é 1. A simetria da matriz de covariância é levada em consideração.

Dois modelos Gaussianos multidimensionais se adequam melhor aos dados X se $\Delta CIB(i)$ tiver um valor negativo, o que significa que ocorreu mudança de locutor no tempo i, de forma que

$$\left\{\max_i \Delta CIB(i)\right\} < 0.$$

O Critério de Informação Bayesiano apresenta vantagens como a robustez, já que calcula a variação no tempo i usando todas as amostras (o logaritmo da razão e a divergência de Kullback-Leibler, por exemplo, utilizam apenas as amostras limitadas contidas em dois segmentos). Outra vantagem é que esse critério evita o uso de qualquer limiar, pois a estimativa de um limiar é um ponto crítico em muitos processos de detecção de mudança de locutor e é geralmente deixado para que o usuário o selecione. O papel da penalidade λ é

equivalente à definição de um limiar. Há ainda a vantagem de a estimativa do Critério de Informação Bayesiano convergir para o ponto em que efetivamente ocorreu a mudança de locutor conforme o tamanho das amostras aumenta.

Múltiplas mudanças de locutor podem ser detectadas por meio do Critério de Informação Bayesiano através de três passos:

1º **Determinar a localização aproximada das mudanças de locutor**: o valor de ΔCIB é calculado entre dois segmentos adjacentes $[a, b]$ e $[b, c]$, nos quais os limiares a e c são determinados, e b assume seus valores em $[a, c]$ e é incrementado a cada iteração. Quando nenhum valor negativo é encontrado para ΔCIB, a distância $d(a, c)$ é aumentada e quando um valor negativo é encontrado, a mudança se torna o novo valor de a.

2º **Refinar o conjunto pontos de mudanças de locutor:** utilizando o mesmo procedimento definido no primeiro passo, com intervalos de exploração $[a, c]$ bem menores são escolhidos e centrados em torno dos pontos anteriormente selecionados como candidatos.

3º **Validar os pontos de mudanças de locutor estimados:** assumindo que $\{s_1, \dots, s_N\}$ é o conjunto de candidatos a mudanças de locutor encontrado no segundo passo, um valor ΔCIB é calculado para cada par de segmentos $[s_{i-1}, s_i][s_i, s_{i+1}]$. Uma mudança de locutor é identificada no tempo i caso o valor seja negativo. Caso contrário, o ponto s_i é descartado do conjunto de candidatos, de forma que o valor ΔCIB seja computado para o novo par de segmentos $[s_{i-1}, s_{i+1}][s_{i+1}, s_{i+2}]$ (com os índices velhos).

O método – necessário para uma estimativa correta dos parâmetros Gaussianos, já que a precisão do modelo depende muito da quantidade de informação disponível – consiste em combinar segmentos quando forem encontrados valores positivos para o Critério de Informação Bayesiano.

Um dos problemas do Critério de Informação Bayesiano é que o uso de apenas esse algoritmo para a detecção de mudança de locutor não é adaptado para segmentos pequenos, pois o algoritmo não consegue detectar duas mudanças de locutor mais próximas uma à outra do que a duração do segmento do segundo passo, que é de aproximadamente dois segundos. O ajuste do fator penalidade λ, que é dependente do tipo de dado analisado, é outro problema (CHEN; GOPALAKRISHNAN, 1998; DELACOURT; WELLEKENS, 2000).

4.2 Algoritmo DISTBIC

O algoritmo DISTBIC, proposto por Delacourt e Wellekens (2000), é constituído por dois passos: no primeiro passo, um cálculo de distância é utilizado para determinar os tempos *t*, que representam os pontos candidatos a mudança de locutor. No segundo passo, esses pontos candidatos são validados ou descartados através do terceiro passo do algoritmo do Critério de Informação Bayesiano para detecção de mudança de locutor. Esse fluxo pode ser observado na Figura 3.

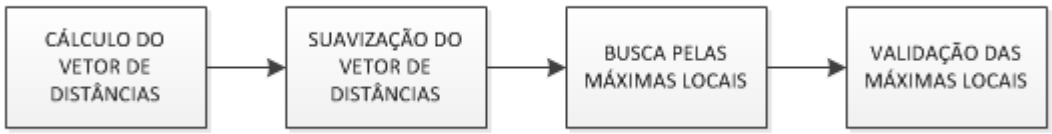

Figura 3: Etapas do algoritmo DISTBIC.

A detecção de pontos candidatos a mudança de locutor, etapa inicial do DISTBIC, confia em uma segmentação baseada em distância definida a partir das verossimilhanças de segmentos adjacentes. A concatenação de dois segmentos é considerada como um terceiro, e, em cada segmento, assume-se que os dados resultam de um único processo Gaussiano multidimensional.

Decidir se os dados no segmento maior se ajustam melhor com uma única Gaussiana multidimensional ou se uma representação em dois segmentos justifica melhor os dados é um problema. O tamanho dos segmentos é o resultado de um balanço entre o número de vetores dentro dos segmentos requerido para a estimação estatística significativa e a homogeneidade de locutor. Geralmente, o tamanho usado para cada segmento é de dois segundos (DELACOURT; WELLEKENS, 2000). As janelas (que têm o tamanho de um segmento) são deslocadas em passos de cem milissegundos, para os quais o critério é calculado, de forma que a resolução de mudanças de locutor é de cem milissegundos.

Delacourt e Wellekens utilizam seis critérios para realizar a detecção de mudança de locutor: a Razão da Verossimilhança Generalizada, a divergência de Kullback-Leibler e outras quatro medidas de similaridade propostas por Bimbot, Magrin-Chagnolleau e Mathan (1995).

Gish e Schmidt (em 1994) e Gish, Siu e Rohlicek (GISH et al, 1991) utilizaram a Razão da Verossimilhança Generalizada para identificação de locutor. O seguinte teste hipotético pode ser considerado para apresentar as principais características da Razão da Verossimilhança Generalizada:

- H_0: o mesmo locutor proferiu os segmentos. Nesse caso, a união dos dois segmentos é produzida por um único processo Gaussiano multidimensional.
- H_1 : vários locutores proferiram os segmentos. Nesse caso, assume-se que os segmentos foram gerados por processos Gaussianos multidimensionais diferentes.

A razão da verossimilhança associada com esse teste é

$$L = \frac{(z;\ \mu;\ \Sigma)}{(x, \mu_1;\ \Sigma_1)\ (y, \mu_2;\ \Sigma_2)},$$

em que, dado o processo Gaussiano multidimensional $N(\mu_X, \Sigma_X)$, $L(X, N(\mu_X, \Sigma_X))$ representa a verossimilhança da sequência de vetores acústicos X. O valor do logaritmo desta razão é considerado para obter uma distância entre dois segmentos (DELACOURT; WELLEKENS, 2000):

$$d_R = -log\, R.$$

A divergência de Kullback-Leibler mede a distância entre duas distribuições:

$$KL(X,Y) = E_X < (log\, P(X) - log\, P(Y)) >,$$

em que $E_{X<.>}$ denota a expectativa calculada com a densidade P de X.

Dessa divergência pode-se obter uma medida simétrica (i. e. uma distância):

$$KL2(X,Y) = KL(X,Y) + KL(Y,X).$$

A distância KL2 pode ser escrita da seguinte forma para as variáveis unidimensionais Gaussianas X e Y:

$$KL2(X,Y) = \frac{\sigma_X^2}{\sigma_Y^2} + \frac{\sigma_Y^2}{\sigma_X^2} + (\mu_X - \mu_Y)^2 \left(\frac{1}{\sigma_X^2} + \frac{1}{\sigma_Y^2}\right) - 1,$$

e se torna, para vetores Gaussianos,

$$\begin{aligned} KL2(X,Y) = {} & \frac{1}{2}\, log \frac{|\Sigma_Y|}{|\Sigma_X|} + \frac{1}{2} tr(\Sigma_Y^{-1}\Sigma_X) + \frac{1}{2}(\mu_Y - \mu_X)^T \Sigma_Y^{-1}(\mu_Y - \mu_X) - \frac{1}{2}\mathrm{N} + \\ & \frac{1}{2}\, log \frac{|\Sigma_X|}{|\Sigma_Y|} + \frac{1}{2} tr(\Sigma_X^{-1}\Sigma_Y) + \frac{1}{2}(\mu_X - \mu_Y)^T \Sigma_X^{-1}(\mu_X - \mu_Y) - \frac{1}{2}\mathrm{N} \\ = {} & \frac{1}{2}\Big(\mathrm{tr}(\Sigma_Y^{-1}\Sigma_X + \Sigma_X^{-1}\Sigma_Y) + (\mu_X - \mu_Y)^T(\Sigma_X^{-1} + \Sigma_Y^{-1})(\mu_X - \mu_Y)\Big) - \mathrm{N}, \end{aligned}$$

em que tr denota o traço de uma matriz (i. e. a soma dos elementos da diagonal principal) e N é a dimensão dos vetores de características (HERSHEY; OLSEN, 2007).

As medidas de similaridade baseiam-se na hipótese de que dois segmentos X e Y de um sinal parametrizado deveriam ter matrizes de covariância similares, respectivamente Σ_X e Σ_Y, se são gerados pelo mesmo locutor. A matriz $\Gamma = \Sigma_X \Sigma_Y^{-1}$ é levada em consideração para medir a similaridade de dois segmentos de locutor X e Y. Se os dois segmentos pertencem ao

mesmo locutor, então $\Sigma_X = \Sigma_Y$, de forma que Γ é a matriz identidade. Essas medidas de similaridade têm custo computacional mais baixo que a Razão da Verossimilhança Generalizada e o KL2, mas seus resultados não são tão bons (DELACOURT; WELLEKENS, 2000).

