UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Raoni Luar de Freitas Alcântara

Efeito da Reverberação na Inteligibilidade e na Identificação Acústica de Indivíduos

Campinas

2017

UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Raoni Luar de Freitas Alcântara

Efeito da Reverberação na Inteligibilidade e na Identificação Acústica de Indivíduos

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica, na Área de Telecomunicações e Telemática.

Orientador: Prof. Dr. Bruno Sanches Masiero

Co-orientadora Prof. Dr. Rosângela Fernandes Coelho

Este exemplar corresponde à versão final da dissertação defendida pelo aluno Raoni Luar de Freitas Alcântara, e orientada pelo Prof. Dr. Bruno Sanches Masiero

---

Campinas

2017

Agência(s) de fomento e nº(s) de processo(s): Não se aplica.
ORCID: https://orcid.org/0000-0003-3037-691

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Luciana Pietrosanto Milla - CRB 8/8129

AL16e

Alcântara, Raoni Luar de Freitas, 1989-
Efeito da reverberação na inteligibilidade e na identificação acústica de indivíduos / Raoni Luar de Freitas Alcântara. – Campinas, SP : [s.n.], 2018.

Orientador: Bruno Sanches Masiero.
Coorientador: Rosângela Fernandes Coelho.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Fala - Inteligibilidade. 2. Acústica. I. Masiero, Bruno Sanches, 1981-. II. Coelho, Rosângela Fernandes. III. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: The effect of reverberation in intelligibility and speaker identification
Palavras-chave em inglês:
Speech - Intelligibility
Acoustics
Área de concentração: Telecomunicações e Telemática
Titulação: Mestre em Engenharia Elétrica
Banca examinadora:
Bruno Sanches Masiero [Orientador]
Lee Luang Ling
Miguel Arjona Ramírez
Data de defesa: 10-08-2018
Programa de Pós-Graduação: Engenharia Elétrica

Comissão Julgadora – Dissertação de Mestrado

Candidato: Raoni Luar de Freitas Alcântara **RA:** 160103
Data da defesa: 10 de agosto de 2018

Título da Dissertação: "Efeito da Reverberação na Inteligibilidade e Identificação Acústica de Indivíduos".

Prof. Dr. Bruno Sanches Masiero (Presidente, FEEC/UNICAMP)
Prof. Dr. Lee Luan Ling (FEEC/UNICAMP)
Prof. Dr. Miguel Arjona Ramírez (Poli/USP)

A ata de defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no processo de vida acadêmica do aluno.

A todos que constroem a ciência para tornar o mundo melhor.

Agradecimentos

Agradeço ao Professor Bruno Masiero, meu orientador, por todo o prestado e a oportunidade de realizar o Mestrado na Unicamp.

À minha amiga e orientadora Professora Rosângela Coelho, que não mediu esforços para me apoiar à distância e que desde a graduação e me é um grande exemplo de professora e cientista.

Aos colegas do Laboratório de Processamento de Sinais Acústicos Zão, Zucatelli e Marília, por toda a ajuda e companheirismo nos momentos finais da dissertação.

Aos professores e funcionários da Unicamp, por toda a contribuição para a minha formação.

Aos colegas e ex-colegas do CPqD, pelo apoio e liberação parcial para cursar as disciplinas do Mestrado.

À minha família, Solange, Alcântara e Ramó, por todo o apoio, carinho e amor à distância e por toda a base que puderam me proporcionar.

À minha esposa Bruna, por todo o amor, companheirismo e compreensão neste período de dedicação durante o período do Mestrado.

Resumo

Nessa Dissertação, é estudado o impacto do efeito da reverberação em sistemas de identificação de locutor com casamento e descasamento de reverberações entre treinamento e teste. No desenvolvimento do trabalho, uma análise do efeito é realizada a partir das medidas espectrograma, cocleograma, INS (índice de não-estacionariedade) e distância Bhattacharyya do sinal de voz reverberado em diferentes condições. O estudo mostrou que o aumento do valor de RT_{60} em uma sala causa uma diminuição da não-estacionariedade do sinal de voz reverberado. Em seguida, experimentos com medidas objetivas indicam como a reverberação é capaz de degradar a inteligibilidade do sinal de voz e que a utilização de máscaras acústicas pode atenuar estes efeitos. Por fim, é proposto o emprego de máscaras acústicas para identificação de locutor em ambientes com reverberação. Experimentos de identificação de locutor indicaram que o uso de máscaras acústicas melhora os resultados de identificação para casamento de reverberação entre treinamento e teste. Também foi proposta a utilização do atributo acústico GFCC (*Gammatone Frequency Cepstral Coefficients*) e do classificador α-GMM para a identificação de locutor com reverberação. Estas técnicas se mostraram eficazes em recuperar as taxas de acerto em casos de descasamento de reverberação em uma mesma sala.

Palavras-chaves: reverberação; inteligibilidade; identificação de locutor.

Abstract

This work presents a study about the effect of reverberation in speaker identification systems with mismatch between training and testing phases. Spectrograms, cochleograms, INS (Index of Non-Stationarity) and the Bhattacharyya distance are used to analyze the reverberated speech signal under several conditions. This study show that an improvement in RT_{60} in a room can reduce the non-stationarity of the reverberated speech. Also, objective measures indicate that reverberation degrades speech intelligibility, that can be improved by binary masks. Then, binary masks are proposed to improve speaker identification systems under reverberant conditions. Results show that the binary masks improved the identification rates for reverberation matches between training and testing. The classifier α-GMM and the acoustic feature GFCC (Gammatone Frequency Cepstral Coefficients) are also proposed in this work for speaker identification in reverberant conditions. Those techniques were capable of improve the correct rates under reverberation mismatch in a room.

Keywords: reverberation; intelligibility; speaker identification.

Lista de ilustrações

Lista de tabelas

Lista de Siglas

γ Distribuição Gama

RT_{60} *Reverberation Time*

C_B Coeficiente Bhattacharyya

d_B Distância Bhattacharyya

d_{FM} Distância fonte-microfone

D_{KL} Distância de Kullback-Leibler

α-GMM α-*integraded Gaussian Mixture Model*

BRM *Binary Reverberant Mask*

CSII *Coherence and Speech Intelligibility Index*

DFT *Discrete Fourier Transform*

DRR *Direct-to-Reverberant energy Ratio*

FFT *Fast Fourier Transform*

GFCC *Gammatone Frequency Cepstral Coefficients*

GMM *Gaussian Mixture Model*

IBM *Ideal Binary Mask*

ICA *Independent Component Analysis*

INS *Index of Non-Stationarity*

IRM *Ideal Reverberant Mask*

MFCC *Mel-Frequency Cepstral Coefficients*

RIR *Room Impulse Response*

SII *Speech Intelligibility Index*

SNR *Signal-to-Noise Ratio*

SRMR *Signal-to-Reverberation Modulation energy Ratio*

SRR *Signal-to-Reverberation Ratio*

STFT *Short-Time Fourier Transform*

STOI *Short-Time Objective Intelligibility*

Sumário

1 Introdução

O efeito da reverberação é causado pelas múltiplas reflexões que ocorrem com uma onda sonora em superfícies e objetos antes desta ser captada por um microfone ou um ouvinte. No dia-a-dia, este efeito é principalmente notado em locais fechados como salas de aula, auditórios, igrejas ou teatros. A reverberação pode ser aplicada para amplificar sinais sonoros, como construção de igrejas e anfiteatros (CAPORALE, 1933) (SCHROEDER *et al.*, 1974) (BERANEK, 2008). No entanto, em sinais de voz, o efeito pode provocar um impacto severo na sua qualidade e inteligibilidade (BOLT; MACDONALD, 1949), afetando particularmente idosos e usuários de implantes cocleares. Esta degradação tem diversas consequências indesejáveis, como o agravamento do desempenho escolar (RABELO *et al.*, 2014), além de reduzir as taxas de acerto de sistemas de reconhecimento de voz e de locutor (GOLD; MORGAN, 1999).

Hoje em dia, com um número cada vez maior de aplicações embarcadas em dispositivos portáteis, as situações em que o sinal de voz será captado e as possíveis interferências que poderão corrompê-lo compõem um grande desafio para o processamento de sinais de voz. Em muitos casos, é interessante que estes efeitos sejam atenuados ao máximo, para que o sinal processado seja semelhante ao sinal limpo.

Na área de processamento de sinais, diferentes abordagens têm sido propostas para solucionar esta questão com o aprimoramento da qualidade e da inteligibilidade. As técnicas para melhorar a qualidade incluem métodos de realce de sinais de voz (LIM, 1983) (O'SHAUGHNESSY, 2000) (ZÃO *et al.*, 2014) (TAVARES; COELHO, 2016), filtragem espacial (VEEN; BUCKLEY, 1988) (KRIM; VIBERG, 1996) e separação cega de fontes com ICA (*Independent Component Analysis*) (LEE, 1998). Entretanto, apesar do desempenho apresentado, estas soluções possuem suas limitações. Por exemplo, técnicas de realce de voz são geralmente definidas para interferências específicas como ruídos acústicos e não apresentam significante aprimoramento de inteligibilidade (LOIZOU, 2013) (TAAL *et al.*, 2011).

As máscaras acústicas (WANG, 2005) (LOIZOU, 2013) surgem então como uma solução para prover inteligibilidade baseada no sistema auditivo. No cenário do "*cocktail party*" (CHERRY, 1953) (BRONKHORST, 2000), um ouvinte é capaz de selecionar e compreender uma única fonte sonora em meio a diversas interferências. Isto ocorre pois o sistema auditivo humano é capaz de realizar uma separação dos sinais correspondentes a fontes diferentes mesmo de maneira mono-auricular (BREGMAN, 1990). As máscaras acústicas ideais (WANG, 2005) (LOIZOU, 2013) são métodos que se baseiam nesta

capacidade auditiva e foram inicialmente definidas para aprimorar a inteligibilidade de sinais de voz corrompidos por interferências ou ruídos acústicos. O procedimento é realizado por uma divisão do sinal corrompido em quadros tempo-frequência com exclusão dos quadros considerados dominados pela interferência. As máscaras ideais são consideradas pela literatura como um limite superior do desempenho das máscaras acústicas. Nelas, são utilizadas informações a priori para se preservar os quadros em que o valor de SNR (*Signal-to-Noise Ratio*) é superior a um limiar predeterminado e excluir os demais quadros.

Máscaras acústicas foram apresentadas para alcançar uma melhora da inteligibilidade. A IRM (*Ideal Reverberant Mask*) apresentou ganhos de até 72% em testes subjetivos de inteligibilidade realizados com usuários de implantes cocleares (KOKKINAKIS *et al.*, 2011). A máscara BRM (*Binary Reverberant Mask*) cega (não-ideal), com foco na reverberação (HAZRATI *et al.*, 2012), mostrou melhorar a inteligibilidade em testes subjetivos. As máscaras não-ideais têm a vantagem de não serem limitadas à necessidade do conhecimento prévio do sinal limpo. Por estes aspectos, estas máscaras são mais adaptadas a situações reais. As máscaras acústicas também se mostraram eficientes em outras aplicações como reconhecimento de voz (HARTMANN; FOSLER-LUSSIER, 2011) e de locutor (NARAYANAN; WANG, 2013) com sinal de voz corrompido com ruídos acústicos e reverberação (ZHAO *et al.*, 2014).

Nesta Dissertação, é realizado um estudo da reverberação que incluem análise do espectrograma e cocleograma da voz reverberada por diferentes tipos de salas, bem como de diferentes respostas ao impulso de uma mesma sala. Além disso, a medida INS (*Index of Non-Stationarity*) é utilizada para investigar a estacionariedade do sinal de voz reverberado e a sua relação com o RT_{60}. A medida distância Bhattacharyya (d_B) também é utilizada para avaliar a o efeito da reverberação no histograma do sinal de voz. Experimentos de identificação de locutor serão realizados para avaliar o impacto da reverberação nesses sistemas, incluindo situações de descasamento de reverberação em uma mesma sala e em salas diferentes.

As máscaras acústicas e a sua eficiência em recuperar a inteligibilidade do sinal de voz reverberado são avaliadas através da utilização de três medidas objetivas de inteligibilidade, sendo uma delas não-intrusiva. Por fim, é proposta a utilização de máscaras acústicas como método de melhorar as taxas de identificação de locutor em situações de descasamento de reverberação.

O cenário dos experimentos de identificação neste trabalho incluiu o uso de sinais de voz selecionados da base TIMIT (GAROFOLO *et al.*, 1993) de 168 locutores, onde cada um forneceu duas locuções de duração média de 3 s para testes, amostradas a 16 kHz. Os testes de inteligibilidade objetiva foram realizados a partir de 240 locuções

obtidas de 20 locutores. De duas bases de reverberações, MARDY (WEN *et al.*, 2006) e AIR (JEUB *et al.*, 2009), foram escolhidas 9 respostas ao impulso de quatro salas distintas, que foram convoluídas com os sinais de voz para os experimentos de identificação e de inteligibilidade.

1.1 Objetivo Principal

O objetivo principal deste trabalho é o estudo do efeito da reverberação e a análise do seu impacto na inteligibilidade do sinal de voz e nos sistemas de identificação de locutor. Os objetivos específicos desse trabalho são:

- Analisar o efeito da reverberação de salas distintas.
- Investigar o impacto de diferentes salas e reverberações segundo a não-estacionariedade do sinal de voz.
- Estudar e avaliar de que forma a reverberação afeta os sistemas de identificação de locutor, incluindo situações de descasamento entre treinamento e teste.
- Examinar a eficiência de máscaras acústicas em recuperar a inteligibilidade de sinais de voz em situação de reverberação considerando diversas medidas objetivas de inteligibilidade.
- Propor um novo modelo de classificação de locutor, α-GMM para aprimorar as taxas de acerto em ambientes com reverberação.

1.2 Resultados obtidos

Os principais resultados e contribuições alcançados nessa Dissertação são:

- O estudo e a análise do sinal de voz reverberado em situações distintas considerando medidas nos domínios temporal e espectral, que incluiu a utilização da medida INS (*Index of Non-Stationarity*) para examinar a não-estacionariedade do sinal de voz reverberado. Este estudo concluiu que, em uma mesma sala, o aumento do valor de RT_{60} de uma resposta ao impulso implica em uma diminuição da não-não-estacionariedade do sinal de voz reverberado.
- Experimentos de identificação de locutor demonstraram que em situações onde há o casamento de reverberação entre treinamento e teste, as taxas são menos impactadas do que quando há o descasamento, mesmo entre uma mesma sala. Em situações de descasamento de reverberação em uma mesma sala, resultados apontaram quedas

de 69,05 p.p. (pontos percentuais). Quando os experimentos ocorreram com descasamento de salas entre treinamento e teste, as taxas de acerto apresentaram quedas de até 81,25 p.p.. Por outro lado, em situações de casamento de reverberação, a maior diminuição de taxa de acerto obtida foi de 9,22 p.p..

- Os resultados obtidos com as medidas objetivas de inteligibilidade CSII, STOI e SRMR acusaram que a reverberação impactou a inteligibilidade do sinal de voz. Os resultados mostraram quedas de até 31,30 p.p. na previsão da taxa de reconhecimento.

- O uso de máscaras acústicas ideais e de uma máscara não-ideal (cega) se mostrou efetivo em recuperar a inteligibilidade do sinal de voz reverberado. Os experimentos indicaram aumento de inteligibilidade de até 20,74 p.p. para máscaras ideais e 11,06 p.p. para a máscara cega.

- Experimentos de identificação de locutor com o atributo acústico GFCC com classificador GMM apresentaram melhores valores de taxa de acerto com descasamento de reverberação em uma mesma sala que o atributo clássico MFCC. Os resultados mostraram melhoras de até 17,56 p.p.

- Os resultados de identificação de locutor indicaram que a utilização de máscaras acústicas ideais melhora as taxas de identificação de locutor em situações de casamento de reverberação. Os resultados indicaram aumento médio de 4,30 p.p. nas taxas de acerto para identificação.

- A proposta do modelo α-GMM para classificação de locutores foi capaz de aprimorar os resultados de identificação com descasamento reverberação em uma mesma sala entre treinamento e teste. Os resultados mostraram um aumento médio até 5,31 p.p.

1.3 Organização da Dissertação de Mestrado

Os demais Capítulos deste manuscrito estão organizados da seguinte forma:

- **Capítulo 2:** Neste Capítulo são introduzidos os conceitos de reverberação e o seus principais parâmetros. Em seguida, a voz reverberada é analisada a partir dos seus espectrogramas e cocleogramas, além das medidas de INS e distância Bhattacharyya. Em seguida, são apresentados os atributos MFCC e modelo GMM, que representam o estado da arte em um sistema de identificação de locutor. Por fim, são mostrados resultados de identificação de locutor com descasamento de reverberação entre treinamento e teste a partir de duas bases de reverberação.

- **Capítulo 3:** O conceito de máscara acústica é apresentado e são mostradas as técnicas para a obtenção de duas máscaras acústicas ideais e uma máscara não-ideal (cega). Depois, são apresentadas as medidas objetivas de inteligibilidade CSII, STOI e SRMR. Por fim, são apresentados os resultados dos experimento de inteligibilidade com sinal de voz reverberado.

- **Capítulo 4:** Neste Capítulo, são mostradas duas propostas para melhorar a identificação de locutor na presença de reverberação, os atributos GFCC e o classificador α-GMM. Por fim, são apresentados os resultados de identificação utilizando as duas propostas e as máscaras acústicas introduzidas no Capítulo 4.

- **Conclusão:** Por fim, as principais conclusões do trabalho são apresentadas, além das propostas e sugestões de atividades futuras.

2 A Reverberação e seu Impacto na Identificação de Locutor

O avanço das tecnologias de aparelhos portáteis como celulares, *tablets* e *smartwatches* traz novos desafios relacionados à captação de áudio nos mais diversos tipos de cenário. Em situações onde o usuário encontra-se em salas, a presença de reverberação degrada a qualidade e a inteligibilidade do sinal de voz (BOLT; MACDONALD, 1949). Este efeito também influencia no desempenho de diferentes tipos de aplicações e sistemas, como, por exemplo, comunicações, reconhecimento de voz e de locutor (PAN; WAIBEL, 2000) (DELFARAH; WANG, 2017). Nos sistemas de identificação de locutor, a reverberação pode degradar as taxas de acerto. Isso pode ocorrer principalmente em situações onde a etapa de treinamento é realizada em ambiente anecoico (CASTELLANO *et al.*, 1996) ou quando há descasamento de salas entre treinamento e teste (SADJADI; HANSEN, 2014). A causa é o fato de a reverberação afetar diversos aspectos espectro-temporais do sinal de voz, preenchendo lacunas nas envoltórias temporais e aumentando a proeminência das baixas frequências (ASSMANN; SUMMERFIELD, 2004). Estes efeitos fazem com que características do sinal que são importantes para esta aplicação sejam "mascaradas". Consequentemente, acarretando na degradação da inteligibilidade do sinal.

Diversas soluções foram propostas para atenuar os efeitos da reverberação nos sistemas de reconhecimento de locutor. Entre elas, estão técnicas que atuam na captação do sinal de voz (GONZALEZ-RODRIGUEZ *et al.*, 1996) (JIN *et al.*, 2007), pré-processamento (BORGSTRÖM; MCCREE, 2012) e classificação ou decisão (PEER *et al.*, 2008) (GARCIA-ROMERO *et al.*, 2012). Propostas que tratam do sinal de voz reverberado na sua captação com o uso de arranjos de microfones (WANG *et al.*, 2012) se mostraram capazes de tornar os sistemas de identificação de locutor mais robustos a reverberação e/ou ruídos. Porém, estas técnicas não são aplicáveis em casos onde se tem acesso a somente um microfone (como na captação por um telefone) ou com sinais de voz obtidos previamente. Algoritmos que atuam nos atributos da voz utilizados para o reconhecimento de locutor, como o CMS (*Cepstral Mean Subtraction*) (ATAL, 1974), também se mostraram eficazes, mas somente em casos em que os valores de RT_{60}[1] das salas são mais baixos.

Em um ambiente com reverberação, os sinais sonoros chegam ao receptor através do caminho direto e de múltiplos percursos produzidos através das reflexões do sinal

[1] RT_{60} (*Reverberation Time*): Tempo necessário para que a resposta ao impulso da sala decaia em 60 dB.

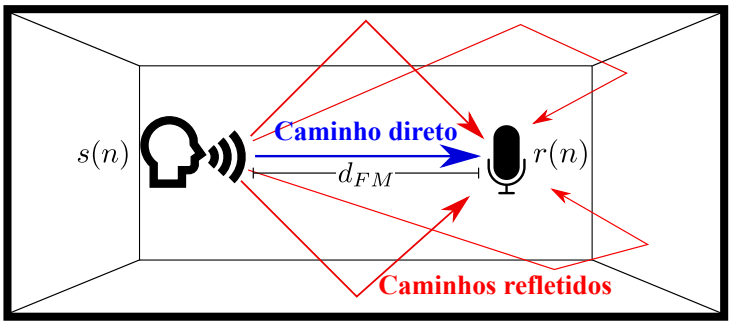

Figura 2.1 – Caminhos direto e refletidos na reverberação de um sinal em uma sala.

nas superfícies presentes na sala. Estes caminhos estão ilustrados na Figura 2.1, onde $s(n)$ representa o sinal produzido pela fonte, d_{FM}, a distância entre fonte e microfone e $r(n)$, o sinal reverberado. Estas reflexões são divididas entre primárias (*early*) e tardias (*late*) se chegam antes ou depois de um limiar definido geralmente entre 50 e 80 ms após a chegada o sinal direto (GÖLZER; KLEINSCHMIDT, 2003). As componentes tardias são responsáveis pelo fenômeno chamado de "mascaramento sonoro", em que os sinais refletidos sobrepõem o sinal direto. Este fenômeno é considerado como a causa principal da degradação da inteligibilidade acústica e da identificação de locutor tanto para a audição humana quanto para sistemas computacionais (NÁBĚLEK *et al.*, 1989).

