6
Projeto de um Sistema de Reconhecimento de Voz Contínua através de Técnicas de pós-extração de Atributos

Visando aumentar a robustez dos sistemas de reconhecimento, foram analisados e testados no capítulo 5, métodos de *realce de voz*, os quais estimaram a fala limpa a partir do sinal degradado. Esses métodos foram denominados de *pré-extração de atributos* e têm como principal função agir sobre o sinal de voz, ou seja, antes da parametrização do sinal.

Neste capítulo descrevem-se dois métodos baseados na *compensação de atributos*, a fim de aumentar ainda mais a robustez do reconhecedor. A ideia destes métodos é restaurar o sinal depois de sua parametrização, ou seja, compensar os atributos do sinal de voz formando um novo vetor restaurado, o qual será fornecido diretamente ao reconhecedor.

A seguir, serão analisados os métodos propostos neste projeto e se avaliará experimentalmente sua eficiência de robustez frente ao ruído.

Cabe salientar que a avaliação dos métodos utilizará os mesmos bancos de dados e configurações de treino e teste do capitulo 5.

6.1
Mapeamento de Histogramas

A corrupção do sinal de voz causada pelo ruído aditivo degrada o rendimento dos sistemas de reconhecimento, devido à distorção do espaço de representação, modificando as médias e as variâncias das gaussianas utilizadas para representar o sinal limpo (Fig. 6.1). Isso gera um descasamento entre as condições de treinamento e de teste, diminuindo assim as taxas de reconhecimento.

Um método que compensa as características não lineares dos coeficientes cepstrais produzidas pelo ruído e que provocam esses descasamentos é o Mapeamento de Histograma (MAP).

O MAP[51] [62] [63] está baseado na técnica conhecida em processamento de imagens como Equalização de Histogramas [6], a qual tem por objetivo estabelecer uma transformação não linear que transforme o histograma da imagem

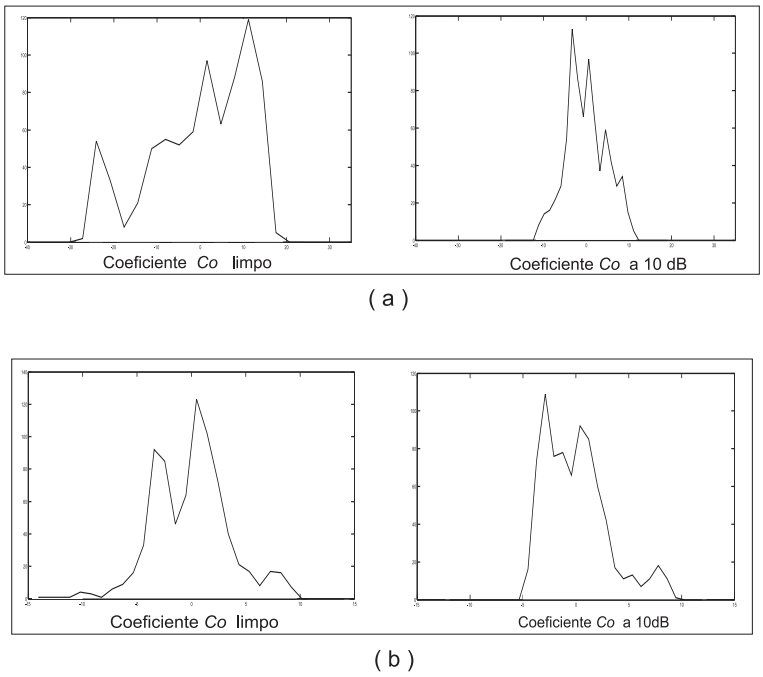

Figura 6.1: Distorção do espaço de representação com ruído branco a 10 db. *(a)* MFCC *(b)* PNCC.

original em histograma de referência, melhorando o brilho e o contraste, otimizando a faixa dinâmica da escala de cinzas.