A primeira das medidas de similaridade propostas por Bimbot, Magrin-Chagnolleau e Mathan (1995) é denominada "medida de verossimilhança Gaussiana", e é definida como

$$\mu_G(X,Y) = a - log\ g + \frac{1}{p}(\mu_X - \mu_Y)\,\Sigma_X^{-1}\,(\mu_X - \mu_Y)^T - 1,$$

em que a é a média aritmética dos autovalores λ_i de Γ e g é a média geométrica. Se $\Sigma_X = \Sigma_Y$ (ou seja, se $X = Y$), então $\mu_G = 0$. Caso contrário, $\mu_G > 0$.

A segunda medida de similaridade é uma variante da primeira. Ela é baseada no fato de que os vetores de média podem ser afetados pelo canal de transmissão e não deveriam ser levados em conta para a segunda medida, ao passo que as matrizes de covariância são mais robustas a variações entre as condições de gravação e linhas de transmissão. Essa medida de similaridade é definida como

$$\mu_{GC}(X,Y) = a - log\ g - 1.$$

A terceira medida de similaridade é um teste de esfericidade para a matriz Γ, e é denominada "medida de esfericidade aritmético-geométrica":

$$\mu_{SC}(X,Y) = log\frac{a}{g}.$$

A última medida de similaridade é denominada "medida de desvio absoluto", e é baseada no desvio absoluto dos autovalores de Γ quando comparado a 1,

$$\mu_{DC}(X,Y) = \frac{1}{p}\sum_{i=1}^{p}|\lambda_i - 1|.$$

Essas medidas de similaridade não satisfazem a propriedade de simetria de uma distância e, por isso, devem ser tornadas simétricas. Para torná-las simétricas, soma-se a divergência de X para Y com a divergência de Y para X:

$$\mu_S(X,Y) = \mu(X,Y) + \mu(Y,X),$$

em que μ representa μ_G, μ_{GC}, μ_{SC} ou μ_{DC}.

Para todas as medidas de distância descritas, uma mudança de locutor é indicada por um valor alto, ao passo que valores baixos indicam que os dois segmentos do sinal correspondem ao mesmo locutor.

As mudanças de locutor podem ser detectadas por meio do DISTBIC através das seguintes etapas (as etapas 1-3 correspondem ao primeiro passo do DISTBIC, enquanto que a etapa 4 corresponde ao segundo passo):

1. **Cálculo do vetor de distâncias entre todos os pares de janelas adjacentes:** o critério (ou "distância") é calculado para um par de segmentos adjacentes de mesmo tamanho (geralmente esses segmentos têm aproximadamente dois segundos), e as janelas são então deslocadas em um passo fixo (aproximadamente um décimo de segundo) ao longo de todo o sinal de fala parametrizado. Este processo tem como saída uma série temporal da distância.
2. **Suavização do vetor de distâncias**: o vetor de distâncias é suavizado por uma operação de filtragem.
3. **Busca pelas máximas locais**: todas as máximas locais significantes (i. e. as máximas locais cujas diferenças entre seus valores e aqueles das mínimas adjacentes estão acima de determinado limiar – que é calculado como uma fração do gráfico da variância –, e quando não há uma máxima local mais alta em sua vizinhança) são procuradas. Dessa forma, a seleção da máxima local é feita considerando o "fator forma" dos picos, e não seu valor absoluto. Considerando que σ e μ denotam, respectivamente, o desvio padrão e a média das distâncias no gráfico, um pico é significante se

   $$\left|d(max) - d\left(\underset{r}{min}\right)\right| > \alpha\sigma$$

   e

   $$\left|d(max) - d\left(\underset{l}{min}\right)\right| > \alpha\sigma,$$

   em que α é real, e min_r e min_l são, respectivamente, as mínimas direita e esquerda em torno do pico max (cf. Figura 4). Impõe-se, ainda, uma duração mínima entre as duas máximas: se duas máximas estão muito próximas, a mais baixa é descartada. Essa restrição resulta na fronteira mais baixa para segmentos de curta duração (que geralmente são de um segundo). Esse tipo de detecção tem os seguintes requisitos:

   - Não depende do tipo de dados de fala (telejornais, conversas telefônicas, estúdio).
   - Nesse passo, a ênfase é colocada em minimizar a perda de detecções (ou seja, a não detecção de uma mudança que de fato aconteceu), resultando

em um alto número de alarmes falsos (isto é, detectar mudanças de locutor que, na realidade, não ocorreram). Esse número será reduzido pela combinação de segmentos contíguos durante o segundo passo, através do Critério de Informação Bayesiano.

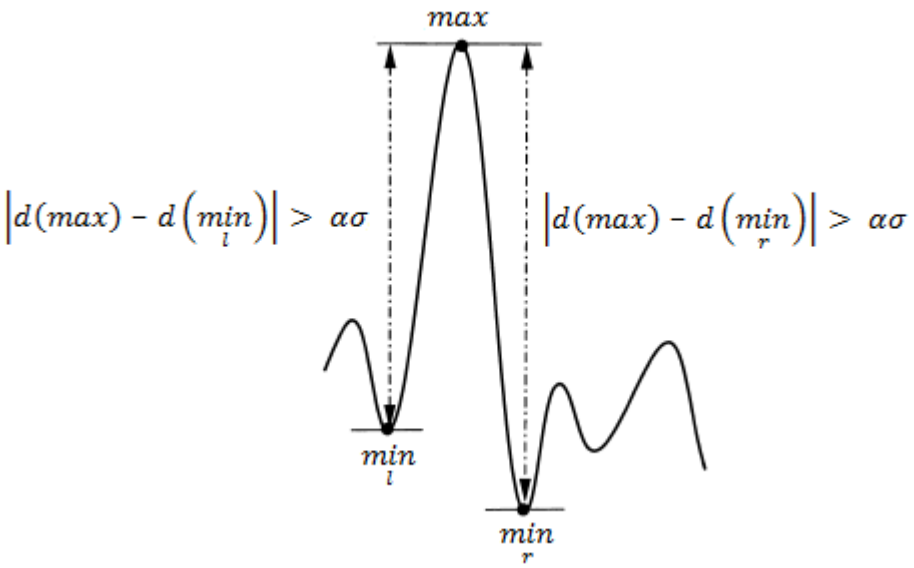

Figura 4. Caracterização de uma mudança de locutor.

4. **Validação das máximas locais:** essa etapa é uma cópia exata do terceiro passo do algoritmo do Critério de Informação Bayesiano para detecção de mudança de locutor: um valor ΔCIB é calculado para cada candidato a mudança de locutor para validar o resultado do primeiro passo. O valor do fator empírico λ tem que ser ajustado para reduzir o número de alarmes falsos sem que ocorra aumento do número de novas detecções perdidas. O uso do algoritmo do Critério de Informação Bayesiano para detecção de mudança de locutor é muito mais apropriado nesse momento, uma vez que o tamanho dos segmentos considerados é agora grande o bastante para uma boa estimativa de parâmetros (DELACOURT; WELLEKENS, 2000).

4.3 Resultados e Discussão

As características utilizadas nos experimentos conduzidos com a implementação deste trabalho foram os Pares de Linhas Espectrais e os Coeficientes Cepstrais de Predição Linear. Os Pares de Linhas Espectrais foram escolhidos por serem bastante utilizados para compressão de voz, com notável extensão para o campo de reconhecimento de locutor

(McLOUGHLIN, 2008). Já os Coeficientes Cepstrais de Predição Linear foram escolhidos por sua aplicação na identificação de locutor (REYNOLDS, 1994). Esses vetores de características foram extraídos com 24 dimensões.

A medida de distância utilizada na etapa de detecção de mudança de locutor (através do algoritmo DISTBIC) é o Critério de Informação Bayesiano. Assumiu-se que a duração mínima entre duas máximas na terceira etapa do DISTBIC é de dois segundos (ou seja, assume-se que cada locutor fala por no mínimo 2 segundos). O tamanho das janelas é 2 segundos, e o deslocamento dessas janelas é de 0.1 segundo. O tamanho da janela de filtragem é de 1.7 segundos.

A Tabela 2 mostra os resultados da detecção de mudança de locutor com variações nos parâmetros λ (penalidade) e α (valor multiplicado pelo desvio padrão do vetor de distâncias para determinar se os picos são válidos na terceira etapa do DISTBIC). Pares de Linhas Espectrais foram usados nestes experimentos.

Tabela 2. Resultados do DISTBIC com Pares de Linhas Espectrais

λ	α	Detecções Perdidas	Falso Alarme	Erro de Rotulação de Locutor
0,5	0,1	0,0%	0,0%	41,7%
0,6	0,1	0,0%	0,0%	39,6%
0,7	0,1	0,0%	0,0%	36,8%
0,8	0,1	0,3%	0,0%	32,6%
0,9	0,1	1,6%	0,1%	29,0%
1,0	0,1	4,6%	0,3%	25,3%
1,1	0,1	8,8%	0,5%	21,2%
1,2	0,1	14,3%	0,4%	16,8%
1,3	0,1	18,8%	0,4%	12,9%
1,4	0,1	23,0%	0,2%	9,5%
1,5	0,1	26,0%	0,2%	7,0%
0,5	0,5	0,0%	0,0%	40,3%
0,6	0,5	0,0%	0,0%	38,6%
0,7	0,5	0,1%	0,0%	36,2%
0,8	0,5	0,2%	0,1%	32,4%
0,9	0,5	1,3%	0,2%	29,1%
1,0	0,5	3,9%	0,2%	25,1%
1,1	0,5	7,9%	0,5%	21,8%
1,2	0,5	12,8%	0,4%	17,3%
1,3	0,5	17,6%	0,4%	13,5%
1,4	0,5	21,3%	0,2%	10,2%
1,5	0,5	24,3%	0,2%	7,9%

Na Tabela 3 são apresentados os resultados da detecção de mudança de locutor com variações nos parâmetros λ e α. Coeficientes Cepstrais de Predição Linear foram utilizados nestes experimentos.