A reverberação pode ser vista um efeito convolutivo no sinal de voz, ou seja, o sinal reverberado $r(n)$ pode ser matematicamente representado por

$$r(n) = s(n) * h(n), \tag{2.1}$$

onde $s(n)$ é o sinal direto e $h(n)$ é a resposta ao impulso da sala (RIR - *Room Impulse Response*). Esta resposta ao impulso depende de diversos fatores como tamanho da sala, objetos no seu interior, material das superfícies e até a posição do locutor e do receptor. Dois importantes parâmetros que se extraem da RIR, e que ajudam a entender a intensidade da reverberação e seu impacto no sinal de voz, são o RT_{60} e a DRR (*Direct-to-Reverberant energy Ratio*) (JEUB *et al.*, 2011). A Figura 2.2 mostra uma representação no tempo da resposta ao impulso de uma sala de aula de dimensões $5,0 \times 6,4 \times 2,9$ m^3 com $d_{FM} = 7,1$ m, $\mathrm{RT}_{60} = 0,85$ s e DRR $= -1,34$ dB. Nela, o primeiro pico, mais alto, é relativo ao sinal direto. Em seguida, percebe-se que a amplitude do sinal decai ao longo do tempo, evidenciando a maior atenuação das reflexões mais atrasadas.

O efeito de mascaramento é interpretado como uma interferência aditiva (LEBART *et al.*, 2001) e a RIR considerada como duas componentes, de forma que

$$h(n) = \begin{cases} 0, & n < 0, \\ h_e(n), & 0 \leq n \leq n_e, \\ h_l(n), & n_e \leq n \leq L. \end{cases} \tag{2.2}$$

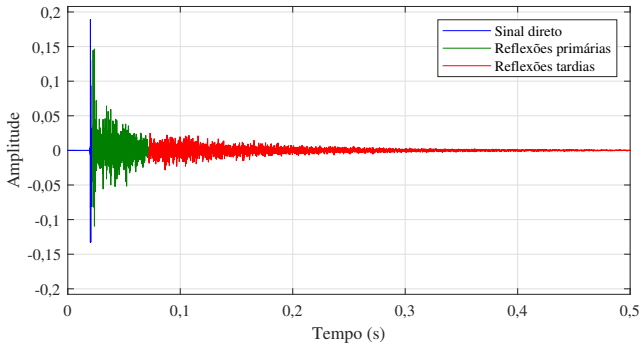

Figura 2.2 – Resposta ao impulso de uma sala de aula com dimensões $5,0 \times 6,4 \times 2,9$ m^3, $\mathrm{RT}_{60} = 0,85$ s e $\mathrm{DRR} = -1,34$ dB.

Onde n_e corresponde ao número de amostras da RIR relativas às reflexões primárias e L o comprimento de $h(n)$. Desta forma, o sinal de voz reverberado $r(n)$ pode ser descrito como

$$r(n) = \sum_{j=0}^{n_e-1} s(n-j)h_e(j) + \sum_{j=n_e}^{L-1} s(n-j)h_l(j), \tag{2.3}$$

em que o primeiro somatório representa as reflexões primárias e o segundo às reflexões tardias. A intensidade das reflexões primárias se reflete na medida de DRR, enquanto as reflexões tardias são caracterizadas pelo RT_{60}. Outros estudos importantes (ASSMANN; SUMMERFIELD, 2004) (LOLLMANN; VARY, 2009) (DELFARAH; WANG, 2017) demonstraram que a reverberação é quem mais impacta a inteligibilidade sonora.

Neste Capítulo, é apresentado um estudo do efeito da reverberação nas características tempo-espectrais de um sinal de voz. Para isto, foram consideradas medidas espectrais (espectrograma, cocleograma) e temporais (histograma). A investigação avalia também o impacto dos distintos graus ou índices de não-estacionariedade da voz a partir da medida objetiva INS (*Index of Non-Stationarity*) e um estudo da diferença entre as RIR através da distância Bhattacharyya. Em seguida, é abordado este efeito em um sistema de identificação de locutor, além de apresentado o estado da arte para esses sistemas, que são o atributo MFCC (*Mel-frequency cepstrum coefficients*) e o classificador GMM (*Gaussian mixture model*). Por fim, os resultados de identificação de locutor são apresentados em diferentes condições de reverberação para treinamento e teste, com variedade de salas e de RT_{60}. O impacto da reverberação nesses sistemas é analisado, bem como do efeito do descasamento de salas e RT_{60}.

2.1 O Efeito da Reverberação

Neste trabalho, é apresentado um estudo temporal e espectral do efeito da reverberação. Para isto, os espectrogramas e os cocleogramas do sinal de voz reverberado

foram utilizados. Além disso, a medida INS (*Index of Non-stationarity*) para análise da estacionariedade e a distância Bhattacharyya para comparação das distribuições do sinal reverberado por diferentes RIR foram empregados. O estudo também contou com a representação do sinal de voz no tempo e o seu histograma e foi realizado em duas partes:

1. **Reverberações em uma mesma sala**: Foram selecionadas 3 RIR medidas de diferentes configurações entre fonte e receptor em com valores de RT_{60} distintos da base MARDY (WEN *et al.*, 2006) para análise da voz reverberada com diferentes RIR de uma mesma sala.

2. **Reverberações de salas diferentes**: Foram selecionadas 2 RIR de 3 salas diferentes da base AIR (JEUB *et al.*, 2009) para análise da voz reverberada em salas distintas.

O sinal de voz utilizado nesta análise foi uma locução feminina extraída da base de voz TIMIT (GAROFOLO *et al.*, 1993). A seguir, uma breve descrição de cada medida:

- Espectrograma:

  O espectrograma (O'SHAUGHNESSY, 2000) é uma forma clássica de análise tempo-frequência de um sinal onde os eixos horizontal e vertical representam o tempo e a frequência, enquanto a magnitude em cada região é mostrada com cores diferentes. A decomposição é realizada a partir de uma STFT (*Short-time Fourier Transform*) com quadros de 20 ms.

- Cocleograma:

  Similar ao espectrograma, o cocleograma (BROWN; COOKE, 1994) é também uma forma de mostrar uma decomposição tempo-frequência de um sinal. Porém, realiza a decomposição em frequências através de um banco de filtros *Gammatone* (JOHANNESMA, 1972) (PATTERSON; MOORE, 1986) (COOKE, 1993), que emulam a filtragem da cóclea humana. Além disso, no cocleograma as frequências no eixo vertical são apresentadas logaritmicamente, de forma a ressaltar as regiões de frequência em que a voz está presente.

- Índice de Não-Estacionariedade (INS):

  Um sinal é dito estacionário quando as suas principais estatísticas são preservadas ao longo do tempo de observação. Quanto mais estas medidas variam, maior é o grau de não-estacionariedade do sinal. O Índice de Não-Estacionariedade (INS) (BORGNAT *et al.*, 2010) é uma medida objetiva capaz de mensurar este comportamento. Isto é feito a partir da comparação entre as componentes espectrais do sinal original e

de referenciais espectrais estacionários (*surrogates*), obtidos a partir da substituição da fase original do sinal por uma sequência aleatória uniformemente distribuída no intervalo $[-\pi, \pi]$.

Em seguida, é utilizada a distância de Kullback-Leibler (D_{KL}) simétrica (BASSEVILLE, 1989) entre o sinal analisado e seu referencial estacionário. Por fim, o INS é definido como a razão entre a variância das distâncias observadas ($\Theta_0(j)$) e a média das variâncias obtidas dos sinais referenciais (Θ_1), ou seja,

$$\text{INS} := \sqrt{\frac{\Theta_1}{\langle \Theta_0(j) \rangle_j}}. \tag{2.4}$$

Em (BORGNAT *et al.*, 2010), é definido um limiar a partir de uma distribuição γ para o teste da não-estacionariedade do sinal analisado considerando uma precisão de 95%. O índice é então definido por,

$$\text{INS} \begin{cases} \leq \gamma, & x(t) \text{ é estacionário;} \\ > \gamma, & x(t) \text{ é não-estacionário.} \end{cases} \tag{2.5}$$

Neste trabalho, o INS é apresentado na escala T_h/T, que significa a razão entre o tamanho da janela (T_h) e o tamanho total do sinal analisado (T). O índice é adotado neste trabalho para análise do efeito da reverberação no sinal de voz em diferentes situações ou salas. A duração dos sinais de voz analisados é de 3 s.

- Distância Bhattacharyya:

  A distância Bhattacharyya (d_B) (KAILATH, 1967) foi proposta como critério de seleção ou separação de sinais segundo suas distribuições estatísticas. A medida é baseada no Coeficiente Bhattacharyya (C_B). Dadas funções densidade de probabilidade com histogramas $p_1(x)$ e $p_2(x)$, C_B é dado por

$$C_B = \int_{-\infty}^{\infty} \sqrt{p_1(x) p_2(x)} \mathrm{d}x. \tag{2.6}$$

  A distância Bhattacharyya (d_B) é então calculada por

$$d_B(X_1, X_2) = -\ln C_B. \tag{2.7}$$

  em que $0 < d_B < \infty$

  Neste estudo, a distância Bhattacharyya é utilizada para comparação entre distribuições do sinal de voz limpo e reverberado.

Tabela 2.1 – Parâmetros RT_{60}, d_{FM} e DRR das RIR selecionadas da base MARDY.

RT_{60} (s)	d_{FM} (m)	DRR (dB)	Volume (m^3)
0,55	1,00	13,04	
0,60	2,00	7,24	208,80
0,65	3,00	3,19	

2.1.1 Análise em uma mesma sala - Base MARDY

A base MARDY (WEN *et al.*, 2006) foi utilizada para a análise da reverberação em uma mesma sala com diferentes configurações de posição de fonte e receptor. Os parâmetros (RT_{60}, d_{FM} e DRR) referentes às três composições de RIR extraídas desta base estão apresentados na Tabela 2.1.

A Figura 2.3 apresenta as medidas espectrograma, cocleograma e INS obtidas usando a voz limpa (sem reverberação). Nas imagens é possível notar a importância do cocleograma para a análise espectral. Para regiões de frequência semelhantes ao do espectrograma, o cocleograma melhor destaca as frequências onde o sinal de voz está presente. Note também que o INS tem seu pico próximo de 340 em $T_h/T = 0,03$ e está

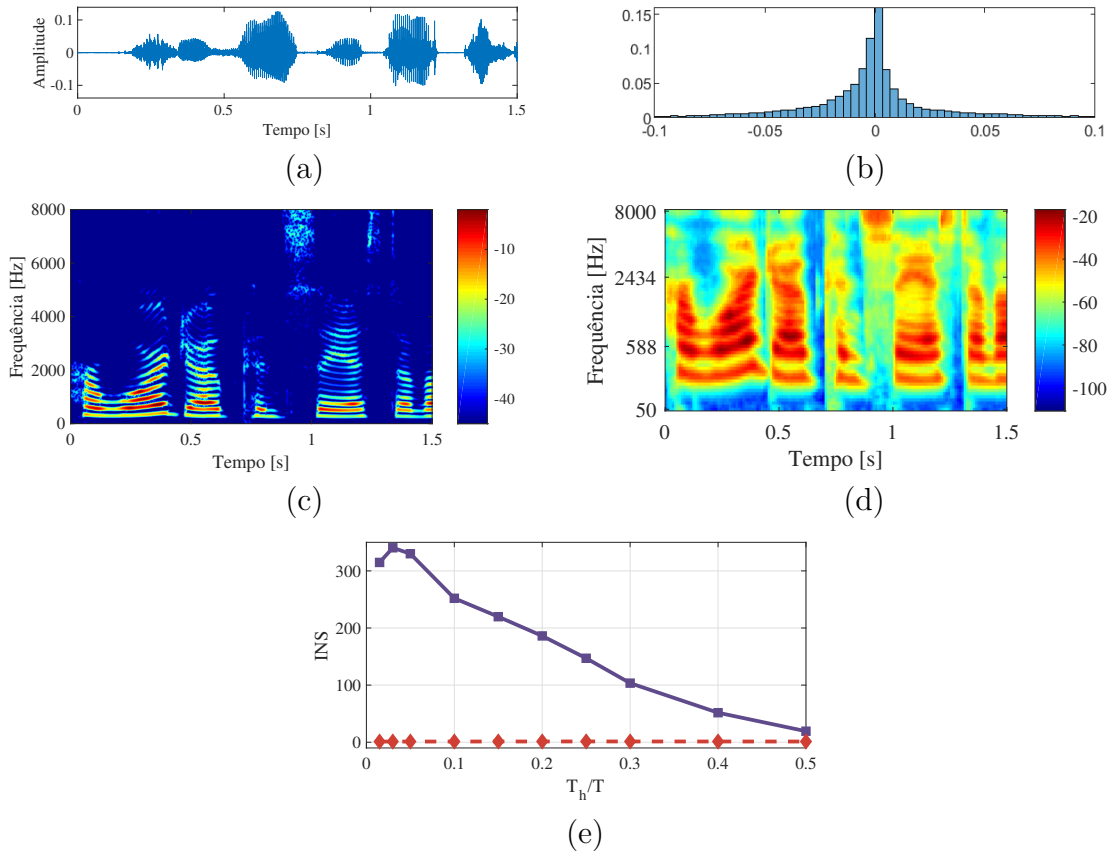

Figura 2.3 – Sinal de voz limpo: (a) Sinal do tempo, (b) histograma, (c) espectrograma, (d) cocleograma e (e) INS.

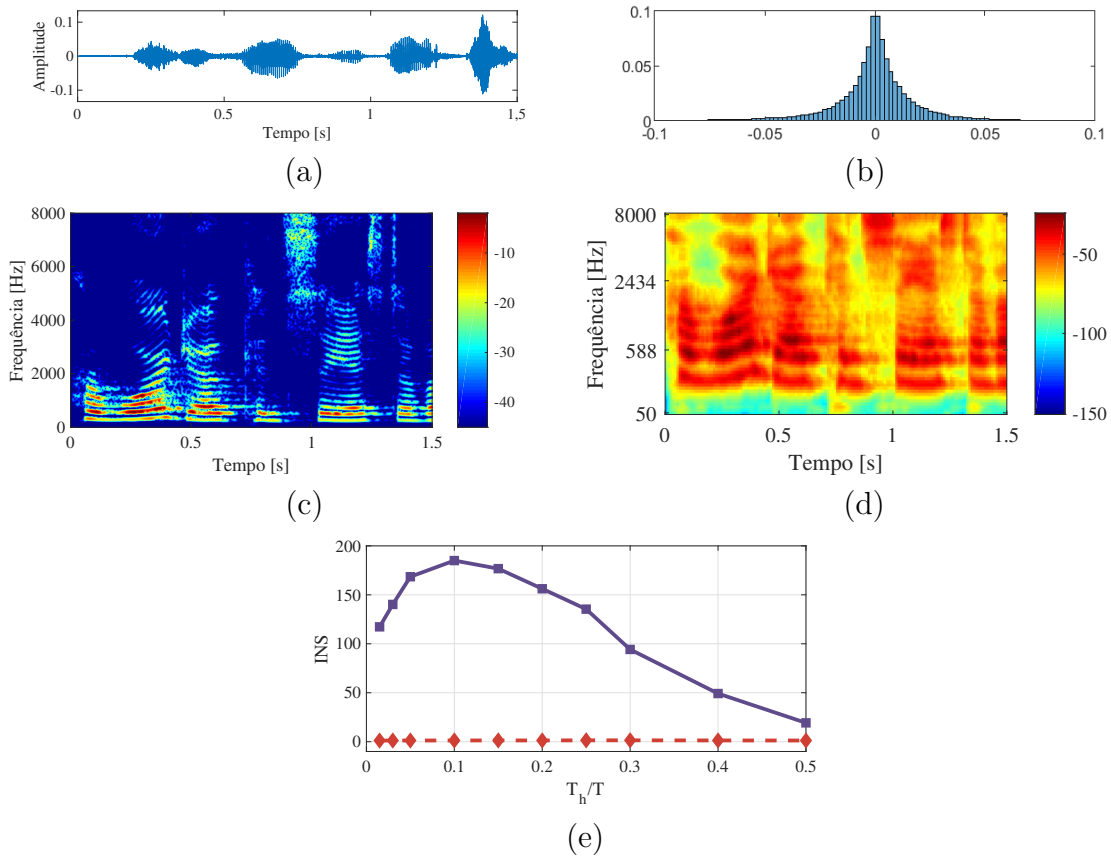

Figura 2.4 – Sinal de voz com reverberação $\mathrm{RT}_{60} = 0,55$ s: (a) Sinal do tempo, (b) histograma, (c) espectrograma, (d) cocleograma e (e) INS.

acima do limiar de não-estacionariedade em todo o domínio analisado, ou seja, para estes valores, o sinal de voz é considerado altamente não-estacionário.

Na Figura 2.4 estão ilustradas as medidas para o sinal de voz com reverberação $\mathrm{RT}_{60} = 0,55$ s. Note que o sinal de voz reverberado, quando comparado com o sinal "limpo" (vide Figura 2.3), apresenta períodos de silêncio preenchidos pelas reflexões deslocadas no tempo em referência ao sinal direto. Estas reflexões também podem ser notadas principalmente nas frequências mais baixas, abaixo de 2 kHz, no espectrograma e no cocleograma. Neste último, a região de atuação do efeito da reverberação está mais destacada. Nas frequências a partir de 2 kHz, a reverberação não está tão presente quanto nas demais. Isto ocorre pois os materiais construtivos das paredes tendem a ser mais absorventes nas altas frequências, resultando em um menor número de reflexões nesta faixa (VORLÄNDER, 2007). O maior valor de INS ocorreu para $\mathrm{T}_h/\mathrm{T} = 0,1$ aproximadamente em 184, abaixo dos 340 do sinal limpo, mesmo assim o sinal permanece acima da linha pontilhada, sendo ainda considerado não-estacionário. O histograma mostrou sinais mais próximos da média que no sinal sem reverberação e o valor de d_B entre o sinal limpo e o sinal reverberado com $\mathrm{RT}_{60} = 0,55$ s foi $d_B = 0,12$.

As medidas realizadas a partir do sinal de voz com reverberação $\mathrm{RT}_{60} = 0,60$ s

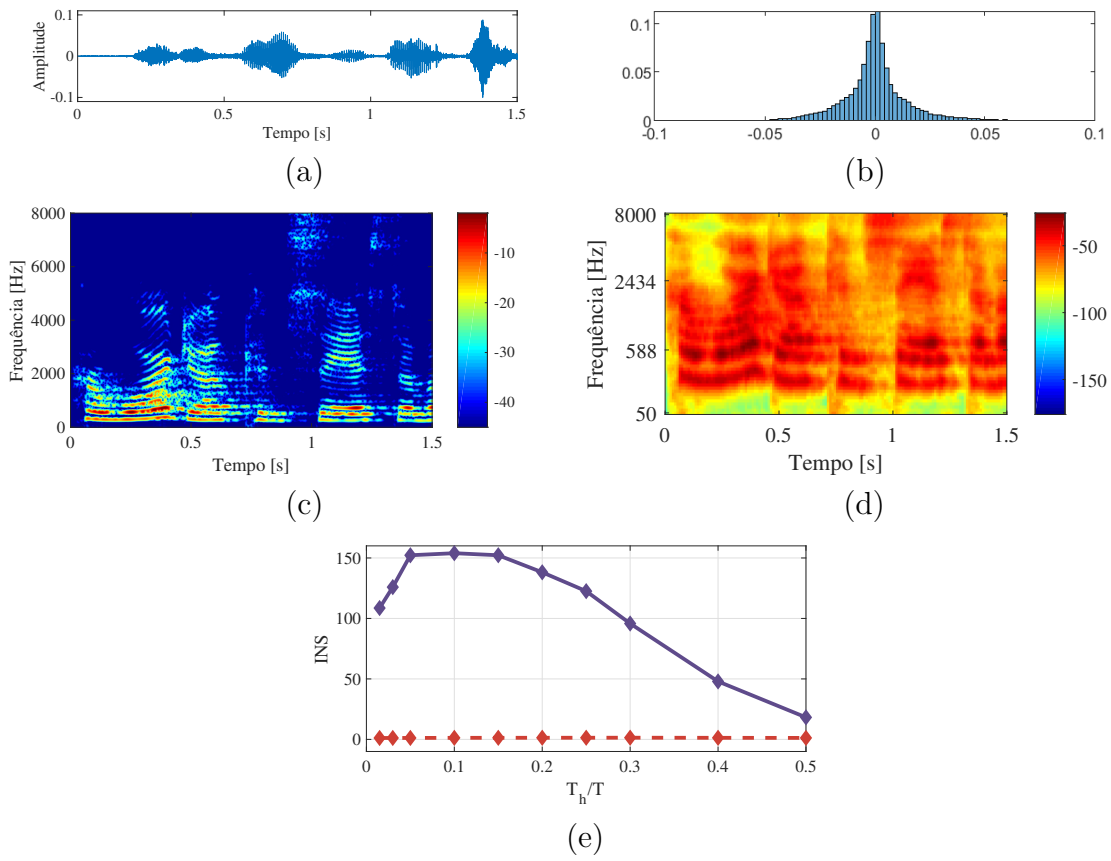

Figura 2.5 – Sinal de voz com reverberação $\mathrm{RT}_{60} = 0,60$ s: (a) Sinal do tempo, (b) histograma, (c) espectrograma, (d) cocleograma e (e) INS.

podem ser observadas na Figura 2.5. Perceba que para esse valor de RT_{60}, o impacto da reverberação no sinal de voz é mais intenso que quando comparado ao caso $\mathrm{RT}_{60} = 0,55$ s. O sinal de voz no tempo mostra que este efeito distorce ainda mais o sinal direto. Isto também pode ser visto no espectrograma e no cocleograma, com as reflexões tardias preenchendo ainda mais os períodos de silêncio do sinal direto, ou seja, distorcendo seus detalhes espectrais e temporais, principalmente nas frequências mais baixas. O maior valor de INS ficou próximo de 153, indicando uma não-estacionariedade menor que quando comparada ao caso $\mathrm{RT}_{60} = 0,55$ s, porém ainda acima do limiar de estacionariedade. Este resultado demonstra quantitativamente o efeito da reverberação na não-estacionariedade. O histograma do sinal reverberado mostrou um sinal mais próximo da média que o anterior e a distância Bhattacharyya entre o sinal reverberado e o sinal limpo foi de $d_B = 0,21$, maior que a registrada para $\mathrm{RT}_{60} = 0,55$ s.