A equalização de histogramas no reconhecimento de voz foi proposta por Balchandran e Mammone em 1998, onde, através de transformações não lineares dos parâmetros da voz, tanto no treinamento como no reconhecimento, procurou-se ajustar os parâmetros dentro de uma faixa comum, a fim de remover as distorções não lineares no cepstrum LPC de um sistema de identificação de locutor usando como distribuição de referência os dados de treinamento limpos. Assim, o reconhecimento é levado para um domínio onde os dados são idealmente invulneráveis às transformações lineares e não lineares que o ruído aditivo possa provocar.

6.1.1 MAP dos parâmetros da voz

O mapeamento de histogramas é um método realizado sobre os vetores de atributos, tanto na fase de treinamento como no reconhecimento, onde cada componente do vetor é mapeada independentemente. Este processo é feito da seguinte forma:

Considere-se um sinal de voz composto por N vetores, um por cada quadro, cada um deles composto por $C(n)$ coeficientes cepstrais. Além disso, cada coeficiente n-ésimo composto como um fluxo de valores independentes,

ou seja, $c(n)^i, i = 1, 2, ..., L$, onde L é o número de amostras por quadro. O objetivo do MAP, é obter para cada componente n do vetor $c(n)$, um histograma que representará uma aproximação discreta da *fdp* daquele fluxo de valores, denotada por $f(y_n)$, a fim de mapeá-la para uma *fdp* de referência $g(x_n)$ correspondente à voz limpa, de modo que sejam removidos os efeitos que o canal e o ruído introduziram na função distribuição original. Isto é, fazer uma transformação não linear, que transforme a função densidade de probabilidade da voz contaminada, em uma função densidade de probabilidade de referência, como é mostrado na Fig.6.2, eliminando assim os efeitos do ruído e proporcionando uma estimativa da voz limpa a partir de cada observação da voz ruidosa.

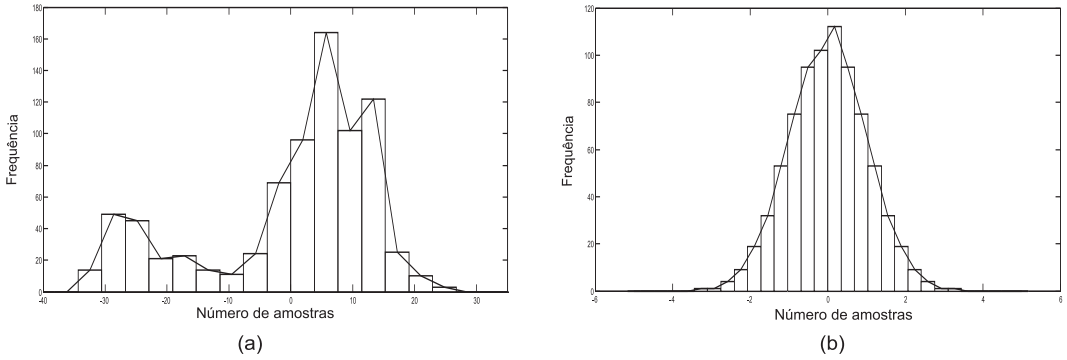

Figura 6.2: Mapeamento de histogramas do coeficiente C_0 dos atributos MFCC *(a) fdp* do coeficiente cepstral original *(b) fdp* do coeficiente cepstral mapeado.

Esta transformação obtém-se a partir dos histogramas cumulativos dos vetores contaminados $CDF_f(y_n)$ e os histogramas cumulativos de referência correspondentes à *fdp* gaussiana de referência $CDF_g(x_n)$

$$CDF_g(x_n) = CDF_f(y_n) \tag{6-1}$$

$$x_n = CDF_g^{-1}(CDF_f(y_n)) \tag{6-2}$$

onde CDF_g^{-1} representa a função inversa do histograma cumulativo de referência $CDF_g(x_n)$. O processo da transformação é detalhado em [64].