Tabela 3. Resultados do DISTBIC com Coeficientes Cepstrais de Predição Linear

λ	α	Detecções Perdidas	Falso Alarme	Erro de Rotulação de Locutor
0,5	0,1	0,0%	0,0%	41,0%
0,6	0,1	0,0%	0,0%	39,9%
0,7	0,1	0,0%	0,0%	36,6%
0,8	0,1	0,3%	0,0%	32,0%
0,9	0,1	1,7%	0,2%	27,3%
1,0	0,1	5,1%	0,3%	22,1%
1,1	0,1	10,6%	0,5%	17,3%
1,2	0,1	16,4%	0,4%	12,9%
1,3	0,1	21,8%	0,3%	8,9%
1,4	0,1	26,1%	0,3%	6,1%
1,5	0,1	29,0%	0,2%	3,9%
0,5	0,5	0,0%	0,0%	40,0%
0,6	0,5	0,0%	0,0%	38,9%
0,7	0,5	0,1%	0,1%	35,5%
0,8	0,5	0,3%	0,0%	31,6%
0,9	0,5	1,3%	0,2%	27,3%
1,0	0,5	4,2%	0,3%	21,9%
1,1	0,5	9,4%	0,4%	17,4%
1,2	0,5	14,8%	0,4%	13,3%
1,3	0,5	20,1%	0,3%	9,7%
1,4	0,5	23,9%	0,3%	7,2%
1,5	0,5	10,3%	1,5%	19,2%

Com base nesses resultados, é possível observar que valores mais altos para a penalidade (λ) implicam no aumento da taxa de detecções perdidas. Isso se deve ao fato que uma penalidade alta resulta na diminuição de pontos candidatos a mudança de locutor aceitos na quarta etapa do DISTBIC (consequentemente, pontos em que de fato ocorreu uma mudança de locutor são ignorados). O inverso ocorre com a taxa de erro de rotulação de locutor: valores mais altos para a penalidade resultam em menor erro na rotulação dos locutores. Essa diminuição na taxa de erro se deve ao fato de haver menos segmentos (de forma que menos segmentos são rotulados incorretamente). A taxa de falso alarme não é tão afetada quanto a de detecções perdidas: com Pares de Linhas Espectrais, a diferença é praticamente nula, ao passo que com Coeficientes Cepstrais de Predição Linear ocorreu leve aumento dessa taxa com valores mais altos para a penalidade a para o coeficiente α. O valor de α afetou levemente os resultados: valores mais altos combinados com um valor mais alto

para a penalidade implicam na diminuição da taxa de detecções perdidas. Isso porque valores mais baixos para o coeficiente α permitem que máximas com distâncias menores em relação às mínimas vizinhas sejam considerados pontos candidatos a mudança de locutor na terceira etapa do DISTBIC.

Delacourt e Wellekens (2000) reportam que conversas telefônicas contêm longos segmentos, porém, constituídos por fala espontânea (i. e. a conversa contém inúmeras palavras pequenas que, quando proferidas ao mesmo tempo em que outra pessoa fala, anulam a presunção de que os dois locutores não falam ao mesmo tempo. A detecção de mudança de locutor é influenciada negativamente, pois essas palavras são pequenas demais para serem detectadas corretamente. Palavras pequenas (e. g. um locutor pode simplesmente reconhecer a fala do outro locutor através de interjeições ou confirmações curtas) também podem implicar em duas rápidas mudanças de locutor, uma das quais não será detectada, devido à conjectura de que cada locutor fala por pelo menos dois segundos.

Além do desempenho global, foram realizadas avaliações em que (1) todas as conversas ocorrem entre duas pessoas do sexo masculino, (2) todas as conversas ocorrem entre duas pessoas do sexo feminino e (3) todas as conversas ocorrem entre uma pessoa do sexo masculino e outra do sexo feminino. Os resultados dessas avaliações ficaram bem próximos do resultado global. A Tabela 4 mostra os resultados das avaliações em que todas as conversas ocorrem entre duas pessoas do sexo masculino. A Tabela 5 mostra os resultados das avaliações em que todas as conversas ocorrem entre duas pessoas do sexo feminino. A Tabela 6 mostra os resultados das avaliações em que todas as conversas ocorrem entre uma pessoa do sexo masculino e outra do sexo feminino. Pares de Linhas Espectrais foram utilizados nos três casos.

Tabela 4. Resultados do DISTBIC com Pares de Linhas Espectrais (conversas entre dois homens)

λ	α	Detecções Perdidas	Falso Alarme	Erro de Rotulação de Locutor
0,5	0,1	0,0%	0,0%	41,2%
0,6	0,1	0.0%	0.0%	38.7%
0,7	0,1	0.1%	0.0%	36.3%
0,8	0,1	0.5%	0.1%	32.8%
0,9	0,1	2.1%	0.3%	29.2%
1,0	0,1	5,7%	0,6%	25,3%
1,1	0,1	10.4%	1.0%	19.6%
1,2	0,1	14.6%	0.6%	16.4%
1,3	0,1	20.0%	0.5%	12.5%
1,4	0,1	23.9%	0.1%	8.9%
1,5	0,1	26,5%	0,1%	6,9%

0,5	0,5	0,0%	0,0%	40,1%
0,6	0,5	0,0%	0,0%	38,3%
0,7	0,5	0,1%	0,1%	35,9%
0,8	0,5	0,3%	0,2%	32,8%
0,9	0,5	1,8%	0,3%	28,9%
1,0	0,5	5,2%	0,6%	24,9%
1,1	0,5	9,4%	1,0%	20,6%
1,2	0,5	13,4%	0,5%	17,1%
1,3	0,5	18,4%	0,4%	12,6%
1,4	0,5	22,0%	0,1%	9,4%
1,5	0,5	24,6%	0,1%	7,5%

Tabela 5. Resultados do DISTBIC com Pares de Linhas Espectrais (conversas entre duas mulheres)

λ	α	Detecções Perdidas	Falso Alarme	Erro de Rotulação de Locutor
0,5	0,1	0,0%	0,0%	42,4%
0,6	0,1	0,0%	0,0%	40,5%
0,7	0,1	0,1%	0,0%	36,9%
0,8	0,1	0,5%	0,0%	32,5%
0,9	0,1	2,1%	0,1%	29,6%
1,0	0,1	5,5%	0,3%	25,9%
1,1	0,1	10,3%	0,5%	21,2%
1,2	0,1	17,2%	0,4%	15,7%
1,3	0,1	21,5%	0,4%	11,9%
1,4	0,1	25,1%	0,3%	8,8%
1,5	0,1	28,5%	0,4%	5,8%
0,5	0,5	0,0%	0,0%	41,0%
0,6	0,5	0,0%	0,0%	38,9%
0,7	0,5	0,2%	0,0%	36,9%
0,8	0,5	0,4%	0,0%	32,5%
0,9	0,5	1,6%	0,2%	30,0%
1,0	0,5	4,4%	0,2%	26,0%
1,1	0,5	8,9%	0,4%	22,1%
1,2	0,5	15,4%	0,3%	16,5%
1,3	0,5	20,6%	0,4%	12,7%
1,4	0,5	24,0%	0,4%	9,6%
1,5	0,5	27,2%	0,4%	6,8%

Tabela 6. Resultados do DISTBIC com Pares de Linhas Espectrais (conversas entre um homem e uma mulher)

λ	α	Detecções Perdidas	Falso Alarme	Erro de Rotulação de Locutor
0,5	0,1	0,0%	0,0%	41,6%
0,6	0,1	0,0%	0,0%	39,5%
0,7	0,1	0,0%	0,0%	37,2%

0,8	0,1	0,1%	0,0%	32,5%
0,9	0,1	0,8%	0,0%	28,4%
1,0	0,1	3,1%	0,1%	24,9%
1,1	0,1	6,6%	0,3%	22,4%
1,2	0,1	11,7%	0,4%	18,0%
1,3	0,1	16,0%	0,2%	14,0%
1,4	0,1	20,7%	0,2%	10,5%
1,5	0,1	23,7%	0,2%	8,2%
0,5	0,5	0,0%	0,0%	40,1%
0,6	0,5	0,0%	0,0%	38,7%
0,7	0,5	0,0%	0,0%	36,0%
0,8	0,5	0,1%	0,0%	32,2%
0,9	0,5	0,6%	0,1%	28,5%
1,0	0,5	2,6%	0,1%	24,7%
1,1	0,5	6,3%	0,2%	22,4%
1,2	0,5	10,4%	0,3%	18,2%
1,3	0,5	14,8%	0,4%	14,7%
1,4	0,5	18,6%	0,2%	11,2%
1,5	0,5	21,9%	0,2%	8,9%

A Tabela 7 mostra os resultados das avaliações em que todas as conversas ocorrem entre duas pessoas do sexo masculino. A Tabela 8 mostra os resultados das avaliações em que todas as conversas ocorrem entre duas pessoas do sexo feminino. A Tabela 9 mostra os resultados das avaliações em que todas as conversas ocorrem entre uma pessoa do sexo masculino e outra do sexo feminino. Coeficientes Cepstrais de Predição Linear foram utilizados nos três casos.

