Pontifícia Universidade Católica
do Rio de Janeiro

Marcelo de Campos Niero

Estudo Comparativo de Técnicas de Diarização de Locutor

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio.

Orientador: Prof. Álvaro de Lima Veiga Filho
Co-orientador: Prof. André Gustavo Adami

Rio de Janeiro

Setembro de 2014

Pontifícia Universidade Católica
do Rio de Janeiro

Marcelo de Campos Niero

Estudo Comparativo de Técnicas de Diarização de Locutor

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Álvaro de Lima Veiga Filho
Orientador
Departamento de Engenharia Elétrica – PUC-Rio

Prof. André Gustavo Adami
Co-Orientador
UCS

Prof. Marco Antonio Grivet Mattoso Maia
Centro de Estudos em Telecomunicações - PUC-Rio

Prof. Fernando Gil Vianna Resende Junior
UFRJ

Prof. Dante Augusto Couto Barone
UFRGS

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico

Rio de Janeiro, 9 de setembro de 2013

Marcelo de Campos Niero

Graduou-se em Ciência da Computação em 2010, tendo desenvolvido projetos junto ao Departamento de Ciência da Computação da UFF, como bolsista de iniciação científica do CNPQ. Durante a graduação também participou de projetos em conjunto com o laboratório ADDLabs para a Petrobrás. Bolsista CAPES do Programa de Mestrado do Departamento de Engenharia Elétrica da PUC-Rio, desenvolvendo pesquisa sobre processamento de voz aplicado a diarização de locutores.

Ficha Catalográfica

Niero, Marcelo de Campos

Estudo comparativo de técnicas de diarização de locutor / Marcelo de Campos Niero; orientador: Álvaro de Lima Veiga Filho; co-orientador: André Gustavo Adami. – 2014.

80 f. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2014.

Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Agrupamento de locutores. 3. Diarização de locutor. 4. Problemas no agrupamento de locutores. I. Veiga Filho, Álvaro de Lima. II. Adami, André Gustavo. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

Dedicado àquele que muda o tempo e as estações, o único digno de toda honra, glória e louvor.

Agradecimentos

Agradeço a Deus pela oportunidade de estudar em uma instituição de qualidade e também, por me capacitou e renovar minhas forças ao longo desta jornada. Além disso, por ter colocado grandes professores em meu caminho.

Aos meus pais, Renato e Iraci, e à minha irmã Cristiane que além de serem interessados pelo meu sucesso, me incentivaram em mais um conquista. Da mesma forma, agradeço aos meus queridos amigos e companheiros de jornada Luig Monteiro, Wendell Galdino, Breno Franco, Pedro Franco e João Almeida, que sempre estiveram presente me apoiando e torcendo por esta vitória.

Também sou grato aos professores Álvaro Veiga e André Adami por me orientarem neste trabalho. Sem o conhecimento adquirido através deles este trabalho não teria êxito.

Não poderia deixar de agradecer meu grande amigo Dirceu Silva por suas contribuições ao longo das pesquisas. Seu interesse pelo rumo deste foi mais uma fonte de motivação.

Finalmente, gostaria de agradecer à Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) que disponibilizaram recursos para a conclusão deste trabalho.

Resumo

Niero, Marcelo de Campos; Veiga Filho, Álvaro de Lima (Orientador); Adami, André Gustavo (Co-orientador). **Estudo Comparativo de Técnicas de Diarização de Locutor**. Rio de Janeiro, 2013. 80p. Dissertação de Mestrado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

A tarefa de diarização de locutor surgiu como forma de otimizar o trabalho do homem em recuperar informações sobre áudios, com o objetivo de realizar, por exemplo, indexação de fala e locutor. De fato, realizar a diarização de locutor consiste em, dado uma gravação de ligação telefônica, reunião ou noticiários, deve responder a pergunta "Quem falou quando?" sem nenhuma informação prévia sobre o áudio. A resposta em questão nos permite saber as referências temporais das atividades de cada locutor participante na gravação. Computacionalmente falando, o processamento da diarização ocorre através de quatro etapas principais: extração de características do sinal, detecção de fala e não fala, segmentação e agrupamento. Neste trabalho realiza-se um estudo sobre a etapa de agrupamento, comparando o desempenho e comprovando problemas de algumas técnicas do estado da arte. Todos os experimento foram executados em uma base controlada, originada do corpus TIMIT, e outra real utilizada no concurso NIST-SRE 2002.

Palavras-chave

Agrupamento de Locutores; Diarização de locutor; Problemas no Agrupamento de Locutores.

Abstract

Niero, Marcelo de Campos; Veiga Filho, Álvaro de Lima (Advisor); Adami, André Gustavo (Co-advisor). **Comparative Study of Techniques to Speaker Diarization**. Rio de Janeiro, 2013. 80p. MSc Dissertation - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

The speaker diarization task emerged as a way to optimize audio information retrieval processing by detecting and tracking speech and speaker information. Actually, speaker diarization consists in answering the question "Who spoke when" for a given conversation in a telephone call, meeting, or broadcast news, without any prior information about neither the audio nor the speakers. This answer allows us to know the time references for each speaker in a recording. Computationally speaking, the diarization processing occurs through four main steps: feature extraction of the signal, speech and non-speech detection, segmentation and clustering. In this work, the clustering step is analyzed by comparing the performance of some methods used in the state of the art and showing some of their problems. All experiments are performed on an excerpt from the TIMIT corpus and the diarization task database used in the 2002 NIST Speaker Recognition Evaluation

Keyword

Speaker Clustering; Speaker Diarization; Problems of Speaker Clustering.

Sumário

Lista de tabelas

Lista de figuras

1
Introdução

Realizar a diarização em áudio significa indexá-lo de acordo com classes de áudio. Em outras palavras significa informar, por exemplo, onde há música, fala, silêncio, música com fala, falas sobrepostas. O número e o tipo das classes podem variar de acordo com a aplicação. No entanto, quando a classe de interesse é locutor, esta é conhecida como diarização de locutor. Neste caso o objetivo da tarefa consiste em responder a pergunta: "Quem falou quando?", ou seja, dada uma conversação, deve-se identificar os trechos de falas pertencentes aos seus respectivos locutores, sem a prévia informação sobre os locutores. A Figura 1.1 exemplifica a resposta.

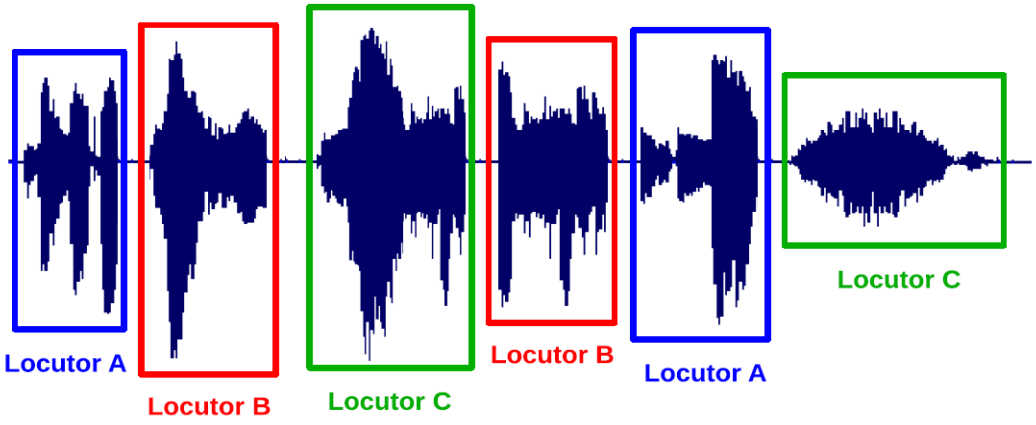

Figura 1.1: Exemplo de diarização de locutor - "Quem falou quando?"

A Figura 1.1 mostra um sinal de áudio com seis segmentos de falas de três locutores distintos. O resultado da diarização de locutor neste sinal deve informar que há três locutores participantes. Além disso, deve associar as falas aos seus respectivos locutores, como mostrado pelos retângulos azuis, vermelhos e verdes.

As pesquisas sobre diarização de locutor têm sido impulsionadas pela necessidade de recuperar informações a partir de grandes volumes de áudio. Uma vez que, nos dias de hoje, este volume cresce rapidamente, torna-se inviável para o homem ouvir milhares de horas de áudio a fim de obter os dados necessários. Estas informações podem ser, por exemplo, saber o número de locutores participantes em uma gravação e seus respectivos tempos de atividade, quando há

mais de um locutor envolvido, o conteúdo falado, o ambiente aonde coletado, o tópico do que é falado, entre outros.

Saber estas informações permite realizar a indexação de fala e locutor em diversos cenários, tais como, gravações de reuniões, conferências, programas televisivos, ligações telefônicas, debates, programas de rádio, entre outros. A indexação do locutor facilita a recuperação de informações de bases de áudio, manipulação de informações e torna os sistemas de reconhecimento de fala mais robustos à troca de locutores (já que os mesmos tendem a se adaptar ao locutor para melhorar o desempenho do reconhecimento).

1.1

Objetivo

O desempenho dos sistemas do estado da arte de diarização de locutor ainda pode melhorar, pois há diversos problemas, na segmentação e agrupamento, que devem ser abordados. Sendo assim o objetivo deste trabalho consiste em realizar um estudo detalhado sobre a fase de agrupamento, levantando e comprovando os problemas existentes, além de comparar diversas metodologias do estado da arte.

1.2
Estrutura da Dissertação

Este trabalho está dividido em cinco capítulos. No Capítulo 2 descrevemos o que é a diarização de locutor, esclarecendo as etapas principais dos sistemas de diarização, assim como, a forma de avaliar seu desempenho e os problemas conhecidos da tarefa em questão. Nos capítulos 3 e 4 realizamos experimentos com o objetivo de analisar a capacidade de discriminação de algumas métricas, assim como, critérios de parada apresentados na Seção 2.1. No entanto, utilizamos uma base artificial para realizar os experimentos no Capítulo 3, e uma base real no Capítulo 4. Por último, no Capítulo 5 aprestaremos as conclusões deste trabalho e sugestão para trabalhos futuros.

2
Diarização de Locutor

Diarização de locutor refere-se ao conjunto de técnicas que visam segmentar um áudio em regiões de fala de locutores homogêneos, respondendo a pergunta "Quem falou quando?". Normalmente os sistemas que realizam esta tarefa trabalham de forma não supervisionada, e sem nenhuma informação a respeito da identidade e do número de locutores participantes no sinal de áudio.

A tarefa em questão se tornou mais popular a partir do ano 2000 com as avaliações anuais *Speaker Recognition Evaluation* (SRE), realizadas pelo *National Institute of Standards and Technology* (NIST), onde foram promovidos grandes avanços no estado da arte das tecnologias de reconhecimento de fala. Entretanto, a partir de 2002, as avaliações passaram a ser conduzidas pelo NIST *Rich transcription* (RT), onde o objetivo é explorar a integração do reconhecimento automático de fala, em inglês *Automatic Speaker Recognition* (ASR), geração de texto a partir da fala (STT), e geração automática de metadados em diversos domínios. São eles: gravações de noticiários, ligações telefônicas e reuniões. Em geral estas competições englobam as seguintes tarefas:

- Fala para texto: converte as palavras faladas em texto.
- Diarização de locutor ou geração de metadados: encontrar os segmentos de tempo nos quais cada locutor participante fala.
- Fala para texto atribuído ao locutor: as palavras faladas são convertidas para texto e associadas ao seu locutor de origem.

2.1
Sistema de Diarização de Locutor

Conceitualmente a diarização de locutor pode ser vista como uma tarefa mais específica da diarização de áudio, a qual tem como objetivo segmentar e classificar o sinal sonoro em classes homogêneas. Estas classes podem ser: fala, silêncio, eventos acústicos, ou diferenças acústicas no ambiente, tais como

música, ruído, entre outras. Desta forma, diarização de locutor consiste no caso em que cada classe representa um único locutor.

Embora existam diversas formas de se abordar a diarização de locutor, a maioria delas segue o diagrama de blocos da Figura 2.1.

Figura 2.1: Diagrama de blocos do sistema de diarização de locutor tradicional

O processo de diarização é composto por quatro módulos: extração de características, detecção de fala, segmentação, e agrupamento. Na primeira etapa, o sinal de áudio é processado pelos módulos de detecção de fala e não fala, e pelo extrator de características. Enquanto o detector visa encontrar as regiões do áudio que possuem apenas fala, o extrator parametriza o sinal. Na sequência, as saídas do extrator e detector são processadas pelo segmentador, onde o sinal de áudio é dividido em segmentos que contenham no máximo um locutor. Finalmente, os segmentos são agrupados por locutor. Cabe ressaltar que alguns sistemas realizam a segmentação e agrupamento de uma só vez, ou seja, no mesmo módulo Wooters et al. [20], Anguera et al. [14], Leeuwen et al.[113], Friedland et al. [110], e Luque et al. [114].

2.1.1
Extração de Características

Primeiramente, um conjunto de características é obtido a partir de um áudio de entrada. Este conjunto, por sua vez, deve conter informações suficientes para distinguir os locutores. Portanto, a escolha do conjunto de características de fala pode afetar o desempenho do sistema.

As características mais usadas em diarização de locutor são: *Mel Frequency Cepstral Coefficients* (MFCC), *Linear Frequency Cepstral Coefficients* (LFCC), entre outras. O MFCC consiste em coeficientes cepstrais baseados na percepção

da audição humana e é extraído através de quatro etapas. Primeiro aplica-se a transformada discreta de *Fourier*, depois um banco de filtros de acordo com a escala *Mel*, em seguida calcula-se o logaritmo da energia sobre o resultado da filtragem, e por fim aplica-se a transformada discreta do cosseno à sequência de logaritmos, que descorrelaciona os coeficientes cepstrais. Cabe ressaltar que descorrelacionar os coeficientes cepstrais permite o uso de matrizes de covariâncias diagonais. A extração do LFCC difere apenas no banco de filtros, enquanto que no MFCC os filtros são espaçados de acordo com a escala *Mel*, no LFCC os filtros são igualmente espaçados.

Com o objetivo de aumentar o desempenho dos sistemas de diarização de locutor, alguns autores propuseram novas técnicas de extração de características. Yamaguchi et al. [1] atestaram que novas características baseadas em correlação cruzada espectral podem ajudar na discriminação da origem do sinal de áudio. Eles propuseram características como estabilidade espectral, similaridade de ruído branco e forma espectral para detectar se o áudio era proveniente de noticiários ou reuniões. Feito isso, as mesmas características propostas foram também utilizadas na etapa de indexação do sinal.

Com o objetivo de evitar influência de ruídos ou qualquer tipo de som que não seja fala, Pelecanos et al. [2] e Ouellet et al. [3] propuseram o uso de técnicas de *feature warping* para fazer um mapeamento da distribuição das características cepstrais observadas para uma distribuição previamente conhecida (gaussiana), com o objetivo de robustecer a distribuição das características cepstrais. Esta técnica foi utilizada com sucesso por Sinhaet al. [4] e Zhu et al.[5] em diarização de locutor em dois cenários: noticiário e reuniões.

Alguns trabalhos mostram que o uso de vários microfones para gravar uma reunião pode trazer ganho de desempenho. Nestes casos pode-se usar características de atraso temporal de chegada, como é apresentado por Pardo et al. [6], Pardo et al. [7], ICSI [8] e Lathoud et al. [9]. Com esta técnica é possível estimar a localização dos locutores participantes na gravação. Além disso, Ajmera et al. [15] combinou a característica de atraso temporal com MFCC a fim de se obter um aumento de desempenho na diarização.

Ferras et al. [10] alcançaram bons resultados com técnicas baseadas em Análise Conjunta de Fatores, em inglês *Joint Factor Analysis* (JFA), quando a variabilidade de locutor ou canal são abordados. A ideia principal desta

abordagem é explorar o conhecimento prévio sobre o espaço dos locutores e do canal. Esta técnica visa modelar os locutores como uma transformação linear entre vetores, de menor dimensionalidade, de fatores do locutor e do canal. Além disso, JFA faz parte do estado da arte em reconhecimento de voz e locutor. Já em diarização, o primeiro sistema utilizando JFA foi proposto por Castaldo em [11] e depois por Kenny em [12].

2.1.2
Detecção de Fala

Detecção de atividade de fala, em inglês *Spech Activity Detection* (SAD), consiste em rotular os segmentos em fala e não fala. Considera-se como não fala trechos de silêncio e ruídos do ambiente como o bater de portas, arrastar de cadeiras, buzinas e músicas de fundo. Algumas aplicações também incluem ruídos não léxicos, como risadas, tosses e espirros. O rotulamento incorreto de fala e não fala tem dois impactos na diarização de locutor. O primeiro relacionado ao desempenho das etapas de segmentação e agrupamento, pois estas devem processar apenas fala. O segundo é referente à taxa de erro da diarização, que leva em conta a falsa detecção (rotular não fala como fala) e a não detecção (rotular fala como não fala). A literatura relata diversas formas de atacar esta questão [41].

Inicialmente, a detecção de não fala era resolvida ao longo das etapas de segmentação e agrupamento. No entanto, é melhor ter uma etapa específica de detecção de fala e não fala antes da segmentação e agrupamento. Dentre as abordagens dedicadas a solucionar esta questão, pode-se destacar as baseadas em modelos por Zhu et al. [13], Anguera et al. [14], Fredouille et al. [16], e Wooters et al. [42]. Estas consistem em um classificador binário que utiliza modelos pré-treinados a partir de dados externos de fala e não fala. Como classificador tem-se utilizado a análise de discriminante linear, em inglês *Linear Discriminant Analysis* (LDA) combinado com a característica MFCC por Rentzeperis et al. em [17], e máquinas de vetor suporte, em inglês *Support Vector Machine* (SVM) por Temko et al. em [19]. O problema principal desta abordagem é a sensibilidade às variações das condições acústicas.

Wooters et al. [20], Anguera et al. [18], Nwe et al. [21], e El-Khoury et al. [22] optaram por uma abordagem de dois passos considerada híbrida, pois utiliza

técnicas baseadas em energia e baseadas em modelo. Na primeira passada, aplica-se um detector baseado em energia que rotula, com alta confiança, uma quantidade limitada de dados. Na etapa seguinte, os trechos rotulados são usados para treinar os modelos de fala e não fala, que serão empregados por um detector baseado em modelo. Este segundo detector tem como objetivo rotular os segmentos restantes e fornecer o resultado final do processo para a segmentação.

2.1.3
Segmentação

Tendo somente as características dos trechos com fala, deve-se então segmentar o sinal em regiões homogêneas. Para isso, deve-se detectar os pontos onde ocorrem mudança das condições acústicas, denominados pontos de trocas de locutor, e então associá-los a uma marca temporal.

Ajmera [47], Kemp et al. [56], Chen et al. [29] e Perez-Freire et al. [37] categorizam os algoritmos de segmentação em três grupos: baseados na métrica, modelo ou em silêncio. Além destas, Anguera [47] acrescentou uma quarta categoria, denominada outros, para os algoritmos que não são baseados em métricas, silêncio ou modelo.

2.1.3.1
Segmentação Baseada em Métrica

Os algoritmos baseados em métricas são provavelmente os mais populares. Estes são baseados no cálculo da distância entre dois segmentos acústicos adjacentes a fim de verificar se eles pertencem ou não ao mesmo locutor. Caso pertençam a locutores diferentes pode-se afirmar que há um ponto de mudança de locutor entre os segmentos.

Além disso, estas distâncias podem ser classificadas em dois tipos. O primeiro, chamado de distância baseada em estatísticas, o qual compara a estatística suficiente entre um par de segmentos sem considerar o uso de qualquer modelo. Esta pode ser calculada rapidamente por diversas formas e obter bons resultados se os segmentos grandes o suficiente de forma que tenham boa representação estatística. Já o segundo tipo, conhecidos por técnicas baseadas em

verossimilhança, são baseados na avaliação da verossimilhança entre um conjunto de dados e os modelos que o representam. Embora este tipo possa obter resultados melhores que o primeiro é mais lento de se calcular.

A maioria das distâncias usadas para detectar mudança acústica também é aplicada à fase de agrupamento, onde o foco é atestar se dois grupos pertencem a um mesmo locutor. Sendo assim serão apresentados os detalhes sobre a notação a ser utilizada, seguido das métricas do estado da arte.