De acordo com o exposto acima, o valor cepstral $c(n) = y_0$ mapeado ao valor de referência g_0 é feito de modo a manter os histogramas cumulativos, ou seja,

$$\int_{y=-\infty}^{y_0} f(y)\,dy = \int_{x=-\infty}^{x_0} g(x)dx \tag{6-3}$$

Porém, esta estimativa da transformação a partir dos histogramas cumulativos pode ser tediosa e demorar mais tempo. Uma forma mais simples e computacionalmente mais eficiente para obter a transformação é através da estatística ordenada [65].

6.1.2 MAP através da estatística ordenada

Uma forma mais eficaz de realizar a transformação é estabelecer uma tabela de pontos de busca.

Seja L o comprimento do fluxo de valores independentes de $c(n)$

$$c(n) = \{c(n)^1, \quad c(n)^2, \quad c(n)^3, \quad ..., \quad c(n)^L\} \tag{6-4}$$

Este valor é reordenado de forma descendente (Equação 6-5), a fim de criar uma tabela de busca ordenada, onde cada valor de $c(n)$ é etiquetado por um índice i de ordenamento, assim ao maior valor de $c(n)$ corresponde-lhe o índice $i(1)$ e ao menor o índice $i(L)$, ou seja,

$$c(n)^{i(1)} \geq c(n)^{i(2)} \geq c(n)^{i(3)} \geq, ..., \geq c(n)^{i(L)} \tag{6-5}$$

Uma vez etiquetados e ordenados todos os valores de $c(n)$ na tabela, pode-se determinar facilmente o valor ao qual se mapeia o coeficiente cepstral n-ésimo do ponto a da Fig. 6.3, simplesmente calculando o índice que ocupa esse coeficiente dentro da tabela.

A tabela de busca através da qual pode se obter o estimador puntual do valor da função $g(x)$ é construída a partir da seguinte equação

$$\int_{x=-\infty}^{x_0} g(x)dx = \frac{L + 0,5 - i}{L} \tag{6-6}$$

onde $g(x)$ é a função densidade de probabilidade de referência e x_0 o valor mapeado correspondente ao índice i da tabela de busca.

A partir da equação 6-6 pode se obter o valor transformado de x_0, isto é

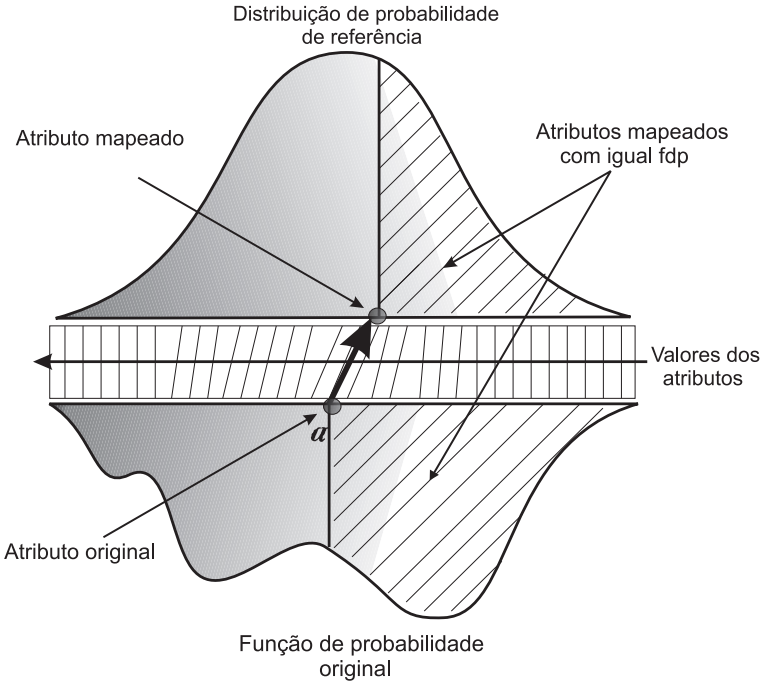

Figura 6.3: Processo de mapeado por através da estatística ordenada.

$$x_0 = g(x)^{-1}\left(\frac{L + 0.5 - i}{L}\right) \tag{6-7}$$

onde $g(x)^{-1}$ representa a função referência inversa, obtendo assim o valor mapeado x_0 por indexação simples.