Por fim, a voz reverberada com o maior tempo de reverberação, $\mathrm{RT}_{60} = 0,65$ s, teve suas medidas ilustradas na Figura 2.6. O impacto causado pela reverberação mostra-s maior neste caso. Isto pode ser percebido observando o sinal de voz no tempo, mais distorcido com a reverberação. O espectrograma e cocleograma mostram um sinal onde as reflexões do sinal direto estão mais presentes e preenchendo mais regiões tempo-frequência

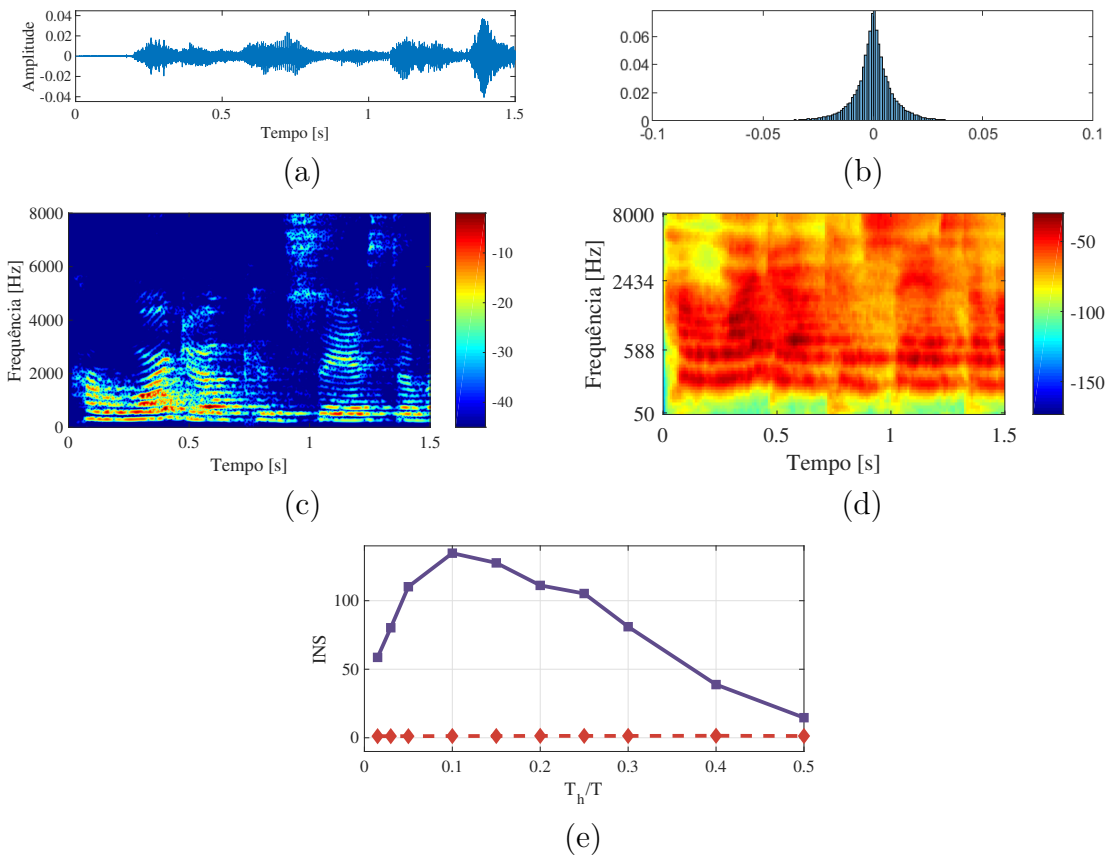

Figura 2.6 – Sinal de voz com reverberação $\mathrm{RT}_{60} = 0,65$ s: (a) Sinal do tempo, (b) histograma, (c) espectrograma, (d) cocleograma e (e) INS.

nas frequências mais baixas. Note que a reverberação agora também pode ser vista nas frequências mais altas, ainda que em menor intensidade. Além disso, a curva do INS indica um sinal com pico próximo de 134, abaixo do sinal reverberado com $\mathrm{RT}_{60} = 0,60$ s, mostrando uma não-estacionariedade menor mas se mantendo acima do limiar de estacionariedade, sendo considerado como não-estacionário. O histograma do sinal reverberado mostra que este está mais próximo da média que os demais. A distância Bhattacharyya medida entre o sinal com reverberação $\mathrm{RT}_{60} = 0,65$ s e o sinal limpo foi a maior das três, de 0,41.

A Tabela 2.2 apresenta os resultados de d_B medidos entre o sinal de voz limpo e os reverberados pelas RIR selecionadas da base MARDY, além de suas classificações de acordo com a sua não-estacionariedade. Os resultados indicam que o histograma dos

Tabela 2.2 – Reverberações da base MARDY com seus d_B medidos entre o sinal de voz limpo e sinal reverberado, RT_{60} e classificação segundo estacionariedade.

RT_{60} (s)	DRR	Não-estacionariedade	d_B
0,55	13,04	Altamente não-estacionário	0,12
0,60	7,24	Moderadamente não-estacionário	0,21
0,65	3,19	Não-estacionário	0,41

Tabela 2.3 – Salas selecionadas da base AIR e parâmetros de RT_{60}, d_{fm} e DRR das RIR

Sala	RT_{60} (s)	d_{sm} (m)	DRR (dB)	Dimensões (m)
Escritório	0,64	3,00	-0,04	$5,0 \times 6,4 \times 2,9$
Sala de aula	0,87	5,00	-4,52	$10,8 \times 6,4 \times 2,9$

sinais de voz reverberados se distanciou do sinal limpo de acordo com o seu valor de RT_{60}, se distanciando mais para as reverberações mais longas. Além disso, apesar de todos os sinais, através do INS, serem considerados não-estacionários, notou-se que esta não-estacionariedade diminuiu com o aumento do valor de RT_{60}. Os sinais foram classificados então como Altamente não-estacionário, Moderadamente não-estacionário e Não-estacionário de acordo com o seu pico de INS nos valores de T_h/T analisados.

2.1.2 Análise em salas diferentes - Base AIR

Nesta Seção, a análise das reverberações foi realizada entre salas diferentes utilizando a base de reverberações AIR (JEUB *et al.*, 2009). Os parâmetros referentes às salas adotadas se encontram na Tabela 2.3.

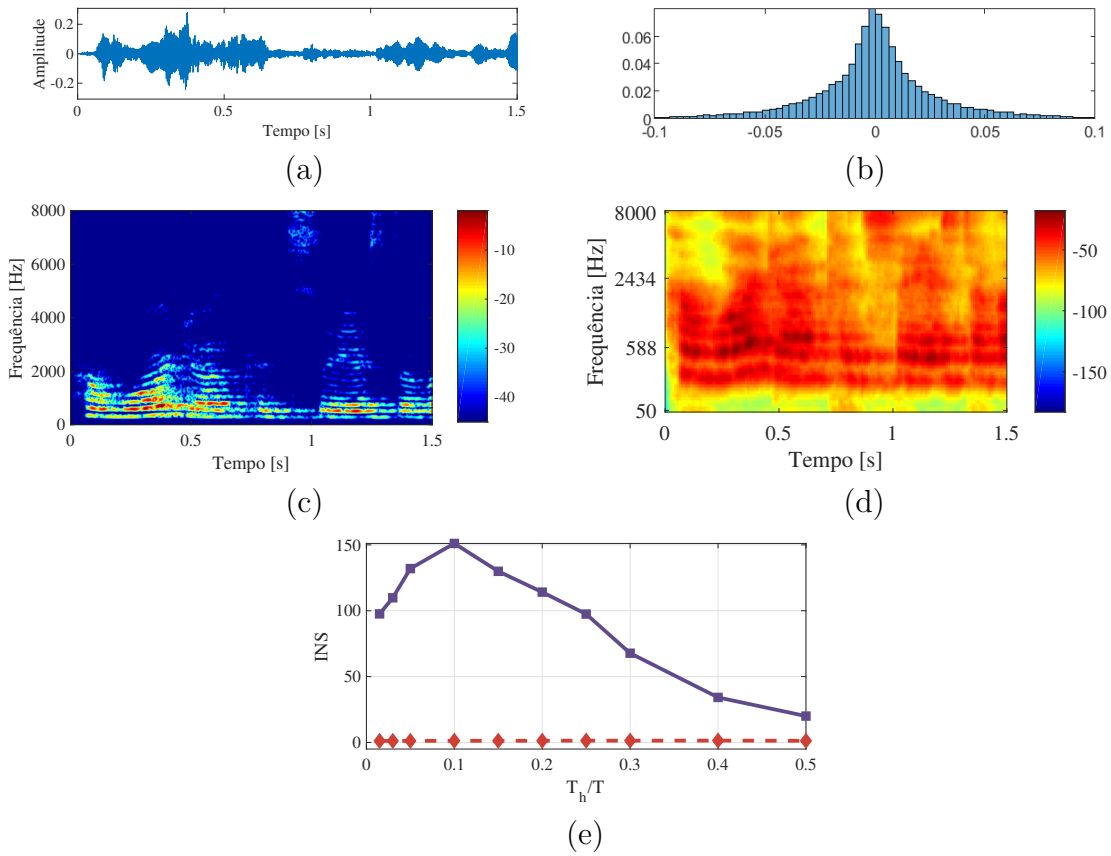

Figura 2.7 – Sinal de voz com reverberação de Escritório (base AIR) com $RT_{60} = 0,64$ s, $d_{FM} = 3,00$ m e DRR $= -0,04$ dB: (a) Sinal do tempo, (b) histograma, (c) espectrograma, (d) cocleograma e (e) INS.

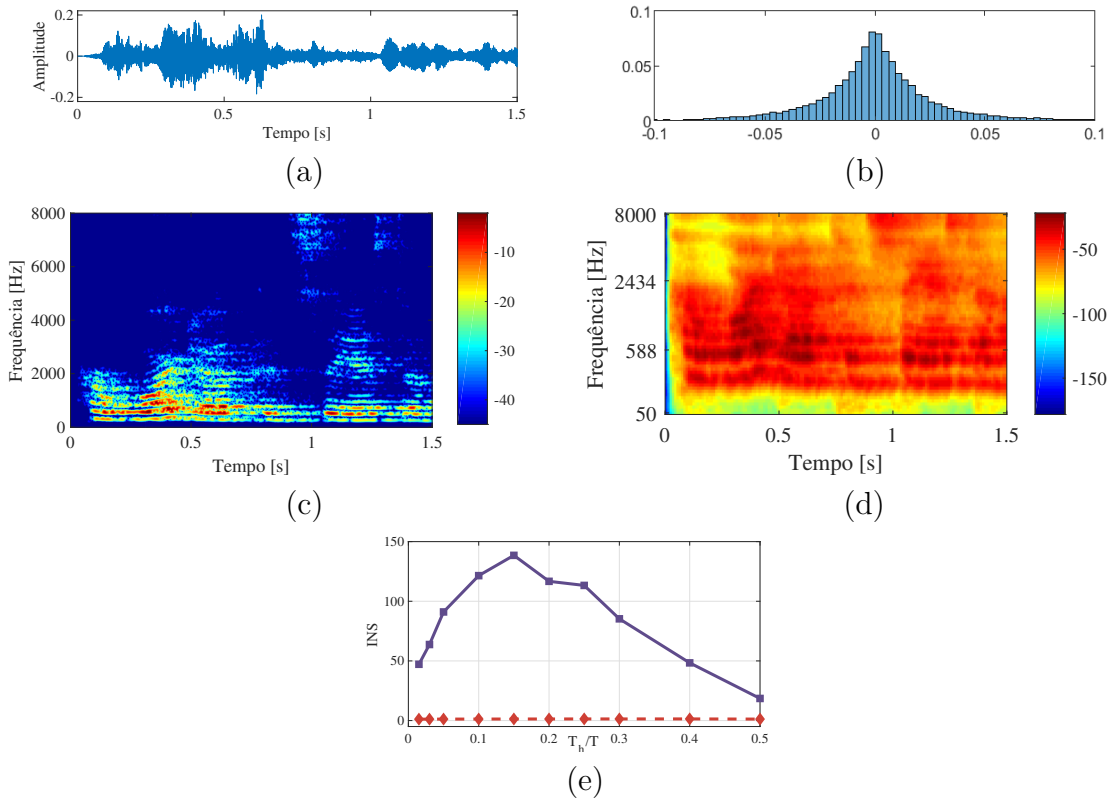

Figura 2.8 – Sinal de voz com reverberação de Sala de aula (base AIR) com $\mathrm{RT}_{60} = 0,87$ s, $d_{FM} = 5,00$ m e DRR $= -4,52$ dB: (a) Sinal do tempo, (b) histograma, (c) espectrograma, (d) cocleograma e (e) INS.

Na Figura 2.7, podem-se verificar as medidas obtidas para o sinal de voz reverberado com a RIR de um Escritório da base AIR. Note que, apesar de um valor de $\mathrm{RT}_{60} = 0,64$ s, semelhante à reverberação de $\mathrm{RT}_{60} = 0,65$ s da base MARDY, o espectrograma e o cocleograma mostram uma maior presença da reverberação que na Figura 2.6. Isto se deve ao fato de que, apesar de valores de RT_{60} semelhantes, as salas apresentam dimensão e volume diferentes, sendo a sala MARDY com volume de 208,80 m^3 e a sala Escritório de 92,80 m^3. Também é possível perceber que as frequências mais altas também foram mais afetadas pela reverberação quando comparadas à sala MARDY com RT_{60} próximo. Além disso, a curva de INS mostrou-se diferente, com pico bem definido próximo de 151, acima do valor encontrado para a base MARDY e também mantendo-se acima do limiar de não-estacionariedade. O valor de d_B em relação ao sinal limpo obtido para esta reverberação foi de 0,02.

A Figura 2.8 apresenta as medidas para o sinal de voz reverberado obtido pela RIR de uma sala de aula. Esta RIR possui $\mathrm{RT}_{60} = 0,87$ s e apresenta a maior predominância das reflexões no sinal reverberado, inclusive nas altas frequências, como mostram o espectrograma e o cocleograma. Nota-se, neste caso, que o pico de INS encontrado foi próximo a 138, em em $\mathrm{T}_h/\mathrm{T} = 0,15$. O valor de d_B encontrado para este efeito de reverberação foi de 0,03.

Tabela 2.4 – Salas da base AIR com medidas de d_B medidas entre o sinal de voz limpo e sinal reverberado, RT_{60} e classificação segundo não-estacionariedade.

Sala	RT_{60} (s)	INS (pico)	Não-estacionariedade	d_B
Escritório	0,64	151	Moderadamente não-estacionário	0,02
Sala de aula	0,84	138	Não-estacionário	0,03

A Tabela 2.4 apresenta os valores de d_B medidos para a voz reverberada no caso da base AIR e sua classificação de acordo com a não-estacionariedade. Os resultados encontrados estiveram abaixo dos medidos com a voz reverberada pela base MARDY (vide 2.2). Isto condiz com os histogramas encontrados nos experimentos com a base AIR, que estão mais afastados da média que medidos com $RT_{60} = 0,55$ s com a MARDY e mais próximos da voz sem reverberação.De acordo com os valores de INS, os sinal reverberado por Escritório foi considerado Moderadamente não-estacionário e o da Sala de aula como de Não-estacionário.

2.2 Identificação de Locutor

Por ser o sinal acústico resultante do sistema de produção da fala (OSHAUGHNESSY, 1987), a voz é uma das principais características biométrica do ser humano. As informações contidas na voz incluem identidade, sexo, idioma e condições físico-emocionais do locutor. O objetivo de um sistema de identificação automática de locutor (IL) é diferenciar os seres humanos por suas características biométricas acústicas ou impressão vocal. Isto é possível devido ao fato de que cada indivíduo possuir um trato vocal distinto (RABINER; JUANG, 1993).

O sistema IL é geralmente operado em duas fases: treinamento e teste. A Figura 2.9 apresenta um diagrama que mostra o processo de IL. Cada fase é dividida em três etapas: pré-processamento, extração de atributos e classificação (CAMPBELL, 1997). Na primeira etapa, são executadas a aquisição e a digitalização do sinal de voz. Na etapa seguinte, o sinal de voz é janelado e passa por filtros e estimadores para a extração dos seus atributos acústicos, ou características. Na etapa de classificação ocorre a produção do modelo do locutor e a decisão do sistema.

Na fase de testes, a etapa de classificação pode executar duas possíveis tarefas: a identificação (ROSE *et al.*, 1994) ou a verificação de locutor (BIMBOT *et al.*, 2004). Na identificação, o sistema decide a qual dos locutores cujos modelos foram previamente calculados pertence aquela locução de teste. Já na tarefa de verificação, o sistema deve responder se o locutor que realiza o teste é o mesmo locutor declarado. Neste trabalho é abordada a tarefa de identificação, e os resultados foram calculados a partir da taxa de

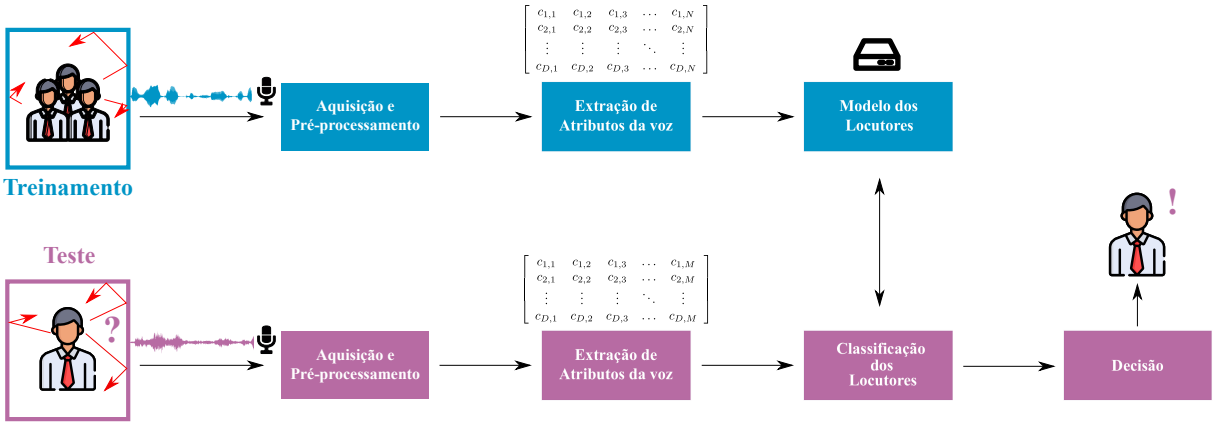

Figura 2.9 – Diagrama ilustrativo de um sistema de identificação.

acertos de cada experimento.

As características da voz são informações extraídas do sinal de voz de um locutor e podem ser de alto ou de baixo nível. As características de alto nível compreendem o dialeto, contexto e estado emocional. Os atributos de baixo nível são o *pitch* (frequência fundamental), as magnitudes espectrais e a energia, que são derivadas da natureza acústica da voz.

Os principais critérios para que uma característica seja considerada interessante para os sistemas de identificação de locutor são uma alta capacidade discriminatória, grande variabilidade entre indivíduos e baixa variabilidade em um mesmo locutor. Para o reconhecimento do locutor, os atributos da voz são geralmente baseados em suas características espectrais (REYNOLDS; ROSE, 1995). Isto se deve à capacidade do espectro de representar a estrutura do trato vocal de um indivíduo, possibilitando uma melhor diferenciação dos locutores.

Neste Capítulo, são utilizados os coeficientes MFCC (*Mel-Frequency Cepstral Coefficients*) (FURUI, 1981) (DAVIS; MERMELSTEIN, 1980) como característica biométrica da voz e o classificador GMM (*Gaussian Mixture Model*) (REYNOLDS; ROSE, 1995). Este conjunto é adotado na literatura como referência de bom desempenho de sistemas de identificação de locutor.

2.2.1 O vetor de MFCC

A Figura 2.10 ilustra a extração dos atributos MFCC de acordo com (DAVIS; MERMELSTEIN, 1980). Após a etapa de pré-processamento, o sinal de voz é dividido em quadros de curta duração, geralmente de 20 ms a 32 ms. Cada quadro é transformado para o domínio da frequência através de uma transformada rápida de Fourier (FFT - *Fast Fourier Transform*) e filtrado por um banco de filtros triangulares na escala MEL (IMAI, 1983).

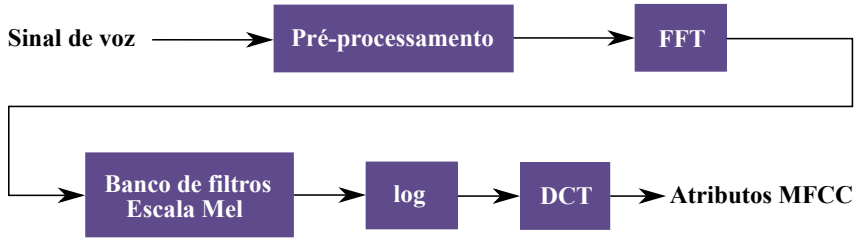

Figura 2.10 – Extração de coeficientes MFCC.