Consideremos dois segmentos de áudio (i,j) e seus respectivos vetores de parâmetros acústicos X_i e X_j de comprimento N_i e N_j , com médias e variâncias μ_i , σ_i e μ_j, σ_j, respectivamente. Cada um destes segmentos pode ser modelado por um modelo gaussiano $\mathcal{M}_i(\mu_i,\sigma_i)$ e $\mathcal{M}_j(\mu_j,\sigma_j)$, o qual pode ser uma única gaussiana ou por um modelo de misturas gaussianas; em inglês, *Gaussian Mixtures Models* (GMM). Além disso, consideremos o segmento, X , gerado a partir a união de X_i e X_j, com média μ , variância σ , modelado por um modelo gaussiano $\mathcal{M}(\mu,\sigma)$, e $\mathcal{L}(X,\mathcal{M}(\mu,\sigma))$ a função de verossimilhança. Desta forma as métricas do estado da arte são:

2.1.3.1.1
Generalized Likelihood Ratio (GLR)

Introduzido por Willsky et al. em [48] e por Appel et al. em [49] para detectar mudança de locutor. Consiste em uma razão de verossimilhanças para comparar duas hipóteses. Dado dois segmentos, a hipótese H_0 considera ambos os segmentos provenientes do mesmo locutor, portanto $X = X_i \cup X_j \sim \mathcal{M}(\mu,\sigma)$ representa melhor os dados. Por outro lado, H_1 assume que os segmentos foram falados por locutores diferentes, assim $X_i \sim \mathcal{M}_i(\mu_i,\sigma_i)$ e $X_j \sim \mathcal{M}_j(\mu_j,\sigma_j)$ representam os dados de maneira mais adequada. Sendo assim, a GLR é dada por

$$GLR(i,j) = \frac{H_0}{H_1} = \frac{\max\limits_{\mu_i,\mu_j,\mu_i=\mu_j,\,\sigma_i,\sigma_j,\sigma_i=\sigma_j} \mathcal{L}(X,\mathcal{M}(\mu,\sigma))}{\max\limits_{\mu_i,\mu_j,\sigma_i,\sigma_j} \mathcal{L}(X_i,\mathcal{M}_i(\mu_i,\sigma_i))\,\mathcal{L}(X_j,\mathcal{M}_j(\mu_j,\sigma_j))},$$

EQUAÇÃO 2-1

onde a distância entre os segmentos é definida como $D_{GLR}(i,j) = -log(GLR(i,j))$ e um limiar, configurado previamente, é adotado para decidir se os segmentos possuem ao mesmo locutor ou não. Esta métrica é semelhante ao *log-likelihood ratio* (LLR), enquanto que o LLR considera as funções de distribuições de probabilidades e os modelos conhecidos a priori, o GLR assume que as *fdps* são desconhecidas, logo os modelos são estimados diretamente dos dados a serem analisados [64]. Além desta diferença, quando aplicado em segmentação o GLR é geralmente calculado sobre segmentos adjacentes de mesmo tamanho ao longo do sinal de áudio, e a decisão é baseada em um limiar pré-fixado ou adaptado dinamicamente.

Bonastre et al. [50] utilizaram a métrica GLR em um sistema de rastreamento de locutor para encontrar os pontos de troca de locutores em apenas uma passada pelo áudio. Neste sistema, o limiar de decisão foi configurado de forma a minimizar a probabilidade de não detecção, pois cada segmento encontrado é considerado um potencial locutor a ser rastreado. Já Gangadharaiah et al.[51] aplicaram uma segmentação de dois locutores em duas etapas, onde a primeira foi feita com a GLR e a segunda utilizando a decodificação *Viterbi*.

Adami et al. [52] fizeram uma segmentação, em duas etapas, para dois locutores. Na primeira, assumiu-se que o primeiro segundo de fala era do primeiro locutor e o segundo locutor foi encontrado determinando o ponto de troca através da GLR. No passo seguinte, os segmentos encontrados eram associados aos seus respectivos locutores comparando-se o resultado da GLR entre cada locutor e todos os segmentos encontrados.

Han et al. [53] apresentam uma variação da GLR denominada *Information Change Rate* (ICR). Esta, por sua vez, consiste em normalizar o valor da GLR pelo somatório do número de observações do par de segmentos em análise. A partir do conhecimento de entropia, Han et al. [53] mostram que a medida ICR calcula a taxa de mudança de informação após a união de dois grupos. Desta forma, o valor ICR é próximo de zero quando dois grupos homogêneos são agrupado, e elevado para o agrupamento de um par heterogêneo. Sendo assim, a métrica ICR é definida como

$$ICR(i,j) = -\frac{1}{N_i + N_j} ln\, GLR(i,j).$$

EQUAÇÃO 2-2

2.1.3.1.2
Bayesian Information Criterion (BIC)

É provavelmente a métrica mais utilizada em segmentação e clusterização devido a sua simplicidade. Consiste em um critério de verossimilhança penalizado pela complexidade do modelo, ou quantidade de parâmetros livre, que foi introduzido na literatura por Schwarz et al. em [23] e [24] como um critério de seleção de modelos. Dado um segmento X_i, o BIC de um modelo $\mathcal{M}_i$ determina quão bem $\mathcal{M}_i$ representa os dados, e é descrito como

$$\mathrm{BIC}(\mathcal{M}_i) = log\,\mathcal{L}(X_i, \mathcal{M}_i) - \lambda\frac{1}{2}$$

EQUAÇÃO 2-3

onde λ é um parâmetro livre e depende dos dados sendo modelados, N_i é o número de observações do segmento X_i, e o número de parâmetros do modelo $\mathcal{M}_i$. Esta expressão é uma aproximação do Fator Bayesiano (Kass et al. [25] e Chickering et al. [26]) no qual os modelos acústicos são treinado pelo método de Máxima Verossimilhança e N_i é considerado grande.

Para usar o BIC como forma de verificar se existe um ponto de troca entre dois segmentos é preciso avaliar a hipótese de que $\mathcal{M}$ modela melhor os dados X contra a hipótese de que $\mathcal{M}_i$ e $\mathcal{M}_j$ representam melhor os dados X_i e X_j. Esta avaliação é dada por

$$\Delta\mathrm{BIC}(i,j) = -R(i,j) + \lambda P.$$

EQUAÇÃO 2-4

O termo $R(i,j)$ pode ser escrito para o caso dos modelos serem representados por apenas uma gaussiana como

$$R(i,j) = \frac{N}{2}log|\Sigma_X| - \frac{N_i}{2}log|\Sigma_{X_i}| - \frac{N_j}{2}log\left|\Sigma_{X_j}\right|,$$

EQUAÇÃO 2-5

onde Σ_X, Σ_{X_i} e Σ_{X_j} são respectivamente as matrizes de covariâncias dos dados X, X_i e X_j; P é uma penalidade que é função da dimensão p do vetor de características, e $N = N_i + N_j$. Para uma matriz de covariâncias completa, tem-se

$$P = \frac{1}{2}\left(p + \frac{1}{2}p(p+1)\right) log(N).$$

EQUAÇÃO 2-6

Já para o caso dos modelos serem representados por GMM, a Equação 2-4 pode ser escrita como

$$\Delta BIC(\mathcal{M}_i) = log\,\mathcal{L}(X, \mathcal{M}) - (log\,\mathcal{L}(X_i, \mathcal{M}_i) + log\,\mathcal{L}(X_j, \mathcal{M}_j)) - \lambda\Delta$$

EQUAÇÃO 2-7

onde Δ é a diferença do número de parâmetros entre modelo conjunto $\mathcal{M}$ e os modelos separados $\mathcal{M}_i$ e $\mathcal{M}_j$. Entretanto Stafylakis et al. [109] propuseram uma expressão para substituir o termo $\lambda\Delta$ da Equação 2-7. Enquanto que na formulação convencional do BIC a penalidade é multiplicada por λ, em uma das propostas mencionadas, o parâmetro λ foi introduzido na penalidade. Esta expressão é definida como

$$\frac{\frac{\wp_i}{2} log\,N_i + \wp_j}{2} log\,N_j - \wp_{iuj} log\,N,$$

EQUAÇÃO 2-8

onde $\wp_{iuj} = (\mathcal{M}_i + \mathcal{M}_j)\left[d\left(1 + \lambda\sqrt{N}\right) + 1\right] - 1$, $\wp_i = \mathcal{M}_i\left[d\left(1 + \lambda\sqrt{N}\right) + 1\right] - 1$, e $\mathcal{M}_i$ e $\mathcal{M}_j$ o número de misturas dos respectivos modelos. Deste modo, pode-se substituir o termo λP, da equação convencional do BIC (Equação 2-4), pela penalidade alternativa (Equação 2-8), assim chegando na expressão

$$\frac{\Delta BIC(i,j) = -R(i,j) + \frac{\wp_i}{2} log\,N_i + \wp_j}{2} log\,N_j - \wp_{iuj} log\left(N_i + N_j\right).$$

EQUAÇÃO 2-9

O BIC foi introduzido como técnica de segmentação em [27], [28] e [29] por Chen et al., onde cada modelo possuía uma matriz de covariâncias completa. Embora na formulação original do BIC não haja o parâmetro λ, este foi introduzido por Chen et al. em [27] como forma de ajustar a penalidade, assim assumindo a função de um limiar que deve ser previamente configurado para os dados de teste. Devido à dependência entre o λ e os dados, Tritschler et al. [30], Delacourt et al. [31][35], Mori et al. [32], Lopez et al. [33], e Vandecatseye et al. [34] propuseram formas automáticas de calibrá-lo.

Para contornar a questão da calibragem, Ajmera et al.[36] propôs uma formulação na qual a penalidade foi anulada. Nesta abordagem os modelos ($\mathcal{M}_i$, $\mathcal{M}_j$, $\mathcal{M}$) foram representados por GMMs. Além disso, a complexidade do modelo $\mathcal{M}$, formado pela soma dos segmentos $\mathcal{M}_i$ e $\mathcal{M}_j$, foi restrita em ser sempre igual à soma das complexidades de $\mathcal{M}_i$ e $\mathcal{M}_j$, ou seja, $\#(i \cup j) = \#(i) + \#(j) = 0$. Desta forma, $\Delta\#(i,j) = \#(i \cup j) - (\#(i) + \#(j)) = 0$, assim anulando o termo $\lambda\Delta$. O resultado desta restrição é equivalente à métrica GLR, porém com restrições impostas aos modelos.

Além da calibragem, há outras limitações. Na formulação original do BIC, Schwarz et al. [24] dizem que à medida que o tamanho da amostra tende a infinito, o valor BIC converge para menos infinito. Entretanto, nas aplicações reais de voz grandes amostragens podem se tornar um problema quando um dos grupos possui poucas observações. A fim de contornar este problema Perez-Freire et al. [37] realizou uma modificação na penalidade, e Vandecatseye et al. [38] contornou esta questão através do valor global.

Ainda sobre os problemas do BIC, pode-se citar a imprecisão quando os grupos possuem poucas amostras, que foi atacado por Roch et al. em [39] através de uma adaptação máxima a priori (MAP) dos modelos, e a questão do esforço computacional. Esta última foi solucionada por diversos autores utilizando o BIC apenas em uma segunda etapa da segmentação. Na primeira, um algoritmo denominado DISTBIC, exibido por Delacourt et al. em [31],[35] e [40] realizou uma análise dos possíveis candidatos a ponto de quebra através da métrica (GLR). De maneira semelhante, Wooters et al. [42], Kim et al. [43] e Tranter et al. [44] propuseram o uso da distância *Hotelling* T^2; Lu et al. [45] subsistiu a GLR pela

distância de Kullback-Leibler (KL2), e Vandecatseye et al. [34] optou pela GLR normalizado (NGLR).

2.1.3.1.3
Divergência de *Kullback-Leibler* (KL)

É uma medida de similaridade entre duas funções de distribuição de probabilidade (*fdp*). Utilizada por Siegler et al. [57] nas etapas de segmentação e agrupamento, e por Hung et al. [58] somente na segmentação. Dado duas distribuições aleatórias X e Y, a divergência KL é definida como

$$KL(X;Y) = E_X\left(\log\frac{P_X}{P_Y}\right),$$

EQUAÇÃO 2-10

onde E_X é o valor esperado em relação a *fdp* de X. Quando as duas distribuições são Gaussianas, Campbell [59] apresentou uma forma fechada da Equação 2-10 em função das matrizes de covariâncias e das médias como

$$KL(X;Y) = \frac{1}{2}tr\left[(C_X - C_Y)\left(C_Y^{-1} - C_X^{-1}\right)\right] + \frac{1}{2}tr\left[(C_Y^{-1} - C_X^{-1})(\mu_X - \mu_Y)(\mu_X - \mu_Y)^T\right].$$

EQUAÇÃO 2-11

Entretanto, a divergência KL tem um comportamento assimétrico, o que levou Zochova et al. [60], Delacourt et al. [31] e Siegler et al. [57] a utilizarem a distância KL2 que pode ser obtida através de uma simetria de KL, definida como

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

EQUAÇÃO 2-12

Delacourt et al. [31] utilizou a distância de KL2 na primeira de duas etapas de um sistema de detecção de mudança de locutor. Em trabalhos posteriores, Zochova et al. [60] usaram a distância de KL2 em uma versão aprimorada do algoritmo de Delacourt et al. [31].

No trabalho [58] por Hung et al., os vetores de características passaram inicialmente por uma análise de componentes principais (ACP) e então foram

submetidos aos cálculos para encontrar a mudança de locutor. Além da distância KL também foram usadas as distâncias de *Mahalanobis* e *Bhattacharyya*.

2.1.3.1.4
Cross BIC (XBIC)

Introduzida em [67] e [68] por Anguera et al., consiste na avaliação de verossimilhanças cruzadas entre dois segmentos adjacentes.

$$\mathrm{BICX}(\boldsymbol{X_1}; \boldsymbol{X_2}) = \mathcal{L}(\boldsymbol{X_1}, \mathcal{M}_2(\mu_2, \sigma_2)) + \mathcal{L}(\mathrm{X}_2, \mathcal{M}_1(\mu_1, \sigma_1)).$$

EQUAÇÃO 2-13

Em [69] os autores propuseram uma métrica similar, além de um estudo sobre formas de normalizar a verossimilhança a fim de tornar mais robusta a métrica em questão, que por sua vez apresentou resultados melhores que a segmentação baseado no BIC.

2.1.3.2
Segmentação Baseada em Modelo

Técnicas baseadas em modelo fazem o uso de um conjunto de modelos a fim de classificar observações acústicas em classes. Estas podem ser locutores participantes no áudio, assim como classes mais genéricas, por exemplo, vozes masculinas e femininas. Geralmente as classes são representadas por GMM, e as observações são verificadas através de um teste de Máxima Verossimilhança (MV) como em [56] por Kemp et al. e [80] por Kubala et al., ou por decodificação de *Viterbi* em [81] por Gauvain et al.. Um dos problemas desta técnica é a necessidade de ter uma base de dados, na mesma condição do sinal de teste, para treinar previamente os modelos.

Com o intuito de eliminar essa necessidade, outras abordagens utilizam uma segmentação inicial para treinar os modelos dos locutores com o próprio sinal de teste. Neste caso, realiza-se uma segmentação iterativa para refinar os modelos. Contudo, quando o número de locutores é conhecido e o objetivo é unicamente refinar os modelos este processo é chamado de re-segmentação, exibido em Reynolds et al.[82]. Caso o número de locutores não seja conhecido este tipo de

segmentação deve rodar juntamente com alguma técnica de agrupamento para re-estimar o número de locutores e refinar os modelos iterativamente, como mostrado em Ajmera et al. [83].

2.1.3.3
Segmentação Baseada em Silêncio

Estas técnicas são baseadas na teoria de que a maior parte das trocas de locutor ocorre após um segmento de silêncio. Sendo assim, alguns trabalhos na literatura, tais como Kemp et al. [76], Ali et al. [77], e Bertrand et al. [78], usam um detector de energia a fim de encontrar os trechos de silêncio. Além disso, Huang et al. [79] apresentou diversas características para a segmentação de fala e não fala. Por outro lado, Kubala et al. [80] fizeram o uso de sistemas de reconhecimento de voz com o objetivo de encontrar segmentos de silêncio.

Embora esta técnica possa obter bons resultados, os sistemas baseados em silêncio não são os mais populares na tarefa de diarização de locutor. Isto se dá ao fato de que muitos segmentos de silêncio não correspondem aos pontos de troca; logo se necessita da aplicação de outras técnicas para verificar se o silêncio encontrado está ou não associado ao ponto de mudança de locutor. Muitas trocas não ocorrem seguidas de silêncio, mas sim de falas sobrepostas, o que degrada o desempenho desta técnica.

2.1.3.4
Outras Técnicas de Segmentação

Há algumas técnicas de segmentação de locutor que não se adequam a nenhuma das especificadas anteriormente, portanto cabe destacá-las nesta seção. Em [84], Vescovi et al. propôs o uso de programação dinâmica para detectar os pontos de troca de locutor. Já Pwint et al. [85] tratou a segmentação um problema de otimização, onde as fronteiras dos segmentos foram detectados através de um algoritmo genético. Por fim Lathoud et al.[86] utilizou múltiplos microfones para propor uma segmentação baseada na estimação do posicionamento dos locutores. No trabalho de Lathoud et al. [86] a diferença entre duas localizações foi

empregada como características e técnicas de rastreamento de locutor para encontrar os pontos de mudança de locutor.

2.1.4
Agrupamento

Também conhecido como clusterização, tem como objetivo agrupar os segmentos acusticamente homogêneos gerados na segmentação, sem nenhuma informação prévia sobre as classes. Cada segmento é associado a um rótulo que representa um grupo ou um locutor único. Sendo assim, o conjunto de registros de tempo obtidos na segmentação, combinados com seus rótulos gerados na clusterização compõe a saída da diarização, chamada na Figura 2.1 de rótulos da diarização.

O método de clusterização mais utilizado é denominado agrupamento hierárquico (AH). Os dois tipos de aplicação hierárquica são *Bottom-up* e *Top-down*, ilustrados na Figura 2.1.

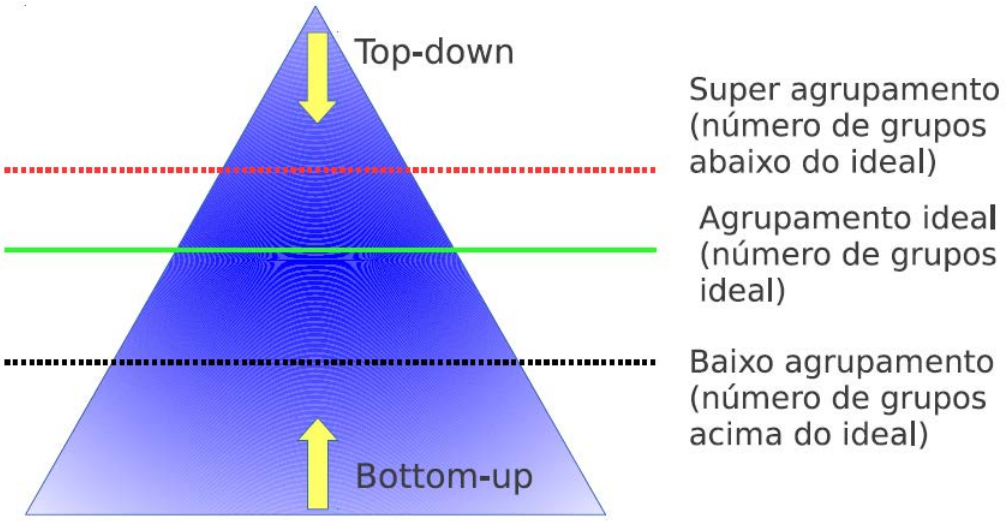

Figura 2.2: Esquema de agrupamento *Bottom-up* e *Top-down*.

2.1.4.1
Bottom-up

A clusterização se inicia considerando cada segmento um grupo distinto e procede iterativamente agrupando os segmentos mais próximos até que algum

critério de parada seja alcançado. Esta técnica é a mais utilizada para agrupar locutores em diarização de locutor devido a sua simplicidade de aplicação nos segmentos gerados pela segmentação.