6.1.3 MAP através de função de referência

No processo de mapeamento, a função de referência $g(x)$ tem o papel mais importante, já que sua fdp representa as estatísticas globais da voz. Uma característica importante que deve ter esta função é ser inversível.

Segundo [66], a função de referência gaussiana de média zero e variância um, tem uma vantagem fundamental frente a outras funções, devido ao fato de que na maior parte dos sistemas de reconhecimento, as distribuições de saída dos HMM serão modeladas com mistura de gaussianas.

Considerando isso, neste trabalho decidiu-se realizar o mapeamento através de funções de referência gaussiana com função densidade de probabilidade dada por

$$g(x) = \frac{1}{\sqrt{2\Pi}} exp\left(\frac{-x^2}{2}\right) \tag{6-8}$$

6.2 Filtro com Redes Neurais

De acordo ao mencionado anteriormente, a relação de mapeamento entre os espaços dos atributos dos sinais degradados e os dos sinais limpos precisa ser não linear. Por isso, aproveitando a complexidade operacional das redes neurais (operações não lineares), propõe-se um outro método para transformar atributos corrompidos em atributos limpos, através de uma filtragem das informações indesejáveis depois da extração de atributos, permitindo assim o mapeio direto entre os espaços de forma não linear, o que constituirá um verdadeiro *filtro não linear*.

A seguir, serão detalhados alguns fundamentos das redes neurais, analisando como estas contribuem à compensação de atributos do sistema.

6.2.1 Fundamentos das redes Neurais Artificias

Tentando simular a estrutura e funções do cérebro humano, surgiu um método computacional inspirado nos neurônios biológicos e na estrutura massivamente paralela do cérebro, chamado *redes neurais artificias* [67].

Nas últimas décadas, tem-se dado um particular interesse por este tipo de método, já que oferece os recursos necessários para modelar de maneira eficiente problemas complexos. Mesmo assim, este método computacional não supera a capacidade do cérebro. A Tabela 6.1 apresenta as principais características de cada um deles.

CÉREBRO	REDES NEURAIS ARTIFICIAIS
Neurônio biológico	Neurônio artificial
Rede de neurônios	Estrutura em camadas
10 bilhões neurônios	Centenas/milhares
Aprendizado	Aprendizado
Generalização	Generalização
Associação	Associação
Reconhecimento de padrões	Reconhecimento de padrões

Tabela 6.1: Relação entre o cérebro e as redes neurais artificiais.

Além desta comparação, pode-se estabelecer uma outra relação que compare o processamento da informação do cérebro com os computadores convencionais. As diferenças apresentam-se a seguir.

- Com relação ao tempo de processamento, o cérebro é 1 milhão de vezes mais lento que qualquer "gate" digital, enquanto para um computador, o processamento é extremadamente mais rápido e preciso na execução de sequências de instruções.
- Processamento extremadamente rápido do cérebro no reconhecimento de padrões, enquanto o computador é muito mais lento nesta tarefa.
- O cérebro humano possui um número estimado de 10^{11} a 10^{14} neurônios, cada um com cerca de 10^3 a 10^4 conexões, representando uma unidade completa de computação, enquanto os computadores convencionais possuem até 7 unidades de processamento central.
- O cérebro não possui endereço de memória e não processa a informação sequencialmente, armazenando sob uma forma dispersa e adaptativa, enquanto o computador armazena em um local endereçável.

Desta forma, as redes neurais são utilizadas com a intenção de explorar a característica de paralelismo e processamento altamente distribuído do cérebro.

Na literatura existem diversos tipos de redes neurais, no entanto, todas possuem três características fundamentais:

- Unidade processadora
- Funções de ativação
- Arquitetura de redes neurais

Estas características vão ser descritas a seguir.