A escala Mel representa a percepção auditiva humana das variações em frequência. Os valores da escala Mel (f_{Mel}) são relacionados com a escala linear (f_{Hz}) por

$$f_{Mel} = 1127 \ln\left(1 + \frac{f_{Hz}}{700}\right) . \tag{2.8}$$

Em seguida, é calculado o logaritmo das potências nas saídas dos filtros e, então, é aplicada uma transformada discreta do cosseno (DCT - *discrete cosine transform*). Assim, os MFCC (c_j) são obtidos de forma que

$$c_j = \sum_{k=1}^{F} (\log S_k) \cos\left[j\left(k - \frac{1}{2}\right)\frac{\pi}{F}\right], \qquad j = 1, 2, \ldots, D, \tag{2.9}$$

onde F é o número de filtros na escala Mel, S_k é a potência na saída do k-ésimo filtro e D é o número de coeficientes MFCC. Assim, a cada quadro é obtido um vetor $\vec{x}$ com D componentes. Ou seja,

$$\vec{x} = [\, c_1 \, c_2 \, \cdots \, c_D \,]^T . \tag{2.10}$$

Se um sinal de voz possui Q quadros, então uma matriz de atributos $X_{D\times Q}$, é formada pela concatenação destes vetores. Logo,

$$X = [\, \vec{x_1} \, \vec{x_2} \, \cdots \, \vec{x_Q} \,] . \tag{2.11}$$

2.2.2 Modelo para classificação: GMM

Sistemas de identificação de locutor que utilizam GMM apresentam ótimo desempenho em termos de taxa de acerto e acurácia, mas podem ser severamente degradados na presença de interferências como a reverberação ou ruídos. O modelo GMM (λ) (REYNOLDS; ROSE, 1995) é definido como uma soma ponderada de M distribuições Gaussianas e foi proposto para classificação de locutores. A soma de gaussianas é dada por

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}), \tag{2.12}$$

Cada componente do modelo GMM $b_i(\vec{x})$ é uma função Gaussiana de dimensão D e podem ser descritas por

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}}\sqrt{|K_i|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu_i})^T K_i^{-1}(\vec{x}-\vec{\mu_i})}. \tag{2.13}$$

onde $\vec{\mu_i}$ é o vetor média e K_i a matriz de covariância, com determinante $|K_i|$. Os pesos p_i satisfazem a condição $\sum_{i=1}^{M} p_i = 1$. O modelo GMM pode ser representado pelo conjunto de parâmetros

$$\lambda = \{p_i, \vec{\mu}_i, K_i\} \quad i = 1, \ldots, M. \tag{2.14}$$

Durante a fase de treinamento, os modelos para classificação dos locutores (λ) são gerados a partir da matriz de atributos X através do algoritmo EM (*expectation-maximization*). O objetivo é obter o modelo λ que maximize a verossimilhança entre seus parâmetros e a matriz de atributos, ou seja,

$$\log p(X|\lambda) = \frac{1}{Q}\sum_{t=1}^{Q} \log p(\vec{x}_t|\lambda). \tag{2.15}$$

Na identificação, dada uma matriz de atributos X, o locutor $\hat{L}$ que será escolhido pelo decisor é aquele cujo modelo $\lambda_{\hat{L}}$ maximiza a probabilidade *a posteriori*. Logo, dado um conjunto de L locutores representados pelos modelos λ_1, λ_2, ..., λ_L, de acordo com a regra de Bayes:

$$\hat{L} = \arg\max_{1\leq k\leq L} P(\lambda_k|X) = \arg\max_{1\leq k\leq S} \frac{p(X|\lambda_k)\cdot P(\lambda_k)}{p(X)}. \tag{2.16}$$

Assumindo que a escolha de cada locutor é equiprovável ($P(\lambda_k) = \frac{1}{L}$), consegue-se simplificar a regra de decisão para:

$$\hat{L} = \arg\max_{1\leq k\leq L} p(X|\lambda_k). \tag{2.17}$$

Considerando as probabilidades do vetor de atributos de cada quadro independentes, conclui-se que:

$$\hat{L} = \arg\max_{1\leq k\leq L} \sum_{t=1}^{T} \log p(\vec{x}_t|\lambda_k), \tag{2.18}$$

onde $p(\vec{x}_t|\lambda_k)$ é dado pela Equação 2.12.

2.3 Resultados de identificação de locutor

Para analisar o impacto da reverberação na identificação de locutor, foram realizados experimentos em dois cenários diferentes. No primeiro, o treinamento e teste

foram realizados em uma mesma sala, com variação de RT_{60}. No segundo, foram utilizadas salas diferentes. Um subconjunto de 168 locutores da base de voz TIMIT (GAROFOLO *et al.*, 1993) amostrados a 16 kHz foi selecionado para condução dos testes de identificação. Cada locutor possui 10 sinais com duração média de 3 s cada. Para os modelos foram utilizados 24 s (8 locuções) e 6 s para os testes. De cada locutor, 2 locuções foram utilizadas para teste, totalizando 336. Cada 6 s de teste dos 168 locutores contou com 188 quadros resultando em 31.584 testes por quadro para cada RIR. Ou seja, 94.752 testes para os experimentos com a base MARDY e 189.504 para os experimentos com a base AIR.

Para a extração dos vetores com MFCC, foram selecionados quadros de 32 ms, com sobreposição de 50% para a extração de 12 coeficientes. Os modelos GMM foram compostos com 32 gaussianas. As reverberações foram obtidas das bases MARDY (WEN *et al.*, 2006) e AIR (JEUB *et al.*, 2009).

O desempenho dos sistemas foram avaliados pela taxa de acertos da identificação (t_a), definida pela razão

$$t_a = \frac{\text{número de identificações corretas}}{\text{número total de testes realizados}} \times 100\%. \tag{2.19}$$

2.3.1 Resultados de Cenário 1: reverberações de uma mesma sala

As RIR selecionadas para este Cenário foram obtidas a partir de uma única sala, variando-se a distância entre fonte e microfone (d_{FM}) e a posição de painéis absortivos em suas superfícies. Na Tabela 2.1 encontram-se os parâmetros RT_{60}, d_{FM} e DRR das RIR selecionadas da base MARDY. Para analisar os resultados do efeito da reverberação da identificação de locutor, foram consideradas 3 situações:

- Treinamento e testes com sinal de voz limpo,
- Treinamento e testes com mesmo RT_{60} e
- Treinamento e testes com RT_{60} diferentes.

A Tabela 2.5 apresenta os resultados dos experimentos de identificação que utilizam atributos MFCC e o classificador GMM. Nela se encontram os resultados obtidos em diferentes condições de treinamento e teste, incluindo com sinal sem reverberação (limpo). Para treinamento e teste sem reverberação, a taxa de acerto encontrada foi de 99,70 p.p. Os experimentos indicam que os resultados com mesmo RT_{60} foram melhores que os realizados em condição de descasamento de reverberação. Com descasamento, os resultados chegaram a cair até o valor de 27,68 p.p. (pontos percentuais) para treinamento de $RT_{60} = 0,60$ s e teste de $RT_{60} = 0,55$ s. Enquanto isso, os experimentos onde não há o

Tabela 2.5 – Resultados de identificação (%) com a base MARDY utilizando atributos MFCC e classificador GMM.

Treinamento - RT_{60} (s)	Teste - RT_{60} (s)			
	Limpo	0,55	0,60	0,65
Limpo	**99,70**	37,20	74,40	55,36
0,55	56,85	**97,62**	41,07	51,49
0,60	71,13	30,65	**93,45**	70,24
0,65	51,79	41,96	66,07	**90,48**

descasamento de reverberação conseguiram até 97,40 p.p. para treinamento e teste com $RT_{60} = 0,55$ s.

2.3.2 Cenário 1: Discussão

Os resultados mostraram que nas situações onde houve o descasamento, as taxas de acerto foram mais impactadas. Além disso, foi mostrado que nos experimentos de identificação com reverberação e descasamento, os testes com sinais menos não-estacionários obtiveram melhores resultados. Isto pode ser observado para as três condições de treinamento. Nas situações com casamento entre treinamento e teste, apesar de menos impactados pela reverberação, o efeito fez com que as taxas de acerto fossem mais degradadas para experimentos com sinais de voz reverberados menos não-estacionários. As melhores taxas de acerto de identificação de locutor em situações onde há o casamento de reverberações entre treinamento e teste condizem com o fato de que um sistema auditivo anteriormente exposto a uma reverberação é menos impactado pelo efeito, como mostrado em (TRAER; MCDERMOTT, 2016) (BRANDEWIE; ZAHORIK, 2010) (WATKINS, 2005).

2.3.3 Resultados de Cenário 2: treinamento e teste em salas diferentes

A base AIR foi escolhida para os experimentos para analisar o impacto da reverberação da identificação de locutor em situações de descasamento de salas entre trei-

Tabela 2.6 – Salas selecionadas da base AIR e parâmetros de RT_{60}, d_{FM} e DRR das RIR

Sala	RT_{60} (s)	d_{FM} (m)	DRR (dB)	Dimensões (m)
Cabine 1	0,24	1,00	7,44	$3,0 \times 1,8 \times 2,2$
Cabine 2	0,37	1,50	5,27	
Escritório 1	0,61	2,00	2,25	$5,0 \times 6,4 \times 2,9$
Escritório 2	0,64	3,00	-0,04	
Sala de aula 1	0,77	7,10	-1,34	$10,8 \times 6,4 \times 2,9$
Sala de aula 2	0,87	5,00	-4,52	

Tabela 2.7 – Resultados de identificação (%) de experimentos com a base AIR em situação de descasamento de salas.

Treinamento	Teste	Taxa de acerto	Média
Cabine 1	Escritório 1	24,40	**21,13**
	Escritório 2	18,45	
	Sala de aula 1	23,21	
	Sala de aula 2	18,45	
Cabine 2	Escritório 1	56,55	51,13
	Escritório 2	61,31	
	Sala de aula 1	43,15	
	Sala de aula 2	44,05	
Escritório 1	Cabine 1	29,76	56,03
	Cabine 2	52,38	
	Sala de aula 1	69,64	
	Sala de aula 2	72,32	
Escritório 2	Cabine 1	25,60	55,36
	Cabine 2	54,17	
	Sala de aula 1	68,15	
	Sala de aula 2	73,51	
Sala de aula 1	Cabine 1	30,95	47,99
	Cabine 2	38,10	
	Escritório 1	64,29	
	Escritório 2	58,63	
Sala de aula 2	Cabine 1	23,21	48,96
	Cabine 2	34,23	
	Escritório 1	70,54	
	Escritório 2	67,86	

namento e teste. Desta base, foram escolhidas 3 salas diferentes, de quais foram utilizadas 2 RIR, com valores de (d_{FM}) diferentes. A Tabela 2.6 contém informações complementares sobre o cenário dos experimentos com a base AIR que incluem os parâmetros das suas RIR selecionadas.

Com a base AIR, foram realizados experimentos de identificação de locutor com descasamento de salas entre treinamento e teste . A Tabela 2.7 apresenta o resultado dos experimentos. Os resultados mostraram que o descasamento de salas impactou mais as taxas de acerto que o descasamento de RIR em uma mesma sala. Os experimentos mais impactados pelo descasamento foram os realizados com treinamento em Cabine 1. Estes tiveram média de taxa de acerto de 21,13 p.p., com os piores resultados para testes realizados em Escritório 2 e Sala de aula 2, com taxas de 18,45 p.p. Os experimentos com treinamento nas salas Escritório 1 e 2 e em Sala de aula 1 e 2 foram mais impactados com os testes em Cabine 1 e 2. Além disso, os resultados com treinamento em Escritório 1 conseguiram taxas de acerto bem próximas dos resultados com treinamento em Escritório 2 e o mesmo ocorreu para o treinamento em Sala de aula 1 e Sala de aula 2.

2.3.4 Cenário 2: Discussão

Os resultados para experimentos com descasamento de salas entre treinamento e teste mostraram que os resultados pouco variaram quando se compara os treinamentos com Escritório 1 e Escritório 2, ou entre Sala de aula 1 e Sala de aula 2. O mesmo não ocorre para Cabine 1 e 2, onde o primeiro apresenta taxa de acerto média de 21,13 p.p., enquanto o segundo apresenta 51,13 p.p. Isto pode ser justificado pela maior diferença relativa nos valores de RT_{60} entre as RIR da sala Cabine, enquanto das outras salas foram selecionadas RIR com RT_{60} próximos. Além disso, os resultados com teste em Cabine 1 e 2 foram mais impactados nos experimentos com descasamento de salas. Isto se explica devido às RIR de Sala de aula e de Escritório terem uma não-estacionariedade mais próxima. O experimentos de não-estacionariedade para Cabine 1 e 2 encontraram, respectivamente, picos de INS de 256 e 242, sendo considerados altamente não-estacionários.

2.4 Resumo

Este Capítulo apresenta o efeito da reverberação em um sinal de voz, que foi analisado a partir de espectrogramas e cocleogramas, além da medida de não-estacionariedade INS. Em seguida, foi introduzido o funcionamento de um sistema de identificação de locutor, além do atributo MFCC e do classificador GMM utilizados nos experimentos de identificação. Por fim, foram mostrados os resultados dos experimentos com casamento e descasamento de salas entre treinamento e teste. Os experimentos mostraram que o reconhecimento de locutor, quando realizado com a mesma RIR no treinamento e no teste, é menos impactado pela reverberação que quando há o descasamento. Testes com descasamento em uma mesma sala mostraram que os melhores resultados foram obtidos quando os testes foram realizados com um sinal reverberado menos não-estacionário. Experimentos de identificação de locutor onde há o descasamento de salas entre treinamento e teste mostraram que quando treinamento e teste são realizados com sinais de voz reverberados Não-estacionários ou Moderadamente não-estacionários, o impacto é menor do que quando envolvem sinais de voz reverberados Altamente não-estacionários.

3 Redução do Efeito da Reverberação Baseada em Máscaras Acústicas

A captação do sinal de voz, em diversos tipos de ambientes, está sujeita a interferências como ruídos acústicos ou reflexões do próprio sinal, como a reverberação. Estes efeitos provocam severa degradação da qualidade e da inteligibilidade dos sinais. Consequentemente, afetam a acurácia de diversas aplicações como identificação de locutor, reconhecimento de voz, localização de fontes acústicas, entre outras.

Diferentes soluções foram propostas na literatura para aprimorar a qualidade dos sinais de voz em presença de distorção por ruídos acústicos. As principais técnicas são de realce de sinais (ZÃO *et al.*, 2014) (TAVARES; COELHO, 2016), separação cega (LEE, 1998) e de filtragem espacial (VEEN; BUCKLEY, 1988) (KRIM; VIBERG, 1996). No entanto, estas soluções não proveem um aprimoramento da inteligibilidade sonora (LOIZOU, 2013) (TAAL *et al.*, 2011). O tratamento de inteligibilidade é fundamental para as aplicações e sistemas que envolvam o reconhecimento de padrão e classificação de sinais. Os principais métodos propostos na literatura para aprimoramento da inteligibilidade são baseados em máscaras acústicas (LOIZOU, 2013).

Este Capítulo apresenta um estudo com medidas objetivas de inteligibilidade que avalia o impacto causado pela reverberação e o desempenho de máscaras acústicas ideais (WANG, 2005) (KOKKINAKIS *et al.*, 2011) e não-ideais (HAZRATI *et al.*, 2012) nestas condições. Para a avaliação objetiva da inteligibilidade , são adotadas três medidas: CSII (*Coherence and Speech Intelligibility Index* (KATES; AREHART, 2005)), STOI (*Short-Time Objective Intelligibility* (TAAL *et al.*, 2011)) e SRMR (*Speech to Reverberation Modulation Energy ratio* (FALK *et al.*, 2010)). Estas medidas foram aplicadas na literatura com sucesso para investigar situações de distorção por ruídos (TAAL *et al.*, 2011) (TAVARES; COELHO, 2016) (ZÃO *et al.*, 2014). Os resultados indicam que a reverberação degradou a inteligibilidade da voz. Em uma mesma sala, esta degradação ocorreu em maior magnitude com o aumento da d_{fm} (distância fonte-microfone) e de RT_{60} (*Reverberation Time*).

3.1 Máscaras Acústicas

As máscaras acústicas foram propostas para simular a capacidade perceptual humana de separar uma fonte de suas interferências como no problema do "*Cocktail party*" (BRONKHORST, 2000). Nesta Seção, é descrito o processo de obtenção de uma máscara

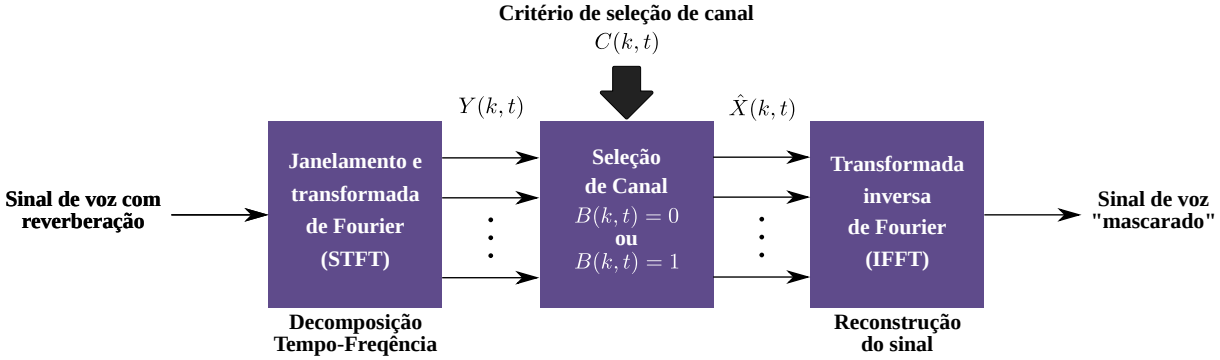

Figura 3.1 – Diagrama genérico de uma máscara acústica (LOIZOU, 2013).

acústica ideal (que utiliza informações do sinal limpo) e suas variações (utilizando FFT ou filtros *Gammatone*). Além disso, é descrita uma máscara cega (não-ideal), ou seja, que não utiliza informações do sinal sem reverberação. O objetivo principal do emprego das máscaras acústicas é a redução dos efeitos da reverberação no sinal de voz, e consequentemente, o aprimoramento da qualidade e inteligibilidade do sinal.

3.1.1 Máscaras acústicas ideais

As máscaras acústicas ideais geralmente estão definidas pelos seguintes passos (LOIZOU, 2013) (WANG, 2005), ilustrados na Figura 3.1:

1. *Decomposição tempo-frequência (TF)*: O sinal reverberado é janelado e, em seguida, é aplicada a transformada de Fourier em cada um dos quadros. O sinal $Y(k,t)$ representa o espectro da sub-banda k do t-ésimo quadro do sinal reverberado. Além da STFT (*Short-Time Fourier Transform*) como método de separação em frequências, também podem ser utilizados bancos de filtros como, por exemplo, *Gammatone* (JOHANNESMA, 1972) (PATTERSON; MOORE, 1986) (COOKE, 1993).

2. *Critério de seleção*: Define-se um critério $C(k,t)$ que determinará se o quadro $Y(k,t)$ será considerado dominante pelo sinal de voz ou pela reverberação. No caso da máscara ideal, além da representação tempo-frequência do sinal reverberado, também é necessário o conhecimento do sinal sem reverberação para a obtenção de $C(k,t)$.

3. *Mascaramento*: Os quadros que comporão o sinal "mascarado" $\hat{X}(k,t)$ são definidos por:

$$\hat{X}(k,t) = \begin{cases} Y(k,t), & \text{se} \quad C(k,t) \geq \gamma, \\ 0, & \text{caso contrário,} \end{cases} \tag{3.1}$$

onde γ é o limiar de seleção.

A Equação 3.1 também pode ser descrita da seguinte maneira:

$$\hat{X}(k,t) = B(k,t) \cdot Y(k,t) \tag{3.2}$$

onde $B(k,t)$ é uma função binária denominada máscara binária, dada por:

$$B(k,t) = \begin{cases} 1 & \text{se} \quad C(k,t) \geq \gamma, \\ 0, & \text{caso contrário.} \end{cases} \tag{3.3}$$

4. *Reconstrução do sinal*: A transformada inversa de Fourier é aplicada em $\hat{X}(k,t)$ para reconstruir os quadros no domínio do tempo. Em seguida, os quadros reconstruídos são concatenados para obter o sinal mascarado, mantendo as sobreposições utilizadas inicialmente. Se o método de decomposição das sub-bandas escolhido no primeiro passo for de banco de filtros, as saídas dos filtros após a máscara são somadas para a obtenção do sinal mascarado.

Neste trabalho, duas máscaras ideais que utilizam diferentes métodos de decomposição em frequência foram utilizadas. As máscaras estão descritas abaixo:

- IBM - *Ideal Binary Mask*

  Na máscara IBM (LI; LOIZOU, 2007) é empregada a STFT como forma de decomposição em frequência dos quadros do sinal. O janelamento é realizado com quadros de Hanning de 20 ms de duração com 50% de sobreposição. O critério de seleção adotado é a razão sinal-reverberação $\text{SRR}(k,t) \geq -5$ dB.