Tabela 7. Resultados do DISTBIC com Coeficientes Cepstrais de Predição Linear (conversas entre dois homens)

λ	α	**Detecções Perdidas**	**Falso Alarme**	**Erro de Rotulação de Locutor**
0,5	0,1	0,0%	0,0%	41,0%
0,6	0,1	0,0%	0,0%	39,7%
0,7	0,1	0,0%	0,0%	36,2%
0,8	0,1	0,3%	0,0%	32,8%
0,9	0,1	1,5%	0,1%	28,7%
1,0	0,1	5,6%	0,4%	21,5%
1,1	0,1	10,8%	0,8%	17,3%
1,2	0,1	17,5%	0,6%	12,3%
1,3	0,1	23,0%	0,5%	8,1%
1,4	0,1	27,1%	0,3%	5,2%
1,5	0,1	29,1%	0,1%	3,8%
0,5	0,5	0,0%	0,0%	40,2%
0,6	0,5	0,0%	0,0%	38,9%
0,7	0,5	0,2%	0,2%	36,1%
0,8	0,5	0,4%	0,0%	32,1%

0,9	0,5	1,4%	0,3%	28,1%
1,0	0,5	4,6%	0,4%	21,6%
1,1	0,5	10,7%	0,6%	16,5%
1,2	0,5	16,9%	0,6%	12,2%
1,3	0,5	22,2%	0,5%	8,3%
1,4	0,5	25,7%	0,3%	5,9%
1,5	0,5	12,0%	1,9%	18,0%

Tabela 8. Resultados do DISTBIC com Coeficientes Cepstrais de Predição Linear (conversas entre duas mulheres)

λ	α	**Detecções Perdidas**	**Falso Alarme**	**Erro de Rotulação de Locutor**
0,5	0,1	0,0%	0,0%	41,8%
0,6	0,1	0,0%	0,0%	39,7%
0,7	0,1	0,0%	0,0%	36,7%
0,8	0,1	0,5%	0,0%	31,9%
0,9	0,1	2,6%	0,2%	28,1%
1,0	0,1	6,9%	0,3%	23,0%
1,1	0,1	13,7%	0,4%	16,7%
1,2	0,1	19,3%	0,4%	12,8%
1,3	0,1	24,6%	0,2%	8,1%
1,4	0,1	28,8%	0,3%	5,0%
1,5	0,1	31,3%	0,3%	3,0%
0,5	0,5	0,0%	0,0%	40,9%
0,6	0,5	0,0%	0,0%	39,6%
0,7	0,5	0,0%	0,0%	36,1%
0,8	0,5	0,3%	0,0%	31,0%
0,9	0,5	1,8%	0,2%	27,9%
1,0	0,5	5,5%	0,3%	23,4%
1,1	0,5	11,6%	0,4%	17,2%
1,2	0,5	17,0%	0,4%	13,2%
1,3	0,5	23,0%	0,2%	8,8%
1,4	0,5	27,0%	0,3%	5,6%
1,5	0,5	9,5%	1,0%	21,3%

Tabela 9. Resultados do DISTBIC com Coeficientes Cepstrais de Predição Linear (conversas entre um homem e uma mulher)

λ	α	**Detecções Perdidas**	**Falso Alarme**	**Erro de Rotulação de Locutor**
0,5	0,1	0,0%	0,0%	41,6%
0,6	0,1	0,0%	0,0%	40,4%
0,7	0,1	0,0%	0,0%	36,8%
0,8	0,1	0,1%	0,0%	31,8%
0,9	0,1	1,0%	0,2%	26,0%
1,0	0,1	3,2%	0,3%	21,8%

1,1	0,1	7,9%	0,3%	17,9%
1,2	0,1	13,4%	0,4%	13,4%
1,3	0,1	18,8%	0,3%	10,2%
1,4	0,1	23,4%	0,3%	7,7%
1,5	0,1	27,1%	0,3%	4,7%
0,5	0,5	0,0%	0,0%	39,4%
0,6	0,5	0,0%	0,0%	38,4%
0,7	0,5	0,1%	0,1%	34,7%
0,8	0,5	0,2%	0,1%	31,8%
0,9	0,5	0,7%	0,2%	26,3%
1,0	0,5	2,9%	0,3%	21,0%
1,1	0,5	6,7%	0,3%	18,2%
1,2	0,5	11,7%	0,2%	14,2%
1,3	0,5	16,4%	0,4%	11,5%
1,4	0,5	20,2%	0,4%	9,4%
1,5	0,5	9,9%	1,6%	18,2%

As taxas de falso alarme e de detecções perdidas obtidas com o DISTBIC com valores baixos para a penalidade foram melhores que as taxas obtidas com o algoritmo baseado em energia do sistema de referência, mas as taxas de erro de rotulação de locutor foram muito mais altas. Com valores mais altos para a penalidade, a taxa de detecções perdidas obtida com o DITSTBIC fica muito acima da taxa obtida com o algoritmo baseado em energia, e a taxa de erro de rotulação só fica baixa porque o número de segmentos obtidos é muito pequeno.

5 Agrupamento

Como a detecção de mudança de locutor pode detectar pontos em que não ocorre de fato uma mudança, a hipótese de que uma conversa é realizada através do uso intercalado de tempo entre os locutores participantes é invalidada. Assim, diferentemente do que foi proposto para o método DISTBIC, pode-se utilizar métodos de agrupamento para dirimir tais problemas de falso alarme na detecção de mudança de locutor. A Seção 5.1 aborda o agrupamento hierárquico aglomerativo. A Seção 5.2 apresenta medidas de distância que podem ser utilizadas para encontrar os segmentos mais próximos no processo de agrupamento. Os resultados obtidos são discutidos na Seção 5.3.

5.1 Agrupamento Hierárquico Aglomerativo

O agrupamento hierárquico aglomerativo é, de longe, a abordagem mais utilizada para o agrupamento dos segmentos de locutor, uma vez que essa aproximação pode usar os segmentos obtidos na fase de detecção de mudança de locutor como ponto de partida (MIRÓ, 2006). Esse tipo de agrupamento também é chamado "agrupamento *bottom-up*", e tem sido usado por muitos anos em classificação de padrões. A fase de agrupamento dos segmentos requer, em qualquer tipo de implementação, a definição de duas informações:

1. Uma distância entre os segmentos para determinar a similaridade acústica. Geralmente é utilizada uma matriz de distâncias ao invés de um valor individual para o par de segmentos. Essa matriz contém as distâncias entre todas as combinações possíveis de pares de segmentos.
2. Um critério de parada para interromper o algoritmo quando o número de segmentos desejado for atingido (MIRÓ, 2006).

Geralmente, os agrupamentos são representados por uma única Gaussiana, mas Modelos de Misturas Gaussianas também têm sido usados (TRANTER; REYNOLDS, 2006). A distância comumente usada para medir a distância entre dois agrupamentos é a Razão da Verossimilhança Generalizada, embora outras distâncias possam ser utilizadas. O critério de parada compara a estatística do Critério de Informação Bayesiano dos dois agrupamentos que estão sendo considerados com a do agrupamento-pai.

Se o par de agrupamentos for melhor descrito por uma única Gaussiana, então ΔCIB será baixo, enquanto que se houverem duas distribuições separadas (o que implica em dois locutores) o valor de ΔCIB será alto. A cada iteração o par de agrupamentos com os menores

valores para ΔCIB são combinados em um novo agrupamento e as estatísticas são recalculadas. Geralmente, o processo é repetido até que o menor valor de ΔCIB seja maior que um limiar especificado (normalmente com valor zero). O uso do número de vetores no agrupamento-pai (no fator penalidade) representa uma decisão "local" com base no Critério de Informação Bayesiano, ou seja, uma decisão considerando apenas os agrupamentos que estão sendo combinados (TRANTER; REYNOLDS, 2006).

A busca global do melhor valor do Critério de Informação Bayesiano, todavia, é muito custosa, uma vez que o agrupamento é executado para obter diferentes números de conjuntos. Apesar disso, é possível aperfeiçoar o Critério para métodos de agrupamento hierárquicos através de um algoritmo guloso (CHEN; GOPALAKRISHNAN, 1998).

Os métodos aglomerativos assumem que cada sinal é um nó inicial, e então sucessivamente combinam os dois nós mais próximos de acordo com a medida de distância (CHEN; GOPALAKRISHNAN, 1998). O processo é repetido até que o critério de parada seja alcançado. Esse processo é ilustrado na Figura 5.

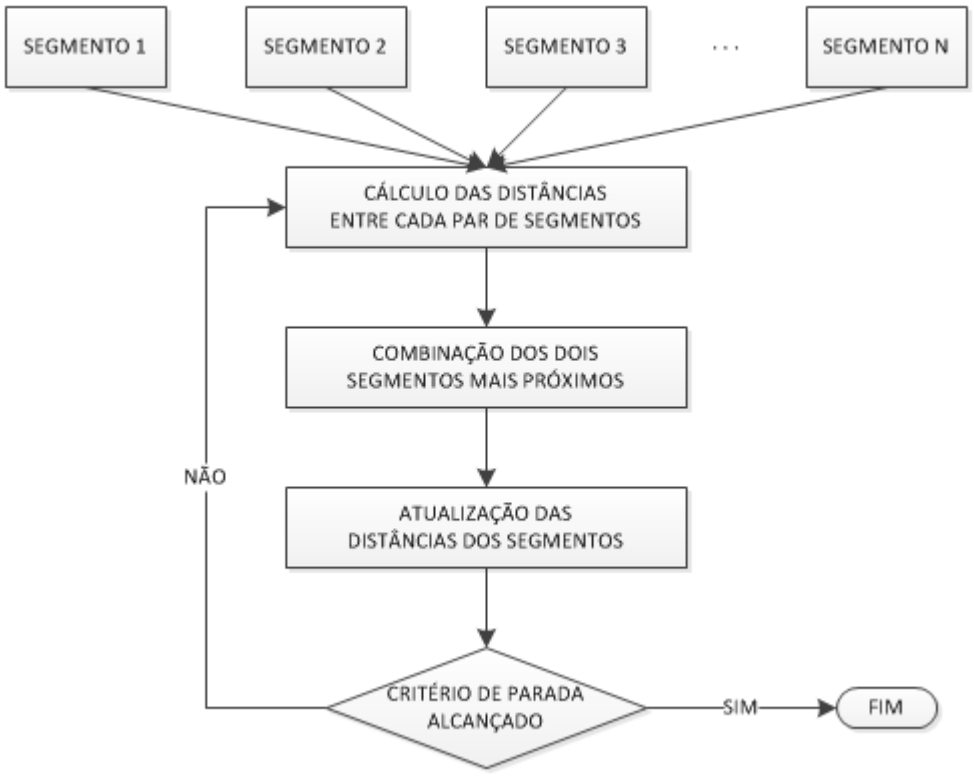

Figura 5. Processo de agrupamento hierárquico aglomerativo.