Unidade processadora (Neurônio Artificial)

Toda rede neural contém um número considerável de unidades de processamento simples, as quais se assemelham aos neurônios do cérebro. Estas unidades são consideradas as bases estruturais do processo artificial. Sua principal função é receber as unidades vizinhas e calcular seu valor de saída, enviando-o a todas as unidades conectadas. Cada unidade terá um número n de entradas e um número m de saídas, onde cada saída é conectada com camadas posteriores.

A Fig. 6-9 mostra o modelo para um neurônio artificial, no qual as entradas $x_(1)...x_n$ são os estímulos que o neurônio recebe do ambiente e a saída y_k é a resposta a esse estímulo. Além disso, podem se identificar dois elementos básicos do modelo

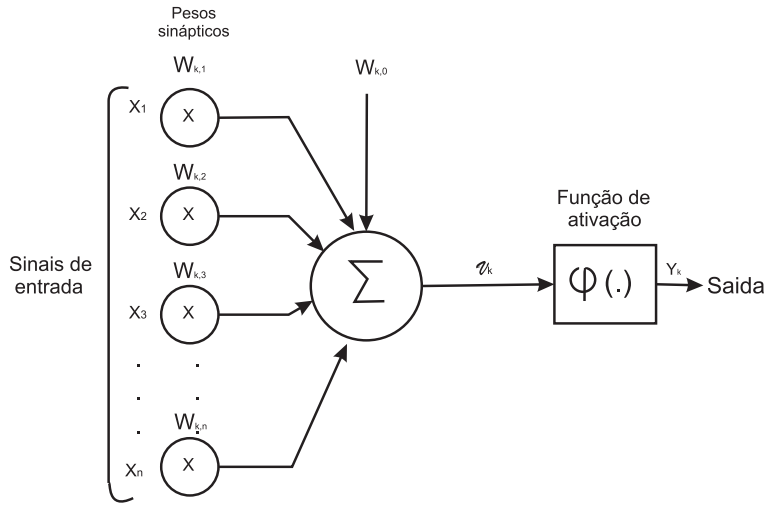

Figura 6.4: Modelo não linear de um neurônio.

– Um conjunto de conexões ou sinapses, em que cada uma delas é caracterizada por seu peso, cujo valor pode ser positivo ou negativo, conforme as sinapses, determinando quando o neurônio deve considerar sinais de disparo que ocorrem naquela conexão. Por exemplo, o sinal x_n na entrada da sinapse n, conectado ao neurônio k, é multiplicado por um peso $w_{k,n}$, onde k é o neurônio em questão e n é a sinapse à qual o peso refere-se.
– Uma função de ativação $\varphi_.$, responsável por ligar a informação de entrada do neurônio com o seguinte estado de ativação que tenha esse neurônio, habilitando ou não a saída, dependendo se o valor da soma ponderada das suas entradas supera o limiar.

Matematicamente, o neurônio artificial é representado por

$$v_k = \sum_{n=1}^{l} w_{kn} x_n \tag{6-9}$$

onde l é o número de entradas do neurônio, $x(n)$ são os sinais de entrada, $w_{k,n}$ são os pesos sinápticos do neurônio k e v_k é o estado de ativação do n-ésimo neurônio.

Função de ativação

A função de ativação é usada para limitar o intervalo de valores da resposta do neurônio. Geralmente os intervalos de valores limitam-se a $[0, 1][-1.1]$, ou seja, estas funções definem as saída y_k de um determinado neurônio de acordo com seu nível de ativação de entrada. Existem diversas funções de

ativação que podem ser aplicadas aos nós para produzir uma saída qualquer. A decisão entre uma ou outra dependerá da aplicação ou o problema a resolver.

Afim de imitar a transmissão de um neurônio biológico, as funções de ativação devem ter um comportamento crescente. Dois exemplos bastante usados são

$$y_k = \frac{2}{1 + e^{-2v_k}} - 1 \tag{6-10}$$

$$y_k = \frac{1}{1 + e^{-v_k}} \tag{6-11}$$

Arquitetura de redes neurais

As redes neurais organizam-se estabelecendo o número de camadas, o tipo de conexão entre nós e como ocorre o fluxo do sinal dentro da rede.