Geralmente cria-se uma matriz contendo as distâncias entre todas as possíveis duplas de segmentos. Feito isso, procura-se e une o par de segmentos com menor distância. Após a união, a dupla unida é removida da matriz, o novo grupo é inserido e são calculadas as distâncias entre o novo grupo e todos os outros segmentos. Este se repete até que o critério de parada definido seja alcançado, o qual pode ser um número de grupos ou alguma métrica. Rougui et al. [75], propuseram a distância entre dois GMM baseada na métrica KL. Assim, sendo dois modelos M_1 e M_2, com K_1 e K_1 gaussianas cada, com pesos $W_1(i), i = 1..K_1$ e $W_2(j), j = 1..K_2$, respectivamente, a distâncias de M_1 para M_2 foi definida como

$$d(M_1, M_2) = \sum_{i=1}^{K_1} \boldsymbol{W}_1(\boldsymbol{i}) \min_{j=1}^{K_2} KL(N_1(i), N_2(j)),$$

EQUAÇÃO 2-14

onde $N_1(i)$ é a gaussiana i do GMM. No entanto, Ben et al. [66] e Moraru et al. [90] optaram por gerar os GMMs através de uma adaptação MAP das médias (os pesos e variâncias permanecem iguais) e calcular a distância entre eles através de uma métrica derivada da distância KL2, expressa como

$$D(M_i, \boldsymbol{M}_j) = \sqrt{\sum_{i=1}^{M} \sum_{d=1}^{D} W_m \frac{(\mu_i(m,d) - \mu_2(m,d))^2}{\boldsymbol{\sigma}_{m,d}^2}},$$

EQUAÇÃO 2-15

onde $\mu_i(m,d)$ e $\mu_2(m,d)$ são as médias da d-ésima componente da m-ésima mistura, $\sigma_{m,d}^2$ a variância da d-ésima componente da m-ésima mistura e, M e D o número de misturas e a dimensão, respectivamente, do GMM.

Utilizando um critério mais simples, Ben et al. [66] optou por apenas um limiar, enquanto que Moraru et al. [90] substituiu o limiar pela aplicação do BIC em todo o sistema, levando em consideração todas as etapas da clusterização, denominado Global BIC. No entanto, é mais comum utilizar a BIC na matriz de distâncias e como critério de parada, como apresentado por Chen et al. em [27] e

[28]. Primeiro procura-se o par de segmentos na matriz de distância com maior ΔBIC ; em seguida une-se a dupla e atualiza-se a matriz com o novo grupo formado. Estas ações são repetidas até que todos os pares tenham ΔBIC maiores do que zero. Em trabalhos posteriores, Chen et al. em [29], Tritschler et al. em [93][30], Tranter et al. em [44], Cettolo et al. em [94], e Meinedo et al. em [95] propuseram modificações na penalidade da BIC e diferenças na segmentação. No entanto, as modificações de Cettolo et al. [94] foram aplicadas ao italiano e as de Meinedo et al. [95] ao português de Portugal.

Em outra pesquisa, a escolha pela métrica GLR foi feita e apresentada por Solomonov et al. em [91] e comparada com a KL2. Ambas as medidas foram usadas na matriz de distância, de forma que a clusterização se deu iterativamente até maximizar a média das estimações da pureza de cada grupo, calculada como

$$\bar{\rho} = \frac{1}{2}\sum_{m=1}^{M} n_m \cdot \rho_m ,$$

EQUAÇÃO 2-16

$$\rho_m = \sum_{p=1}^{P} \left(\frac{n_{mp}}{n_{m^*}}\right)^2 ,$$

EQUAÇÃO 2-17

onde ρ_m é a pureza do grupo m , n_{m^*} é o número total de segmentos do grupo m , n_{mp} o número de segmentos do grupo m produzidos pelo locutor S_p, e P é o número de locutores. Este mesmo critério de parada foi utilizado por Tsai et al. em [92], onde vários métodos foram apresentados para criar um espaço de referência das características que represente melhor a similaridade entre dois locutores. Além disso, a métrica cosseno foi usada na matriz de distâncias.

Barras et al. em [96] e [97], e Zhu et al. [5] propuseram um sistema de diarização utilizando técnicas abordadas em identificação de locutor. Nestes três trabalhos a modelagem dos dados foi feita através de uma adaptação MAP de um modelo universal, popularmente conhecido como *Universal Background Model* (UBM) e como distância adotou-se a *Cross Likelihood Ratio* (CLR), definida por Reynolds et al. em [98] como

$$D(X_1, X_2) = \frac{1}{N_1} \log \frac{p(X_1|\mathcal{M}_{2-UBM})}{p(X_1|\mathcal{M}_{UBM})} + \frac{1}{N_2} \log \frac{p(X_2|\mathcal{M}_{1-UBM})}{p(X_2|\mathcal{M}_{UBM})},$$

EQUAÇÃO 2-18

onde $\mathcal{M}_{1-UBM}$ e $\mathcal{M}_{2-UBM}$ representam os modelos gerados pela adaptação MAP do UBM, e $\mathcal{M}_{UBM}$ o UBM propriamente dito.

Por fim, pode-se citar Leeuwen et al. [88] onde os resultados de testes de um sistema de verificação de locutor foram adotados diretamente como matriz de distâncias. Esta clusterização contou com uma base de dados contendo diversas gravações, cada uma com apenas um locutor. Sendo assim, o objetivo do sistema foi agrupar os segmentos através de testes de verificação de locutor. As tecnologias de verificação utilizadas foram GMM-*Support Vector Machine* (GMM-SVM) e técnicas de compensação de canal por Campbell et al. [99]. É importante esclarecer que qualquer sistema de verificação de locutor pode ser usado para obter as distâncias para um sistema de agrupamento hierárquico de locutores.

2.1.4.2
Top-Down

Esta abordagem inicia seu processamento com apenas um grupo o qual vai iterativamente sendo dividido até que um critério de parada seja alcançado. Primeiramente, foi aplicada em clusterização de locutor para reconhecimento automático de locutor (RAL) por Johnson et al. em [102]. Posteriormente Johnson et al. [103] e Tranter et al. [44] aplicaram em diarização de locutor, onde o algoritmo interativamente divide os dados em quatro subgrupos e permite que os mais similares sejam unidos. Pode-se citar Meignier et al. [104] e Anguera et al. [105] que propuseram o uso de um modelo oculto de *Markov* Evolutivo, em inglês, *evolutive hidden Markov model* (E-HMM), e um UBM gerado a partir de todo o dado de entrada para gerar os GMMs de novos locutores através de adaptações MAP. Esta abordagem é bem menos utilizada do que a *Bottom-Up* devido às características do problema.

2.1.5
Critério de Parada

O agrupamento hierárquico aglomerativo é uma estratégia não supervisionada de classificação que une iterativamente os pares mais próximos. Devido ao fato de ser iterativo e não supervisionado, surge a questão crítica de como parar o processo no ponto em que houver a menor taxa de erro possível. Em clusterização de locutor, o ponto ótimo de parada ocorre quando o número de grupos atual é igual ao número de classes, geralmente o número de locutores.

A forma mais simples de controlar o agrupamento e encontrar seu ponto de parada é através de um limiar. Neste caso, deve-se interromper o agrupamento quando todos os pares de grupos possuírem distâncias acima de um valor pré-estabelecido. Esta abordagem pode ser utilizada com qualquer métrica. Ben et al. utilizou esta técnica em [66] com a distância KL2 para calcular a simetria entre dois GMMs.

O critério de parada Delta BIC por Chen et al. em [27] é amplamente utilizado nos sistema de diarização. Este verifica, antes de cada etapa do agrupamento, se os grupos são homogêneos ou não em termos de identidade do locutor. Se a distância BIC for menor que um limiar considera-se que os grupos devem ser agrupados, caso contrário, devem permanecer separados. O processo de agrupamento é interrompido quando o par mais próximo possui distância maior que o limiar; pois pela lógica, todos os outros pares também terão distância maior que o limiar. Kenny et al. [12] e Iso et al. [100] adotaram este critério no agrupamento.

Por último, pode-se citar o critério Global BIC adotado por Stafylakis et al. em [101] e Moraru et al. em [90]. Este critério avalia todas as iterações do agrupamento, diferentemente do Delta BIC, que analisa somente a iteração corrente. O agrupamento baseado no Global BIC deve agrupar os segmentos até obter um único grupo, calculando após cada união o $\mathrm{BIC}(\mathcal{M}_s)$ definido como

$$\mathrm{BIC}(\mathcal{M}_s) = \sum_{i=1}^{N_{sp}} \log P(X_i|\mathcal{M}_i) - \lambda \frac{m}{2} N_{sp} \log \sum_{i=1}^{N_{sp}} N_{X_i},$$

EQUAÇÃO 2-19

onde $\mathcal{M}_s$ é o modelo que representa um estágio s do agrupamento, $\mathcal{M}_i$ é o GMM que modela o grupo X_i, λ um parâmetro de calibragem, N_{sp} o número de locutores ou grupos, N_{X_i} o número de observações de X_i, e m a complexidade do modelo que foi definido para o caso de matriz de covariâncias completa e diagonal como

$$m_{completa} = \left(p + \frac{1}{2}p(p+1)\right) \cdot n,$$

EQUAÇÃO 2-20

$$m_{diag} = 2 \cdot p \cdot n,$$

EQUAÇÃO 2-21

sendo p a dimensão do vetor de características e n o número de misturas do GMM. Por fim, o produto final do agrupamento é $\arg\max_s BIC(\mathcal{M}_s)$.

2.2
Avaliação do Desempenho da Diarização de Locutor

A principal medida de desempenho utilizada nas competições do NIST-RT é denominada Taxa de Erro de Diarização, ou, em inglês *Diarization Error Rate* (DER) [106]. Esta medida consiste na fração de tempo das regiões que não foram associadas corretamente aos locutores ou a não fala.

Esta medição é feita através da comparação entre as hipóteses do sistema de diarização e as referências, que possuem o gabarito da diarização. Por definição da tarefa da competição, as hipóteses não precisam identificar os locutores pelos respectivos nomes ou IDs, portanto os IDs associados nas hipóteses ou nas referências não precisam ser os mesmos. Além disso, os trechos de silêncio não devem aparecer na hipótese do sistema, pois estas regiões são identificadas pelo intervalo de tempo entre rótulos consecutivos. Desta forma, o *script* de avaliação otimiza o mapeamento de um para um entre cada ID das hipóteses com cada ID das referências. O DER é definido como

$$DER = \frac{\sum_{s=1}^{S} dur(s) \cdot \left(max\left(N_{ref}(s), N_{hip}(s)\right) - N_{correto}(s)\right)}{\sum_{s=1}^{S} dur(s) \cdot N_{ref}(s)}.$$

EQUAÇÃO 2-22

onde S é o conjunto de todos os segmentos onde mesmo locutor aparece tanto na referência quanto na hipótese do sistema, $dur(s)$ a duração do segmento s, $N_{ref}(s)$ e $N_{hip}(s)$ representam respectivamente o número de locutores da referência e da hipótese que falam no segmento s, e por fim $N_{correto}(s)$ indica o número de locutores da referência que estão falando no segmento s, os quais coincidem com os locutores da hipótese. O numerador da Equação 2-22 representa o tempo de erro de diarização, que também pode ser definido através do tempo de locutor, tempo de locutor não detectado, e tempo de falsa detecção de locutor.

O tempo de locutor que é associado ao locutor incorreto, denominado em inglês de *Speaker Error Time* (SET) é definido como

$$E_{SET} = \sum_{s=1}^{S} dur(s) \cdot \left(min\left(N_{ref}(s), N_{hip}(s)\right) - N_{correto}(s)\right).$$

EQUAÇÃO 2-23

Já o tempo de locutor não detectado, em inglês *Miss Speaker Time*, é a soma dos segmentos nos quais há um locutor de referência falando, mas não há hipótese do sistema. Este por sua vez é definido como

$$E_{miss} = \sum_{s=1}^{S} dur(s) \cdot \left(N_{ref}(s) - N_{hip}(s)\right).$$

EQUAÇÃO 2-24

Por último, o tempo de detecção errada, ou falso alarme, é a soma dos segmentos nos quais há um locutor de hipótese falando, mas não há atividade de nenhum locutor do sistema, e é expresso como

$$E_{fa} = \sum_{s=1}^{S} dur(s) \cdot \left(N_{hip}(s) - N_{ref}(s)\right).$$

EQUAÇÃO 2-25

Desta forma, pode-se reescrever a Equação 2-22 como

$$\mathrm{DER} = \frac{\mathrm{E_{SET}} + \mathrm{E_{miss}} + \mathrm{E_{fa}}}{\sum_{s=1}^{S} \mathrm{dur}(s) \cdot \mathrm{N_{ref}}(s)}.$$

Neste trabalho, o objetivo é avaliar somente a fase de agrupamento. Sendo assim, assume-se que as etapas de detecção de fala e não fala, e segmentação são realizadas sem erros, ou seja, não há não detecção nem falso alarme. Além disso, desconsideram-se os trechos de falas sobrepostas, os quais ocorrem quando há mais de um locutor falando no mesmo trecho. Com estas especificações, o erro de diarização torna-se igual ao erro de locutor (SET), que é a medida de desempenho usada neste trabalho. No entanto esta será expressa sempre em porcentagem, ou seja, dividida pelo tempo total avaliado.

2.3
Problemas Conhecidos

Um dos grandes problemas da diarização está relacionado às etapas de segmentação e clusterização. Durante estas, é necessário saber se dois segmentos são homogêneos ou não. Em outras palavras, a discriminação entre segmentos provenientes do mesmo locutor e de locutores diferentes deve ser eficiente.

As locuções de curta duração, menor do que 3 segundos, também são problemas, como relatado por Tranter et. al [117]. Estas merecem atenção tanto na detecção de fala, como na segmentação e clusterização. A pouca quantidade de observações presentes nestas, diminui a qualidade da estimação de parâmetros pertinentes às três etapas citadas, assim aumentando a probabilidade de erro do sistema.

Também, pode-se citar a dificuldade de tratar falas sobrepostas, como relatado por Vipperla et al. [118]. Em outras palavras, trechos onde há mais de uma pessoa falando ao mesmo tempo. Ainda que estas sejam detectadas, pode ser necessário saber quantos participantes estão ativos nesta região. Em muitas aplicações, estes segmentos são descartados nas fases de segmentação e agrupamento; ou classificados como falas sobrepostas devido a incapacidade de associá-los a somente um locutor.

Outra questão importante, relatada por Lapidot el al. [116], está relacionada ao número de participantes, relatado por que é conhecido em alguns cenários. Neste caso, devem-se agrupar os segmentos até que o número de grupos coincida com o número de participantes. No entanto, em muitas aplicações esta informação não é conhecida, e por isso, deve-se utilizar um critério de parada que não seja o número de participantes.

E por último, na maioria das aplicações todos os participantes são desconhecidos. Entretanto, o conhecimento total ou parcial da identidade dos participantes pode ajudar na diarização. Esta informação implica na existência de modelos que representem os participantes conhecidos. Com isso, pode-se utilizar técnicas de identificação e verificação tanto na segmentação como na clusterização, como é relatado por Lapidot el al. [116].

3
Análise Comparativa dos Métodos para Agrupamento de Locutor

Este capítulo tem como objetivo relatar os experimentos sobre a capacidade de discriminação de métricas e critério de parada que foram realizados em uma base controlada. Cada relato é seguido de uma análise dos resultados obtidos. Além disso, assume-se a segmentação ideal para todos os áudios, ou seja, sabem-se exatamente os tempos de início e fim das falas de cada locutor. Este capítulo está dividido em três seções. Na Seção 3.1 descrevemos a base de dados controlada. Na Seção 3.2, experimentos sobre a capacidade de discriminação das métricas de agrupamentos são analisados. Finalmente, na Seção 3.3, experimentos sobre critérios de parada do agrupamento são avaliados. Cabe ressaltar que o teste binomial para diferenças de proporções é usado para verificar se a diferença entre os valores de SET, EER e o número de conversações agrupadas corretamente, são estatisticamente significantes (Gillick and Cox, [115]). Quando não especificado, o nível de significância é de $\alpha = 0.05$.

3.1
Base de Dados

O corpus TIMIT [111] foi desenvolvido com a finalidade de prover dados de fala para estudos da fonética acústica, e desenvolvimento e avaliação de sistemas de reconhecimento de locutor. O TIMIT contém gravações, em Inglês, de 630 locutores, tanto homens como mulheres. Todas as gravações estão quantizados a uma taxa de 16 bits e amostrados à 16KHz.

A partir do corpus TIMIT, são criadas diversas conversas utilizando locuções dos locutores. As conversas possuem diferentes configurações, as quais são baseadas na variação do número de locutores, da duração das locuções, e do número de locuções por locutor. No entanto, todas as locuções de uma conversa possuem a mesma duração. A variação da configuração procede da seguinte forma:

- O número de locutores varia de 2 a 7.
- A duração das locuções recebe os valores: 0.5, 1, 2, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 25, 30, 35, e 40 segundos.
- Em cada conversa há quatro locuções de cada locutor presente.

Para as combinações em que as locuções possuem duração de meio segundo foram geradas 400 conversas, de um a dez segundos foram geradas 200 arquivos de áudio. Já para falas mais longas foram geradas 150 conversações. Além disso, os locutores participantes de cada áudio foram escolhidos aleatoriamente. Pode-se considerar esta base como limpa, pois em nenhuma gravação há ruído. Para todos os experimentos foram extraídos o mesmo conjunto de características, os primeiros 19 coeficientes MFCC, com janelas de 25 ms e deslocamento de 10 ms.

3.2
Capacidade de Discriminação das Métricas

O teste de capacidade de discriminação consiste em avaliar se uma métrica possui boa capacidade para discriminar segmentos provenientes do mesmo locutor e de locutores diferentes. Para uma melhor precisão dos resultados deve-se considerar a duração das locuções e o ruído presente. Baseado nestas considerações, este teste foi realizado somente na base controlada originada do TIMIT.

A análise em questão foi feita através da geração de dois gráficos. O primeiro exibe as funções de distribuição de probabilidade (*fdp*) para distância entre locuções de mesmo locutor e de locutores diferentes. Já o segundo, consiste em calcular a curva *Detection Error Tradeoff* (DET) [108] das duas distribuições, a qual mostra a relação entre as probabilidades de falso alarme e falsa rejeição, além do ponto onde elas são iguais, conhecido como *Equal Error Rate* (EER). Falso alarme ocorre quando o sistema aponta que locuções de pessoas diferentes pertencem a um mesmo locutor, enquanto que, na falsa rejeição segmentos originados do mesmo são considerados de locutores diferentes.

Dentre as métricas apresentadas na Seção 2.1.3, foram selecionadas para esta avaliação as distâncias: BIC convencional (BIC_c), BIC alternativa (BIC_a),

GLR, ICR, KL2, e por fim, a CLR. As seções a seguir mostram suas respectivas análises.

3.2.1
Bayesian Information Criterion

A distância BIC foi testada usando a formulação convencional (Equação 2-4) e também utilizando a penalidade alternativa (Equação 2-8). A motivação para se utilizar diferentes penalidades é baseada no comportamento desta métrica quando há variação do número de observações dentro de cada grupo em análise. A métrica BIC tende a assumir valores muito pequenos à medida em que o número de observações do par de segmentos em análise cresce. Com a finalidade de demonstrar este comportamento, foram gerados dois grupos de dados C_i e C_j, modelados por uma gaussiana bidimensional com médias $\mu_i = (0,0)$, $\mu_j = (2,2)$, e matrizes de covariâncias $\Sigma_i = \Sigma_j = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Com estes modelos, variou-se o número de observações de cada grupo de 50 a 10000, simulando de 0.5 a 100 segundos, de acordo com o tamanho do janelamento realizado. Em seguida, foi calculada a $\Delta BIC(i,j)$ para cada combinação. A Figura 3.1 mostra o comportamento da distância.