O procedimento de término do Critério de Informação Bayesiano é que dois nós não devem ser combinados se a penalidade do Critério de Informação Bayesiano resultar um valor

negativo. Uma vez que o valor do Critério é aumentado a cada combinação, busca-se uma árvore de agrupamento "ótimo" aperfeiçoando o Critério de forma gulosa. A penalidade do Critério é usada apenas para o término. Embora o Critério de Informação Bayesiano possa ser utilizado como medida de distância no processo aglomerativo, em muitas aplicações é provavelmente melhor usar medidas de distância mais sofisticadas. Esse critério também pode ser aplicado em métodos *top-down* (CHEN; GOPALAKRISHNAN, 1998).

Diversos autores propuseram variações dessa técnica. Exemplos são os sistemas descritos por Wooters, Fung, Peskin e Anguera (2004) e por Ajmera e Wooters (2003) que removem a necessidade de ajustar o peso da penalidade assegurando-se que o número de parâmetros nas distribuições combinadas e separadas é igual, embora o número-base de Gaussianas (e daí o número de parâmetros livres) precise ser cuidadosamente escolhido para obter efeito ótimo (TRANTER; REYNOLDS, 2006).

Ben, Betser, Bimbot e Gravier (2004) apresentam uma técnica alternativa que usa uma distância Euclidiana entre Modelos de Misturas Gaussianas adaptadas – uma técnica relacionada à estimativa de Monte-Carlo da Divergência Gaussiana (que é o KL-2 simétrico) ao mesmo tempo que é um limite superior a ela. Um limiar fixo é usado como critério de parada. Seu desempenho é comparável ao Critério de Informação Bayesiano convencional.

5.2 Medidas de Distância

Diversas medidas de distância podem ser utilizadas para calcular a matriz de distâncias para o agrupamento dos segmentos. Essas distâncias são distâncias entre conjuntos de dados, e não distâncias entre pontos, pois os segmentos são partes da matriz de características (logo, são matrizes também).

Algumas medidas utilizadas para esse cálculo são o Critério de Informação Bayesiano, descrito em 4.1, a Razão da Verossimilhança Generalizada e o KL2, descritos na Seção 4.2 (MIRÓ, 2006). Há também medidas baseadas no Logaritmo da Razão da Verossimilhança Cruzada (LRVC), que é uma medida que pode ser utilizada para medir a distância entre dois modelos de locutor. Três de suas variações são apresentadas nas seções que seguem: o Logaritmo da Razão da Verossimilhança Cruzada Normalizada (LRVCN), o Logaritmo da Razão da Verossimilhança Cruzada Normalizada pelo Modelo de Impostores (LRVCI) e o Logaritmo da Razão da Verossimilhança Cruzada Adaptada (LRVCA) (LE et al, 2007; ZHU, 2006).

5.2.1 Logaritmo da Razão da Verossimilhança Cruzada Normalizada

Dados dois modelos de locutor M_i e M_j, logaritmo da razão da verossimilhança cruzada normaliza (LRVCN) é definido como

$$LRVCN(M_i, M_j) = \frac{1}{N_i} log\left(\frac{L(X_i|M_i)}{L(X_i|M_j)}\right) + \frac{1}{N_j} log\left(\frac{L(X_j|M_j)}{L(X_j|M_i)}\right)$$

$$= \frac{1}{N_i}[logL(X_i|M_i) - log\, L(X_i|M_j)]$$

$$+ \frac{1}{N_j}[logL(X_j|M_j) - log\, L(X_j|M_i)]$$

em que a razão da verossimilhança cruzada $\frac{L(X_i|M_i)}{L(X_i|M_j)}$ mede quão bem M_j, que é o modelo do locutor j, combina com X_i, que são os dados do locutor i, em relação a quão bem M_i combina com seu próprio conjunto de dados (X_i). N_i é o tamanho do segmento X_i. Cada logaritmo da verossimilhança é normalizado pelo tamanho do segmento de dados correspondente (LE et al, 2007).

5.2.2 Logaritmo da Razão da Verossimilhança Cruzada Normalizada pelo Modelo de Impostores com Adaptação

Dados dois modelos adaptados de locutor M_i^* e M_j^*, a distância Logaritmo da Razão da Verossimilhança Cruzada Normalizada pelo Modelo de Impostores com Adaptação (LRVCA) é definida como

$$LRVCA(M_i^*, M_j^*) = \frac{1}{N_i} log\left(\frac{L(X_i|M_j^*)}{L(X_i|UBM)}\right) + \frac{1}{N_j} log\left(\frac{L(X_j|M_i^*)}{L(X_j|UBM)}\right)$$

$$= \frac{1}{N_i}[logL(X_i|M_j^*) - log\, L(X_i|UBM)]$$

$$+ \frac{1}{N_j}[logL(X_j|M_i^*) - log\, L(X_j|UBM)]$$

em que a razão da verossimilhança cruzada $\frac{L(X_i|M_j^*)}{L(X_i|UBM)}$ mede quão bem o modelo adaptado do locutor j combina com os dados do locutor i em relação a quão bem o modelo de impostores combina com o conjunto de dados do locutor i (LE et al, 2007).

5.2.3 Logaritmo da Razão da Verossimilhança Cruzada Normalizada pelo Modelo de Impostores sem Adaptação

Uma variação do LRVCA é a não utilização da adaptação. A motivação desta abordagem é que como o locutor pode ter pouquíssimos dados disponíveis (segundos, por exemplo), a adaptação geraria modelos muito mais voltados para a modelagem dos dados dos impostores do que o próprio locutor. Assim, a distância logaritmo da razão da verossimilhança cruzada normalizada pelo modelo de impostores sem adaptação (LRVCI) pode ser definida como

$$LRVCI(Mi, Mj) = \frac{1}{N_i} log\left(\frac{L(X_i|M_j)}{L(X_i|UBM)}\right) + \frac{1}{N_j} log\left(\frac{L(X_j|M_i)}{L(X_j|UBM)}\right)$$

$$= \frac{1}{N_i}[logL(X_i|M_j) - log\, L(X_i|UBM)]$$

$$+ \frac{1}{N_j}[logL(X_j|M_i) - log\, L(X_j|UBM)]$$

em que a razão da verossimilhança cruzada $\frac{L(X_i|M_j)}{L(X_i|UBM)}$ mede quão bem o modelo do locutor *j* combina com os dados do locutor i em relação a quão bem o modelo de impostores combina com o conjunto de dados do locutor i (ZHU, 2006).

5.2.4 Modelos de Misturas Gaussianas

Modelos de misturas gaussianas são funções paramétricas de densidade probabilística representadas como a soma ponderada das densidades de componentes Gaussianos. Esses modelos são usados geralmente como modelos paramétricos da distribuição da probabilidade de medidas contínuas ou características em sistemas biométricos, como o trato vocal em relação às características espectrais em um sistema de reconhecimento de locutor (REYNOLDS, 2008). A utilização frequente de modelos de misturas Gaussianas em sistemas biométricos – especialmente em sistemas de reconhecimento de locutor – se deve à sua capacidade de representar uma grande classe de distribuições de amostras. Os parâmetros de um modelo de misturas Gaussianas são estimados a partir de dados de treinamento usando o algoritmo iterativo de maximização da expectativa ou a estimativa *maximum a posteriori* a partir de um modelo anterior bem treinado (REYNOLDS, 2008).

O modelo de misturas Gaussianas completo é parametrizado pelos vetores de média, pelas matrizes de covariância e pelos pesos das misturas de todas as densidades dos componentes. As matrizes de covariância podem ser completas ou diagonais. Além disso,

alguns parâmetros podem ser compartilhados entre os componentes Gaussianos (e. g. uma matriz de covariância pode ser usada para todos os componentes). A escolha da configuração do modelo (como o número de componentes e se a matriz de covariância é completa ou diagonal) geralmente é determinada pelo tamanho do conjunto de dados disponível para a estimativa dos seus parâmetros e por como o modelo é usado em uma aplicação biométrica (REYNOLDS, 2008).

Dados vetores de treinamento e uma configuração para o modelo, os parâmetros de um modelo de misturas Gaussianas podem ser estimados através de várias técnicas, sendo a da verossimilhança máxima a mais utilizada. O objetivo da estimativa da verossimilhança máxima é encontrar os parâmetros para o modelo que maximizam a verossimilhança do modelo de misturas Gaussianas, dado um conjunto de dados para treinamento (REYNOLDS, 2008). A estimativa desses parâmetros pode ser obtida iterativamente através de um caso especial do algoritmo de maximização da expectativa (DEMPSTER et al, 1977).

A idéia básica do algoritmo de maximização da expectativa é estimar, a partir de um modelo inicial λ, um novo modelo $\bar{\lambda}$, de forma que a verossimilhança do novo modelo seja maior que a do primeiro. A cada iteração, o modelo estimado na etapa anterior se torna o modelo inicial, e o processo é repetido até que um limiar de convergência seja alcançado. O modelo inicial geralmente é obtido através de alguma forma de estimativa de quantização vetorial binária. Os pesos das misturas, suas médias e suas matrizes de covariância são re-estimados a cada iteração do algoritmo, buscando o aumento da verossimilhança do modelo (REYNOLDS, 2008).

Outra forma de estimar os parâmetros de um modelo de misturas Gaussianas é a estimativa *maximum a posteriori*. Essa estimativa é usada, por exemplo, para derivar modelos de locutor através da adaptação de um modelo de impostores (também chamado modelo universal) em aplicações de reconhecimento de locutor (REYNOLDS et al, 2000). Também é usada em outras tarefas de reconhecimento de padrões em que dados para treinamento rotulados limitados são usados para adaptar um modelo anterior, genérico. A estimativa *maximum a posteriori* é dividida em duas etapas: a primeira consiste do cálculo das estimativas de estatísticas suficientes para cada mistura no modelo anterior; a segunda é a adaptação, em que ocorre a combinação das novas estimativas de estatísticas suficientes com as antigas (REYNOLDS, 2008).