Proporcionais à quantidade de neurônios e de camadas, serão a complexidade e o tempo de processamento da rede. Por exemplo, uma rede de camada única só pode resolver problemas linearmente separáveis, tornando-se sistemas de pouca utilidade. No entanto, se a rede possui mais conexões de neurônios e mais camadas ocultas, o sistema conseguirá representar um comportamento matemático complexo, incluindo os efeitos não-lineares, que no caso do ruído, são os que afetam o reconhecimento.

O uso de um neurônio artificial por si só não é de grande utilidade. É por isso que visando conseguir melhores resultados é preciso definir uma rede que interligue um grupo de neurônios que estão distribuídos entre camadas, fazendo que os nós de uma camada se conectem a cada nó na camada seguinte, formando pontes $W_{(ij)}$ que são elementos de união e propagação dos sinais de entrada da primeira camada até obter uma resposta na camada n da saída.

Uma das arquiteturas mais utilizadas para realizar esta conexão é através das redes neurais *feedforward*[68]. Esta arquitetura, composta por múltiplas camadas de processadores, têm definido um fluxo de dados em uma única direção. Isso quer dizer que a saída de cada neurônio é propagada como uma entrada nos neurônios das camadas seguintes, como mostrado na Fig. 6.2.

Não existe um limite para fixar a quantidade de camadas deste tipo de redes, porém segundo [69], com só uma camada oculta e um certo número de nós, pode ser solucionado a maioria dos problemas.

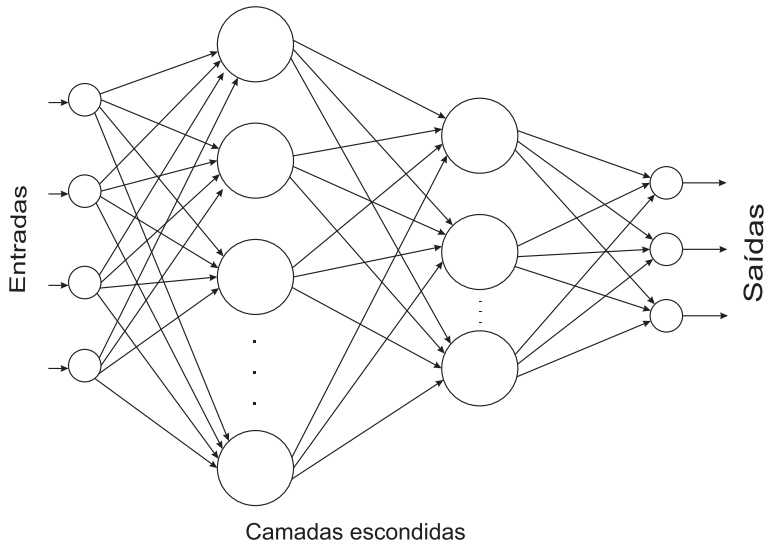

Figura 6.5: Rede neural *feedforward*, com 4 camadas formadas por conexões entre neurônios artificiais.

6.2.2 Filtragem não linear através de redes neurais

Inicialmente, estabelece-se o tipo de rede que seja capaz de transformar qualquer padrão de entrada em qualquer saída desejada. Esta rede é submetida a uma fase de aprendizado, através da qual serão capturadas as características particulares do ambiente, modificando os pesos em cada camada, de tal forma que coincida a saída desejada com a saída obtida, mediante a apresentação de um determinado dado de entrada.

Esses dados são representados por todos os atributos do sinal na entrada, com o fim de capturar todos os atributos limpos na saída, tudo isso em uma única rede. Porém, isto complicaria o treinamento, já que precisaria da restauração do peso w de milhares de conexões ao tempo, tornando a rede complexa no seu desenvolvimento. A partir disso, [59] propõe dividir a rede em várias redes com as mesmas características, com o objetivo de que a entrada de cada rede tenha os atributos de cada quadro, fornecendo só um atributo restaurado daquele quadro na saída.