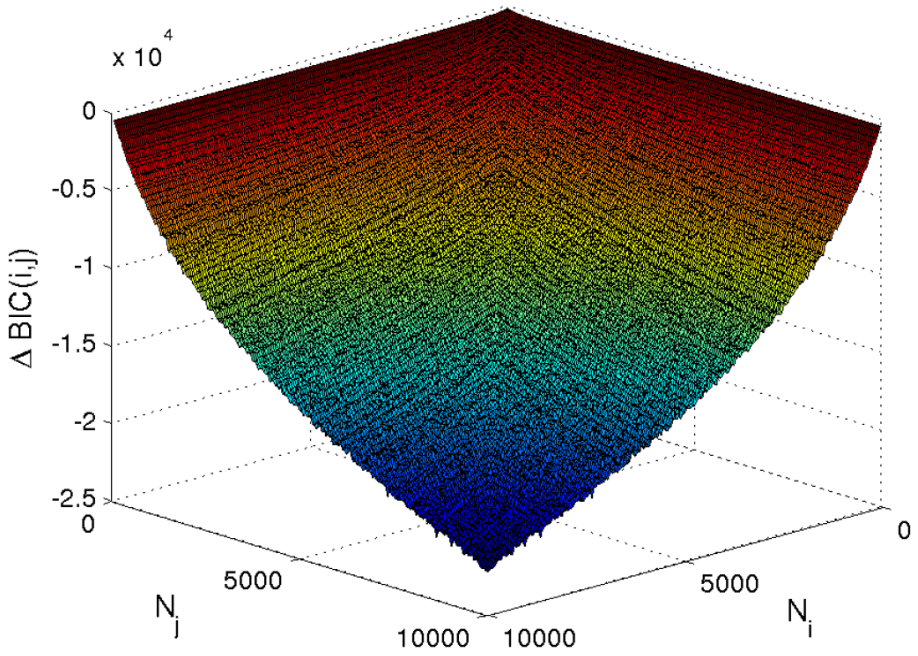

Figura 3.1: Relação entre o número de observações de dois grupos com a distância BIC.

Nota-se que a distância diminui com o aumento do número de observações dentro de cada grupo. Como resultado deste comportamento, pares de grupos heterogêneos que possuem muitas observações (região azul) são considerados mais próximos do que aqueles com poucas observações (região vermelha). Por outro lado, pares de grupos homogêneos constituídos de poucas observações possuem valores mais altos de $\Delta BIc(i,j)$, e podem ser considerados mais distantes do que aqueles com grandes quantidades de amostras, assim, podendo ser classificados erroneamente como heterogêneos. Este comportamento é indesejado na etapa de agrupamento, pois, deteriora a capacidade de discriminação dos locutores, e consequentemente, torna o sistema mais suscetível a agrupamento de pares heterogêneos. Desta forma, a penalidade é empregada como forma de reduzir o decrescimento mostrado na Figura 3.1.

A Figura 3.2 mostra os valores de EER, da distância BIC convencional, para cada duração de segmento. Observa-se que o EER é inversamente proporcional à duração das locuções, assim mostrando que a precisão da capacidade de discriminação é maior para segmentos mais longos, os quais possuem mais dados para estimar um modelo adequado.

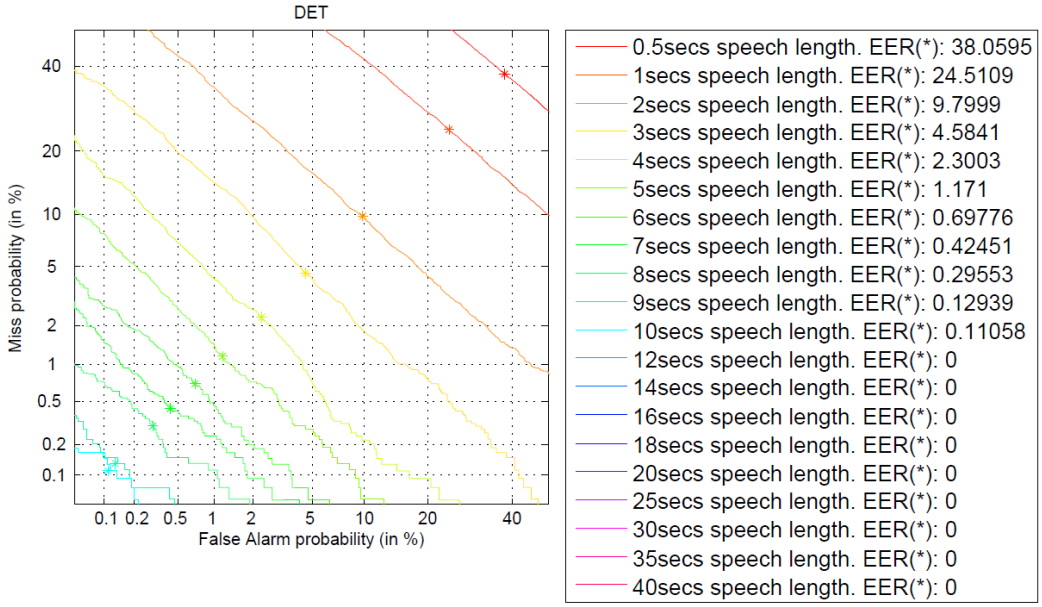

Figura 3.2: Curvas DET BIC convencional.

Com o objetivo de verificar o efeito de diferentes formulações de penalidade, a Figura 3.3 exibe os valores de EER, da medida BIC alternativa, para diversas durações de locuções. Analisando os valores de EER da Figura 3.3,

verifica-se que o aumento da duração das locuções, também promove uma melhoria na capacidade de discriminação da formulação alternativa da BIC. Nas falas de 0.5 segundos o erro é de aproximadamente 38%, porém, à medida em que a duração do segmento de voz aumenta, o EER diminui até chegar a zero. Em outras palavras, quanto menor a duração da locução, menor será a quantidade de dados disponíveis para estimar as variâncias e covariâncias, assim gerando um modelo mal representado, o que consequentemente aumenta a probabilidade de erro de discriminação.

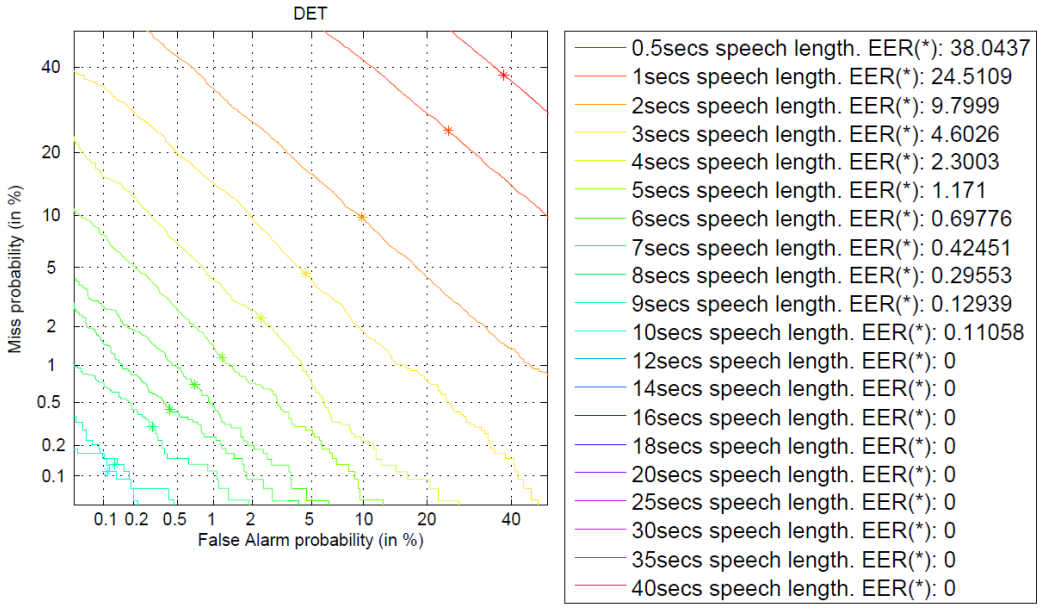

Figura 3.3: Curvas DET BIC com a penalidade alternativa.

Comparando-se os valores de EER da Figura 3.2 com os da Figura 3.3, verifica-se que a variação da formulação da penalidade, ou do parâmetro λ, não interfere na capacidade de discriminação para locuções de mesmo tamanho, pois os valores de EER são os mesmo para cada tamanho de locução. Nesta condição, a penalidade torna-se uma constante que é somada a todas as distâncias; logo pode ser descartada. Além disso, ela atua indiretamente no limiar de decisão, assim interferindo apenas no desempenho do agrupamento. Pode-se também comprovar este fato através da Equação 3-1a e Equação 3-1b. A Equação 3-1a, mostra que as verossimilhanças são normalizadas pelo logaritmo do tamanho dos grupos em análise. Pelo fato de todos os segmentos serem do mesmo tamanho, N será o mesmo para todos os pares de locuções. Também nota-se na Equação 3-1b, que o termo $\lambda\Delta$ funciona como um limiar de decisão.

$$\log \mathcal{L}(X, \mathcal{M}) - \log \mathcal{L}(X_i, \mathcal{M}_i) - \log \mathcal{L}\left(X_j, \mathcal{M}_j\right) - \lambda\Delta$$

EQUAÇÃO 3-1a

$$\frac{1}{\log(N)}\left(\log \mathcal{L}(X, \mathcal{M}) - \log \mathcal{L}(X_i, \mathcal{M}_i) - \log \mathcal{L}\left(X_j, \mathcal{M}_j\right)\right) < \lambda\Delta$$

EQUAÇÃO 3-1b

A relação entre o tamanho da fala e a capacidade de discriminação também pode ser visto através das fdps das distâncias entre locuções de mesmos locutores e locutores diferentes. A Figura 3.4 exibe o comportamento das curvas de distância com o aumento da duração dos segmentos de voz associadas às curvas DET da Figura 3.2 através da duração dos segmentos de voz. Estas se afastam à medida em que os modelos dos locutores são mais bem estimados, assim, aumentando a precisão da discriminação e diminuindo o EER. Nota-se que a variância e a média da distribuição das distâncias interlocutoras crescem com o aumenta do número de observações dos segmentos. Por outro lado, a média e variância da distribuição interlocutora diminuem.

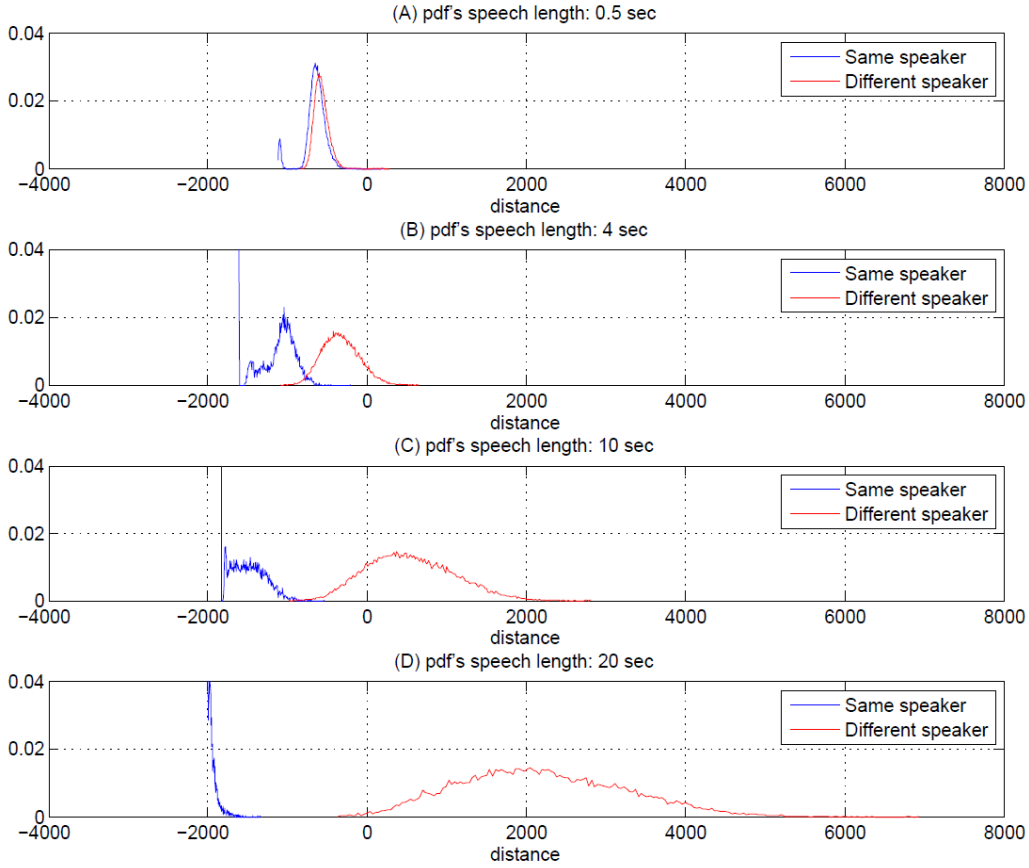

Figura 3.4: Comportamento das fdps das distâncias entre locuções de mesmo locutor e locutores diferente com a variação da duração da locução.

3.2.2
Generalized Likelihood Ratio

Os modelos avaliados pela GLR podem ser representados por GMMs. Contudo o aumento do número de gaussianas causa um crescimento na complexidade do modelo, aumentando o número de parâmetros a serem estimados, e a demanda por locuções mais longas a fim de gerar modelos bem representados. Por este fato, a capacidade de discriminação do GLR foi avaliada modelando os grupos de dados X_i, X_j, e a concatenação destas partes (X) por um modelo gaussiano com matriz de covariâncias diagonal. Desta forma, as dimensões são consideradas independentes e identicamente distribuídas (iid).

Além disso, a GLR tende a assumir valores grandes à medida em que o número de observações dentro de cada grupo cresce. A fim de demonstrar este comportamento, foram gerados dois grupos de dados C_i e C_j, modelados por uma gaussiana bidimensional com médias $\mu_i = (0,0)$, $\mu_j = (2,2)$, e matrizes de covariâncias $\Sigma_i = \Sigma_j = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Com estes dados fixos, variou-se o número de observações de cada grupo de 50 a 10000, simulando de 0.5 a 100 segundos, de acordo com o tamanho do janelamento realizado. Em seguida, foi calculado a $D_{GLR}(i,j)$ para cada combinação. A Figura 3.5 mostra o comportamento da distância.

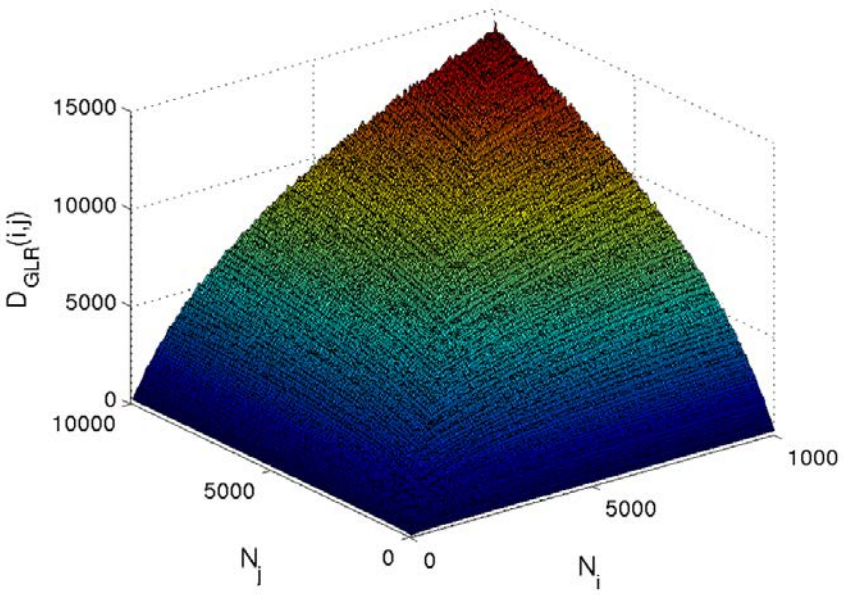

Figura 3.5: A relação entre o número de observações de dois grupos com a distância GLR.

A Figura 3.5 mostra explicitamente o crescimento do valor da distância, à medida em que o número de observações aumenta. Como consequência deste fato, pares de grupos homogêneos que possuem poucas observações (região azul) são considerados mais próximos do que aqueles com grande número de observações. Por outro lado, pares de grupos heterogêneos constituídos de poucas observações possuem um baixo valor da $D_{GLR}(i,j)$, e podem ser considerados mais próximos do que aqueles com grandes quantidades de amostras, assim, podendo ser classificados erroneamente como homogêneos. Este comportamento é indesejado na clusterização, pois, deteriora a capacidade de discriminação dos locutores, e consequentemente, torna o sistema mais suscetível a agrupamentos de pares heterogêneos. Este resultado também foi constatado por Han et al.[53].

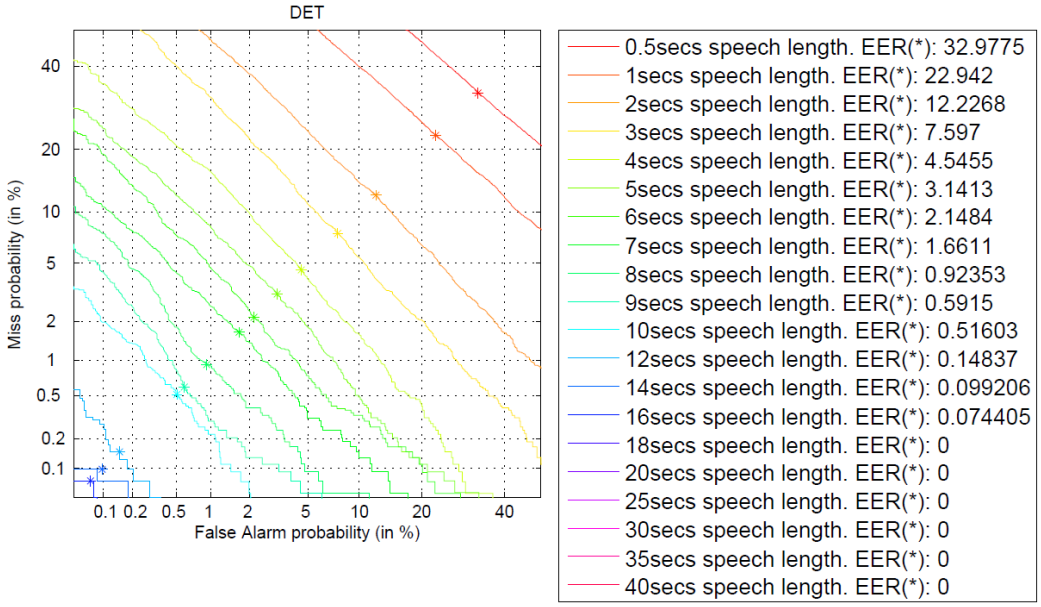

Figura 3.6: Curvas DET da distância GLR.

Comparando os valores de EER da BIC e com os da GLR, apresentados na Figura 3.6, é possível observar que a GLR supera a BIC para segmentos bem curtos. Esta diferença se deve ao número de parâmetros a serem estimados pelas métricas. Enquanto o GLR necessita estimar as médias e uma matriz de covariâncias diagonal, contabilizando $2d$ parâmetros, sendo d a dimensionalidade da gaussiana, o BIC deve estimar uma matriz de covariâncias completa (d^2 parâmetros). Por outro lado, a GLR necessita de mais dados de voz do que a BIC para obter EER zero, assim mostrando que uma matriz de

covariâncias completa possui melhor capacidade de discriminação do que distribuições estatisticamente independentes.

Estas diferenças mostram que as fdps das distâncias de mesmo locutor e locutores diferentes iniciam um pouco mais afastadas que as da Figura 3.4. A, porém se afastam mais lentamente à medida em que a duração das locuções aumentam.

3.2.3
Information Change Rate

A distância ICR (Equação 2-2) surgiu com como tentativa de controlar o crescimento da distância, à medida em que aumenta o número de observações, e o efeito indesejado relatado anteriormente. Este controle é realizado através da normalização da distância pelo número de observações utilizadas no teste de verossimilhança. Os dados que originaram a Figura 3.5 foram devidamente normalizados e são mostrados da Figura 3.7.