- IRM - *Ideal Reverberant Mask*

  Os filtros *Gammatone* (JOHANNESMA, 1972) (PATTERSON; MOORE, 1986) (COOKE, 1993) foram propostos para descrever o comportamento da função de resposta ao impulso do sistema auditivo humano no domínio do tempo. Sendo assim, na literatura este banco de filtros é geralmente aplicado para modelar ou simular o sistema auditivo atingindo a inteligibilidade sonora. Na máscara IRM é utilizado um banco de 128 filtros *Gammatone* de quarta ordem para realizar a decomposição tempo-frequência. As frequências centrais são espaçadas entre si de acordo com a escala ERB (*Equivalent rectangular bandwidth*) distribuída entre 50 Hz e 8 kHz. Em seguida, os sinais de cada sub-banda são divididos em quadros de 20 ms[1] com 50% de sobreposição. Este processo é realizado com o sinal reverberado e com o sinal direto para a obtenção da SRR de cada quadro $Y(k,t)$. O critério de seleção utilizado é $\text{SRR}(k,t) \geq -5$ dB para a escolha dos quadros dominados pela voz. Em (KOKKINAKIS *et al.*, 2011), foi mostrado que o valor de -5 dB como limiar de

[1] A análise do sinal de voz é aplicada em tempo curto (QUATIERI, 2002), ou seja, em quadros de cerca de 20-30 ms de duração. Durante este intervalo, o sinal de voz é considerado estável (processo estacionário). Isso permite que vetores de atributos possam ser extraídos ou estimados de cada quadro de voz.

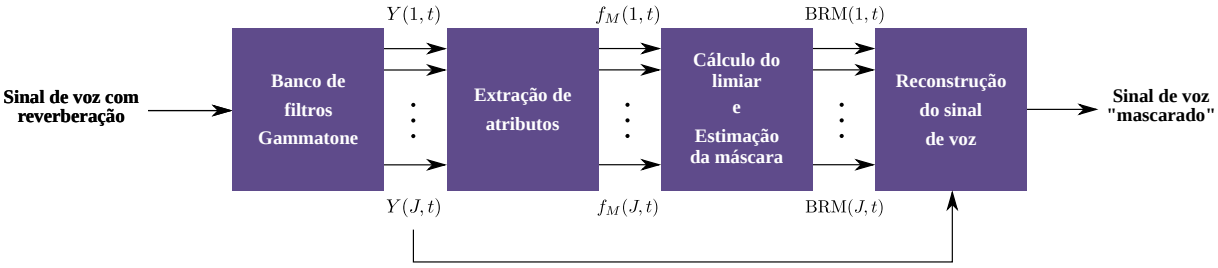

Figura 3.2 – Diagrama da máscara acústica BRM (HAZRATI *et al.*, 2012)

SRR obteve o maior incremento na taxa de reconhecimento de palavras em testes perceptuais de inteligibilidade em que foram testados limiares de -15 a 5 dB.

Para reconstruir o sinal, as 128 sub-bandas são obtidas a partir da concatenação das janelas após a máscara $\hat{X}(k,t)$ e invertidas no tempo. Em seguida, é aplicado um filtro *Gammatone* em cada sub-banda correspondente à sua frequência central e estas são invertidas no tempo novamente. Ao final, o sinal de voz com redução do efeito de reverberação, ou "mascarado", é obtido da soma das sub-bandas resultantes.

3.1.2 Máscara acústica não-ideal ou Cega: BRM - *Binary Reverberant Mask*

A principal limitação das máscaras acústicas ideais deve-se à necessidade de informações do sinal de voz limpo (sem reverberação) para o cálculo de SRR(k,t). A máscara cega não-ideal BRM (HAZRATI *et al.*, 2012) foi proposta para evitar a exigência das informações do sinal sem reverberação. Para isto, é necessário um critério de seleção diferente da SRR.

O diagrama da máscara BRM(k,t) está exemplificado na Figura 3.2 e nas etapas a seguir.

1. *Decomposição tempo-frequência (TF) e cálculo dos coeficientes*: Na BRM a representação tempo-frequência aplica um banco de 64 filtros *Gammatone* de quarta ordem espaçados logaritmicamente entre 50 Hz e 8 kHz. Em seguida, para cada quadro tempo-frequência $r(k,t)$ é calculado um coeficiente pela razão da variância do sinal elevada a uma dada potência sobre a sua variância absoluta, de acordo,

$$f_M(k,t) = 10\log_{10}\left(\frac{\sigma^2_{r'}(k,t)}{\sigma^2_{|r|}(k,t)}\right), \tag{3.4}$$

onde $r'(t,j) = |r(k,t)|^\alpha$, $|r(t,j)|$ é o valor absoluto do quadro no tempo t e sub-banda j e $\sigma^2_{r'}(k,t)$ e $\sigma^2_{|r|}(k,t)$ são suas respectivas variâncias dos quadros correspondentes. O objetivo deste atributo é identificar picos e vales (HAZRATI *et al.*, 2012) em cada sub-banda. Posteriormente, os valores de f_M são suavizados no tempo através de um filtro mediana de ordem 3.

2. *Critério de seleção*: O critério de seleção da máscara é baseado no histograma $f_{hist}(k,t)$, calculado a partir dos valores de f_M dos Q_p quadros anteriores a t, até os seus Q_f quadros seguintes. Cada histograma $f_{hist}(k,t)$ normalizado possui L classes com pesos p_i $(i = 1, ..., L)$. A partir destes valores, são calculadas a média global m_G, a média cumulativa $m(l)$ e a soma cumulativa $P_s(l)$, definidos por:

$$m_G = \sum_{i=1}^{L} i.p_i \qquad m(l) = \sum_{i=1}^{l} i.p_i \qquad P_s(l) = \sum_{i=1}^{l} p_i. \tag{3.5}$$

   O limiar ótimo l^* é definido como o valor de l que maximiza a variância entre classes $\sigma_B^2(l)$, dada por:

$$\sigma_B^2(l) = \frac{(m_G P_s(l) - m(l))^2}{P_s(l)(1 - P_s(l))}. \tag{3.6}$$

3. *Mascaramento*: O valor l^* é empregado como critério de seleção para definir se o conteúdo do quadro $r(k,t)$ é predominante pela voz e será mantido após o mascaramento. Logo, a máscara é expressa por:

$$\mathrm{BRM}(k,t) = \begin{cases} 1 & \text{se} \quad f_M(k,t) > \max(l^*(k,t), l_0), \\ 0, & \text{caso contrário,} \end{cases} \tag{3.7}$$

   onde l_0 é o limiar de silêncio. A máscara é aplicada em $Y(k,t)$ segundo

$$\hat{X}(k,t) = Y(k,t)\,\mathrm{BRM}(k,t). \tag{3.8}$$

4. *Reconstrução do sinal*: Finalmente, a reconstrução do sinal mascarado é realizada em cada sub-banda. Os quadros são concatenados de acordo com as suas sobreposições iniciais e invertidas no tempo. Um filtro *Gammatone* é aplicado em cada sub-banda. Em seguida, o sinal é invertido no tempo novamente. Por fim, os sinais são somados para a obtenção do sinal reconstruído ou "mascarado".

3.2 Medidas de inteligibilidade Acústica

A inteligibilidade acústica de um sinal de voz reflete o quão compreensível é a mensagem transmitida para o sistema auditivo humano. Estudos em (BOLT; MACDONALD, 1949) indicaram uma relação entre a inteligibilidade da voz reverberada e o valor de RT_{60} da RIR. Em (NÁBĚLEK; ROBINSON, 1982), diversos testes perceptuais monoaurais e binaurais mostraram que a inteligibilidade na presença de reverberação é afetada pelo RT_{60} da sala e pela idade do ouvinte. Testes perceptuais são a forma mais precisa de se avaliar estes fenômenos, porém são frequentemente substituídos por medidas objetivas devido aos menores custos e maior rapidez na execução dos experimentos

(QUACKENBUSH *et al.*, 1988) (RIX *et al.*, 2001), (HU; LOIZOU, 2008) (BISPO *et al.*, 2016).

Neste trabalho, foram adotadas para o estudo do fenômeno três medidas objetivas de inteligibilidade: CSII (KATES; AREHART, 2005), STOI (TAAL *et al.*, 2011) e SRMR (FALK *et al.*, 2010). Estas medidas permitem avaliar o impacto do efeito da reverberação causado nos sinais de voz e a eficiência das máscaras acústicas em recuperar a sua inteligibilidade.

3.2.1 CSII

A CSII (*coherence and speech intelligibility index*) (KATES; AREHART, 2005) é um aprimoramento da medida SII (*speech intelligibility index*). Esta medida utiliza a coerência quadrática (MSC - *magnitude-squared coherence*) para calcular a SRR usada na computação dos índices, que variam entre zero e um. A técnica tem como diferencial o fato de levar em consideração as distorções causadas por "*center-clipping*" (redução da amplitude próximas ao nível de referência em regiões com atividade de voz) e "*peak-clipping*" (redução da amplitude por saturação). Primeiramente, o sinal de referência sem reverberação $x(t)$ e o sinal resultante do uso das máscaras $y(t)$ são janelados com tamanho de quadro de 16 ms com 50% de sobreposição. A partir da aplicação de uma DFT (*Discrete Fourier Transform*), são obtidos os respectivos espectros $X_j(f)$ e $Y_j(f)$, com $f = 0, ..., F$, referentes ao quadro j. A medida MSC é dada por,

$$\text{MSC}(f) = \frac{|\sum_{j=0}^{Q-1} X_j(f) Y_j^*(f)|^2}{(\sum_{j=0}^{Q-1} |X_j(f)|^2)(\sum_{j=0}^{Q-1} |Y_j(f)|^2)}, \tag{3.9}$$

onde Q é o número total de quadros. A MSC representa o quanto da potência do sinal de saída é linearmente dependente da entrada (KATES; AREHART, 2005). O complemento $1 - \text{MSC}(f)$ representa a presença de distorção e ruído. Em seguida, a SRR é calculada por,

$$\text{SRR}(j) = \frac{\sum_{f=0}^{F} I_j(f) MSC(f) S_y(f)}{\sum_{f=0}^{F} I_j(f)[1 - MSC(f)] S_y(f)}, \tag{3.10}$$

onde $S_y(f)$ é a amostra f da densidade espectral de potência de $y(t)$ e $I_b(f)$ é um filtro que atribui um peso à frequência f relativo à inteligibilidade.

Em (KATES; AREHART, 2005), foi mostrado que uma alta correlação com testes subjetivos de acertos de palavras é alcançada quando se realiza a obtenção do índice CSII em três níveis de amplitude diferentes. Assim, o cálculo de CSII neste trabalho foi realizado com a divisão do sinal de voz em três regiões diferentes de amplitude. O $\text{CSII}_{\text{alto}}$ é obtido a partir das regiões com amplitude acima do valor RMS (*root mean square*). O $\text{CSII}_{\text{médio}}$ é calculado com as regiões entre 0 e 10 dB abaixo de RMS. A partir dos quadros

entre 10 e 30 dB abaixo do nível RMS, é obtido $\text{CSII}_{\text{baixo}}$. O janelamento é feito com quadros de Hamming com 50% de sobreposição e duração de 16 ms.

Neste trabalho, o resultado da composição dos três índices foi adaptado para a base de voz TIMIT (GAROFOLO *et al.*, 1993) e é dado por:

$$c = -3,47 + 1,84\text{CSII}_{\text{baixo}} + 9,99\text{CSII}_{\text{médio}} + 0,00\text{CSII}_{\text{alto}}. \tag{3.11}$$

Uma função logística de mapeamento em entre o índice CSII e testes perceptuais de inteligibilidade foi utilizada neste trabalho. Ela é descrita por

$$f(c) = \frac{100\%}{1 + \exp(ac + b)}, \tag{3.12}$$

onde a e b são constantes. Neste trabalho, foram utilizados os valores $a = -6,13$ e $b = -0,75$, calculados a partir dos resultados de testes perceptuais em ambientes com reverberação de (NÁBĚLEK; ROBINSON, 1982).

3.2.2 STOI

A medida STOI (*short-time objective intelligibility*) (TAAL *et al.*, 2011) estima a degradação da inteligibilidade de algoritmos de redução de ruídos. Diferentemente da medida clássica AI (*Articulation Index*) (KRYTER, 1962) e suas derivadas (STEENEKEN; HOUTGAST, 1980) (RHEBERGENT; VERSFELD, 2005) (LOIZOU; HU, 2011), esta medida não se baseia no cálculo da SRR para avaliar a inteligibilidade dos sinais de voz. Na STOI, o coeficiente de correlação entre os espectros dos sinais limpo e realçado é utilizado para avaliar a degradação da inteligibilidade de algoritmos de redução de ruídos. A medida apresenta forte relação monotônica com testes subjetivos de inteligibilidade em que a voz corrompida por ruídos é realçada a partir de ponderações tempo-frequência [2].

Primeiramente, o sinal de voz limpo $x(t)$ é reamostrado a 10 kHz e dividido em janelas de Hamming de 256 amostras com 50% de sobreposição. Esta taxa de amostragem é a mesma utilizada em (TAAL *et al.*, 2011). Em seguida, aplica-se uma DFT de 512 pontos em cada quadro, formando a matriz X, onde $X(\kappa, \tau)$ representa o κ-ésimo ponto da DFT do quadro τ. Os pontos $X(\kappa, \tau)$ são então agrupados em 15 sub-bandas de frequência cujos centros variam entre 150 Hz e 4300 Hz. A norma para cada sub-banda é definida por

$$\bar{X_j}(\tau) = \sqrt{\sum_{\kappa=\kappa_l(j)}^{\kappa_u(j)-1} |X(\kappa, \tau)|}, \tag{3.13}$$

[2] Uma medida de inteligibilidade é capaz de realizar predições de testes subjetivos de inteligibilidade (e.g., taxa de reconhecimento de palavras). Entretanto, uma relação monotônica entre a medida e a inteligibilidade da voz é suficiente para uma análise de métodos de realce (TAAL *et al.*, 2011).

onde $\kappa_l(j)$ e $\kappa_u(j)$ são, respectivamente, os limites inferior e superior da sub-banda j ($j = 1, 2, ..., 15$). Com os valores das normas, define-se a envoltória temporal de cada sub-banda pelo seguinte vetor:

$$\vec{x}_{(j,\tau)} = [\bar{X}_j(\tau - 29), \bar{X}_j(\tau - 28), ..., \bar{X}_j(\tau)]^T. \tag{3.14}$$

A partir do mesmo processo com o sinal de voz corrompido $y(t)$ obtém-se $\vec{y}_{(j,\tau)}$. Este é normalizado segundo,

$$\bar{\vec{y}}_{(j,\tau)} = \min\left(\frac{\|\vec{x}_{(j,\tau)}\|}{\|\vec{y}_{(j,\tau)}\|}\vec{y}_{(j,\tau)}, (1 + 10^{-\frac{\beta}{20}})\vec{x}_{(j,\tau)}(n)\right), \tag{3.15}$$

onde $||\cdot||$ representa a norma ℓ^2 e $\beta = -15$ dB indica o limite inferior de SDR (*Signal-to-Distortion Ratio*). O valor de $\text{STOI}_{(j,\tau)}$ é dado por:

$$\text{STOI}_{(j,\tau)} = \frac{(\vec{x}_{(j,\tau)} - \mu_{\vec{x}_{(j,\tau)}})^T(\bar{\vec{y}}_{(j,\tau)} - \mu_{\bar{\vec{y}}_{(j,\tau)}})}{\|\vec{x}_{(j,\tau)} - \mu_{\vec{x}_{(j,\tau)}}\|\|(\bar{\vec{y}}_{(j,\tau)} - \mu_{\bar{\vec{y}}_{(j,\tau)}}\|}, \tag{3.16}$$

sendo μ a média do vetor correspondente. Por fim, a medida STOI é calculada a partir da média de todos os valores de $\text{STOI}_{(j,\tau)}$, dados por:

$$\text{STOI} = \frac{1}{15Q}\sum_{j=1}^{15}\sum_{\tau=1}^{Q}\text{STOI}_{(j,\tau)}, \tag{3.17}$$

onde Q é o número total de quadros.

Além da proposta de medida objetiva, os autores também propuseram uma função de mapeamento do índice STOI para predição da taxa de acerto de palavras em testes subjetivos. Esta função de mapeamento é dada por

$$f(\text{STOI}) = \frac{100\%}{1 + \exp(a\text{STOI} + b)}, \tag{3.18}$$

onde a e b são constantes. Para a adaptação da medida à experimentos com sinais de voz reverberados, foram utilizados os resultados de testes perceptuais obtidos em (NÁBĚLEK; ROBINSON, 1982) para se definirem os valores $a = -12,47$ e $b = 5,90$.

Além da função de mapeamento, a inteligibilidade também foi avaliada segundo a variação percentual do índice através do ΔSTOI (DELFARAH; WANG, 2017), de forma que

$$\Delta\text{STOI}(\%) = 100\% \times (\text{STOI}_{\text{voz mascarada}} - \text{STOI}_{\text{voz reverberada}}). \tag{3.19}$$

3.2.3 SRMR

A SRMR (*Speech to reverberation modulation energy ratio*) (FALK *et al.*, 2010) é uma medida objetiva não-intrusiva de qualidade e inteligibilidade da voz voltada para

situações de reverberação e voz desreverberada. A característica não-intrusiva diz respeito ao fato de a medida não ter a limitação de precisar do sinal limpo para ser calculada. A SRMR se baseia no fato de que a voz reverberada concentra a energia do sinal modulado nas frequências mais altas que a voz sem reverberação.

O cálculo desta medida é realizado a partir da divisão em sub-bandas do sinal de voz $x(n)$ através de um banco de 23 filtros *Gammatone* com frequências centrais distribuídas de 125 Hz até a metade da taxa de amostragem de acordo com a escala de Bandas Retangulares Equivalentes (GLASBERG; MOORE, 1990). A partir do sinal de cada sub-banda $x_j(n)$, é utilizada a transformada de Hilbert $\mathcal{H}(\cdot)$ para se calcular a envoltória temporal $e_j(n)$ de forma que

$$e_j(n) = \sqrt{x_j(n)^2 + \mathcal{H}\{x_j(n)\}^2} \quad j = \{1, \ldots, 23\}. \tag{3.20}$$

Em seguida, cada envoltória temporal $e_j(n)$ é janelada com quadros de Hamming de 256 ms com 32 ms de sobreposição, representados por $e_j(m, n)$, onde m é o índice do quadro no domínio do tempo. Utiliza-se então uma transformada discreta de Fourier para o cálculo da potência espectral do sinal modulado de cada quadro $e_j(m, n)$, logo

$$E_j(m, f) = |\mathcal{F}(e_j(m, n))|^2, \tag{3.21}$$

onde f representa o índice do *bin* de frequência. Estas frequências são então agrupadas em oito bandas de forma a emular um banco de filtros modulados inspirado no sistema auditivo humano (DAU *et al.*, 1996). A partir da média temporal da potência espectral $E_j(m, f)$, é definida $\bar{\varepsilon}_{j,k}$ a potência média do sinal modulado da j-ésima sub-banda e k-ésimo filtro modulado, onde $j = 1, \ldots, 23$ e $k = 1, \ldots, 8$. Quando se calcula o valor médio de $\bar{\varepsilon}_{j,k}$ para cada sub-banda j, é obtida a potência média por frequência modulada $\bar{\varepsilon}_k$, dada por

$$\bar{\varepsilon}_k = \frac{1}{23} \sum_{j=1}^{23} \bar{\varepsilon}_{j,k}. \tag{3.22}$$

Por fim, considera-se que a qualidade e a inteligibilidade do sinal reverberado e desreverberado está relacionada à energia nas frequências moduladas a partir do quinto filtro. O índice SRMR é então definido pela razão

$$\text{SRMR} = \frac{\sum_{k=1}^{4} \bar{\varepsilon}_k}{\sum_{k=5}^{K^*} \bar{\varepsilon}_k} \tag{3.23}$$

onde K^* é a banda modulada até onde se encontra 90% da energia.

3.3 Resultados dos experimentos de inteligibilidade

No Capítulo 2, foi apresentado o efeito da reverberação e seu impacto nos resultados de identificação de locutor. Neste Capítulo, o objetivo é mostrar como as máscaras

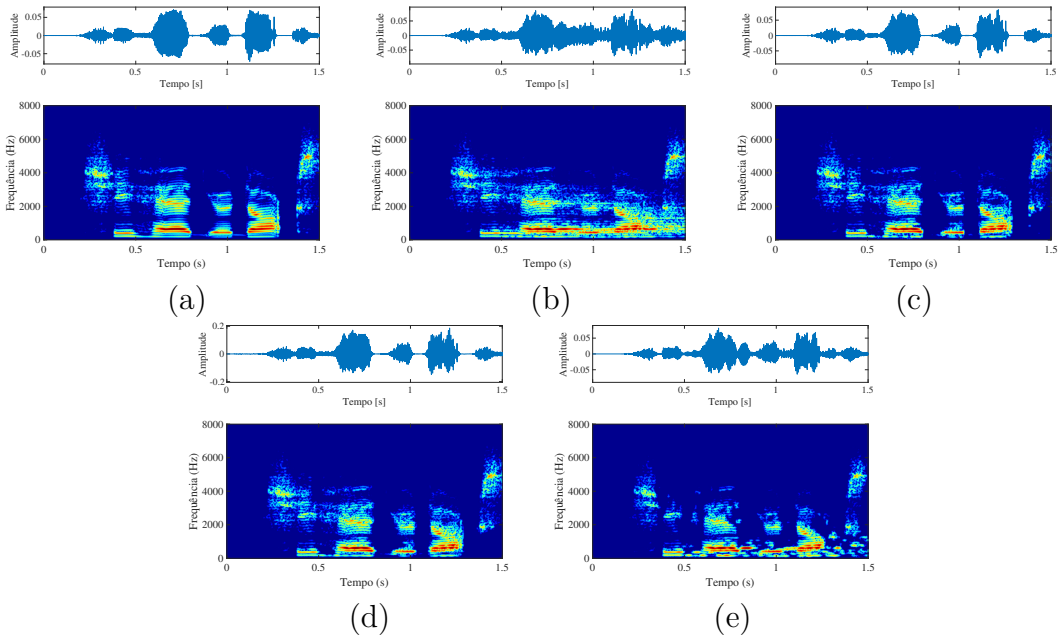

Figura 3.3 – Sinais de voz e seus respectivos espectrogramas (a) sinal direto, (b) reverberado, (c) IBM, (d) IRM e (e) BRM.

acústicas podem melhorar a inteligibilidade do sinal de voz na presença de reverberação. Nos experimentos conduzidos e apresentados neste Capítulo, sinais de voz foram reverberados e receberam a aplicação das máscaras acústicas ideais e a não-ideal apresentadas. Em seguida, foram empregadas as medidas objetivas de inteligibilidade (CSII, STOI e SRMR) para análise do impacto da reverberação nos sinais de voz e da capacidade das máscaras acústicas de recuperar a sua inteligibilidade. Além disso, foram mostrados os espectrogramas (O'SHAUGHNESSY, 2000) do sinal de voz com reverberação e após a aplicação das máscaras acústicas.