A Fig. 6.6 apresenta um modelo simplificado do modelo de aprendizado,

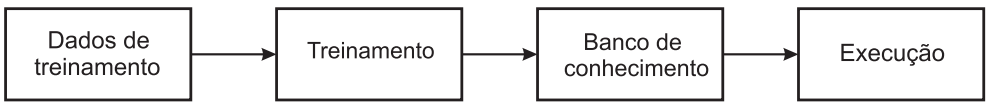

Figura 6.6: Modelo de aprendizado.

onde

- **Dados de treinamento:** são os dados para aprendizagem, tomados do ambiente onde a rede vai se desenvolver. No caso deste trabalho são os atributos dos sinais limpos.
- **Treinamento:** é o método utilizado para que a rede capture as características dos dados. Neste trabalho utilizou-se o treinamento supervisionado da seguinte maneira:
  - Selecionam-se alguns sinais de voz limpos.
  - Esses sinais são corrompidos acrescentando-lhes ruído branco.
  - Os atributos dos sinais corrompidos são calculados e apresentados na entrada das redes.
- **Banco de conhecimentos:** é o resultado do treinamento, onde se armazena o conhecimento da rede, ajustando seus pesos, para que a saída seja mais parecida com o atributo limpo respectivo.

A configuração completa da rede é apresentada na Tabela 6.2

Tipo de rede	*Feedforward*
Tipo de treinamento	Supervisionado
Numero de camadas ocultas	1
Numero de neurônios	10
Função de ativação	*Tansing*

Tabela 6.2: Configuração da rede neural.

6.3 Avaliação das técnicas de pós-extração de atributos

A seguir, descrevem-se os experimentos realizados e os resultados obtidos a partir dos mesmos, a fim de realizar uma avaliação, análise e comparação dos distintos algoritmos de robustez de ruído propostos no sistema de reconhecimento implementado.

Para avaliar estes métodos, utilizaram-se os mesmos bancos de dados e configurações do capítulo 5.

6.3.1 Resultados Experimentais

Nesta seção descrevem-se as experiências feitas para completar com sucesso os métodos que compõem o sistema de *pós-extração de atributos*. Em primeiro lugar, compara-se os resultados obtidos do sistema *Ref* da seção 5.3.3 com os resultados após aplicar cada um dos métodos deste capítulo, a fim de determinar se as melhoras são ou não são significativas. Depois analisa-se a mistura dos métodos apresentados, a fim de determinar se a mistura traz melhoras às taxas de reconhecimento.

Teste das técnicas de pós-extração de atributos

Utilizando os resultados obtidos no sistema *Ref* da seção anterior, se faz um primeiro teste que compara esses resultados com os obtidos depois de utilizar cada um dos métodos de robustez propostos neste capítulo. Os testes foram feitos da seguinte forma:

- *Ref + Filtro com Redes Neurais* (FRN).
- *Ref + Mapeamento de Histogramas* (MAP).

Na Tabela 6.3 são apresentadas as taxas de acerto do reconhecimento para cada um dos métodos de robustez propostos para diferentes relações sinal ruído.

SNR[dB]	**MFCC**			**PNCC**		
	Ref	**Ref + FRN**	**Ref + MAP**	**Ref**	**Ref + FRN**	**Ref + MAP**
limpo	89,42%	57,68%	93,97%	86,58%	77,36%	89,19%
15db	34,47%	59,50%	82,37%	80,89%	71,56%	83,50%
10db	7,85%	39,93%	64,62%	74,52%	61,55%	78,73%

Tabela 6.3: Taxas de acerto utilizando as técnicas pós-extração de atributos.

A partir da Tabela 6.3 pode-se observar o efeito positivo dos métodos de robustez FRN e MAP quando são utilizados após a extração de atributos MFCC, melhorando as taxas de acerto para amostras corrompidas com ruído branco em diferente razões sinal-ruído, provando ser eficazes na redução de efeitos aditivos do canal.

Observando os resultados dos métodos de compensação, pode-se ver que a aplicação do MAP melhora significativamente as taxas de acerto do reconhecedor em comparação ao sistema *Ref* e ao método FRN, para todos os níveis de SNR.