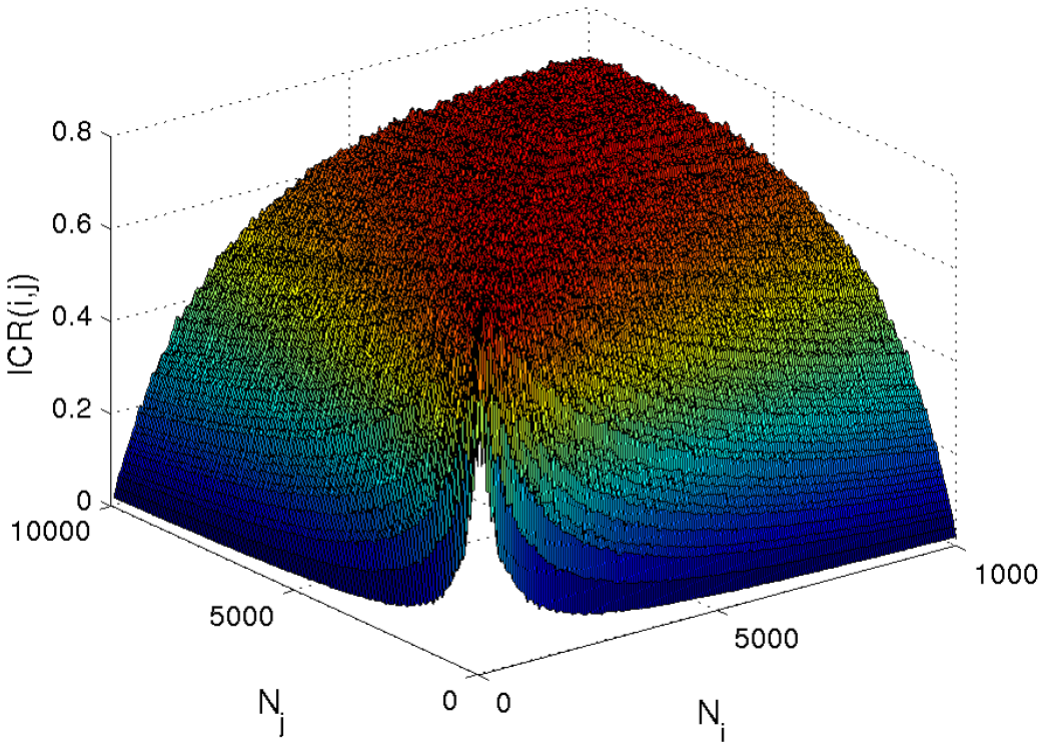

Figura 3.7: Relação entre o número de observações de dois grupos com a distância ICR.

Embora seja notável que o crescimento da distância foi estabilizado, o efeito indesejado ainda permanece, porém em ordem de grandeza menor. Este fato é provado pela Figura 3.8, a qual exibe valores de SET bem similares aos da distância GLR. Com isso pode-se afirmar que a normalização empregada não influencia na capacidade de discriminação.

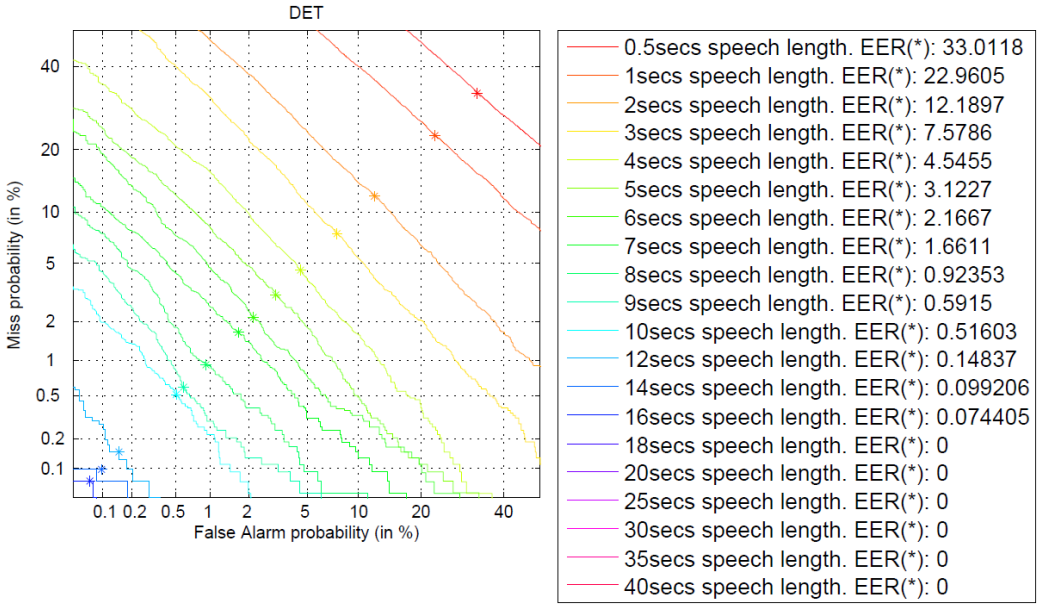

Figura 3.8: Curvas DET da distância ICR.

3.2.4
Cross Likelihood Ratio

O intuito desta análise é verificar se o uso de modelos externos, neste caso um UBM, traz ganhos na discriminação de locutores. Este ganho pode ocorrer devido ao fato do UBM ser um modelo de misturas gaussianas que deve representar completamente o espaço dos locutores; em outras palavras, deve conter a variabilidade entre os indivíduos. Esse espaço bem representado pode enriquecer estatisticamente as locuções em análise, assim podendo aumentar a capacidade de discriminação. No entanto, o número de gaussianas do modelo está relacionado com a quantidade de dados disponíveis para estimar o UBM. Se o número de misturas for muito grande, a quantidade de dados pode ser insuficiente para estimar adequadamente o modelo. Em contrapartida, um número pequeno de

componentes também pode gerar um modelo não adequado, já que cada uma representará uma grande quantidade de dados.

Sendo assim, foram gerados UBMs, com 32 e 128 misturas, a partir de todos os áudios da base TIMIT, somando aproximadamente 3,5 horas de fala. Além disso, consideraram-se matrizes de covariâncias diagonais, e a adaptação MAP foi feitas apenas das médias.

Os gráficos da Figura 3.9 e Figura 3.10 exibem as curvas DET do desempenho da métrica CLR com UBM de 32 e 128 misturas respectivamente.

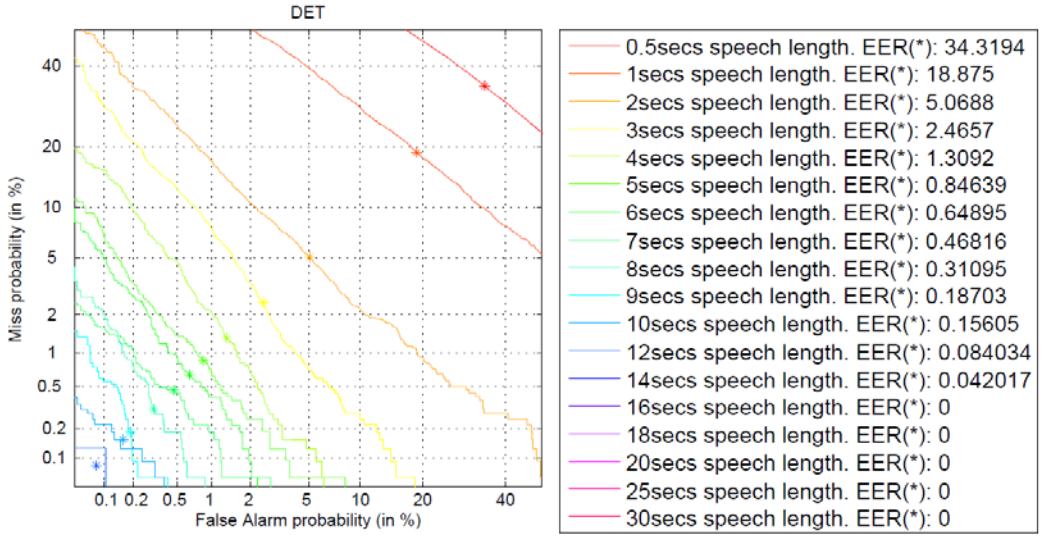

Figura 3.9: Curvas DET da métrica CLR utilizando um UBM de 32 componentes.

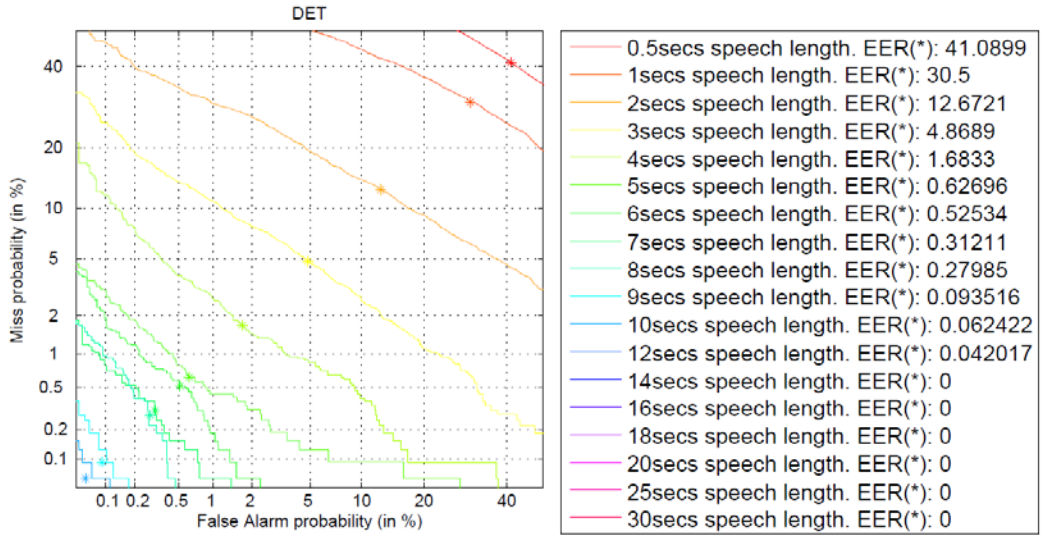

Figura 3.10: Curvas DET da métrica CLR utilizando um UBM de 128 componentes.

Em primeiro lugar, o UBM com 32 componentes superou o com 128, com significância de 0,01, para segmentos menores de cinco segundos. Já para

locuções com duração de pelo menos 5 segundos de duração, o UBM com 128 misturas contribui mais com a capacidade de discriminação, superando o modelo com 32 gaussianas, com significância de 0,025. Além disso, observa-se que a diferença entre os EERs diminui com o aumento da duração das locuções, isto mostra que segmentos mais ricos estaticamente sofrem menos influência de modelos externos. Estas diferenças da capacidade de discriminação com a variação da duração dos segmentos atestam que para locuções curtas, um UBM com menos misturas contribui mais com a capacidade de discriminação dos locutores, porém, para segmentos mais longos necessita-se de um UBM mais bem definido para compensar a redução da influência de modelos externos nos dados em análise.

3.2.5
Kullback-Leibler 2

Verifica-se na, Figura 3.11, que a distância KL2 não se destaca entre as demais métricas. Além disso, quando comparada com a medida BIC, nota-se que as diferenças de EER não são estatisticamente significantes.

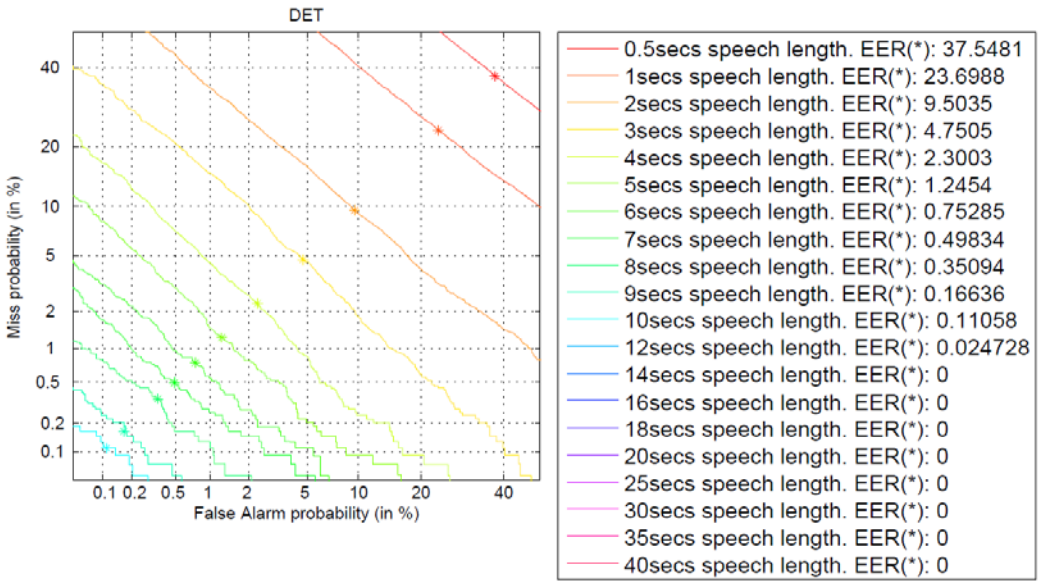

Figura 3.11: Curvas DET distância KL2.

3.2.6
Considerações sobre as Métricas

A Tabela 3.1 exibe todos os valores de EER obtidos no teste de capacidade de discriminação. A duração das locuções afeta o desempenho da capacidade de discriminação, já que quanto maior a duração, melhor será a estimação das estatísticas das métricas. Desta forma, para segmentos bem curtos, deve-se escolher uma métrica com poucos parâmetros a serem estimados, como é o caso da medida GLR. Além disso, à medida em que o tamanho das locuções aumenta, o uso de modelos externos pode contribuir com uma melhor capacidade de discriminação, como é o caso da métrica CLR. No entanto, o número de gaussianas do UBM também deve ser levado em consideração. Para locuções curtas, a Tabela 3.1 mostra que um UBM com 32 misturas produz menores EERs, enquanto que para segmentos maiores, um UBM com mais componentes contribui mais com a capacidade de discriminação.

Tabela 3.1: Relação entre o EER de cada métrica e a duração dos segmentos, em segundos. A métrica CLR é representada por CLR 32 quando esta utiliza o UBM com 32 misturas e CLR 128 quando usa o UBM com 128 componentes

Duração	BIC	GLR	ICR	KL2	CLR 32	CLR128
0.5s	38.04	32.98	33.01	37.55	34.32	41.09
1s	24.51	22.94	22.96	23.7	18.88	30.5
2s	9.8	12.23	12.19	9.5	5.07	12.67
3s	4.6	7.6	7.58	4.75	2.47	4.667
4s	2.3	4.55	4.55	2.3	1.31	1.68
5s	1.17	3.14	3.12	1.25	0.85	0.63
6s	0.7	2.15	2.17	0.75	0.65	0.53
7s	0.42	1.66	1.66	0.5	0.47	0.31
8s	0.3	0.92	0.92	0.35	0.31	0.28
9s	0.13	0.59	0.59	0.17	0.19	0.09
10s	0.11	0.51	0.52	0.11	0.16	0.062
12s	0	0.15	0.15	0.02	0.08	0.042
14s	0	0.1	0.1	0	0.04	0
16s	0	0.07	0.07	0	0	0
18s	0	0	0	0	0	0
30s	0	0	0	0	0	0
40s	0	0	0	0	0	0

3.3
Critério de Parada

O critério de parada tem como meta encerrar automaticamente o processo de agrupamento de modo a atingir a menor taxa de erro possível

Desta forma, os critérios avaliados foram: Delta BIC e Global BIC. Estes por sua vez, foram combinados com algumas das métricas avaliadas anteriormente. Devido à variedade de combinações entre o número de locutores e a duração dos segmentos, optou-se por analisar os critérios para as conversas com quatro locutores, e a duração das locuções de um a doze segundos.

3.3.1
Delta BIC

Ao utilizar este critério, a clusterização deve continuar até que todas as distâncias $\Delta \mathrm{BIC}(i,j)$ sejam positivas, ou seja, enquanto a Equação 3-1a for satisfeita. No entanto, há a possibilidade de se utilizar uma métrica diferente da BIC para procurar os pares mais próximos. Por este fato, as métricas de clusterização avaliadas com este critério foram: BIC, GLR, ICR e KL2. Além disso, analisaram-se as formulações convencional e alternativa do BIC.

Antes de iniciar o agrupamento, configurou-se empiricamente o parâmetro λ_{BIC_c} e λ_{BIC_a} de acordo com a duração inicial das falas. Durante a calibragem dos parâmetros λ_{BIC_c} e λ_{BIC_a}, notou-se que a versão alternativa converge mais rapidamente para os valores que geram os melhores SETs, além de ser mais sensível a pequenas alterações no parâmetro livre. Para cada agrupamento realizado, a Tabela 3.2 e Tabela 3.4 apresentam os valores de SET, enquanto que a Tabela 3.3 e Tabela 3.5 os números de arquivos em que o critério de parada encerrou o agrupamento no ponto correto para cada uma das métricas. Cabe ressaltar que a duração dos segmentos exibidos nas Tabelas 3.2, 3.3, 3.4, e 3.5, indicam apenas a duração das locuções antes de iniciar a clusterização, pois a partir do primeiro par agrupado, haverá cálculo de distâncias entre segmentos de tamanhos diferentes.

Tabela 3.2: Valores de SET obtidos com o critério de parada Delta BIC convencional para diferentes combinações de duração inicial do segmento e métricas

		SET (%)				
Duração	λ	BIC_c	**GLR**	**ICR**	**KL2**	**CLR 32**
1s	1.1	23.9	22.4	41	42.8	30.1
2s	1.3	5.6	7.1	30	29.8	10.4
4s	1.3	0.6	1	7.8	6.9	2.9
8s	1.5	0.1	0.1	0.6	0.2	0.1
12s	1.5	0	0	0	0	0

Embora a medida BIC_c obtenha valores de SET menores do que a distância GLR e a CLR, na Tabela 3.2, a diferença não é estatisticamente significante. Logo, não se pode afirmar qual das medidas trabalha melhor com o critério Delta BIC. No entanto, pode-se dizer que as distâncias BIC e GLR obtém melhor desempenho que a ICR e a KL2 com significância de 0.01, enquanto que a CLR obtêm melhor desempenho que a ICR e KL2 com significância de 0.025.

A Tabela 3.3 mostra que a variação do número de conversas agrupadas corretamente é maior para os arquivos com locuções de 1 segundo. Analisando esta linha da Tabela 3.3, afirma-se que a diferença do número de agrupamentos corretos das métricas GLR e CLR32 não possui significância estatística. Entretanto, pode-se dizer que estas apresentam um ganho de desempenho sobre as métricas BIC_c, ICR e KL2, com significância de 0.025. Cabe ressaltar que o número de acertos das medidas BIC_c, ICR e KL2 não possuem significância estatística. Sobretudo, o desempenho das distâncias GLR e CLR32, para conversas que falas de 1 segundo, é justificado por suas capacidades de discriminação apresentadas na Seção 3.2.

Tabela 3.3: Número de arquivos, por duração inicial das locuções, em que o agrupamento foi realizado corretamente (SET = 0%), utilizando o critério Delta BIC convencional.

	# total	SET (%)				
Duração	**arquivos**	BIC_c	**GLR**	**ICR**	**KL2**	**CLR 32**
1s	200	36	61	40	40	70
2s	200	145	143	60	60	130
4s	200	192	190	172	172	172
8s	200	198	198	197	197	199
12s	150	150	150	150	150	150

Utilizando o critério de parada Delta BIC com a formulação alternativa, nota-se que métrica CLR32 supera os valores de SET das medidas BIC_a, GLR, ICR e KL2, para conversas com locuções bem curtas, com significância de 0.01. Outro fato importante é observado através da comparação dos valores de SET da Tabela 3.2 com a Tabela 3.4, pode-se notar que o uso da formulação alternativa da BIC no critério de parada causou uma deterioração no desempenho das métricas BIC_a e GLR, em conversas que possuem falas curtas. Para o BIC_a esta deterioração no desempenho foi de 72% relativa à BIC_c, enquanto que para a GLR, deterioração no desempenho relativo foi de 96%.

Tabela 3.4: Valores de SET, obtidos com o critério de parada Delta BIC alternativa para diferentes combinações de duração inicial do segmento e métricas.

		SET (%)				
Duração	λ	BIC_c	**GLR**	**ICR**	**KL2**	**CLR 32**
1s	1.2	41.1	44	40.1	42	26.2
2s	0.95	26.7	24.8	25.8	29.4	11.6
4s	0.7	6.2	6.6	8	7.2	3.4
8s	0.6	0.1	0.1	1.2	0.8	0.2
12s	0.4	0	0.5	0	0	0

Tabela 3.5: Número de arquivos, por duração inicial das locuções, em que o agrupamento foi realizado corretamente (SET = 0%), utilizando o critério Delta BIC melhorado.

	# total	SET (%)				
Duração	**arquivos**	BIC_c	**GLR**	**ICR**	**KL2**	**CLR 32**
1s	200	7	3	50	50	84
2s	200	45	22	86	86	130
4s	200	152	152	165	165	172
8s	200	197	197	193	188	199
12s	150	150	150	150	150	150

A deterioração no desempenho das métricas BIC e GLR ao adotar a formulação alternativa do BIC no critério de parada também pode ser visualizada na Tabela 3.5. Para a distância BIC_a a redução do número de agrupamentos corretos é de 80.5% relativa à BIC_c, enquanto que para a GLR a redução relativa é de 95%.