3.3.1 Cenário dos experimentos

Para analisar o efeito da reverberação no sinal de voz, foram realizados testes com 24 locutores selecionados da base de voz TIMIT (GAROFOLO *et al.*, 1993), sendo 8 mulheres e 16 homens. De cada locutor, foram utilizadas 10 gravações com duração média de 3 s e amostradas a 16 kHz, totalizando 240 sinais de voz. Estes sinais foram reverberados a partir de RIR selecionadas das bases AIR (JEUB *et al.*, 2009), de onde foram escolhidas 6 RIR de 3 salas distintas, e MARDY (WEN *et al.*, 2006), de onde foram escolhidas 3 RIR de uma única sala, com valores de RT_{60} diferentes. Os parâmetros das salas e das RIR foram descritos no Capítulo 2, nas Tabela 2.1 e 2.6. As medidas foram aplicadas nos sinais de voz nas situação de reverberação sem aplicação da máscara (SM) e após o uso das máscaras IBM, IRM e BRM.

A Figura 3.3 ilustra os espectrogramas dos sinais limpo e reverberado pela base

Tabela 3.1 – Resultados de predição de inteligibilidade (%) com CSII para os sinais de voz com reverberação da base AIR e após a aplicação das máscaras acústicas.

Sala	RT_{60}	SM	IBM	IRM	BRM
Limpo	-	99,90	99,90	99,90	99,66
Cabine 1	0,24	98,45	99,21	98,93	98,42
Cabine 2	0,37	97,83	99,15	99,04	98,05
Escritório 1	0,61	95,80	98,75	96,17	96,16
Escritório 2	0,64	93,67	98,24	97,89	94,31
Sala de aula 1	0,77	90,33	97,86	98,11	92,36
Sala de aula 2	0,89	86,86	**97,14**	97,01	**89,74**
Média (reverberações)		93,82	**98,39**	97,86	94,84

MARDY e após o tratamento das máscaras acústicas, bem como a representação no tempo destes sinais. Além de mostrar o efeito da reverberação no tempo e na frequência, a figura também apresenta a atuação das máscaras acústicas IBM, IRM e BRM em remover as regiões dominadas pela interferência. A seguir, são apresentados os resultados das medidas objetivas de inteligibilidade.

3.3.2 CSII

A Tabela 3.1 apresenta os resultados de CSII nas situações de reverberação com a base AIR e com a aplicação das máscaras acústicas apresentadas. Os resultados da coluna SM, quando comparados com o a medida de CSII para a situação anecoica, mostram que a reverberação degradou a inteligibilidade do sinal de voz. A comparação dos resultados de CSII em uma mesma sala mostrou que, neste caso, a inteligibilidade diminuiu com o aumento do valor de RT_{60}.

Os resultados também mostram que as máscaras acústicas tiveram êxito em melhorar a inteligibilidade do sinal de voz reverberado. O melhor resultado em média ocorreu para a máscara IBM, com melhora de 4,57 p.p. de predição de inteligibilidade. O maior aumento de inteligibilidade foi de 10,28 p.p. e ocorreu em Sala de aula 2, para a máscara IBM. A máscara não-ideal BRM, mostrou-se capaz de melhorar a inteligibilidade mesmo sem informações do sinal direto, com aumento de 1,01 p.p. na predição de inteligibilidade em média, com maior aumento em Sala de aula 2, de 2,88 p.p..

A Figura 3.4 ilustra os resultados da medida CSII para as situações anecoica e com reverberação da base MARDY. Os resultados para a situação sem máscara (SM) mostram que, para a mesma sala, o aumento do valor de RT_{60} diminuiu a inteligibilidade, chegando a uma redução da predição de taxa de acerto de palavras em 3,93 p.p. para a RIR de $RT_{60} = 0,65$ s. O maior aumento de inteligibilidade ocorreu para a máscara IBM com a reverberação de $RT_{60} = 0,65$ s, com melhora de 3,01 p.p. de inteligibilidade, seguido de

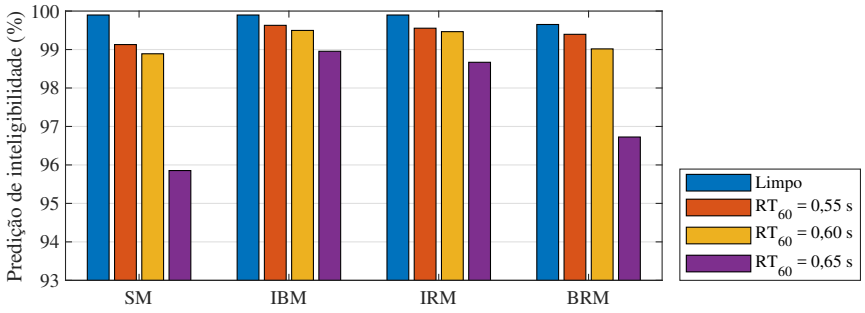

Figura 3.4 – Resultados de CSII para os sinais de voz com reverberação da base MARDY e após a aplicação das máscaras acústicas.

um aumento de 0,51 p.p. para o mesmo RT_{60} com a máscara IRM. A máscara não-ideal BRM também conseguiu uma melhora na inteligibilidade chegando a um aumento de 0,79 p.p. para $RT_{60} = 0,55$ s, com aumento médio de 0,18 P.P. para situações de reverberação.

3.3.3 STOI

Os resultados apresentados na Figura 3.5 mostram os resultados de inteligibilidade com a medida STOI nas situações de reverberação com a base AIR. A coluna SM apresenta os resultados para as situações sem máscara. A medida STOI também mostrou que a inteligibilidade dos sinais de voz diminuiu com o aumento do valor de RT_{60} quando na mesma sala. Isso se repetiu para as três salas da base AIR com maior diferença em Sala de aula 2, de 31,30 p.p. de previsão de inteligibilidade. Os resultados de STOI com as máscaras acústicas mostraram que elas melhoraram a inteligibilidade nas situações de reverberação, com a melhor média de inteligibilidade para a máscara IRM, de 96,68 p.p.. A maior melhora com a máscara IRM ocorreu em Sala de aula 2, de 20,74 p.p.. A máscara IBM conseguiu melhora média de 9,34 p.p., com maior aumento para Sala de aula 2, de 29,38 p.p.. A medida mostrou também que máscara não-ideal conseguiu melhorar também a inteligibilidade, com aumento médio de 3,23 p.p. nas situações de reverberação, com

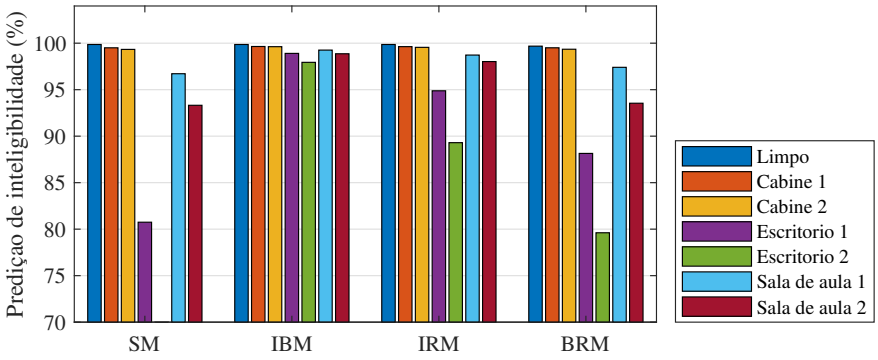

Figura 3.5 – Resultados de predição de inteligibilidade (%) com STOI para os sinais de voz com reverberação da base AIR e após a aplicação das máscaras acústicas.

Tabela 3.2 – Resultados de ΔSTOI para experimentos com a base MARDY.

RT_{60} (s)	IBM	IRM	BRM
Limpo	0,00	-0,11	-6,63
0,55	3,60	3,39	1,12
0,60	4,30	4,12	1,41
0,65	14,02	13,79	4,64
Média (com reverberação)	7,30	7,10	2,39

melhor resultado para Sala de aula 2, com 11,06 p.p. de aumento.

Os resultados apresentados na Tabela 3.2 mostram o desempenho das máscaras acústicas com as reverberações da base MARDY, todas de uma mesma sala. Os melhores resultados também foram para a IBM, com aumento médio de 14,02 p.p. de ΔSTOI, seguido de um aumento de 13,49 p.p. da máscara IRM. A máscara não-ideal BRM melhorou a inteligibilidade do sinal com um aumento médio de 2,39 p.p. de ΔSTOI para as situações reverberadas.

3.3.4 SRMR

A Tabela 3.3 mostra os resultados com experimentos realizados com a medida não-intrusiva SRMR e as reverberações da base AIR. A coluna SM apresenta tendência de diminuição da inteligibilidade com o aumento do valor de RT_{60} também foi refletida com a SRMR. Porém, os maiores valores ocorreram com a máscara não-ideal BRM, com média de SRMR de 6,32. O maior aumento também ocorreu com esta máscara, com a reverberação Cabine 2. As máscaras IBM e IRM conseguiram média de 2,13 e 2,48 de diferença em relação à média das situações sem tratamento.

A Figura 3.6 mostra os resultados de SRMR para com a base MARDY. A medida também mostrou o impacto da reververação na inteligibilidade do sinal de voz

Tabela 3.3 – Resultados de SRMR para os sinais de voz com reverberação da base AIR e após a aplicação das máscaras acústicas.

Sala	RT_{60} (s)	SM	IBM	IRM	BRM
Anecoico	-	6,37	6,37	6,37	8,24
Cabine 1	0,24	4,71	5,97	4,70	7,68
Cabine 2	0,37	3,74	5,40	6,31	**8,34**
Escritório 1	0,77	2,72	4,74	5,79	6,35
Escritório 2	0,89	2,56	5,25	5,98	5,92
Sala de aula 1	0,61	1,68	4,58	3,83	4,63
Sala de aula 2	0,64	1,66	3,94	5,35	5,01
Média		2,85	4,98	5,33	**6,32**

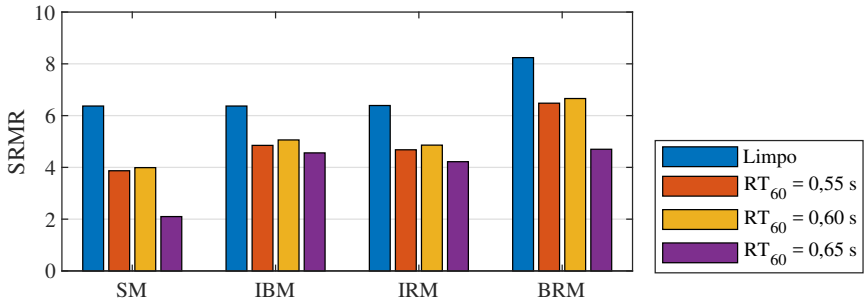

Figura 3.6 – Resultados de SRMR para os sinais de voz com reverberação da base MARDY e após a aplicação das máscaras acústicas.

com diminuição de até 4,27 entre o índice para a voz anecoica e com a reverberação com $\mathrm{RT}_{60} = 0,65$. A máscara BRM apresentou a melhor média de índice SRMR, de 4,70 e o melhor aumento, de 2,67 para $\mathrm{RT}_{60} = 0,60$.

3.4 Resumo

Este Capítulo apresentou uma solução para recuperação da inteligibilidade da voz reverberada baseada no sistema auditivo humano, as máscaras acústicas. Duas máscaras ideais e uma não-ideal foram mostradas. Para avaliar o impacto da reverberação na inteligibilidade e o desempenho das máscaras em reduzir este impacto, foram apresentadas três medidas objetivas de inteligibilidade, sendo uma delas não-intrusiva. Por fim, foram apresentados os resultados obtidos, demonstrando o impacto da reverberação na inteligibilidade do sinal de voz e a capacidade das máscaras acústicas de recuperar a inteligibilidade.

4 Solução para Identificação de Locutor com Reverberação

Os estudos referentes ao efeito da reverberação, bem como seu impacto na inteligibilidade sonora e, consequentemente, na identificação de locutor, realizados neste trabalho inspiraram soluções para o aprimoramento da acurácia dos sistemas de identificação. Ainda que o estudo tenha comprovado que a identificação de locutor é menos afetada em situações de mesma condições de reverberação para as fases de treinamento e teste, os resultados mostraram que, em situações reais de descasamento, as taxas de acerto sofrem um forte impacto.

Os cocleogramas apresentados no Capítulo 2 mostraram que os filtros *Gammatone* (JOHANNESMA, 1972) (PATTERSON; MOORE, 1986) (COOKE, 1993), por sua relação com o sistema auditivo humano, são capazes de destacar as frequências em que a voz humana é mais presente. Esta propriedade destas funções nos inspira a utilizar os coeficientes GFCC (*Gammatone-Frequency Cepstral Coefficients*) (ZHAO *et al.*, 2012) (SHAO *et al.*, 2007) como atributos do sinal de voz na identificação. Na etapa de classificação, a generalização do classificador GMM, o α-GMM, se mostrou eficiente em melhorar resultados de verificação de locutor em situações de interferências ou limitação de canal. Os modelos produzidos a partir deste classificador, por serem mais robustos a interferências, podem ser eficientes em situações de reverberação.

Os resultados obtidos no Capítulo 3 mostraram que as máscaras acústicas foram eficazes em recuperar a inteligibilidade dos sinais de voz reverberados. Estes resultados positivos inspiram a utilização de máscaras acústicas como solução para atenuar o impacto da reverberação na identificação de locutor. A literatura apresenta estudos com máscaras acústicas aplicadas a sinais de voz corrompidos por ruídos ambientais com aplicações para sistemas de reconhecimento de fala (HARTMANN; FOSLER-LUSSIER, 2011) (NARAYANAN; WANG, 2013) e de locutor (SHAO; WANG, 2006). Nestes estudos, máscaras acústicas ideais aumentaram as taxas de acertos de palavras e o desempenho da identificação de locutor.

Neste Capítulo, é proposto um novo modelo de classificação de locutores α-GMM para atenuação do efeito de descasamento entre fases. Além disso, será estudada a eficiência das máscaras acústicas ideais e da máscara não-ideal apresentadas neste trabalho em recuperar as taxas de identificação de locutor degradadas pelo descasamento de reverberação entre treinamento e teste. Situações de descasamento de reverberação

entre salas e dentro de uma mesma sala serão abordadas utilizando, respectivamente, reverberações das bases MARDY 2.1 e AIR 2.6.

4.1 Vetor de Atributos GFCC: *Gammatone-Frequency Cepstral Coefficients*

Os atributos GFCC foram propostos em (SHAO *et al.*, 2007) para reconhecimento de locutor. Assim como no MFCC, estes atributos se baseiam na aproximação computacional do sistema auditivo. Neste caso, é feito uso dos filtros *Gammatone*, que foram propostos para descrever o comportamento da função de resposta ao impulso do sistema auditivo humano no domínio do tempo.

A resposta ao impulso de um filtro *Gammatone* centrado na frequência f_c em Hz é dada por:

$$g(t) = Kt^{(n-1)}e^{-2\pi Bt}\cos(2\pi f_c + \varphi), \qquad t > 0, \tag{4.1}$$

onde K é o fator de amplitude, n a ordem do filtro, φ a fase e B a duração da resposta ao impulso. Neste trabalho foram utilizados filtros de ordem $n = 4$.

A extração dos atributos GFCC ocorre após a filtragem do sinal de voz por um banco de N filtros com frequências centrais espaçadas quasi-logaritmicamente entre 50 Hz e $f_s/2$, onde f_s é a taxa de amostragem do sinal. Em seguida, cada um dos N sinais obtidos é sub-amostrado a 100 Hz. A raiz cúbica do sinal então é definida como o vetor de atributos *Gammatone* GF (*Gammatone Feature*) de acordo com a equação

$$G_m[i] = ||g|_{decimado}[i, m]|^{1/3}, \qquad i = 0, \ldots, N-1, \qquad m = 0, \ldots, M-1. \tag{4.2}$$

No próximo passo, assim como procedido com o MFCC, utiliza-se uma DCT para obtenção dos coeficientes GFCC. Os coeficientes $C[j]$ são calculados por

$$G_m[i] = \sqrt{\frac{2}{N}} \sum_{i=0}^{N-1} \cos\left(\frac{j\pi}{2N}(2i+1)\right), \qquad j = 0, \ldots, N-1. \tag{4.3}$$

4.2 Os modelos de Classificação α-GMM

Uma generalização do classificador GMM cujo objetivo é uma modelagem que integre as interferências ou variações acústicas e a limitação de banda do canal é o α-GMM (α-*integrated* GMM) (WU *et al.*, 2009) (VENTURINI *et al.*, 2014). Este modelo se baseia na utilização de gaussianas α-integráveis. A soma de gaussianas de um modelo GMM convencional é dada por:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}), \tag{4.4}$$

onde $\vec{x}$ é o vetor de atributos, $b_i(\vec{x}), i = 1, \ldots, M$, são as funções densidade e $p_i, i = 1, \ldots, M$ são os pesos de cada Gaussiana.

Cada componente do modelo GMM $b_i(\vec{x})$ é uma função Gaussiana de dimensão D e podem ser descritas por:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}}\sqrt{|K_i|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu_i})^T K_i^{-1}(\vec{x}-\vec{\mu_i})}. \tag{4.5}$$

onde $\vec{\mu_i}$ é o vetor média e K_i a matriz de covariância, com determinante $|K_i|$. Os pesos p_i satisfazem a condição $\sum_{i=1}^{M} p_i = 1$. Para a generalização α-GMM, é definida uma função $q(s)$ tal que:

$$q(s) = c f_\alpha^{-1}\left\{\sum_{i=1}^{K} w_i f_\alpha[p_i(s)]\right\}, \tag{4.6}$$

em que c é uma constante cujo valor permite que $q(s)$ seja uma função densidade de probabilidade. A função $f_\alpha[p_i(s)]$ é um α-representação das funções densidade de probabilidade $p_i(s)$ e é definida por:

$$f_\alpha(p_i(s)) = \begin{cases} \frac{2}{1-\alpha} p_i(s)^{\frac{1-\alpha}{2}} & \alpha \neq 1, \\ \log(p_i(s)) & \alpha = 1. \end{cases} \tag{4.7}$$

Utilizando a α-integração da Equação 4.6 na Equação 4.4, o modelo α-GMM é descrito pela equação:

$$p_\alpha(\vec{x}|\lambda_\epsilon) = \begin{cases} c\left(\sum_{j=1}^{M} p_j (b_j(\vec{x}))^{\frac{1-\alpha}{2}}\right)^{\frac{2}{1-\alpha}} & \alpha \neq 1, \\ c e^{\sum_{j=1}^{M} p_j \log(b_j(\vec{x}))} & \alpha = 1. \end{cases} \tag{4.8}$$

Para $\alpha = -1$, o modelo α-GMM se resume ao GMM clássico. Quando se aplica $\alpha < -1$, enfatizam-se as gaussianas com maiores valores de probabilidade, diminuindo a relevância das que possuem os menores valores.

4.3 Resultados de Identificação de Locutor: GFCC + α-GMM

Experimentos de identificação de locutor foram realizados para avaliar a eficácia das máscaras acústicas, dos atributos GFCC e do classificador α-GMM em melhorar as taxas de acertos degradadas pelo efeito da reverberação. As máscaras utilizadas foram as ideais IBM e IRM, além da máscara não-ideal BRM apresentadas no Capítulo 3. Os experimentos foram realizados em duas situações de descasamento de reverberação entre treinamento e teste: primeiramente com descasamento entre salas com a base AIR (JEUB *et al.*, 2009), cujos parâmetros das RIR selecionadas foram apresentados na Tabela 2.6, e em seguida com treinamento e teste realizados na mesma sala com a base MARDY (WEN *et al.*, 2006), apresentada na Tabela 2.1, porém com variação de posicionamento de fonte e microfone.

Tabela 4.1 – Taxa de acertos de identificação de locutor (%) com atributos MFCC e GFCC e classificador GMM com a base MARDY.

Treinamento	Teste	MFCC	GFCC
0,55	0,55	**95,62**	**83,93**
	0,60	41,07	42,40
	0,65	51,49	53,71
0,60	0,55	30,65	33,18
	0,60	**93,45**	**90,02**
	0,65	70,24	73,68
0,65	0,55	41,96	44,07
	0,60	66,07	83,63
	0,65	**90,48**	**88,68**

As máscaras acústicas utilizadas nos experimentos tiveram janelamento realizado com quadros de 20 ms e 50% de sobreposição. Nas máscaras ideais IBM e IRM, o critério para exclusão do quadro tempo-frequência a ser "mascarado" é de SRR < -5 dB. As máscaras IRM e BRM tiveram a sua decomposição em sub-bandas realizada com 64 filtros *Gammatone* espaçados de acordo com a escala ERB (Bandas Regulares Equivalentes) distribuídos entre 50 Hz e 8 kHz.