Modelos de impostores são modelos usados em sistemas de verificação biométrica para representar características genéricas, que são comparados com modelos de características

específicas de uma pessoa para se aceitar ou rejeitar uma decisão. Em um sistema de verificação de locutor, por exemplo, o modelo de impostores é um modelo de misturas Gaussianas treinado com amostras de fala de um grande conjunto de locutores para representar características genéricas da fala. A partir de um modelo de misturas Gaussianas específico de um locutor e treinado com amostras de fala de um locutor presente na conversa, um teste da razão da verossimilhança com uma amostra de fala desconhecida pode ser feito entre o modelo de locutor específico correspondente e o modelo de fundo universal. Esses modelos podem ser usados para treinamento de um modelo específico de locutor, atuando como modelo anterior na estimativa de parâmetros para a *maximum a posteriori* (REYNOLDS, 2008).

5.3 Resultados e Discussão

Os resultados obtidos com o método DISTBIC (i. e. as hipóteses de onde ocorreram mudanças de locutor) foram submetidos a um processo de agrupamento. Uma vez que cada segmento obtido com o método DISTBIC é rotulado em relação ao anterior (e. g. locutor 1, locutor 2, locutor 1, locutor 2, ...) e nem sempre ocorreu de fato uma mudança de locutor no tempo em que o DISTBIC fez a detecção, é interessante agrupar os segmentos e rotulá-los novamente. A medida de distância utilizada para realizar o agrupamento dos segmentos foi o KL2.

A Tabela 10 mostra os resultados do agrupamento dos segmentos obtidos com o DISTBIC na fase de detecção de mudança de locutor com variações na penalidade (λ) e no coeficiente α. Pares de Linhas Espectrais foram usados nestes experimentos.

Tabela 10. Resultados do agrupamento com Pares de Linhas Espectrais

λ	α	Detecções Perdidas	Falso Alarme	Erro de Rotulação de Locutor
0,5	0,1	9,5%	1,7%	24,5%
0,6	0,1	9,4%	1,7%	24,6%
0,7	0,1	9,3%	1,7%	24,6%
0,8	0,1	8,5%	1,6%	24,7%
0,9	0,1	8,2%	1,5%	24,5%
1,0	0,1	9,5%	1,4%	21,9%
1,1	0,1	11,7%	1,1%	19,6%
1,2	0,1	15,9%	0,8%	15,8%
1,3	0,1	19,7%	0,6%	12,4%
1,4	0,1	23,6%	0,4%	9,2%
1,5	0,1	26,2%	0,3%	6,9%

0,5	0,5	9,9%	1,8%	24,0%
0,6	0,5	9,9%	1,8%	24,1%
0,7	0,5	9,8%	1,8%	24,0%
0,8	0,5	8,1%	1,6%	24,7%
0,9	0,5	7,6%	1,5%	24,5%
1,0	0,5	8,7%	1,3%	22,3%
1,1	0,5	10,7%	1,1%	20,1%
1,2	0,5	14,6%	0,8%	16,4%
1,3	0,5	18,4%	0,6%	13,1%
1,4	0,5	22,0%	0,4%	10,1%
1,5	0,5	24,6%	0,3%	7,8%

Na Tabela 11 são apresentados os resultados do agrupamento dos segmentos obtidos com o DISTBIC na fase de detecção de mudança de locutor com variações nos parâmetros λ e α. Coeficientes Cepstrais de Predição Linear foram utilizados nestes experimentos.

Tabela 11. Resultados do agrupamento com Coeficientes Cepstrais de Predição Linear

λ	α	**Detecções Perdidas**	**Falso Alarme**	**Erro de Rotulação de Locutor**
0,5	0,1	13,3%	2,0%	20,1%
0,6	0,1	13,4%	2,0%	20,0%
0,7	0,1	13,1%	2,0%	20,1%
0,8	0,1	11,8%	1,9%	20,6%
0,9	0,1	10,6%	1,7%	20,3%
1,0	0,1	11,2%	1,4%	18,4%
1,1	0,1	14,5%	1,2%	14,9%
1,2	0,1	18,2%	0,8%	11,8%
1,3	0,1	22,8%	0,5%	8,4%
1,4	0,1	26,4%	0,4%	5,8%
1,5	0,1	29,0%	0,3%	3,8%
0,5	0,5	0,0%	0,0%	40,2%
0,6	0,5	13,3%	2,0%	20,0%
0,7	0,5	12,9%	2,0%	20,2%
0,8	0,5	11,7%	1,9%	20,6%
0,9	0,5	10,0%	1,8%	20,6%
1,0	0,5	10,3%	1,5%	19,2%
1,1	0,5	13,4%	1,2%	15,4%
1,2	0,5	16,7%	0,8%	12,3%
1,3	0,5	21,2%	0,6%	9,1%
1,4	0,5	24,5%	0,5%	6,6%
1,5	0,5	27,0%	0,3%	4,9%

Com base nos resultados apresentados nas Tabelas Tabela 10 e Tabela 11, é perceptível que praticamente todas as taxas de detecções perdidas e de falso alarme aumentaram. Este efeito negativo é devido ao tamanho dos segmentos que o DISTBIC obtém: seu primeiro passo tem como objetivo obter segmentos grandes que serão submetidos ao Critério de

Informação Bayesiano no segundo passo. O segundo passo tem como objetivo descartar pontos candidatos a mudança de locutor encontrados no primeiro passo (i. e. alguns segmentos ficarão ainda maiores ao fim do segundo passo). O agrupamento é mais efetivo quando há mais segmentos e menores. Segmentos maiores têm maior probabilidade de conter fala dos dois locutores (especialmente quando é assumido que cada locutor fala por, no mínimo, dois segundos).

Em contrapartida, praticamente todas as taxas de erro de rotulação de locutor diminuíram, sendo notável a melhoria obtida com valores mais baixos para a penalidade e para o coeficiente α com os dois tipos de características e a melhora alcançada com um valor baixo para a penalidade e um valor mais elevado para o coeficiente α com Pares de Linhas Espectrais. Como valores baixos para a penalidade e para o coeficiente α implicam na detecção de mais pontos em que existe a hipótese de ter havido mudança de locutor, o agrupamento é mais eficiente. Segmentos com fala de dois locutores deixam de ser um problema quando o motivo são as pequenas palavras, pois cada segmento é rotulado novamente, de forma que o que passa a importar não é a ordem em que cada segmento está na conversa, mas sim sua distância em relação aos outros: segmentos com distâncias menores entre si têm maior probabilidade de pertencerem ao mesmo locutor do que segmentos com distâncias maiores entre si.

Assim como na avaliação dos resultados do DISTBIC, avaliações em que (1) todas as conversas ocorrem entre duas pessoas do sexo masculino, (2) todas as conversas ocorrem entre duas pessoas do sexo feminino e (3) todas as conversas ocorrem entre uma pessoa do sexo masculino e outra do sexo feminino foram realizadas. Os resultados dessas avaliações também ficaram bem próximos do resultado global. A Tabela 12 mostra os resultados das avaliações em que todas as conversas ocorrem entre duas pessoas do sexo masculino. A Tabela 13 mostra os resultados das avaliações em que todas as conversas ocorrem entre duas pessoas do sexo feminino. A Tabela 14 mostra os resultados das avaliações em que todas as conversas ocorrem entre uma pessoa do sexo masculino e outra do sexo feminino. Pares de Linhas Espectrais foram utilizados nos três casos.

Tabela 12. Resultados do agrupamento com Pares de Linhas Espectrais (conversas entre dois homens)

λ	α	Detecções Perdidas	Falso Alarme	Erro de Rotulação de Locutor
0,5	0,1	9,6%	2,0%	24,5%
0,6	0,1	9,5%	1,9%	24,3%
0,7	0,1	9,1%	1,9%	24,6%

0,8	0,1	9,0%	2,0%	24,3%
0,9	0,1	9,2%	1,9%	23,7%
1,0	0,1	11,0%	1,7%	20,8%
1,1	0,1	13,1%	1,4%	18,4%
1,2	0,1	15,9%	0,9%	15,9%
1,3	0,1	20,6%	0,7%	12,1%
1,4	0,1	24,4%	0,3%	8,6%
1,5	0,1	26,8%	0,2%	6,7%
0,5	0,5	10,0%	2,0%	24,0%
0,6	0,5	10,0%	1,9%	23,9%
0,7	0,5	10,1%	2,0%	23,7%
0,8	0,5	8,8%	2,0%	24,4%
0,9	0,5	8,3%	1,8%	23,9%
1,0	0,5	10,0%	1,8%	21,4%
1,1	0,5	11,9%	1,4%	19,5%
1,2	0,5	14,7%	0,9%	16,2%
1,3	0,5	19,2%	0,7%	12,2%
1,4	0,5	22,6%	0,2%	9,2%
1,5	0,5	24,9%	0,1%	7,4%

Tabela 13. Resultados do agrupamento com Pares de Linhas Espectrais (conversas entre duas mulheres)

λ	α	**Detecções Perdidas**	**Falso Alarme**	**Erro de Rotulação de Locutor**
0,5	0,1	8,4%	1,2%	26,3%
0,6	0,1	8,5%	1,2%	26,4%
0,7	0,1	8,2%	1,2%	26,4%
0,8	0,1	7,8%	1,2%	26,6%
0,9	0,1	7,5%	1,1%	26,6%
1,0	0,1	8,7%	1,0%	24,1%
1,1	0,1	11,6%	0,8%	20,7%
1,2	0,1	18,2%	0,6%	14,9%
1,3	0,1	22,0%	0,5%	11,6%
1,4	0,1	25,6%	0,4%	8,4%
1,5	0,1	28,5%	0,3%	5,7%
0,5	0,5	9,3%	1,3%	25,7%
0,6	0,5	9,4%	1,4%	25,7%
0,7	0,5	9,0%	1,3%	25,7%
0,8	0,5	7,7%	1,3%	26,5%
0,9	0,5	6,8%	1,1%	26,7%
1,0	0,5	7,9%	0,8%	24,5%
1,1	0,5	10,4%	0,7%	21,3%
1,2	0,5	16,5%	0,6%	16,0%
1,3	0,5	21,1%	0,5%	12,0%
1,4	0,5	24,6%	0,4%	9,3%
1,5	0,5	27,2%	0,4%	6,8%