De modo geral, o desempenho do agrupamento melhora com aumento da duração das falas no início da clusterização. Este comportamento é explicado pelos experimentos realizados na Seção 3.2, nos quais foram mostrados que pares de curta duração são mais difíceis de discriminar, ou seja, as probabilidades de falsa rejeição (não agrupar locuções de mesmo locutor), e falso alarme (agrupar falas de locutores diferentes) são maiores. Nestas circunstâncias o sistema tende a ter alto índice de erro (SET) e, consequentemente, baixo número de conversas nas quais os segmentos foram agrupados corretamente.

3.3.2
Global BIC

Diferentemente do Delta BIC, o global BIC considera todos os estágios do agrupamento, ou seja, a clusterização para quando houver somente um grupo. A cada etapa calcula-se o Global BIC através da Equação 2-19, e no final deste processo, escolhe-se o estágio que possuir o maior $BIC(\mathcal{M})$ como resultado final. Cabe ressaltar que o modelo $\mathcal{M}_i$ foi representado por uma simples gaussiana com matriz de covariâncias diagonal.

Este critério foi combinado com as métricas: BIC convencional, BIC alternativa, GLR, ICR, KL2 e CLR. No entanto, devido ao fato do Global BIC possuir um parâmetro livre, este foi calibrado empiricamente, para cada uma das métricas, de forma que o melhor SET fosse obtido. As configurações do $\lambda_{\text{Global BIC}}$ estão na Tabela 3.6, junto com os valores de SET obtidos. Além disso, a Tabela 3.7 exibe o número de arquivos em que o critério acertou o número de locutores presentes.

Tabela 3.6: Valores de SET, obtidos com o critério de parada Global BIC para diferentes combinações de duração inicial do segmento e métricas.

		SET (%)					
Duração	λ	BIC_c	BIC_α	**GLR**	**ICR**	**KL2**	**CLR32**
1s	1.4	19.4	27.2	18.5	30.9	32.5	46.1
2s	2	5.4	12.8	6.4	17.3	19.1	24.8
4s	2	1.2	3.4	1.5	5.1	4	1.6
8s	2.2	0.2	0.2	0.2	0.4	0.2	0
12s	2.4	0	0	0	0	0	0

Em conversas com locuções curtas, a Tabela 3.6 mostra que as métricas BIC_c e GLR obtêm desempenhos melhores que as métricas BIC_α, KL2, ICR e CLR, com significância de 0.1. Embora a medida GLR apresente um aumento de desempenho relativo à BIC_c de 4.6%, nada se pode afirmar qual delas trabalha melhor com o critério Global BIC, pois a diferença no valor do SET não possui significância estatística. Sobre a distância BIC, a formulação alternativa obtém uma redução no SET de 40.2% relativo à convencional, com significância de 0.1.

Comparando os critérios Delta BIC (tabelas 3.2 e 3.4) e Global BIC (Tabela 3.6), nota-se que a utilização do critério Global BIC proporcionou uma redução no desempenho da métrica CLR, em conversações com falas bem curtas. Esta redução relativa ao uso do critério Delta BIC é 53.2%, com significância de 0.01. Entretanto, o Global BIC aumentou o desempenho das medidas: BIC_α, ICR e KL2. Para a distância BIC_α, a redução do SET relativo ao critério Delta BIC é de 33.8% com significância de 0.01. Enquanto que para a ICR o decrescimento do SET relativo ao critério Delta BIC é de 22.9% com significância de 0.1. A distância KL2, apresenta uma redução do SET relativo ao critério Delta

BIC de 22.6% com significância de 0.05. Já para as métricas BIC_c e GLR, a diferença de SET não possui significância estatística.

Tabela 3.7: Número de arquivos, por duração inicial das locuções, em que o agrupamento foi realizado corretamente (SET = 0%), utilizando o critério Global BIC.

	# total	# acertos					
Duração	**arquivos**	BIC_c	BIC_a	**GLR**	**ICR**	**KL2**	**CLR32**
1s	200	82	34	98	23	12	14
2s	200	150	105	153	57	44	54
4s	200	183	164	182	143	144	182
8s	200	194	194	194	192	194	200
12s	150	150	150	150	150	150	150

Em contrapartida à análise da Tabela 3.6, a Tabela 3.7 mostra que o critério Global BIC proporcionou aumento de desempenho no agrupamento de conversações com locuções curtas para as medidas BIC_c e GLR. Esta melhora consiste no aumento de agrupamentos sem erros, que é observada comparando-se a quantidade de acertos das tabelas 3.3 e 3.5 com a Tabela 3.7. A medida BIC_c combinada com o Global BIC apresenta um aumento de 127.8% relativo à BIC_c combinada com o Delta BIC, com significância de 0.01. Além disso, a medida GLR combinada com o Global BIC apresenta um aumento de 385.7% relativo à GLR combinada com o Delta BIC, com significância de 0.01. No entanto, as distâncias ICR e KL2 combinadas com o Global BIC apresentam uma redução do número de agrupamentos corretos. Enquanto a ICR combinada com o Global BIC, apresenta uma redução de 42.5% relativa à ICR combinada com o Delta BIC, com significância de 0.01. A KL2 combinada com o Global BIC convencional, apresenta uma redução de 70% relativa à KL2 combinada com o Delta BIC convencional, com significância de 0.01.

4
Avaliação do Agrupamento de Locutor na Diarização de Locutor

Este capítulo tem como objetivo refazer os experimentos realizados na Seção 3.2 e Seção 3.3, porém em uma base que expressa condições reais. Este capítulo está dividido em três seções. Na Seção 4.1 descrevemos a base de dados. Na Seção 4.2, experimentos sobre a capacidade de discriminação das métricas de agrupamentos são analisadas. Finalmente, na Seção 4.3, experimentos sobre critérios de parada do agrupamento são avaliados. Cabe ressaltar que o teste binomial para diferenças de proporções é usado para verificar se a diferença entre os valores de SET, EER e o número de conversações agrupadas corretamente, são estatisticamente significantes (Gillick and Cox, [115]). Quando não especificado, o nível de significância é de $\alpha = 0.05$.

4.1
Base de Dados

A base do NIST-SRE 2002 [112] contém áudios em três cenários: encontros, noticiários e ligações telefônicas. Neste trabalho foram utilizados apenas os áudios de noticiários, por possuírem dois ou mais locutores e por serem gravados por apenas um microfone. No conjunto de noticiários há 75 áudios somando aproximadamente 148 minutos. Todos eles são quantizados a uma taxa de 32 bit e amostrados à 16KHz. Ao contrário da base controlada, nesta há músicas de fundo e outros tipos de ruído. Além disso, há locuções de tamanhos variados. A Tabela 4.1 mostra a quantidade de gravações por número de locutores, e a Figura 4.1 exibe as fdp e função de distribuição acumulada (FDA) das durações das locuções desta base.

Tabela 4.1: Quantidade de gravações por número de locutores participantes.

# locutores	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
# gravações	16	18	18	8	10	3	1	1

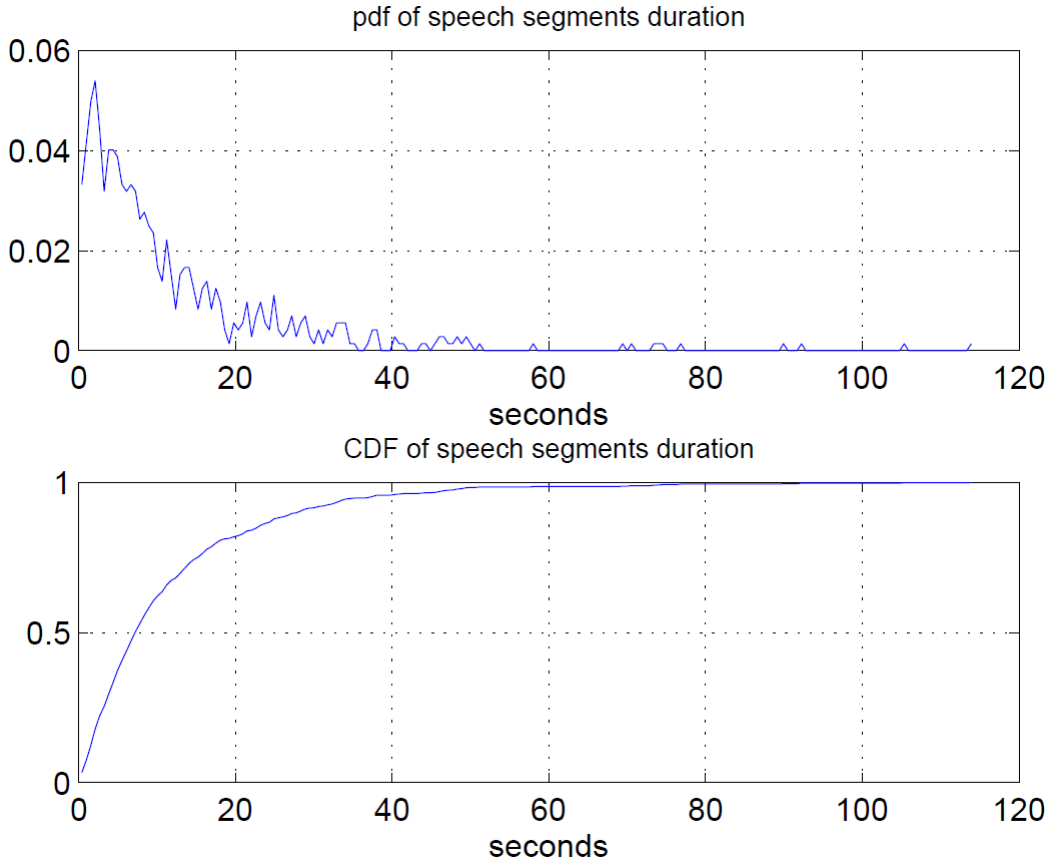

Figura 4.1: Funções de densidade de probabilidade (fdp) e função de distribuição de probabilidade (FDP) da duração das falas da base NIST-SRE 2002 - *broadcast news*.

Analisando os dois gráficos da Figura 4.1, pode-se observar que a maioria dos segmentos possui duração de até 20 segundos (82% dos segmentos), e que há poucas locuções com duração maior que 40 segundos (4% possuem mais de 40 segundos de duração). Baseado nessas informações foi definido 40 segundos como tempo máximo de duração das locuções da base controlada. Por último, as características extraídas nesta base foram as mesmas usadas na TIMIT, os primeiros 19 coeficientes MFCC's, com janelas de 25 ms e deslocamento de 10 ms.

4.2
Capacidade de Discriminação das Métricas

A base do NIST contém gravações reais, logo, as conversações possuem falas de diversos tamanhos, como mostra a Figura 4.2. Por este fato, não há a possibilidade de avaliar a capacidade de discriminação das métricas pela duração das locuções. O gráfico *box plot* da Figura 4.2 exibe a variabilidade da duração das locuções relacionada com o número de locutores participantes nas conversas. Observa-se mais de 50% das locuções, de conversas com dois e três locutores, possuem mais de 10 segundos de duração. Para conversas com quatro, cinco, seis e sete participantes, mais de 50% das falas duram menos que 10 segundos. Nas conversas com mais de seis locutores, mais de 75% dos segmentos são menores que 8 segundos, e mais de 50% duram menos que 5 segundos. Além disso, todas as configurações de número de locutores possuem falas mais curtas que 1 segundo.

As curvas DET e os respectivos SETs das métricas: BIC, GLR, ICR e KL2, serão apresentados sem nenhuma relação com a duração das locuções. Além disso, optou-se também por avaliar a métrica CLR em uma base real, já que esta medida obteve bom desempenho na base controlada. A ideia desta avaliação é verificar a influência de modelos externos, com diferentes configurações, em segmentos de diferentes tamanhos. Desta forma, serão exibidas as curvas DET e o desempenho SET para as diferentes configurações de UBM.

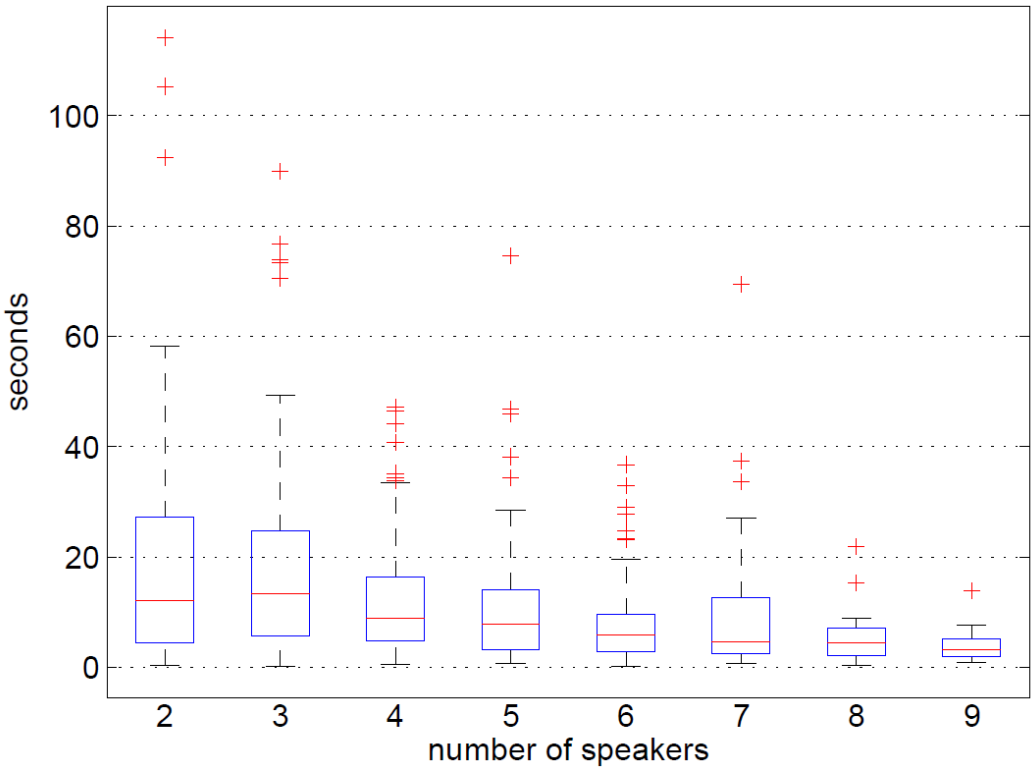

Figura 4.2: Variabilidade do tamanho das locuções de acordo com o número de locutores participantes nas conversas.

4.2.1

Métricas não Baseadas em Modelo

A Figura 4.3 exibe as curvas DET e os valores de EER das medidas: BIC_c, BIC_α, GLR, ICR e KL2. Comparando os valores de EER obtidos nas bases TIMIT e NIST, percebe-se que as métricas BIC_c e BIC_α desempenham de forma diferente. A formulação alternativa da distância BIC (BIC_α) promoveu uma redução relativa de 24.4% sobre a medida BIC com a formulação convencional (BIC_c), com significância de 0.01. Este comportamento é explicado pelo fato das locuções serem de tamanhos diferentes. Nesta circunstância, a penalidade do BIC convencional e alternativa apresentam valores diferentes para cada par de segmentos, assim interferindo na capacidade de discriminação.

A distância GLR apresenta um aumento de desempenho relativo de 7% sobre a ICR, com significância de 0.1. No entanto, quando cada uma delas é comparada à BIC, observa-se, que na base NIST, a diferença entre os EERs aumenta. Este fato é explicado pelo efeito indesejado apresentado nas Seções 3.2.2 e 3.2.3 através da Figura 3.5 e Figura 3.7, onde pares de diferentes tamanhos

são considerados mais próximos do que aqueles com grandes durações, assim degradando a capacidade de discriminação.

Por último, nota-se que os valores de EER da distância KL2 na base TIMIT diferem em aproximadamente 1% dos valores da BIC, porém na NIST esta diferença aumenta em aproximadamente 50% em relação à BIC_c, e 99% em relação à BIC_a. Assim mostrando que a capacidade de discriminação da KL2 torna-se menos robusta quando os pares de segmentos avaliados possuem tamanhos diferentes.

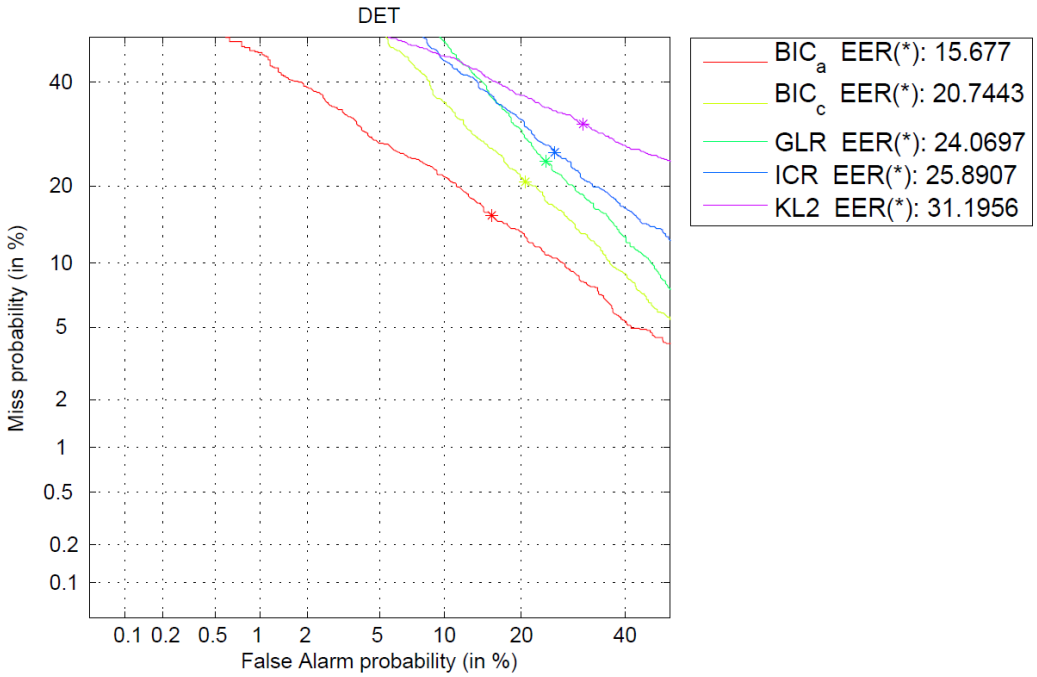

Figura 4.3: Curvas DET e valores de EER das métricas: BIC convencional (BIC_c), BIC alternativa (BIC_a), GLR, ICR e KL2.

4.2.2
Cross Likelihood Ratio

Pelo fato da base NIST ser menor, a criação do UBM foi feita de forma diferente da base TIMIT. Seja $C = a_1, \ldots, a_n$ o conjunto de todos os arquivos a_i da base NIST. Tem-se que o UBM_i do áudio a_i foi gerado a partir do conjunto $C - a_i$. Nesta métrica avaliou-se de UBMs com 16, 32, 64 e 128 gaussianas. Além disso, a adaptação MAP foi feita apenas das médias. A Figura 4.4 exibe as curvas DET e os valores de EER relacionados com a configuração do UBM utilizado.

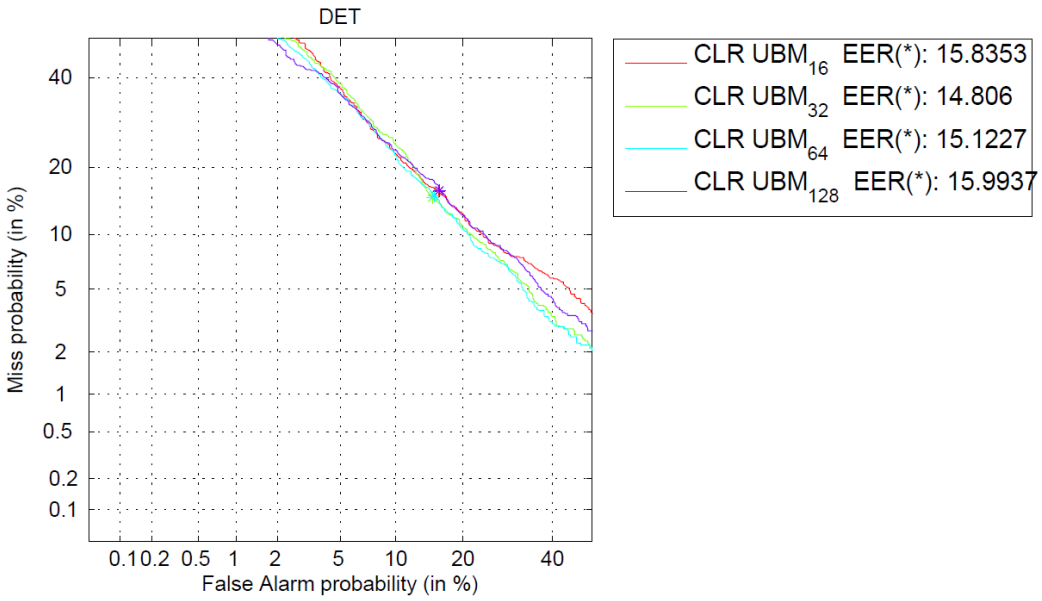

Figura 4.4: Curvas DET e valores de EER das métricas CLR utilizando modelos de 16, 32, 64 e 128 gaussianas.