Os testes foram conduzidos com sinais de voz selecionados da base TIMIT (GAROFOLO *et al.*, 1993), mesma base utilizada nos experimentos do Capítulo 2. As reverberações para os experimentos com descasamento entre salas foram extraídas da base AIR (JEUB *et al.*, 2009). Os testes realizados somente com descasamento de posicionamento de fonte e microfone utilizaram reverberações da base MARDY (WEN *et al.*, 2006). Na identificação de locutor, foram utilizados quadros de 32 ms com 50% de sobreposição. De cada quadro foram extraídos 12 coeficientes MFCC. O modelo α-GMM foi utilizado com $\alpha = -1$ (GMM convencional) e $\alpha = \{-2, -4, -6, -8\}$. Os modelos α-GMM foram compostos a partir de 32 gaussianas.

4.3.1 Treinamento e teste em uma mesma sala - Base MARDY

A Tabela 4.1 apresenta uma comparação de resultados entre testes realizados com os atributos MFCC e os atributos GFCC. Os experimentos mostraram que, comparado aos MFCC, os atributos GFCC apresentaram melhores resultados para situações com descasamento de reverberação. A taxa de acerto teve a sua maior melhora com GFCC em treinamento com reverberação $\mathrm{RT}_{60} = 0,65$ s e teste em $\mathrm{RT}_{60} = 0,60$ s, com melhora de 17,56 p.p.. Nas situações de casamento de reverberação com GFCC, os resultados foram melhores que com descasamento, porém não superaram as taxas de acerto com atributo MFCC.

Os resultados apresentados na Tabela 4.2 mostram as taxas de identificação

Tabela 4.2 – Taxa de acertos de identificação de locutor (%) com a base MARDY utilizando o classificador α-GMM, com $\alpha = \{-1, -2, -4, -6, -8\}$

Treinamento	Teste	GMM	α				Média
			-2	-4	-6	-8	
0,55	0,55	**95,62**	**96,13**	**96,43**	**95,54**	**96,43**	**96,13**
	0,60	41,07	84,82	83,63	83,93	83,63	84,00
	0,65	51,49	50,60	52,68	52,38	52,08	51,93
0,60	0,55	30,65	80,06	80,06	82,44	82,74	81,32
	0,60	**93,45**	**95,54**	**96,73**	**94,94**	**97,32**	**96,13**
	0,65	70,24	41,67	42,86	43,45	44,48	42,11
0,65	0,55	41,96	38,99	38,39	39,88	38,39	38,91
	0,60	66,07	35,12	32,44	36,31	33,04	34,23
	0,65	**90,48**	**88,69**	**86,90**	**87,80**	**86,90**	**87,57**

.

de locutor obtidas nos experimentos com a base MARDY onde treinamento e teste foram realizados na mesma sala com diferentes valores de RT_{60}. A coluna $\alpha = 1$ apresenta os resultados para o método GMM convencional. Para $\alpha < -1$, foi observado que, para cada valor RT_{60} de treinamento, há um valor de α que maximiza os resultados utilizando o modelo α-GMM. Para o treinamento $RT_{60} = 0,55$ s, os melhores resultados ocorreram com $\alpha = -2$. Para os experimentos com treinamento realizado com $RT_{60} = 0,60$ s, o valor que maximiza as taxas de acerto é $\alpha = -8$. Já com treinamento com $RT_{60} = 0,65$ s, $\alpha = -6$ apresentou os melhores resultados de α-GMM. Os resultados mostraram que o único treinamento em que o α-GMM não superou o GMM convencional, em média, foi $RT_{60} = 0,65$ s. Nos demais, os resultados superaram o GMM convencional em todas as combinações de treinamento e teste.

Na Tabela 4.3 encontram-se os resultados de identificação de locutor com reverberação realizados com classificador GMM convencional e α-GMM em que treinamento

Tabela 4.3 – Taxa de acertos de identificação de locutor (%) com a base MARDY utilizando o classificador α-GMM, com $\alpha = \{-1, -2, -4, -6, -8\}$ e aplicação da máscara IBM.

Treinamento	Teste	GMM	α				Média
			-2	-4	-6	-8	
0,55	0,55	**98,21**	**98,81**	**99,11**	**98,81**	**98,81**	**98,88**
	0,60	86,61	86,01	86,31	85,71	84,52	85,64
	0,65	41,79	32,98	41,79	37,89	57,68	42,58
0,60	0,55	85,71	87,50	85,12	84,23	85,12	85,49
	0,60	**98,81**	**99,11**	**97,32**	**99,40**	**99,11**	**98,74**
	0,65	52,08	51,19	52,08	53,27	51,49	52,01
0,65	0,55	23,87	23,27	25,65	24,76	24,17	24,46
	0,60	48,89	51,49	42,98	42,08	42,38	44,73
	0,65	**97,32**	**97,02**	**96,73**	**96,73**	**97,02**	**96,88**

Tabela 4.4 – Taxa de acertos de identificação de locutor (%) com a base MARDY utilizando o classificador α-GMM, com $\alpha = \{-1, -2, -4, -6, -8\}$ e aplicação da máscara IRM.

Treinamento	Teste	GMM	α				Média
			-2	-4	-6	-8	
0,55	0,55	**98,21**	**98,21**	**97,62**	**98,51**	**97,02**	**97,84**
	0,60	82,74	79,76	84,82	84,23	82,14	82,74
	0,65	51,79	41,60	39,89	40,89	41,79	41,04
0,60	0,55	83,04	85,42	83,33	86,90	85,71	85,34
	0,60	**97,02**	**98,51**	**95,54**	**97,92**	**96,73**	**97,17**
	0,65	41,49	50,60	50,89	50,30	51,49	50,82
0,65	0,55	30,27	23,27	23,87	24,46	24,46	24,02
	0,60	51,79	52,98	52,38	51,79	53,57	52,68
	0,65	**83,63**	**80,95**	**83,93**	**83,63**	**82,44**	**82,74**

e teste "mascarados" com uma máscara IBM. Os resultados mostraram que a máscara aumentou as taxas de identificação para as situações onde há o casamento de reverberações entre treinamento e teste. Nas demais, os resultados se mantiveram abaixo dos obtidos sem a utilização da máscara, com exceção para os experimentos com treinamento em $\mathrm{RT}_{60} = 0,55$ s e teste em $\mathrm{RT}_{60} = 0,60$ s e com treinamento em $\mathrm{RT}_{60} = 0,60$ s e $\mathrm{RT}_{60} = 0,65$ s.

Os experimentos foram também realizados com a máscara IRM, com resultados apresentados na 4.4. Os resultados se mostraram abaixo dos obtidos para a máscara IBM. Nas situações sem descasamento de reverberação, os resultados mantiveram a tendência de melhorar os resultados, com exceção para $\mathrm{RT}_{60} = 0,65$ s.

Na Tabela 4.5 são mostrados os resultados com a máscara BRM. Eles mostraram que a máscara não-ideal, apesar de melhorar a inteligibilidade do sinal de voz

Tabela 4.5 – Taxa de acertos de identificação de locutor (%) com a base MARDY utilizando o classificador α-GMM, com $\alpha = \{-1, -2, -4, -6, -8\}$ e aplicação da máscara BRM.

Treinamento	Teste	GMM	α				Média
			-2	-4	-6	-8	
0,55	0,55	**89,29**	**87,80**	**86,31**	**89,88**	**89,58**	**88,39**
	0,60	82,74	77,68	81,85	81,25	80,95	80,43
	0,65	42,26	47,32	47,32	45,83	41,37	45,46
0,60	0,55	78,57	77,38	74,70	76,79	74,11	75,74
	0,60	**89,29**	**86,61**	**88,39**	**88,99**	**83,33**	**86,83**
	0,65	41,67	36,31	35,71	38,10	35,71	36,46
0,65	0,55	44,35	41,07	37,80	44,64	39,29	40,70
	0,60	38,39	38,39	37,80	36,31	36,31	37,20
	0,65	**76,49**	**71,43**	**73,21**	**73,51**	**75,00**	**73,29**

reverberado, não foi capaz de aumentar as taxas de identificação em casamento ou descasamento de reverberação. A exceção ocorreu para os experimentos com treinamento em $\mathrm{RT}_{60} = 0,55$ s e teste em $\mathrm{RT}_{60} = 0,60$ s e com treinamento em $\mathrm{RT}_{60} = 0,60$ s e $\mathrm{RT}_{60} = 0,65$ s.

4.3.2 Treinamento e testes em salas diferentes - Base AIR

Os resultados de identificação de locutor com os atributos GFCC estão apresentados na Figura 4.1. Nelas, os resultados estão separados por médias dos treinamentos de cada sala. Nota-se que, com o descasamento de salas, o atributo MFCC supera em quatro das situações de treinamento, menos em Cabine 1 e Sala de aula 2, onde as taxas de acerto com GFCC aumentam em 4,54 p.p. e 1,71 p.p., respectivamente. A média dos resultados com descasamento de salas dos testes com atributo GFCC é 43,48 p.p., 3,31 p.p. abaixo dos testes com MFCC.

A Figura 4.2 apresenta os resultados de identificação com descasamento de salas separadas pelo fator α e pela condição de treinamento. Para todos os valores de α, os resultados de identificação de locutor diminuíram quando aplicadas as máscaras ideais, porém aumentaram com as máscaras não-ideais. Com o GMM convencional, a máscara BRM melhorou em 4,23 p.p. as taxas de reconhecimento. Nas condições SM (sem máscara), $\alpha = -6$ obteve os melhores resultados, com média de 53,65 p.p. comparado com 46,16 p.p. do GMM convencional. Com a máscara BRM, as salas Escritório 1 e Escritório

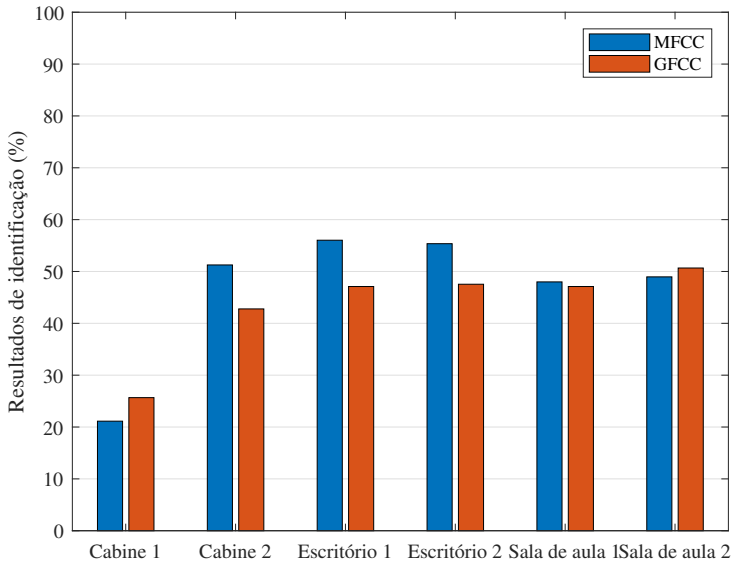

Figura 4.1 – Médias de resultados de experimentos de identificação de locutor com descasamento de salas utilizando a base AIR por sala de teste com atributos MFCC e GFCC.

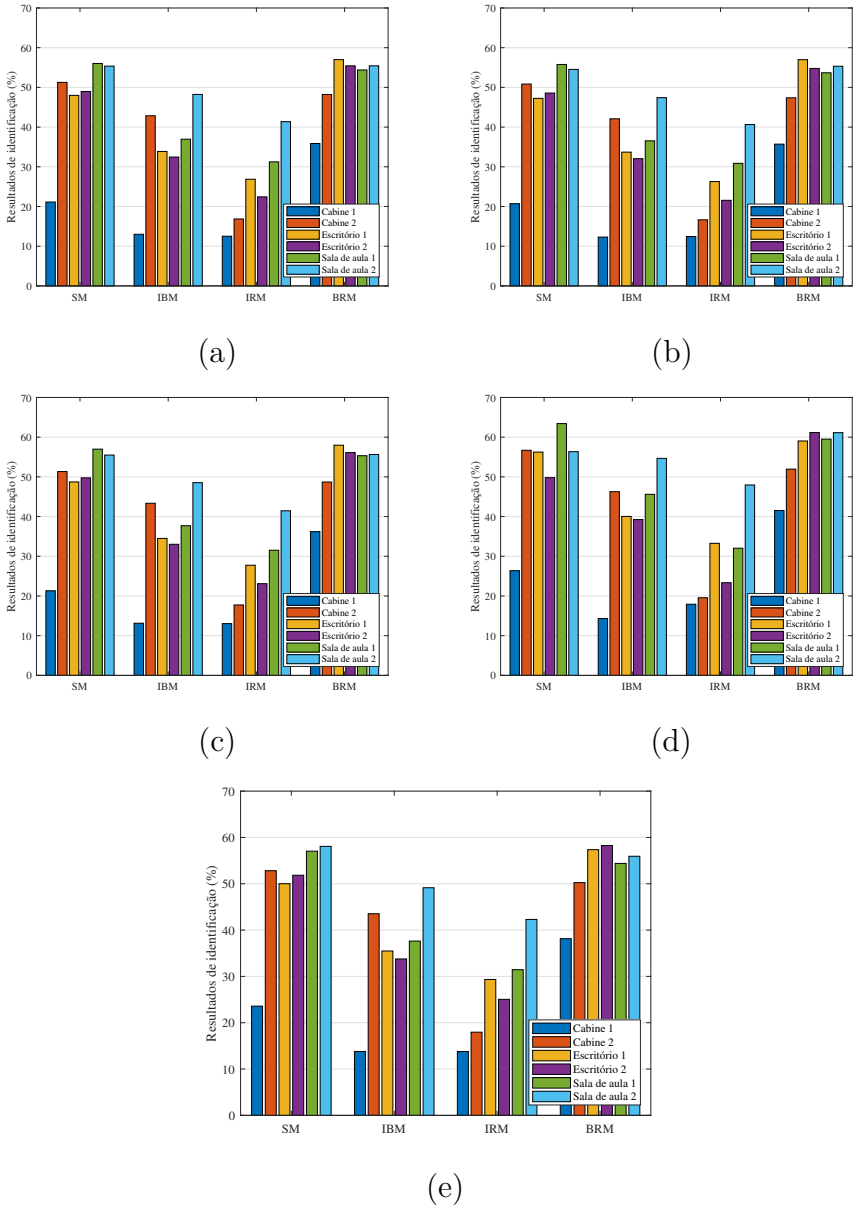

Figura 4.2 – Resultados de identificação com a base AIR e descasamento de salas entre treinamento e teste. Os experimentos com α-GMM foram realizados com (a) $\alpha = -1$, (b) $\alpha = -2$, (c) $\alpha = -4$, (d) $\alpha = -6$ e (e) $\alpha = -8$. A legenda indica a condição de treinamento em que foi realizada a identificação.

2 obtiveram os melhores resultados, com 65,79 p.p. e 66,02 p.p. respectivamente, para $\alpha = -6$.

4.4 Resumo

Neste Capítulo, foi estudada a eficácia das máscaras acústicas em melhorar as taxas de identificação de locutor em situações de descasamento de reverberação entre treinamento e teste. O descasamento foi abordado de duas maneiras: entre salas diferentes e em uma mesma sala, com variação da posição do locutor. Neste estudo, foram utilizadas

duas máscaras acústicas ideais (IBM e IRM) e uma não-ideal (BRM), ou seja, que não utiliza informações do sinal direto. Além do uso das máscaras acústicas, foi investigado o uso do atributo GFCC e a interação das máscaras acústicas com o classificador GMM clássico e com α-GMM com valores de $\alpha < -1$. O estudo concluiu que as máscaras acústicas ideais conseguiram bons resultados nas situações onde não houve o descasamento de reverberação. Isto pode ser observado nos estudos com descasamento em uma mesma sala e em salas diferentes. Para situações de descasamento de salas entre treinamento e teste, a máscara não-ideal BRM foi a única que obteve melhorias nas taxas de identificação de locutor. Os experimentos com o atributo GFCC mostraram que estes superaram o MFCC em situações de descasamento de reverberação em uma mesma sala. Para casamento de reverberação ou descasamento de salas, o atributo MFCC conseguiu melhores resultados que o GFCC.

Conclusão

Nesta Dissertação, foi apresentado um estudo do efeito da reverberação no sinal de voz e de seu impacto na inteligibilidade sonora e em sistemas de identificação de locutor. Para a análise do efeito, foram utilizadas RIR de duas bases de reverberação, MARDY e AIR, além da base de voz TIMIT.

Na pesquisa, foi mostrado que, em uma mesma sala, o aumento do valor de RT_{60} de uma RIR implica na diminuição da não-estacionariedade do sinal de voz reverberado. Além disso, experimentos de identificação de locutor utilizando o atributo MFCC e o classificador GMM mostraram que o sistema é impactado pela presença da reverberação de uma sala e que isto ocorre principalmente em situações de descasamento de reverberação entre treinamento e teste. Nas situações onde não há o descasamento, as taxas de acerto decaem com a diminuição da não-estacionariedade do sinal de voz reverberado. Cenários com descasamento de salas entre treinamento e teste são mais impactados que cenários com descasamento de reverberação em uma mesma sala.

Experimentos em que foram utilizadas medidas objetivas de inteligibilidade indicaram que a reverberação tem impacto na inteligibilidade do sinal de voz. Além disso, as medidas mostraram que o uso das máscaras acústicas ideais IBM e IRM e da máscara não-ideal (cega) BRM foi capaz de recuperar a inteligibilidade do sinal de voz reverberado. Propostas para melhorar o desempenho de sistemas de locutor que incluíram o atributo GFCC, o classificador α-GMM e a aplicação de máscaras acústicas nos sinais de voz reverberados foram apresentadas. O atributo GFCC mostrou melhorar as taxas de acerto de sistemas de identificação de locutor em situações de descasamento de reverberação em uma mesma sala em 4,86 p.p.. A utilização do classificador α-GMM melhorou as taxas de identificação em situações de descasamento em uma mesma sala de 5,31 p.p. para $\alpha = -2$. A aplicação da máscara acústica IBM nos sinais de voz reverberados de treinamento e teste mostrou melhorar os resultados de identificação de locutor em 4,30 p.p. em situações de casamento de reverberação.

Os experimentos com o sinal de voz reverberado contaram com 3 RIR da base MARDY e 6 RIR de 3 salas diferentes da base AIR. A análise do sinal de voz reverberado contou com espectrogramas e cocleogramas, além das medidas distância Bhattacharyya e INS, esta última usada para mesurar a não-estacionariedade de um sinal. Os testes de inteligibilidade contaram com 3 medidas objetivas de alta correlação com testes perceptuais de reconhecimento de palavras: CSII, STOI e SRMR, sendo esta última uma medida não-intrusiva. Os experimentos contaram com 240 sinais de voz de 20 locutores da base

TIMIT. Os resultados de identificação de locutor foram obtidos a partir das locuções de voz de 168 locutores da base TIMIT, com média de 24 segundos de locuções para treinamento e 6 segundos para testes.

As principais contribuições deste trabalho podem ser resumidas da seguinte forma:

- Análise do efeito da reverberação de salas distintas com RIR de diferentes valores de RT_{60} no sinal de voz;
- Demonstração da relação entre o RT_{60} de uma RIR e a não-estacionariedade de um sinal de voz reverberado;
- Estudo do impacto da reverberação em sistemas de identificação de locutor em situações de casamento e descasamento (em uma mesma sala ou em salas diferentes) de reverberação entre treinamento e teste.
- Análise da inteligibilidade do sinal de voz reverberado por diferentes salas através de medidas objetivas de inteligibilidade e estudo do uso de máscaras acústicas ideais e de uma máscara não-ideal em recuperar a inteligibilidade do sinal de voz com reverberação;
- Proposta de uso do atributo GFCC e do classificador α-GMM para melhorar as taxas de acerto de um sistema de identificação de locutor degradadas pela presença da reverberação;
- Proposta do emprego de máscaras acústicas em sistemas de identificação de locutor com treinamento e teste em situações de reverberação.

Sugestões para trabalhos futuros

Algumas sugestões para trabalhos futuros podem ser destacadas:

- investigar o uso do índice de não-estacionariedade como critério para a estimação da reverberação e a seleção de quadros para obtenção do aprimoramento da inteligibilidade pós-mascaramento do sinal reverberado;
- propor novas soluções de máscaras acústicas, ideais e não-ideais , para aprimoramento da inteligibilidade considerando a diversidade do efeito da reverberação nos sinais de voz;
- estudar proposta de nova medida objetiva de inteligibilidade que reflita o efeito perceptual da reverberação;

- avaliar o emprego de atributos acústicos tempo-frequência como os vetores pH (SANT'ANA *et al.*, 2006), MP (CHU *et al.*, 2009) e fusões de atributos para melhora das taxas de acertos em situações de descasamento de reverberação entre as fases de sistemas de reconhecimento automático de locutor;

- Investigar o uso da representação da reverberação em métodos baseados em aprendizado de máquina, por exemplo, SVM (*Supervised Vector Machine*) e DNN (*Deep Neural Networks*), para redução do efeito de descasamento de condições na classificação de locutores.

Comentários finais

Esta Dissertação apresentou uma análise dos efeitos da reverberação no sinal de voz em relação a seu impacto na inteligibilidade e em sistemas de identificação de locutor. O estudo mostrou que a presença reverberação é capaz de reduzir as taxas de acerto de identificação de locutor, que são mais degradadas em situações de descasamento de reverberação entre treinamento e teste. Além disso, foi mostrado através de medidas objetivas que a reverberação tem impacto na inteligibilidade do sinal de voz e que o emprego de máscaras acústicas pode recuperar a inteligibilidade do sinal de voz reverberado. A utilização de máscaras acústicas no sinal de voz reverberado em sistemas de identificação de locutor foi proposta, provendo uma melhora em casos de casamento de reverberação entre treinamento e teste. Também foi proposta a utilização do atributo GFCC e do classificador α-GMM para identificação de locutor em ambientes reverberados. O atributo GFCC melhorou as taxas de acerto para situações de descasamento de reverberação em uma mesma sala. O classificador α-GMM foi capaz de melhorar as taxas de acerto em situações de casamento e descasamento de reverberação em uma mesma sala.