No entanto, o ganho de desempenho não é consistente para todos os métodos de parametrização . No caso dos atributos PNCC as taxas de acerto são sempre inferiores às do sistema baseline quando é usado o método FRN.

Por outro lado, as taxas de acerto do sistema MAP com atributos PNCC mostrou a maior robustez, com ganhos significativos nos cenários mais ruidosos, mostrando assim a melhora que introduz o MAP sobre o sistema de reconhecimento.

Teste misturando as técnicas de pós-extração de atributos

A partir da Motivação em ralação ao bom desempenho dos métodos testados individualmente, fez-se um último teste, onde serão analisados os resultados concatenando os dois métodos de robustez na seguinte ordem.

– *Ref* + *Mapeamento de Histogramas* (MAP) + *Filtro com Redes Neurais* (FRN)

A Tabela 6.4 apresenta os resultados obtidos após de misturar os métodos.

SNR[dB]	**MFCC**		**PNCC**	
	Ref	**Ref + MAP + FRN**	**Ref**	**Ref + MAP + FRN**
limpo	89,42%	83,62%	86,58%	82.03%
15db	34,47%	84,76%	80,89%	78.83%
10db	7,85%	76,22%	74,52%	74.29%

Tabela 6.4: Taxas de acerto utilizando a mistura das técnicas pós-extração de atributos.

Os resultados correspondentes para este teste têm como objetivo averiguar se a mistura dos métodos traz melhoras nas taxas de reconhecimento quando comparada aos testes anteriores. Infelizmente, nos atributos PNCC as taxas de reconhecimento foram inferiores ao sistema *Ref*, mesmo como apresentado na Tabela 6.3. Isto pode indicar que embora o erro seja menor, o FRN remove o ruído à custa de distorcer de alguma forma o sinal limpo. Porém, pode-se ver como o uso do MAP em combinação com FRN melhorou as taxas de acerto em comparação com o FRN puro da Tabela 6.3.

Já para os atributos MFCC, pode-se ver como a combinação de métodos estabelece uma melhora considerável nas taxas de acerto do reconhecedor, mesmo comparando com a Tabela 6.3. No caso do cenário de maior ruído, a combinação apresentou melhores resultados do sistema do que a utilização dos métodos individualmente.

Conclusões dos testes

Este projeto discute duas formas diferentes de melhorar o desempenho do reconhecimento de voz contínua na presença do ruído aditivo através de métodos de compensação denominados *pós-extração de atributos*.

Inicialmente, avaliou-se o *Mapeamento de histogramas*. Este método, quando é usado com atributos MFCC, apresenta resultados com melhoras consideráveis no rendimento do reconhecimento respeito ao sistema *Ref* e aos outros métodos de realce de fala apresentados no capitulo 5, confirmando os testes de outros autores. No entanto, quando é utilizado com atributos PNCC, os resultados não foram muito beneficiados, embora suas taxas tenham se mantido as mais altas.

Em seguida foi testado o *Filtro com redes neurais* que, de acordo ao mencionado, este método permite o mapeamento direto entre os espaços dos atributos do sinal degradado e do sinal limpo, o que constitui um verdadeiro filtro não linear. Porém, os resultados tiveram um desempenho inferior, devido ao fato de que as redes neurais precisam levar em conta muitos fatores para seu projeto e de muitas provas para construir uma configuração de rede que apresente um bom desempenho, já que a maior parte de procedimentos usados para seu treinamento são lentos e muito prováveis a cair em mínimos locais, gerando assim os resultados pouco satisfatórios como pode-se ver nas tabelas acima apresentadas.

O mais notável dos experimentos foi comprovar que independentemente da técnica de extração de atributos que seja utilizada, o método de compensação MAP mostrou diferenças significativas nas taxas de reconhecimento, proporcionando melhoras importantes no rendimento do sistema devido ao fato de ter a capacidade de compensar os efeitos não lineares causados pelo ruído.