Observando os valores de EER da Figura 4.4, nota-se que as diferenças entre eles não possuem significância estatística. Além disso, comparando os EERs da Figura 4.4 com os obtidos pelas métricas não baseadas em modelos (Figura 4.3), verifica-se que a medida CLR32 obtém um aumento de desempenho relativo de 5.5% a BIC_{α}, porém esta diferença não é estatisticamente significante. Entretanto, observa-se que a medida CLR com UBM de 32 misturas obtém um aumento de desempenho relativo de 28.6% sobre a BIC_{c}, com significância de 0.01. Além disso, ao comparar os valores de EER obtidos pela CLR, com os obtidos pelas métricas GLR, ICR e KL2 (Tabela 4.3), verifica-se que o uso de dados externos promove uma redução do EER, com significância de 0.01. Desta forma, nota-se que nas bases TIMIT e NIST, a CLR obtém melhor capacidade de discriminação dos que as medidas: GLR, ICR e KL2.

4.3
Critério de Parada

Como explicado na Seção 3.3, tem como objetivo encontrar o ponto do agrupamento com menor taxa de erro (SET). As combinações entre critério de parada e métricas avaliadas são as mesmas da Seção 3.3, porém o estado inicial da clusterização é diferente. Enquanto que todos os pares de locuções, da base

TIMIT possuem a mesma duração no inicio da clusterização nestes experimentos isto não ocorre. No começo do agrupamento há falas de tamanhos variados.

4.3.1
Delta BIC

Embora este critério já tenha sido avaliado na Seção 3.3.1, necessita-se recalibrar o parâmetro livre para as duas formulações do BIC, pois este depende da base de dados. Desta forma, os valores de calibragem foram: $\lambda_{BIC_c} = 2.3$ e $\lambda_{BIC_a} = 0.75$. Feito isso, a Tabela 4.2 e Tabela 4.3 apresentam os valores de SET do critério Delta BIC convencional e alternativa, respectivamente. Além disso, o número de conversas agrupadas corretamente com critério Delta BIC convencional são exibidos na Tabela 4.3, assim como, os agrupamentos corretos com o Delta BIC alternativa na Tabela 4.5.

Tabela 4.2: Valores de SET, obtidos com o critério de parada Delta BIC convencional para diferentes combinações de número de locutores por conversa e métricas.

	SET (%)				
# Locutores	**BIC_c**	**GLR**	**ICR**	**KL2**	**CLR 32**
2	0.2	0.3	10.6	10.7	10.9
3	4.5	4.5	28	31.4	1.5
4	5	5.5	38.4	38.6	16
5	14.9	12.8	44.8	43.4	20.6
6	13.5	15.2	47.2	50.6	32.5
7	10.8	11.7	48.9	52.9	15.3
8	15.4	12	45.3	67.2	8.1
9	31.2	22.9	34.5	34.5	19.9
Total	**6.7**	**6.7**	**32.5**	**34.2**	**18.8**

Tabela 4.3: Número de arquivos, por locutor e total, em que o agrupamento foi realizado corretamente (SET = 0%), utilizando o critério Delta BIC convencional.

# Locutores	# total arquivos	# acertos				
		BIC_c	GLR	ICR	KL2	CLR 32
2	16	15	15	7	7	7
3	18	13	14	2	2	1
4	18	11	12	0	0	3
5	8	3	4	0	0	0
6	10	3	2	0	0	0
7	3	0	0	0	0	0
8	1	0	0	0	0	1
9	1	0	0	0	0	0
Total	**75**	**45**	**47**	**9**	**9**	**12**

A Tabela 4.2 e Tabela 4.3 mostram que o critério Delta BIC com a formulação convencional obtém melhor desempenho para arquivos com poucos locutores. Enquanto que na Tabela 4.2 nota-se o aumento dos valores de SET com o aumento dos locutores participantes, na Tabela 4.3 verifica-se a redução do número de conversações agrupadas corretamente à medida em que cresce o número de participantes. %Isto é explicado pelo fato destas conversas possuírem locuções mais longas entre homens e mulheres. Já que a discriminação entre homem e mulher é mais fácil do que entre homens ou entre mulheres. Além disso, destaca-se, na Tabela 4.2, o baixo valor de SET das métricas BIC_c e GLR em relação à ICR, KL2 e CLR, com significância de 0.05.

Tabela 4.4: Valores de SET, obtidos com o critério de parada Delta BIC alternativo para diferentes combinações de número de locutores por conversa e métricas.

	SET (%)				
# Locutores	BIC_α	**GLR**	**ICR**	**KL2**	**CLR 32**
2	0.2	0.2	12.1	9.1	7.4
3	2.4	3.5	26.9	30.9	21.4
4	4.6	4.8	30.8	32.1	20.8
5	13.3	13.2	38.4	34.7	22.3
6	10.3	14	38.2	47.4	30.7
7	7.4	6	12.3	42.6	10
8	5.4	4.5	45.3	67.2	48.7
9	17.6	22.9	34.5	34.5	19.9
Total	**5.1**	**5.9**	**27.2**	**30.3**	**19.5**

A Tabela 4.4 mostra que o critério de parada Delta BIC com a formulação alternativa proporcionou uma redução nos valores totais de SET das métricas: BIC, GLR, ICR e KL2. Entretanto a diferença entre dos valores totais de SET por distância das Tabelas 4.2 e 4.4 não possuem significância estatística.

Tabela 4.5: Número de arquivos, por locutor e total, em que o agrupamento foi realizado corretamente (SET = 0%), utilizando o critério Delta BIC alternativo.

	# total	**# acertos**				
# Locutores	**arquivos**	BIC_c	**GLR**	**ICR**	**KL2**	**CLR 32**
2	16	11	14	8	8	6
3	18	12	13	3	3	2
4	18	6	9	1	1	2
5	8	1	2	2	2	1
6	10	2	2	1	1	0
7	3	0	0	0	0	0
8	1	0	0	0	0	0
9	1	0	0	0	0	0
Total	**75**	**32**	**40**	**15**	**15**	**11**

Embora as variações do SET não apresentem significância estatística, a Tabela 4.5 mostra que o uso da formulação alternativa no critério Delta BIC, proporcionou um aumento do número de conversações agrupadas corretamente. Este aumento do número de acerto da BIC_a relativo à BIC_c de 40.6%, com significância de 0.05.

Comparando o comportamento dos valores SET com critério Delta BIC nas bases TIMIT e NIST, nada se pode afirmar a respeito das diferenças, já que a mudança da formulação no critério de parada não promoveu mudanças estatisticamente significantes na base NIST. Por outro lado, comparando o número de conversações agrupadas corretamente nas duas bases, observa-se uma diferença de desempenho. Enquanto que na base TIMIT a utilização da formula alternativa no critério promoveu uma redução do número de agrupamentos corretos das métricas BIC, GLR, ICR e KL2, na base NIST ocorre um aumento do número de acertos.

4.3.2
Global BIC

Este critério foi combinado com as métricas: BIC com a penalidade alternativa, GLR, ICR e KL2. No entanto, devido ao fato do Global BIC possuir um parâmetro livre, este foi calibrado com uma base de testes, para cada uma das métricas, de forma que o melhor SET fosse obtido. As configurações do $\lambda_{\text{Global } BIC}$ estão na Tabela 4.6.

Tabela 4.6: Parâmetro $\lambda_{\text{Global } BIC}$ usado para cada métrica.

Métrica	$\lambda_{\text{Global } BIC}$
BIC_c ($\lambda_{BIC} = 2.3$)	6
BIC_a ($\lambda_{BIC} = 0.75$)	4.4
GLR	4.2
ICR	4.9
KL2	3.98
CLR 32	4.6

Com base no parâmetro estimado, o agrupamento foi realizado para cada métrica. A Tabela 4.7 apresenta os valores de SET, total e por locutor, de cada métrica, e a Tabela 4.8 exibe o número de conversações em que o agrupamento foi realizado corretamente.

Tabela 4.7: Valores de SET, obtidos com o critério de parada Global BIC para diferentes combinações de número de locutores por conversa e métricas.

	SET (%)					
# Locutores	BIC_c	BIC_a	**GLR**	**ICR**	**KL2**	**CLR 32**
2	4.2	0	4.7	5.2	4.2	0.4
3	3.3	2.7	5	5.3	4.6	2.6
4	6.4	3.9	7.9	7.3	7.1	6.2
5	19.1	17.7	14.1	20.8	14.5	14.3
6	15.3	10	18.8	18.7	15.8	13.1
7	14.3	10.1	10.4	10.4	6.9	11.6
8	32.3	21.1	7.9	31.3	10.5	8.1
9	39.5	22.9	28.3	28	34.5	18
Total	**8.7**	**5.8**	**8.9**	**9.9**	**8.2**	**7**

As diferenças entre os valores totais de SET apresentados na Tabela 4.7 não possuem significância estatística. Da mesma maneira, comparando-se os valores totais de SET obtidos pelo critério Global BIC e Delta BIC, para as métricas BIC_c, BIC_a e GLR, pode-se verificar que suas respectivas diferenças também não possuem significância estatística. No entanto, nota-se que o critério Global BIC promoveu um aumento de desempenho das medidas ICR, KL2 e CLR32, em relação ao Delta BIC.

Comparando o desempenho da métrica ICR com os critérios de parada Global BIC e Delta BIC, nota-se que o critério Global BIC apresenta um aumento relativo no desempenho de 63.3% com significância de 0.01. Já a medida KL2, com o Global BIC, apresenta um aumento de desempenho relativo de 72.9% em relação à KL2 com o Delta BIC, com significância de 0.01. Também, a métrica CLR32, com o Global BIC, obtém um aumento de desempenho relativo de 64.1% sobre a CLR32 com o critério Delta BIC, com significância de 0.025.

Tabela 4.8: Número de arquivos, por locutor e total, em que o agrupamento foi realizado corretamente (SET = 0%), utilizando o critério Global BIC.

# Locutores	# total arquivos	# acertos					
		BIC_c	BIC_a	GLR	ICR	KL2	CLR32
2	16	13	16	14	12	11	13
3	18	13	14	11	11	8	11
4	18	7	11	6	5	7	9
5	8	2	1	5	4	1	2
6	10	2	5	3	2	3	4
7	3	0	0	0	1	0	0
8	1	0	0	0	0	0	1
9	1	0	0	0	0	0	0
Total	**75**	**37**	**47**	**39**	**35**	**30**	**40**

Embora a variação do SET da métrica BIC_a entre os dois critérios de parada não seja estatisticamente significante, a diferença entre o número de conversações agrupadas corretamente possui significância estatística de 0.025. Observando a Tabela 4.5 e a Tabela 4.8, nota-se que a BIC_a com o Global BIC obtêm um aumento relativo, do número total de acertos, de 46.9% sobre a BIC_a com o Delta BIC.

Analisando os dois critérios avaliados na base NIST, nenhum deles conseguiu acertar o número correto de locutores nos arquivos com oito e nove participantes (arquivos paby e paaz). Ao analisar os dois áudios, foi constatado que mais de 60% dos segmentos possuem duração menor do que 5 segundos, o que, de acordo com os experimentos da Seção 3.2, dificulta a capacidade de discriminação. Além disso, o fato de haver muitos locutores implica em possuir falar de vários locutores do mesmo sexo, o que também dificulta a discriminação.

Como tentativa de melhorar o desempenho do Global BIC, calculou-se a verossimilhança da Equação 2-19 de duas formas diferentes. Na primeira, as verossimilhanças foram calculadas sobre modelos gerados por adaptações MAP, enquanto que na segunda, foi utilizado a BIC (Equação 2-3). No entanto, nenhum dos experimentos gerou resultados significantes. Em todos eles, ou o sistema agrupou todos os segmentos em um único grupo, ou parou a clusterização muito antes do ponto ideal.

5
Conclusões

Este trabalho teve como objetivo analisar a capacidade de discriminação de algumas métricas usadas no estado da arte de diarização de locutor, assim como, avaliar o desempenho do agrupamento *bottom-up*. As métricas selecionadas foram: BIC, com duas formulações de penalidades, GLR, ICR, KL2 e CLR. Da mesma maneira, selecionaram-se os critérios de parada Delta BIC e Global BIC, utilizados no estado da arte, para avaliar o desempenho da fase de agrupamento de locutores. A capacidade de discriminação e o agrupamento foram avaliados através de experimentos em duas bases de dados: uma controlada, originada do corpus TIMIT, e a utilizada na competição NIST-SRE 2002.

Ao analisar as distâncias BIC e GLR, pode-se concluir que em conversas que possuam pares com grandes durações e pares de segmentos curtos, há maior probabilidade de falso alarme e falsa rejeição. Este fato ocorre devido à distância BIC diminuir à medida que as locuções crescem. De maneira semelhante, a medida GLR cresce com o aumento do número de observações dentro dos segmentos em análise. Além disso, ao realizar os experimentos de capacidade de discriminação na base TIMIT, concluiu-se que a mudança na penalidade da distância BIC não influencia na discriminação de locutores. Quando as locuções possuem a mesma duração, a penalidade torna-se uma constante somada a todas às distâncias.

Quando se avaliou a métrica CLR, baseada em modelo verificou-se que o uso de dados externos influencia na discriminação dos locutores. Entretanto, o número de misturas está relacionado com a duração das locuções em análise. Verificou-se que para segmentos curtos, um UBM com 32 misturas contribui mais na discriminação. Já em pares de segmentos longos, um UBM com 128 gaussianas cooperou mais em termos de capacidade de discriminação.

Ao comparar os EERs obtidos pelas métricas selecionadas, pode-se concluir que a capacidade de discriminação é diretamente proporcional à duração das locuções em análise. A capacidade de discriminação melhora à medida que a duração das locuções em análise aumenta, pois com locuções longas a estimação

do modelo do locutor se torna mais adequada. Além disso, conclui-se que média e matriz de covariâncias diagonal são melhores para discriminar os locutores, em fala bem curtas, do que somente uma matriz de covariâncias completa. Já que uma matriz de covariâncias diagonal possuem menos parâmetros para serem estimados do que uma matriz de covariâncias completa.

A partir dos experimentos com o critério de parada Delta BIC na base TIMIT, conclui-se que a mudança da penalidade pode influenciar no erro SET. A Seção 3.3.1 mostrou que a penalidade alternativa promoveu uma redução no SET da métrica CLR. Esta foi de 13% relativa à medida CLR com o critério Delta BIC convencional. Por outro lado, a formulação alternativa como critério de parada promoveu um aumento relativo na BIC_a de 72% sobre a BIC_c. Da mesma forma, o aumento relativo do SET para medida GLR com o Delta BIC alternativo é de 96% sobre a GLR com o Delta BIC convencional. Ao substituir o critério Delta BIC pelo Global BIC, a distância CLR apresentou um aumento de 53% do SET relativo à CLR com o critério Delta BIC convencional, com significância de 0.01. Já a distância KL2 com o Global BIC, apresenta uma redução do SET relativo ao critério Delta BIC de 22.6% com significância de 0.05.

Foi possível concluir que a mudança da penalidade interfere na discriminação de locutores, quando o par de segmentos em análise possui durações diferentes. A formulação alternativa da distância BIC (BIC_a) promoveu uma redução relativa de 24.4% sobre a medida BIC com a formulação convencional (BIC_c), com significância de 0.01. Esta redução no EER é justificada pelo fato das locuções serem de tamanhos diferentes. Além disso, conclui-se que o uso de modelos externos também contribui para a discriminação de locuções de durações diferentes. Na Seção 4.2.2, foi possível observar que o uso de um UBM com 32 misturas promoveu um aumento de desempenho relativo de 28.6% sobre a BIC_a, com significância de 0.01. Também, pode-se concluir que utilizar o critério Delta BIC junto com uma métrica diferente para procurar os pares mais próximos pode aumentar a probabilidade de erro no agrupamento. Nesta circunstância haverá probabilidades de falsas rejeições e falsos alarmes diferentes para a busca do par mais próximo e para o critério de parada.

Finalmente, na Seção 4.3, conclui-se que o aumento do número de locutores nas conversas pode dificultar o agrupamento. Observa-se nas tabelas 4.3, 4.5 e 4.8, que com exceção da métrica CLR combinada com os critérios Delta BIC

convencional e BIC Global, nenhuma das outras combinações acertou o número de locutores participantes. Isto se deve ao amento de locutores do mesmo sexo, o que dificulta a discriminação. Sobretudo, pode-se concluir que o conhecimento prévio sobre a base de dados pode ajudar na escolha da métrica e o critério de parada adotado no agrupamento de locutores. Visto que, neste trabalho mostrou-se a variação do EER, SET e número de locutores agrupados corretamente, com a mudança da métrica e critério de parada utilizados.

5.1
Trabalhos Futuros

Com base nas conclusões deste trabalho, sugere-se como trabalho futuro, desenvolver um algoritmo de agrupamento que faça o uso de mais de uma métrica para procurar os pares mais próximos. Ao selecionar as medidas, deve-se considerar a duração das falas em análise, assim como, a quantidade de parâmetros a serem estimados. Para segmentos curtos, deve-se usar uma distância que necessite estimar poucos parâmetros, como por exemplo, uma matriz de covariâncias diagonal. No entanto, à medida que os segmentos são unidos, pode-se utilizar outra métrica que necessite estimar mais parâmetros, sendo eles, médias e uma matriz de covariâncias completa.

Finalmente, pode-se sugerir utilizar o critério Global BIC com diferentes formas de se calcular as verossimilhanças. Consideram-se métricas que possuam poucos parâmetros a serem estimados, para calcular as verossimilhanças de segmentos curtos com seus respectivos modelos. Por outro lado, para falas mais longas adota-se medidas que estimem um número maior de parâmetros.

Referências bibliográficas

[1] Yamaguchi, M., Yamashita, M. and Matsunaga, S.: 2005, Spectral cross-correlation features for audio indexing of broadcast news and meetings, Proc. International Conference on Speech and Language Processing.

[2] Pelecanos, J. and Sridharan, S.: 2001, Feature warping for robust speaker verification, ISCA Speaker Recognition Workshop odyssey, Crete, Grece.

[3] Ouellet, P., Boulianne, G. and Kenny, P.: 2005, Flavors of gaussian warping, Proc. International Conference on Speech and Language Processing, Lisbon, Portugal.

[4] Sinha, R., Tranter, S.E., Gales, J. J. F. and Woodland, P. C.: 2005, The cambridge university march 2005 speaker diarisation system, European Conference on Speech Communication and Technology (Interspeech), Lisbon, Portugal, pp. 2437-2440.

[5] Zhu, X., Barras, C., Lamel, L. and Gauvain, J.-L.: 2006, Speaker diarization: from broadcast news to lectures, NIST 2006 Spring Rich Transcrition Evaluation Workshop, Washington DC, USA.

[6] Pardo, J. M., Anguera, X. and Wooters, C.: 2006a, Speaker diarization for multi-microphone meetings using only between-channel differences, MLMI 2006.

[7] Pardo, J. M., Anguera, X. and Wooters, C.: 2006b, Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences, Proc. International Conference on Speech and Language Processing.

[8] ICSI Meeting Recorder Project: Channel skew in ICSI-recorded meetings: 2006. URL: http://www.icsi.berkeley.edu/dpwe/research/mtgrcdr/chanskew.html

[9] Lathoud, G. and McCowan, I. A.: 2003, Location based speaker segmentation, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing.

[10] Ferras, M.; Bourlard, H., "Speaker diarization and linking of large corpora," Spoken Language Technology Workshop (SLT), 2012 IEEE , vol., no., pp.280,285, 2-5 Dec. 2012 doi: 10.1109/SLT.2012.6424236.