Referências

ASSMANN, P.; SUMMERFIELD, A. The perception of speech under adverse conditions. In: GREENBERG, S.; AINSWORTH, W.; POPPER, A.; FAY, R. (Ed.). *Speech Processing in the Auditory System*. New York: Springer-Verlag, 2004. cap. 14, p. 231–308. Citado 2 vezes nas páginas 21 e 23.

ATAL, B. S. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America*, v. 55, n. 6, p. 1304–1312, 1974. Disponível em: <https://doi.org/10.1121/1.1914702>. Citado na página 21.

BASSEVILLE, M. Distance measures for signal processing and pattern recognition. *Signal Processing*, v. 18, n. 4, p. 349 – 369, 1989. ISSN 0165-1684. Disponível em: <http://www.sciencedirect.com/science/article/pii/0165168489900790>. Citado na página 25.

BERANEK, L. L. Concert Hall Acoustics—2008. *Journal of the Audio Engineering Society*, Audio Engineering Society, v. 56, n. 7/8, p. 532–544, aug 2008. Disponível em: <http://www.aes.org/e-lib/browse.cfm?elib=14398>. Citado na página 16.

BIMBOT, F.; BONASTRE, J.; FREDOUILLE, C.; GRAVIER, G.; MAGRIN-CHAGNOLLEAU, I.; MEIGNIER, S.; MERLIN, T.; ORTEGA-GARCIA, J.; PETROVSKA-DELACRETAZ, D.; REYNOLDS, D. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, v. 4, p. 430–451, 2004. Citado na página 32.

BISPO, B.; ESQUEF, P.; BISCAINHO, L.; LIMA, A.; FREELAND, F.; JESUS, R.; SAID, A.; LEE, B.; SCHAFER, R.; KALKER, T. Ew-pesq: A quality assessment method for speech signals sampled at 48 khz. *Journal of the Audio Engineering Society*, v. 58, n. 4, p. 251–268, 2016. Citado na página 45.

BOLT, R. H.; MACDONALD, A. D. Theory of speech masking by reverberation. *The Journal of the Acoustical Society of America*, v. 21, n. 6, p. 577–580, 1949. Disponível em: <https://doi.org/10.1121/1.1906551>. Citado 3 vezes nas páginas 16, 21 e 44.

BORGNAT, P.; FLANDRIN, P.; HONEINE, P.; RICHARD, C.; XIAO, J. Testing stationarity with surrogates: A time-frequency approach. *IEEE Transactions on Signal Processing*, v. 58, n. 7, p. 3459–3470, July 2010. ISSN 1053-587X. Citado 2 vezes nas páginas 24 e 25.

BORGSTRÖM, B. J.; MCCREE, A. The linear prediction inverse modulation transfer function (lp-imtf) filter for spectral enhancement, with applications to speaker recognition. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2012. p. 4065–4068. ISSN 1520-6149. Citado na página 21.

BRANDEWIE, E.; ZAHORIK, P. Prior listening in rooms improves speech intelligibility. *The Journal of the Acoustical Society of America*, v. 128, n. 1, p. 291–299, 2010. Disponível em: <https://doi.org/10.1121/1.3436565>. Citado na página 37.

BREGMAN, A. *Auditory scene analysis: The perceptual organization of sound*. [S.l.]: The MIT Press, 1990. Citado na página 16.

BRONKHORST, A. W. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, S. Hirzel Verlag, v. 86, n. 1, p. 117–128, January 2000. Disponível em: <http://www.ingentaconnect.com/content/dav/aaua/2000/00000086/00000001/art00013>. Citado 2 vezes nas páginas 16 e 40.

BROWN, G. J.; COOKE, M. Computational auditory scene analysis. *Computer Speech and Language*, v. 8, n. 4, p. 297–336, 1994. ISSN 0885-2308. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0885230884710163>. Citado na página 24.

CAMPBELL, J. Speaker recognition: a tutorial. *Proceedings of the IEEE*, v. 85, n. 9, p. 1437–1461, September 1997. Citado na página 32.

CAPORALE, P. Musical acoustics of auditoriums. *Journal of the Society of Motion Picture Engineers*, v. 20, n. 2, p. 119–127, Feb 1933. ISSN 0097-5834. Citado na página 16.

CASTELLANO, P. J.; SRADHARAN, S.; COLE, D. Speaker recognition in reverberant enclosures. In: *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. [S.l.: s.n.], 1996. v. 1, p. 117–120 vol. 1. ISSN 1520-6149. Citado na página 21.

CHERRY, E. C. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, v. 25, n. 5, p. 975–979, 1953. Disponível em: <https://doi.org/10.1121/1.1907229>. Citado na página 16.

CHU, S.; NARAYANAN, S.; KUO, C. . J. Environmental sound recognition with time–frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 17, n. 6, p. 1142–1158, Aug 2009. ISSN 1558-7916. Citado na página 65.

COOKE, M. *Modelling Auditory Processing and Organisation*. New York, NY, USA: Cambridge University Press, 1993. ISBN 0-521-45094-2. Citado 4 vezes nas páginas 24, 41, 42 e 54.

DAU, T.; PüSCHEL, D.; KOHLRAUSCH, A. A quantitative model of the "effective" signal processing in the auditory system. i. model structure. *The Journal of the Acoustical Society of America*, v. 99, n. 6, p. 3615–3622, 1996. Disponível em: <https://doi.org/10.1121/1.414959>. Citado na página 48.

DAVIS, S.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 28, p. 357–366, 1980. Citado na página 33.

DELFARAH, M.; WANG, D. Features for masking-based monaural speech separation in reverberant conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 25, n. 5, p. 1085–1094, May 2017. ISSN 2329-9290. Citado 3 vezes nas páginas 21, 23 e 47.

FALK, T. H.; ZHENG, C.; CHAN, W. Y. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech and Language Processing*, v. 18, n. 7, p. 1766–1774, sep 2010. ISSN 15587916. Disponível em: <http://ieeexplore.ieee.org/document/5547575/>. Citado 3 vezes nas páginas 40, 45 e 47.

FURUI, S. Comparison of speaker recognition methods using statistical features and dynamic features. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 29, n. 3, p. 342–350, Jun 1981. ISSN 0096-3518. Citado na página 33.

GARCIA-ROMERO, D.; ZHOU, X.; ESPY-WILSON, C. Y. Multicondition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2012. p. 4257–4260. ISSN 1520-6149. Citado na página 21.

GAROFOLO, J. S.; LAMEL, L. F.; FISHER, W. M.; FISCUS, J. G.; PALLETT, D. S.; DAHLGREN, N. L.; ZUE, V. *TIMIT Acoustic Phonetic Continuous Speech Corpus*. [S.l.]: NIST, 1993. Citado 6 vezes nas páginas 17, 24, 36, 46, 49 e 57.

GLASBERG, B. R.; MOORE, B. C. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, v. 47, n. 1-2, p. 103–138, aug 1990. ISSN 03785955. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/2228789>. Citado na página 48.

GOLD, B.; MORGAN, N. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. 1st. ed. New York, NY, USA: John Wiley & Sons, Inc., 1999. ISBN 0471351547. Citado na página 16.

GONZALEZ-RODRIGUEZ, J.; ORTEGA-GARCIA, J.; MARTIN, C.; HERNANDEZ, L. Increasing robustness in gmm speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays. In: *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. [S.l.: s.n.], 1996. v. 3, p. 1333–1336 vol.3. Citado na página 21.

GÖLZER, H.; KLEINSCHMIDT, M. Importance of early and late reflections for automatic speech recognition in reverberant environments. *Elektronische Sprachsignalverarbeitung (ESSV)*, 2003. Disponível em: <http://medi.uni-oldenburg.de/projects/asr>. Citado na página 22.

HARTMANN, W.; FOSLER-LUSSIER, E. Investigations into the incorporation of the ideal binary mask in asr. In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2011. p. 4804–4807. ISSN 1520-6149. Citado 2 vezes nas páginas 17 e 54.

HAZRATI, O.; LEE, J.; LOIZOU, P. C. Binary mask estimation for improved speech intelligibility in reverberant environments. In: *INTERSPEECH*. ISCA, 2012. p. 162–165. Disponível em: <http://dblp.uni-trier.de/db/conf/interspeech/interspeech2012.html#HazratiLL12>. Citado 4 vezes nas páginas 8, 17, 40 e 43.

HU, Y.; LOIZOU, P. C. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 16, n. 1, p. 229–238, Jan 2008. ISSN 1558-7916. Citado na página 45.

IMAI, S. Cepstral analysis synthesis on the mel frequency scale. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. [S.l.: s.n.], 1983. v. 8, p. 93–96. Citado na página 33.

JEUB, M.; NELKE, C.; BEAUGEANT, C.; VARY, P. Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals. In: *2011 19th European Signal Processing Conference*. [S.l.: s.n.], 2011. p. 1347–1351. ISSN 2076-1465. Citado na página 22.

JEUB, M.; SCHAFER, M.; VARY, P. A binaural room impulse response database for the evaluation of dereverberation algorithms. In: *2009 16th International Conference on Digital Signal Processing*. [S.l.: s.n.], 2009. p. 1–5. ISSN 1546-1874. Citado 7 vezes nas páginas 18, 24, 30, 36, 49, 56 e 57.

JIN, Q.; SCHULTZ, T.; WAIBEL, A. Far-field speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 15, n. 7, p. 2023–2032, Sept 2007. ISSN 1558-7916. Citado na página 21.

JOHANNESMA, P. I. M. The pre-response stimulus ensemble of neurons in the cochlear nucleus. In: . [S.l.: s.n.], 1972. p. 58–69. Citado 4 vezes nas páginas 24, 41, 42 e 54.

KAILATH, T. The divergence and bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology*, v. 15, n. 1, p. 52–60, February 1967. ISSN 0018-9332. Citado na página 25.

KATES, J. M.; AREHART, K. H. Coherence and the speech intelligibility index. *The Journal of the Acoustical Society of America*, v. 117, n. 4, p. 2224–2237, 2005. Disponível em: <https://doi.org/10.1121/1.1862575>. Citado 2 vezes nas páginas 40 e 45.

KOKKINAKIS, K.; HAZRATI, O.; LOIZOU, P. A channel-selection criterion for suppressing reverberation in cochlear implants. *Journal of the Acoustic Society of America*, v. 129, n. 5, p. 3221–3232, may 2011. Citado 3 vezes nas páginas 17, 40 e 42.

KRIM, H.; VIBERG, M. Two decades of array signal processing research: the parametric approach. *IEEE Signal Processing Magazine*, v. 13, n. 4, p. 67–94, Jul 1996. ISSN 1053-5888. Citado 2 vezes nas páginas 16 e 40.

KRYTER, K. Methods for the calculation and use of the articulation index. *Journal of the Acoustic Society of America*, v. 34, n. 11, p. 1689–1697, november 1962. Citado na página 46.

LEBART, K.; BOUCHER, J.; DENBIGH, P. A new method based on spectral subtraction for speech dereverberation. *Acta Acustica*, v. 87, p. 359–366, 2001. Citado na página 22.

LEE, T.-W. *Independent Component Analysis: Theory and Applications*. boston, USA: Springer, 1998. ISBN 978-0792382614. Citado 2 vezes nas páginas 16 e 40.

LI, N.; LOIZOU, P. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *Journal of the Acoustic Society of America*, v. 123, n. 3, p. 1673–1682, march 2007. Citado na página 42.

LIM, J. *Speech enhancement*. Englewood Cliffs, NJ, USA: Prentice Hall, 1983. ISBN 978-0-7803-3449-6. Citado na página 16.

LOIZOU, P.; HU, Y. Extending the articulation index to account for non-linear distortions introduced by noise-suppression algorithms. *The Journal of the Acoustical Society of America*, v. 130, n. 2, p. 986–995, 2011. Citado na página 46.

LOIZOU, P. C. *Speech Enhancement: Theory and Practice*. 2nd. ed. Boca Raton, FL, USA: CRC Press, Inc., 2013. ISBN 1466504218, 9781466504219. Citado 4 vezes nas páginas 8, 16, 40 e 41.

LOLLMANN, H. W.; VARY, P. A blind speech enhancement algorithm for the suppression of late reverberation and noise. In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2009. p. 3989–3992. ISSN 1520-6149. Citado na página 23.

NÁBĚLEK, A. K.; LETOWSKI, T. R.; TUCKER, F. M. Reverberant overlap- and self-masking in consonant identification. *The Journal of the Acoustical Society of America*, v. 86, n. 4, p. 1259–65, oct 1989. ISSN 0001-4966. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/2808901>. Citado na página 22.

NÁBĚLEK, A. K.; ROBINSON, P. K. Monaural and binaural speech perception in reverberation for listeners of various ages. *The Journal of the Acoustical Society of America*, v. 71, n. 5, p. 1242–1248, 1982. Disponível em: <https://doi.org/10.1121/1.387773>. Citado 3 vezes nas páginas 44, 46 e 47.

NARAYANAN, A.; WANG, D. The role of binary mask patterns in automatic speech recognition in background noise. *The Journal of the Acoustical Society of America*, v. 133, n. 5, p. 3083–3093, 2013. Disponível em: <https://doi.org/10.1121/1.4798661>. Citado 2 vezes nas páginas 17 e 54.

OSHAUGHNESSY, D. *Speech Communications: Human and Machine*. [S.l.]: Addison-Wesley Publishing Co., 1987. Citado na página 32.

O'SHAUGHNESSY, D. D. *Speech communications - human and machine, 2nd Edition*. [S.l.]: IEEE, 2000. ISBN 978-0-7803-3449-6. Citado 3 vezes nas páginas 16, 24 e 49.

PAN, Y.; WAIBEL, A. The effects of room acoustics on MFCC speech parameter. In: *Sixth International Conference on Spoken Language*. [s.n.], 2000. p. 129–132. ISBN 7801501144. Disponível em: <https://pdfs.semanticscholar.org/957f/ca51926fda953acbc61a0afeca53fd00e457.pdf>. Citado na página 21.

PATTERSON, R. D.; MOORE, B. C. J. Auditory filters and excitation patterns as representations of frequency resolution. *Frequency selectivity in hearing*, p. 123–177, 1986. Citado 4 vezes nas páginas 24, 41, 42 e 54.

PEER, I.; RAFAELY, B.; ZIGEL, Y. Reverberation matching for speaker recognition. In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2008. p. 4829–4832. ISSN 1520-6149. Citado na página 21.

QUACKENBUSH, S. R.; BARNWELL, T. P.; CLEMENTS, M. A. *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988. Citado na página 45.

QUATIERI, T. *Discrete-time Speech Signal Processing: Principles and Practice*. Prentice Hall PTR, 2002. (Prentice-Hall signal processing series). ISBN 9780132429429. Disponível em: <https://books.google.com.br/books?id=5KYeAQAAIAAJ>. Citado na página 42.

RABELO, A. T. V.; SANTOS, J. N.; OLIVEIRA, R. C.; MAGALHAES, M. d. C. Effect of classroom acoustics on the speech intelligibility of students. *CoDAS*, scielo, v. 26, p. 360–366, october 2014. ISSN 2317-1782. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2317-17822014000500360&nrm=iso>. Citado na página 16.

RABINER, L.; JUANG, B. *Fundamentals of Speech Recognition*. [S.l.]: Prentice Hall, 1993. Citado na página 32.

REYNOLDS, D.; ROSE, R. Robust text independent speaker identification using gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing*, v. 3, p. 72–82, 1995. Citado 2 vezes nas páginas 33 e 34.

RHEBERGENT, K.; VERSFELD, N. A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, v. 117, n. 4, p. 2181–2192, 2005. Citado na página 46.

RIX, A. W.; BEERENDS, J. G.; HOLLIER, M. P.; HEKSTRA, A. P. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In: *Proceedings of the Acoustics, Speech, and Signal Processing, 200. On IEEE International Conference - Volume 02*. Washington, DC, USA: IEEE Computer Society, 2001. (ICASSP '01), p. 749–752. ISBN 0-7803-7041-4. Disponível em: <http://dx.doi.org/10.1109/ICASSP.2001.941023>. Citado na página 45.

ROSE, R.; HOFSTETTER, E.; REYNOLDS, D. Integrated models of signal and background with application to speaker identification in noise. *IEEE Transactions on Speech and Audio Processing*, v. 2, n. 2, p. 245–257, abr. 1994. Citado na página 32.

SADJADI, S. O.; HANSEN, J. H. L. Blind spectral weighting for robust speaker identification under reverberation mismatch. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 22, n. 5, p. 937–945, May 2014. ISSN 2329-9290. Citado na página 21.

SANT'ANA, R.; COELHO, R.; ALCAIM, A. Text-independent speaker recognition based on the hurst parameter and the multidimensional fractional brownian motion model. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 14, n. 3, p. 931–940, May 2006. ISSN 1558-7916. Citado na página 65.

SCHROEDER, M. R.; GOTTLOB, D.; SIEBRASSE, K. F. Comparative study of european concert halls: correlation of subjective preference with geometric and acoustic parameters. *The Journal of the Acoustical Society of America*, v. 56, n. 4, p. 1195–1201, 1974. Disponível em: <https://doi.org/10.1121/1.1903408>. Citado na página 16.

SHAO, Y.; SRINIVASAN, S.; WANG, D. Incorporating auditory feature uncertainties in robust speaker identification. In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*. [S.l.: s.n.], 2007. v. 4, p. IV–277–IV–280. ISSN 1520-6149. Citado 2 vezes nas páginas 54 e 55.

SHAO, Y.; WANG, D. Robust speaker recognition using binary time-frequency masks. In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. [S.l.: s.n.], 2006. v. 1, p. I–I. ISSN 1520-6149. Citado na página 54.

STEENEKEN, H. J. M.; HOUTGAST, T. A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America*, v. 67, n. 1, 1980. Citado na página 46.

TAAL, C. H.; HENDRIKS, R. C.; HEUSDENS, R.; JENSEN, J. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 19, n. 7, p. 2125–2136, september 2011. ISSN 1558-7916. Citado 4 vezes nas páginas 16, 40, 45 e 46.

TAVARES, R.; COELHO, R. Speech enhancement with nonstationary acoustic noise detection in time domain. *IEEE Signal Processing Letters*, v. 23, n. 1, p. 6–10, Jan 2016. ISSN 1070-9908. Citado 2 vezes nas páginas 16 e 40.

TRAER, J.; MCDERMOTT, J. H. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, National Academy of Sciences, v. 113, n. 48, p. E7856–E7865, 2016. ISSN 0027-8424. Disponível em: <http://www.pnas.org/content/113/48/E7856>. Citado na página 37.

VEEN, B. D. V.; BUCKLEY, K. M. Beamforming: a versatile approach to spatial filtering. *IEEE ASSP Magazine*, v. 5, n. 2, p. 4–24, April 1988. ISSN 0740-7467. Citado 2 vezes nas páginas 16 e 40.

VENTURINI, A.; ZÃO, L.; COELHO, R. On speech features fusion, α-integration gaussian modeling and multi-style training for noise robust speaker classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 22, n. 12, p. 1951–1964, Dec 2014. ISSN 2329-9290. Citado na página 55.

VORLÄNDER, M. *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. Springer Berlin Heidelberg, 2007. (RWTHedition). ISBN 9783540488309. Disponível em: <https://books.google.com.br/books?id=CuXF3JkTuhAC>. Citado na página 27.

WANG, D. On ideal binary mask as the computational goal of auditory scene analysis. In: _____. *Speech Separation by Humans and Machines*. Boston, MA: Springer US, 2005. p. 181–197. ISBN 978-0-387-22794-8. Disponível em: <https://doi.org/10.1007/0-387-22794-6_12>. Citado 3 vezes nas páginas 16, 40 e 41.

WANG, L.; ODANI, K.; KAI, A. Dereverberation and denoising based on generalized spectral subtraction by multi-channel lms algorithm using a small-scale microphone array. *EURASIP Journal on Advances in Signal Processing*, v. 2012, n. 1, p. 12, Jan 2012. ISSN 1687-6180. Disponível em: <https://doi.org/10.1186/1687-6180-2012-12>. Citado na página 21.

WATKINS, A. J. Perceptual compensation for effects of reverberation in speech identification. *The Journal of the Acoustical Society of America*, v. 118, n. 1, p. 249–262, 2005. Disponível em: <https://doi.org/10.1121/1.1923369>. Citado na página 37.

WEN, J.; GAUBITCH, N.; HABETS, E.; MYATT, T.; NAYLOR, P. Evaluation of speech dereverberation algorithms using the mardy database. In: *2006 International Workshop on Acoustic Echo and Noise Control*. [S.l.: s.n.], 2006. Citado 7 vezes nas páginas 18, 24, 26, 36, 49, 56 e 57.

WU, D.; LI, J.; WU, H. α-gaussian mixture modelling for speaker recognition. *Pattern Recognition Letters*, v. 30, n. 6, p. 589 – 594, 2009. ISSN 0167-8655. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167865509000038>. Citado na página 55.

ZHAO, X.; SHAO, Y.; WANG, D. Casa-based robust speaker identification. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 20, n. 5, p. 1608–1616, July 2012. ISSN 1558-7916. Citado na página 54.

ZHAO, X.; WANG, Y.; WANG, D. Robust speaker identification in noisy and reverberant conditions. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2014. p. 3997–4001. ISSN 1520-6149. Citado na página 17.

ZÃO, L.; COELHO, R.; FLANDRIN, P. Speech enhancement with emd and hurst-based mode selection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 22, n. 5, p. 899–911, May 2014. ISSN 2329-9290. Citado 2 vezes nas páginas 16 e 40.