Tabela 14. Resultados do agrupamento com Pares de Linhas Espectrais (conversas entre um homem e uma mulher)

λ	α	Detecções Perdidas	Falso Alarme	Erro de Rotulação de Locutor
0,5	0,1	10,4%	1,9%	23,2%
0,6	0,1	10,1%	1,8%	23,5%
0,7	0,1	10,3%	2,0%	23,2%
0,8	0,1	8,7%	1,7%	23,3%
0,9	0,1	8,2%	1,6%	23,3%
1,0	0,1	9,3%	1,5%	20,8%
1,1	0,1	10,8%	1,2%	19,5%
1,2	0,1	14,1%	1,0%	16,4%
1,3	0,1	17,3%	0,7%	13,3%
1,4	0,1	21,4%	0,4%	10,3%
1,5	0,1	24,0%	0,3%	8,1%
0,5	0,5	10,5%	2,0%	22,9%
0,6	0,5	10,4%	2,0%	23,0%
0,7	0,5	10,3%	2,0%	22,9%
0,8	0,5	8,1%	1,7%	23,5%
0,9	0,5	7,7%	1,5%	23,2%
1,0	0,5	8,6%	1,4%	21,1%
1,1	0,5	10,3%	1,1%	19,6%
1,2	0,5	12,9%	0,9%	17,0%
1,3	0,5	15,8%	0,7%	14,6%
1,4	0,5	19,5%	0,5%	11,3%
1,5	0,5	22,4%	0,4%	8,9%

A Tabela 15 mostra os resultados das avaliações em que todas as conversas ocorrem entre duas pessoas do sexo masculino. A Tabela 16 mostra os resultados das avaliações em que todas as conversas ocorrem entre duas pessoas do sexo feminino. A Tabela 17 mostra os resultados das avaliações em que todas as conversas ocorrem entre uma pessoa do sexo masculino e outra do sexo feminino. Coeficientes Cepstrais de Predição Linear foram utilizados nos três casos.

Tabela 15. Resultados do agrupamento com Coeficientes Cepstrais de Predição Linear (conversas entre dois homens)

λ	α	Detecções Perdidas	Falso Alarme	Erro de Rotulação de Locutor
0,5	0,1	13,8%	2,1%	19,3%
0,6	0,1	14,3%	2,2%	19,0%
0,7	0,1	14,2%	2,2%	18,8%
0,8	0,1	13,2%	2,2%	19,2%
0,9	0,1	12,3%	2,0%	18,9%
1,0	0,1	13,3%	1,9%	17,1%

1,1	0,1	15,4%	1,5%	14,3%
1,2	0,1	19,6%	0,9%	10,7%
1,3	0,1	24,0%	0,7%	7,6%
1,4	0,1	27,5%	0,4%	4,7%
1,5	0,1	29,2%	0,2%	3,7%
0,5	0,5	0,0%	0,0%	40,2%
0,6	0,5	14,6%	2,4%	18,8%
0,7	0,5	13,9%	2,4%	19,1%
0,8	0,5	13,3%	2,3%	19,1%
0,9	0,5	12,0%	2,2%	18,9%
1,0	0,5	12,0%	1,9%	18,0%
1,1	0,5	15,3%	1,5%	14,4%
1,2	0,5	18,9%	1,0%	11,0%
1,3	0,5	23,2%	0,7%	7,9%
1,4	0,5	26,3%	0,4%	5,2%
1,5	0,5	28,2%	0,2%	4,1%

Tabela 16. Resultados do agrupamento com Coeficientes Cepstrais de Predição Linear (conversas entre duas mulheres)

λ	α	**Detecções Perdidas**	**Falso Alarme**	**Erro de Rotulação de Locutor**
0,5	0,1	12,8%	1,7%	21,8%
0,6	0,1	13,1%	1,8%	21,6%
0,7	0,1	12,5%	1,7%	21,7%
0,8	0,1	10,9%	1,6%	22,8%
0,9	0,1	9,9%	1,2%	22,5%
1,0	0,1	10,9%	1,0%	19,9%
1,1	0,1	16,2%	0,8%	15,1%
1,2	0,1	19,9%	0,5%	12,3%
1,3	0,1	24,9%	0,2%	7,9%
1,4	0,1	29,0%	0,4%	5,0%
1,5	0,1	31,4%	0,3%	3,0%
0,5	0,5	0,0%	0,0%	40,9%
0,6	0,5	13,2%	1,8%	21,4%
0,7	0,5	12,3%	1,6%	21,9%
0,8	0,5	10,7%	1,5%	22,9%
0,9	0,5	8,0%	1,2%	23,6%
1,0	0,5	9, 5%	1,0%	21,3%
1,1	0,5	14,3%	0,8%	15,7%
1,2	0,5	18,1%	0,6%	12,7%
1,3	0,5	23,7%	0,4%	8,4%
1,4	0,5	27,5%	0,4%	5,5%
1,5	0,5	29,5%	0,3%	3,9%

Tabela 17. Resultados do agrupamento com Coeficientes Cepstrais de Predição Linear (conversas entre um homem e uma mulher)

λ	α	Detecções Perdidas	Falso Alarme	Erro de Rotulação de Locutor
0,5	0,1	13,3%	2,1%	19,2%
0,6	0,1	13,1%	2,0%	19,4%
0,7	0,1	12,9%	2,0%	19,6%
0,8	0,1	11,7%	1,9%	19,8%
0,9	0,1	10,0%	1,8%	19,5%
1,0	0,1	10,2%	1,5%	18,0%
1,1	0,1	12,6%	1,2%	15,3%
1,2	0,1	15,9%	0,9%	12,1%
1,3	0,1	20,3%	0,6%	9,5%
1,4	0,1	23,7%	0,5%	7,3%
1,5	0,1	27,1%	0,3%	4,5%
0,5	0,5	0,0%	0,0%	39,4%
0,6	0,5	12,6%	2,0%	19,7%
0,7	0,5	12,9%	2,1%	19,5%
0,8	0,5	11,4%	2,0%	19,9%
0,9	0,5	10,4%	2,0%	19,1%
1,0	0,5	9,9%	1,6%	18,2%
1,1	0,5	11,5%	1,2%	15,7%
1,2	0,5	14,3%	0,8%	12,9%
1,3	0,5	17,8%	0,7%	10,7%
1,4	0,5	20,9%	0,6%	8,6%
1,5	0,5	24,2%	0,3%	6,2%

A taxa de detecções perdidas obtida com o agrupamento dos segmentos obtidos com o DISTBIC com valor baixo para a penalidade sobre Coeficientes Cepstrais de Predição Linear foi melhor que a taxa obtida com o algoritmo baseado em energia do sistema de referência, mas a taxa de erro de rotulação de locutor foi muito mais alta. Com Pares de Linhas Espectrais e com valores mais altos para a penalidade, a taxa de detecções perdidas obtida com o agrupamento fica muito acima da taxa obtida com o algoritmo baseado em energia, e a taxa de erro de rotulação só fica baixa porque o número de segmentos obtidos é muito pequeno.

6 Conclusões

Este trabalho apresentou o processo de diarização de locutor, utilidades deste processo (i. e. aplicações que utilizam seus resultados), suas etapas e os pré-requisitos para sua execução. As etapas de detecção de mudança de locutor e de agrupamento foram implementadas (sendo a detecção de mudança de locutor feita com base no algoritmo DISTBIC de Delacourt e Wellekens (2000)) e seus resultados obtidos na avaliação do NIST.

O algoritmo DISTBIC não se mostrou eficiente comparado com o método baseado na energia sobre a base de conversas telefônicas. Uma vez que se assume que cada locutor fala por um tempo mínimo (dois segundos neste trabalho) e há palavras pequenas nas conversas, pontos em que ocorre mudança de locutor são perdidos, e os segmentos obtidos acabam contendo fala de dois locutores. Variações nos valores do coeficiente α (que define a validade de uma máxima local da série de distâncias entre segmentos) e, principalmente, da penalidade, afetam as taxas de falso alarme, de detecções perdidas e de erro de rotulação de locutor: valores mais baixos para a penalidade diminuem as taxas de falso alarme e de detecções perdidas, mas aumentam as taxas de erro de rotulação. Valores maiores para o coeficiente α afetam positivamente as taxas de alarme falso e de detecções perdidas quando o valor da penalidade é mais alto. O oposto ocorre com as taxas de erro de rotulação de locutor, que aumentam quando os valores da penalidade e do coeficiente α são aumentados.

A etapa de agrupamento melhorou a taxa de erro de rotulação de locutor, mas com o ônus do aumento das taxas de detecções perdidas e de falso alarme. Como o objetivo do DISTBIC é gerar segmentos grandes, a etapa de agrupamento não é tão eficiente quanto seria se os segmentos fossem menores e mais puros (i. e. com a fala de apenas um locutor). Além disso, tal objetivo pode ser invalidado quando se trata de conversas espontâneas, em que um locutor pode simplesmente reconhecer a fala do outro locutor através de confirmações curtas ou interjeições.

Uma vez que a base de dados é composta por conversas telefônicas, é necessária a utilização de outro algoritmo para realizar a detecção de mudança de locutor, de forma que os segmentos obtidos fossem menores, melhorando o desempenho do agrupamento. Além disso, operações de detecção de atividade de voz e de re-segmentação poderiam reduzir as taxas de erro, já que essas operações refinam os limites dos segmentos.