[11] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices", in Proc. ICASSP '08, Las Vegas, NV, USA, 2008, pp. 4133-4136.

[12] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis", IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 6, pp. 1059-1070, 2010.

[13] X. Zhu, C. Barras, L. Lamel, and J.-L. Gauvain, "Multi-stage speaker diarization for conference and lecture meetings", in Proc. Multimodal Technol. Perception of Humans: Int. Eval. Workshops CLEAR 2007 and RT 2007, Baltimore, MD, May 8-11, 2007, Revised Selected Papers, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 533-542.

[14] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system", in Proc. NIST MLMI Meeting Recognition Workshop, Edinburgh, U.K., 2005.

[15] J. Ajmera, G. Lathoud, and L. McCowan, "Clustering and segmenting speakers and their locations in meetings", in Proc. ICASSP '04, vol. 1, Montreal, Canada, 2004, pp. I-605-8.

[16] C. Fredouille and G. Senay, "Technical improvements of the E-HMM based speaker diarization system for meeting records", in Proc. MLMI Third Int. Workshop, Bethesda, MD, USA, Revised Selected Paper, Berlin, Heidelberg: Springer-Verlag, 2006, pp. 359-370.

[17] Rentzeperis, A. Stergious, C. Boukis, A. Pnevmatikakis, and L. Polymenakos, "The 2006 Athens information technology speech activity detection and speaker diarization systems", in Proc. Mach. Learn. Multimodal Interaction: 3rd Int. Workshop, MLMI 2006, Bethesda, MD, Revised Selected Paper, Berlin, Heidelberg: Springer-Verlag, 2006, pp. 385-395.

[18] X. Anguera, C. Wooters, M. Anguilo, and C. Nadeu, "Hybrid speech non-speech detector applied to speaker diarization of meetings", in Proc. Speaker Odyssey Workshop, Puerto Rico, Jun. 2006.

[19] Temko, D. Macho, and C. Nadeu, "Enhanced SVM training for robust speech activity detection", in Proc. ICASSP, Honolulu, HI, 2007,pp. 1025-1028.

[20] Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system", in Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 509-519.

[21] T. L. Nwe, H. Sun, H. Li, and S. Rahardja, "Speaker diarization in meeting audio", in Proc. ICASSP, Taipei, Taiwan, 2009, pp.4073-4076.

[22] E. El-Khoury, C. Senac, and J. Pinquier, "Improved speaker diarization system for meetings", in Proc. ICASSP, Taipei, Taiwan, 2009, pp.4097-4100.

[23] Schwarz, G.: 1971, A sequential student test, The Annals of Statistics 42(3), 1003-1009.

[24] Schwarz, G.: 1978, Estimating the dimension of a model, The Annals of Statistics 6, 461-464.

[25] Kass, R. E. and Raftery, A. E.: 1995, Bayes factors, Journal of the American Statistics association 90, 773-795.

[26] Chickering, D. M. and Heckerman, D.: 1997, Efficient approximations for the marginal likelihood of bayesian networks with hidden variables, Machine Learning 29, 181-212.

[27] Shaobing Chen, S. and Gopalakrishnan, P.: 1998, Speaker, environment and channel change detection and clustering via the bayesian information criterion, Proceedings DARPA Broadcast News Transcription and Understanding Workshop, Virginia, USA.

[28] Chen, S. S. and Gopalakrishnan, P.: 1998, Clustering via the bayesian information criterion with applications in speech recognition, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, Seattle, USA, pp. 645-648.

[29] Chen, S. S., Gales, M. J. F., Gopinath, R. A., Kanvesky, D. and Olsen, P.: 2002, Automatic transcription of broadcast news, Speech Communication 37, 69-87.

[30] Tritschler, A. and Gopinath, R.: 1999, Improved speaker segmentation and segments clustering using the bayesian information criterion, Eurospeech'99, pp. 679-682.

[31] Delacourt, P. and Wellekens, C.J.: 2000, DISTBIC: A speaker-based segmentation for audio data indexing, Speech Communication: Special Issue in Accessing Information in Spoken Audio 32, 111-126.

[32] Mori, K. and Nakagawa, S.: 2001, Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, Salt Lake City, USA, pp. 413-416.

[33] Lopez, J. F. and Ellis, D. P W.: 2000a, Using acoustic condition clustering to improve acoustic change detection on broadcast news, Proc. International Conference on Speech and Language Processing, Beijing, China.

[34] Vandecatseye, A., Martens, J.-P. Et al.: 2004, The cost 278 pan-european broadcast news database, LREC'04, Lisbon, Portugal.

[35] Delacourt, P., Kryze, D. and Wellekens, C. J.: 1999a, Detection of speaker changes in an audio document, Eurospeech-1999, Budapest, Hungary.

[36] Ajmera, J., McCowan, I. and Bourlard, H.: 2003, Robust speaker change detection, Technical report, IDIAP.

[37] Perez-Freire, L. and Garcia-Mateo, C.: 2004, A multimedia approach for audio segmentation in TV broadcast news, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Montreal, Canada, pp. 369-372.

[38] Vandecatseye, A. and Martens, J.-P.: 2003, A fast, accurate and stream-based speaker segmentation and clustering algorithm, Eurospeech'03, Geneva, Switzerland, pp. 941-944.

[39] M. Roch and Y. Cheng, "Speaker segmentation using the MAP- Adapted bayesian information criterion", in Proc. Odyssey-2004 The Speaker and Language Recognition Workshop, Toledo, Spain, 2004, pp. 349-354.

[40] Delacourt, P., Kryze, D. and Wellekens, C. J.: 1999b, Speaker-based segmentation for audio data indexing, ESCA Workshop on accessing Information in Audio Data.

[41] J. Ramirez, J. M. Girriz, and J. C. Segura, M. Grimm and K. Kroschel, Eds., Voice activity detection. Fundamentals and speech recognition system robustness, in Proc. Robust Speech Recognit. Understand., Vienna, Austria, Jun. 2007, p. 460.

[42] Wooters, C., Fung, J., Peskin, B. and Anguera, X.: 2004, Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system, Fall 2004 Rich Transcription Workshop (RT04), Palisades, NY.

[43] Kim, H.-G., Ertelt, D. and Sikora, T.: 2005, Hybrid speaker-based segmentation system using model-level clustering, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, USA.

[44] Tranter, S. and Reynolds, D.: 2004, Speaker diarization for broadcast news, ODYSSEY'04, Toledo, Spain.

[45] Lu, L. and Zhang, H.-J.: 2002a, Real-time unsupervised speaker change detection, ICPR'02, Vol. 2, Quebec City, Canada.

[46] Ajmera, J.: 2004, Robust Audio Segmentation, PhD thesis, Ecole Polytechnique Federale de Lausanne.

[47] Anguera, X.: 2006, Robust Seaker Diarization for Meetings, PhD thesis, Universitat Politècnica de Catalunya.

[48] Willsky, A. S. and Jones, H. L.: 1976, A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems, IEEE Transactions on Automatic Control AC-21(1), 108-112.

[49] Appel, U. and Brandt, A.: 1982, Adaptive sequential segmentation of piecewise stationary time series, Inf. Sci. 29(1), 27-56.

[50] Bonastre, J.-F., Delacourt, P., Fredouille, C., Merlin, T. and Wellekens, C.: 2000, A speaker tracking system based on speaker turn detection for NIST evaluation, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Istanbul, Turkey, pp.1177-1180.

[51] Gangadharaiah, R., Narayanaswamy, B. and Balakrishnan, N.: 2004, A novel method for two-speaker segmentation, Proc. International Conference on Speech and Language Processing, Jeju, S. Korea.

[52] Adami, A. G., Kajarekar, S. S. and Hermansky, H.: 2002, A new speaker change detection method for two-speaker segmentation, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Orlando, Florida.

[53] Kyu J. Han, Samuel Kim, and Shrikanth S. Narayanan, Strategies to Improve the Robustness of Agglomerative Hierarchical Clustering Under Data Source Variation for Speaker Diarization, IEEE Transactions on Audio, Speech, and Language Processing, VOL. 16, NO. 8, November 2008.

[54] Gish, H., Siu, M.-H. and Rohlicek, R.: 1991, Segregation of speakers for speech recognition and speaker identification, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, Toronto, Canada, pp. 873-876.

[55] Gish, H. and Schmidt, M.: 1994, Text-independent speaker identification, Signal Processing Magazine, IEEE pp. 18-32.

[56] Kemp, T., Schmidt, M., Westphal, M. and Waibel, A.: 2000, Strategies for automatic segmentation of audio data, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Istanbul, Turkey, pp. 1423-1426.

[57] Siegler, M.A., Jain, U., Raj, B. and Stern, R. M.: 1997, Automatic segmentation, classification and clustering of broadcast news audio, DARPA Speech Recognition Workshop, Chantilly, pp. 97-99.

[58] Hung, J., Wang, H. and Lee, L.: 2000, Automatic metric based speech segmentation for broadcast news via principal component analysis, Proc. International Conference on Speech and Language Processing, Beijing, China.

[59] Campbell, J. P.: 1997, Speaker recognition: a tutorial, Proceedings of the IEEE 1.85(9), 1437-1462.

[60] Zochova, P. and Radova, V.: 2005, Modified DISTBIC algorithm for speaker change detection, Proc. International Conference on Speech and Language Processing, Lisbon, Portugal.

[61] Lu, L. and Zhang, H.-J.: 2002b,Speaker change detection and tracking in real-time news broadcasting analysis, ACM International Conference on Multimedia, pp. 602-610.

[62] Wu, T., Lu, L., Chen, K. and Zhang, H.-J.: 2003a, UBM-based incremental speaker adaptation, ICME'03, Vol. 2, pp. 721-724.

[63] Wu, T., Lu, L., Chen, K. and Zhang, H.-J.: 2003b, UBM-based real-time speaker segmentation for broadcasting news, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing.

[64] John A. Rice, Mathematical Statistics and Data Analysis, 3rd Edition.

[65] Wu, T., Lu, L., Chen, K. and Zhang, H.-J.: 2003c, Universal background models for real-time speaker change detection, International Conference on Multimedia Modeling.

[66] Mathieu Ben, Michael Betser, Frederic Bimbot, Guillaume Gravier, Speaker Diarization using bottom-up clustering based on a Parameter-derived Distance between adapted GMMs, Proc. ICSLP. Vol. 2004. 2004.

[67] Anguera, X. and Hernando, J.: 2004b, XBIC: nueva medida para segmentacion de locutor hacia el indexado automatico de la senal de voz, III Jornadas en Tecnologia del Habla, Valencia, Spain.

[68] Anguera, X., Wooters, C. and Hernando, J.: 2005, Speaker diarization for multi-party meetings using acoustic fusion, IEEE Automatic Speech Recognition and Understanding Workshop, Puerto Rico, USA.

[69] Malegaonkar, A., Ariyaeeinia, A., Sivakumaran, P. and Fortuna, J.: 2006, Unsupervised speaker change detection using probabilistic pattern matching, IEEE Signal Processing Letters 13(8), 509-512.

[70] Kemp, T., Schmidt, M., Westphal, M. and Waibel, A.: 2000, Strategies for automatic segmentation of audio data, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Istanbul, Turkey, pp. 1423-1426.

[71] Lee, K.-F.: 1998, Large vocabulary speaker-independent continuous speech recognition: the SPHINX system, PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA.

[72] Nakagawa, S. and Suzuki, H.: 1993, A new speech recognition method based on VQ-distortion and hmm, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, Minneapolis, USA, pp. 676-679.

[73] Zhou, B. and Hansen, J. H.: 2000, Unsupervised audio stream segmentation and clustering via the bayesian information criterion, Proc. International Conference on Speech and Language Processing, Vol. 3, Beijing, China, pp. 714-717.

[74] Lu, L., Zhang, H.-J. and Jiang, H.: 2002, Content analysis for audio classification and segmentation, IEEE Transactions on Speech and Audio Processing 10(7), 504-516.

[75] Rougui, J., Rziza, M., Aboutajdine, D., Gelgon, M. and Martinez, J.: 2006, Fast incremental clustering of gaussian mixture speaker models for scaling up retrieval in on-line broadcast, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse, France.

[76] Thomas Kemp, Michael Schmidt, Martin Westphal, and Alex Waibel. Strategies for automatic segmentation of audio data. In Proc. ICASSP, pages 1423-1426, 2000.

[77] Ali, I.; Saha, G., "A Robust Iterative Energy Based Voice Activity Detector," Emerging Trends in Engineering and Technology (ICETET), 2010 3rd International Conference on , vol., no., pp.599,604, 19-21 Nov. 2010 doi: 10.1109/ICETET.2010.58.

[78] Bertrand, A.; Moonen, M., "Energy-based multi-speaker voice activity detection with an ad hoc microphone array," Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on , vol., no., pp. 85, 88, 14-19 March 2010 doi: 10.1109/ICASSP.2010.5496183

[79] R. Huang and J. H. L. Hansen. Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora. IEEE Trans. Speech and Audio Processing, 14:907-919, 2006.

[80] Francis Kubala, Hubert Jin, Yspyros Matsoukas, Long Nguyen, Rich Schwartz, and John Makhoul. The 1996 bbn byblos hub-4 transcription system. In Proc. of DARPA Speech Recognition Workshop, pages 90-93, 1996.

[81] Jean-Luc Gauvain, Lori Lamel, Gilles Adda, and Michale Jardino. The limsi 1998 hub-4e transcription system. In PROC. OF THE DARPA BROADCAST NEWS WORKSHOP, pages 99-104, 1999.

[82] D. Reynolds and P. Torres-Carrasquillo. Approaches and applications of audio diarization. In Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP), volume V, pages 953-956, Philadelphia, PA, March 2005.

[83] J. Ajmera, H. Bourlard, and I Lapidot. Improved unknown-multiple speaker clustering using hmm. Technical Report IDIAP, 2002.

[84] Vescovi, M., Cettolo, M. and Rizzi, R.: 2003, A DP algoritm for speaker change detection, Eurospeech'03.

[85] Pwint, M. and Sattar, F.: 2005, A segmentation method for noisy speech using genetic algorithm, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, USA.

[86] Lathoud, G., McCowan, I. and Odobez, J.: 2004, Unsupervised location-based segmentation of multi-party speech, ICASSP-NIST Meeting Recognition Workshop.

[87] Niko; Brummer and Edward De Villiers. The Speaker Partitioning Problem. In Oddyssey Speaker and Language Recognition Workshop, 2010.

[88] David van Leeuwen. Speaker linking in large data sets. In Oddyssey Speaker and Language Recognition Workshop, 2010.

[89] Jin, H., Kubala, F. and Schwartz, R.: 1997, Automatic speaker clustering, DARPA Speech Recognition workshop, Chantilly, USA.

[90] Moraru, D., Ben, M. and Gravier, G.: 2005, Experiments on speaker tracking and segmentation in radio broadcast news, Proc. International Conference on Speech and Language Processing, Lisbon, Portugal.

[91] Solomonov, A., Mielke, A., Schmidt, M. and Gish, H.: 1998, Clustering speakers by their voices, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, Seattle, USA, pp. 757-760.

[92] Tsai, W.-H., Cheng, S.-S. and Wang, H.-M.: 2004, Speaker clustering of speech utterances using a voice characteristic reference space, Proc. International Conference on Speech and Language Processing, Jeju Island, Korea.

[93] Tritschler, A. and Gopinath, R.: 1999, Improved speaker segmentation and segments clustering using the bayesian information criterion, Eurospeech'99, pp. 679-682.

[94] Cettolo, M. and Vescovi, M.: 2003, Efficient audio segmentation algorithms based on the BIC, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing.

[95] Meinedo, H. and Neto, J.: 2003, Audio segmentation, classification and clustering in a broadcast news task, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Hong-Kong, China.

[96] Barras, C., Zhu, X., Meignier, S. and Gauvain, J.-L.: 2004, Improving speaker diarization, Fall 2004 Rich Transcription Workshop (RT04), Palisades, NY.

[97] Zhu, X., Barras, C., Meignier, S. and Gauvain, J.-L.: 2005, Combining speaker identification and bic for speaker diarization, Proc. International Conference on Speech and Language Processing, Lisbon, Portugal.

[98] Douglas A. Reynolds, Elliot Singer, Beth A. Carlson, Gerald C. O'Leary, Jack McLaughlin, and Marc A. Zissman. Blind clustering of speech utterances based on speaker and language characteristics. In ICSLP, 1998.

[99] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff. SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation. In 2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings, volume 1, pages I-97-I-100. IEEE 2006.

[100] Iso, Ken-ichi, "Speaker clustering using vector quantization and spectral clustering," Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on , vol., no., pp.4986,4989, 14-19 March 2010 doi: 10.1109/ICASSP.2010.5495078.

[101] Stafylakis, T.; Katsouros, V.; Carayannis, G., "The Segmental Bayesian Information Criterion and Its Applications to Speaker Diarization," Selected Topics in Signal Processing, IEEE Journal of , vol.4, no.5, pp. 857, 866, Oct. 2010 doi: 10.1109/JSTSP.2010.2048656.

[102] Johnson, S. and Woodland, P.: 1998, Speaker clustering using direct maximization of the MLLR-adapted likelihood, Proc. International Conference on Speech and Language Processing, Vol. 5, pp. 1775-1779.

[103] Johnson, S.: 1999, Who spoke when? - automatic segmentation and clustering for determining speaker turns, Eurospeech-99, Budapest, Hungary.

[104] Meignier, S., Bonastre, J.-F. and Igournet, S.: 2001, E-HMM approach for learning and adapting sound models for speaker indexing, A speaker Oddissey, Chania, Crete, pp. 175-180.

[105] Anguera, X. and Hernando, J.: 2004a, Evolutive speaker segmentation using a repository system, Proc. International Conference on Speech and Language Processing, Jeju Island, Korea.

[106] NIST Rich Transcription evaluations, website: URL: http://www.itl.nist.gov/iad/mig//tests/rt/2009/docs/rt09-meeting-eval-plan-v2-trackchanges.pdf.

[107] Base de dados TIMIT, URL: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

[108] Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M., The DET curve in assessment of detection task performance. In Proceedings of the European Conference on Speech Communication and Technology, 1997, pp. 1895-1898.

[109] T. Stafylakis, X. Anguera, Improvements to the equal-parameter BIC for Speaker Diarization, in Proceedings of Interspeech'10, 2010, September 26-30, 2010, Kaihin Makuhari, Japan.

[110] G. Friedland, O. Vinyals, Y. Huang, C. Muller: Prosodic and other Long-Term Features for Speaker Diarization, IEEE Transactions on Audio, Speech, and Language Processing, Vol 17, No 5, pp 985-993, July 2009

[111] John S. Garofolo, et al. 1993 TIMIT Acoustic-Phonetic Continuous Speech Corpus Linguistic Data Consortium, Philadelphia.

[112] NIST Rich Transcription evaluations, website: URL: http://www.itl.nist.gov/iad/mig/tests/rt/2002/index.html

[113] D.A.V. Leeuwen and M. Koneny, "Progress in the AMIDA speaker diarization system for meeting data, "in Proc. Multimodal Technol. for Percept. of Humans: Int. Eval. Workshops CLEAR 2007 and RT 2007, Baltimore, MD, May 8-11, 2007, Revised Selected Papers, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 475-483.

[114] J. Luque, X. Anguera, A. Temko, and J. Hernando, "Speaker diarization for conference room: The UPC RT07s evaluation system", in Proc. Multimodal Technol. for Perception. of Humans: Int. Eval. Workshops CLEAR 2007 and RT 2007, Baltimore, MD, May 8-11, 2007, Revised Selected Papers, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 543-553.

[115] Gillick, L., Cox, S.J., 1989. Some Statistical Issues in the Comparison of Speech Recognition Algorithms, ICASSP. IEEE, Glasgow, Scotland, pp.532-535.

[116] Lapidot, H. Guterman. "Future chanllenges in speaker diarization".

[117] Tranter, S.E.; Reynolds, D.A., "An overview of automatic speaker diarization systems," Audio, Speech, and Language Processing, IEEE Transactions on , vol.14, no.5, pp.1557,1565, Sept. 2006 doi: 10.1109/TASL.2006.878256.

[118] Ravichander Vipperla, Geiger, Jurgen T., Simom Bonzonnet, Dong Wang, Nicholas Evans, Bjorn Schuller and Gerhard Rigoll, "Speech overlap detection using convolutive non-negative sparse coding: New improvements and insights." Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European. IEEE, 2012.