6.1 Trabalhos Futuros

Trabalhos futuros podem ser desenvolvidos para refinar os segmentos, para aplicar outras medidas de distância e para executar a diarização de locutor sobre outras características, ou ainda uma mistura dessas alternativas. O refinamento dos segmentos pode ser feito através da implementação das etapas de detecção de fala e de re-segmentação. Outras medidas de distância podem ser utilizadas tanto na detecção de mudança de locutor quanto na etapa de agrupamento. Alguns exemplos de distâncias são Mahalanobis, Bhattacharyya e Chernoff. Outros tipos de características que podem ser usados são o Mel-spectrum e Coeficientes Cepstrais de Frequência Mel. O número de dimensões para os Pares de Linhas Espectrais e os Coeficientes Cepstrais de Predição Linear utilizados neste trabalho foi 24. Mais ou menos dimensões podem ser utilizadas para quaisquer características que sejam empregadas na diarização de locutor.

Outros trabalhos também podem verificar o desempenho do DISTBIC com outros tipos de conversas. Uma vez que conversas telefônicas contêm muita fala espontânea e reconhecimentos da fala do outro locutor, esse tipo de conversa prejudica o algoritmo. Entretanto, outro tipo de base de dados pode mostrar melhores resultados.

7 Referências

ADAMI, André G.; KAJAREKAR, Sachin S.; HERMANSKY, Hynek. A New Speaker Change Detection Method for Two-speaker Segmentation. **IEEE ICASSP**, v. 4, n. 4, p. 3908-3911, 2002.

AJMERA, J; WOOTERS, C. A Robust Speaker Clustering Algorithm. **IEEE ASRU**, p. 411-416, 2003.

BEN, M.; BETSER, M.; BIMBOT, F.; GRAVIER, G. Speaker Diarization Using Bottom-up Clustering Based on a Parameter-derived Distance Between Adapted GMMs. **ICSLP**, Jeju Island, Coréia, out. 2004.

BIMBOT, Frédéric; MAGRIN-CHAGNOLLEAU, Ivan; MATHAN, Luc. Second-Order Statistical Measures for Text-Independent Speaker Identification. **Speech Communication**, v. 17, n. 1-2, p. 177-192, ago. 1995.

BULLINGTON, K.; FRASER, J. M. Engineering Aspects of TASI. **The Bell System Technical Journal**, p. 353-364, 1959.

CHEN, Scott Shaobing; GOPALAKRISHNAN, P. S. Speaker Environment and Channel Change Detection and Clustering Via the Bayesian Information Criterion. **DARPA Broadcast News Transcription and Understanding Workshop**, Landsdowne, VA, 1998.

DELACOURT, P.; WELLEKENS, C. J. DISTBIC: A speaker-based segmentation for audio data indexing. **Speech Communication**, v. 32, p. 111-126, 2000.

DEMPSTER, A.; LAIRD, N.; RUBIN, D. Maximum Likelihood from Incomplete Data via the EM Algorithm. **Journal of the Royal Statistical Society**. V. 39, p. 1-38, 1977.

DEMUYNCK, Kris; LAUREYS, Tom. A Comparison of Different Approaches to Automatic Speech Segmentation. **Proceedings of the 5th International Conference on Text, Speech and Dialogue**, p. 277-284, 2002.

GAUVAIN, J.-L.; LAMEL, L.; ADDA, G. Partitioning and Transcription of Broadcast News Data. **ICSLP**, v. 4, p. 1335–1338, dez. 1998.

GISH, H.; SIU, M.-H.; ROHLICEK, R. Segregation of Speakers for Speech Recognition and Speaker Identification. **IEEE ICASSP**, v. 2, p. 873-876, 1991.

HAIN, T.; JOHNSON, S. E.; TUERK, A.; WOODLAND, P. C.; YOUNG, S. J. Segment Generation and Clustering in the HTK Broadcast News Transcription System. **DARPA Broadcast News Transcription and Understanding Workshop**, Lansdowne, VA, p. 133–137, 1998.

HERSHEY, John R.; OLSEN, Peder A. Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models. **IEEE ICASSP**, v. 4, p. 317-320, 2007.

JOHNSON, S.E.; WOODLAND, P.C. A Method for Direct Audio Search with Applications to Indexing and Retrieval. **IEEE ICASSP**, v. 3, p. 1427-1430, 2000.

KEMP, T.; SCHMIDT, M.; WESTPHAL, M.; WAIBEL, A. Strategies for Automatic Segmentation of Audio Data. **IEEE ICASSP**, v. 3, n. 3, p. 1423-1426, 2000.

KIL, D. H.; SHIN, F. B. Pattern Recognition and Prediction with Applications to Signal Characterization. Woodbury: American Institute of Physics, 1996, 418 p.

LE, Viet-Bac; MELLA, Odile; FOHR, Dominique. Speaker Diarization using Normalized Cross Likelihood Ratio. **INTERSPEECH**, 2007.

McLOUGHLIN, Ian Vince. Line Spectral Pairs. **Signal Processing**, v. 88, n. 3, p. 448-467, mar. 2008.

MEIGNIER S.; MORARU, D.; FREDOUILLE, C.; BONASTRE, J.-F.; BESACIER, L. Step-by-Step and Integrated Approaches in Broadcast News Speaker Diarization. **Computer Speech and Language**, v. 20, p. 303-330, 2006.

MIRÓ, Xavier Anguera. **Robust Speaker Diarization for Meetings**. Barcelona: Universitat Politècnica de Catalunya, 2006. 248 p. Tese (Doutorado) – Speech Processing Group, Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona, 2006.

MORARU, Daniel; MEIGNIER, Sylvain; BESACIER, Laurent; BONASTRE, Jean-François; MAGRIN-CHAGNOLLEAU, Ivan. The Elisa Consortium Approaches in Speaker Segmentation During the NIST 2002 Speaker Recognition Evaluation. **IEEE ICASSP**, v. 2, p. 89-92, abr. 2003.

OLSEN, Jesper Ø. Separation of Speakers in Audio Data. **EUROSPEECH-1995**, p. 355-358, 1995.

PARK, A.; GLASS, J.R. A Novel DTW-Based Distance Measure for Speaker Segmentation. **IEEE Spoken Language Technology Workshop**, p. 22-25, dez. 2006.

PFAU, T.; ELLIS, D.; STOLCKE, A. Multispeaker Speech Activity Detection for the ICSI Meeting Recorder. **IEEE ASRU Workshop**, Trento, Itália, dez. 2001.

PIETQUIN, Olivier; COUVREUR, Laurent; COUVREUR, Pierre. Applied Clustering for Automatic Speaker-Based Segmentation of Audio Material. **Belgian Journal of Operations Research, Statistics and Computer Science (JORBEL), Special Issue Operations Research and Statistics in the Universities of Mons**, v. 41, n. 1-2, 2001.

REYNOLDS, D. A. Experimental Evaluation of Features for Robust Speaker Identification. **IEEE Transactions on Speech and Audio Processing**, v. 2, n. 4, p. 639-643, out. 1994.

REYNOLDS, D. A. Gaussian Mixture Models. **Encyclopedia of Biometric Recognition**. 2008.

REYNOLDS, D. A. Universal Background Models. **Encyclopedia of Biometric Recognition**. 2008.

REYNOLDS, D.; DUNN, R. B.; MCLAUGHLIN, J. J. The Lincoln Speaker Recognition System: NIST EVAL2000, **ICSLP**, Pequim, China, 2000.

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.

REYNOLDS, D. A.; TORRES-CARRASQUILLO, P. The MIT Lincoln Laboratory RT-04F Diarization Systems: Applications to Broadcast Audio and Telephone Conversations. **Rich Transcription Workshop (RT-04)**, Palisades, NY, nov. 2004.

SCHWARZ, Gideon E. Estimating the Dimension of a Model. **The Annals of Statistics**, v. 6, n. 2, p. 461-464, mar. 1978.

SIEGLER, M. A.; JAIN, U.; RAJ, B.; STERN, R. M. Automatic Segmentation, Classification and Clustering of Broadcast News. **DARPA Speech Recognition Workshop**, p. 97–99, fev. 1997.

SINHA, R.; TRANTER, S. E.; GALES, M. J. F.; WOODLAND, P. C. The Cambridge University March 2005 speaker diarisation system. **European Conference on Speech Communication and Technology (Interspeech)**, p. 2437–2440, set. 2005.

TANYER, S. Gökhun; ÖZER, Hamza. Voice Activity Detection in Nonstationary Noise. **IEEE Transactions on Speech and Audio Processing**, v. 8, n. 4, p. 478-482, 2000.

TRANTER, S.; REYNOLDS, D. An Overview of Automatic Speaker Diarisation Systems. **IEEE Transactions on Audio, Speech and Language Processing**, v. 14, n. 5, p. 1557-1565, 2006.

TRANTER, S. E.; YU, K.; EVERMANN, G.; WOODLAND, P. C. Generating and Evaluating Segmentations for Automatic Speech Recognition of Conversational Telephone Speech. **ICASSP**, v. 1, p. 753–756, maio 2004.

WOOTERS, Chuck; FUNG, James; PESKIN , Barbara; ANGUERA, Xavier. Towards Robust Speaker Segmentation: the ICSI-SRI Fall 2004 Diarization System. **Rich Transcription Workshop (RT-04)**, Palisades, NY, nov. 2004.

ZHU, Xuan. Speech Activity Detection and Speaker Diarization for Lectures. 2006. Trabalho apresentado ao RT-06s workshop, Bethesda, 2006.

ZHU, X.; BARRAS, C.; MEIGNIER, S.; GAUVAIN, J.-L. Combining Speaker Identification and BIC for Speaker Diarization. **European Conference on Speech Communication and Technology (Interspeech)**, p. 2441–2444, set. 2005.

NIST. **Benchmark Tests: 2000 Speaker Recognition Evaluatuion (SRE)**. Disponível em: <http://www.itl.nist.gov/iad/mig//tests/sre/2000/index.html>. Acesso em 20 jul. 2010.

NIST. **Benchmark Tests : 2002 Speaker Recognition Evaluatuion (SRE)**. Disponível em: <http://www.itl.nist.gov/iad/mig//tests/sre/2002/index.html>. Acesso em 20 jul. 2010.