

# Reconhecimento de Locutor via Rede Neural Artificial

Breno M. C. C. e Souza  
breno.ec@gmail.com

Rondinelli L. Jorge  
rondyleonardo@gmail.com


CEFET-MG

## Resumo

*Breve discussão sobre reconhecimento de locutor. Abordagem via rede neural artificial para reconhecimento de locutor através de trechos de curta duração de áudio de entrada. Sinal de entrada é tratado com finalidade de remover dados não pertinentes, bem como reduzir a complexidade da rede neural que fará a classificação. A entrada da rede neural é o espectrograma do sinal de entrada, adquirido através da transformada de Fourier do áudio tratado. A rede estudada mapeia subespaços não lineares e descontínuos dos possíveis valores de entrada. O desenvolvimento da rede, seu desempenho e possível aplicabilidade em cenários reais são avaliados. O tempo de treinamento da rede analisada é da ordem de 10 segundos, sendo que a classificação de todas as amostras ocorre em tempo inferior a 1 segundo. A rede neural avaliada classificou corretamente todas as amostras de áudios de entrada.*

**Palavras-chave:** reconhecimento de locutor, redes neurais artificiais, classificação.

## I. INTRODUÇÃO

 reconhecimento de locutor por meio automático é um campo de estudo tecnológico que desde o surgimento das máquinas, sempre fez parte das aspirações humanas. Isto se deve às facilidades inerentes a sistemas capazes de obter êxito nessa tarefa, tais como fazer com que um computador identifique seu dono através da fala, o controle de e robôs e máquinas em uma linha de montagem, sem a necessidade de entradas codificadas ou linguagens específicas e de difícil acesso aos seres humanos.

As redes neurais artificiais (RNAs) são aliadas neste sentido, pois através delas é possível implementar modelos de reconhecimento de locutor e tornar viável inúmeras outras comodidades resultantes deste tipo de aplicação. Existem alguns registros na literatura sobre aplicações para as técnicas de reconhecimento de locutor como, por exemplo, o auxílio a portadores de deficiência [1] [2], aplicações na prática forense [3] e automação de ambientes [4].

Por outro lado, verifica-se a inexistência de grande diversidade na literatura da área, visto que os sistemas que permitem uma comunica-

ção mais natural entre homem e máquina ainda não estejam plenamente dominados [1].

Esse trabalho busca demonstrar através de uma abordagem prática, a aplicação de reconhecimento de locutor através de uma rede projetada e analisada para este fim. Para isto, serão selecionados pequenos trechos de áudios tratados e encaminhados à rede para que sejam classificados pela mesma.

A rede neural artificial em questão deve ser capaz de reconhecer a voz de determinado locutor dentre um número de locutores aleatórios. Ruídos serão incrementados aos áudios a fim de oferecer maior poder de generalização. A soma ruído e áudio passa pelo mesmo processo de tratamento que o áudio original.

O restante do artigo está organizado de forma direta, com metodologia, resultados, ameaças à validade e conclusão.

A seção 2 aborda com detalhes a metodologia de trabalho, com ênfase na aquisição e manipulação dos dados de entrada. Se expõe como que informações referentes aos áudios de entrada são utilizadas como entrada da rede projetada. A arquitetura da rede é brevemente discutida.

## II. METODOLOGIA

Foram selecionadas 10 pessoas (tabela 1) para fazer gravações e nos fornecer dados com a finalidade de estudar o tema e propor uma rede neural artificial de classificação para identificar os locutores. Uma vez caracterizada o conjunto de pessoas, a identidade de cada locutor não é importante. Eles serão, pois, enumerados de 1 a 10.

Um conjunto de 10 frases curtas foram selecionadas para que cada locutor as proferisse (tabela 2), totalizando 100 amostras: 10 indivíduos e 10 frases. Cada frase contém 8 sílabas e o conjunto procurou explorar diferentes fonemas da Língua Portuguesa.

Com o intuito de aprimorar a capacidade de generalização da rede projetada e de refletir a impossibilidade de controlar o meio no qual se grava o áudio de entrada em aplicações reais, 5 ruídos são adicionados às amostras. A base de amostras contém 600 elementos (equação 1); portanto.

$$m = s + r \times a, \quad (1)$$

onde  $s = 100$  é o número de amostras sem adição ruído,  $r = 5$  é o número de elementos da base de ruídos e  $m = 600$  é o número total de amostras. A base final contém 100 amostras sem adição ruído e 500 amostras com adição de ruído: um total de 600 amostras.

É desejável que o processo de reconhecimento de locutor não dependa do texto veiculado e, na medida do possível, do tempo de locução. Com a finalidade de remover os aspectos temporais e extrair parte da informação vocal, aplicamos transformadas de Fourier aos sinais de áudio. O algoritmo utilizado foi o *Fast Fourier Transform* (FFT), variante discreta e eficiente da transformada de Fourier.

A taxa de amostragem de gravação escolhida foi de 44100 Hz. Essa escolha, arbitrária, se baseia no fato de que essa taxa de amostragem é a mais comum e acessível, popularizada por ser a taxa de amostragem utilizada em CDs de áudio.

Número	Sexo	Idade
1	M	29
2	M	26
3	F	22
4	M	23
5	F	29
6	M	34
7	F	25
8	M	62
9	M	28
10	F	34

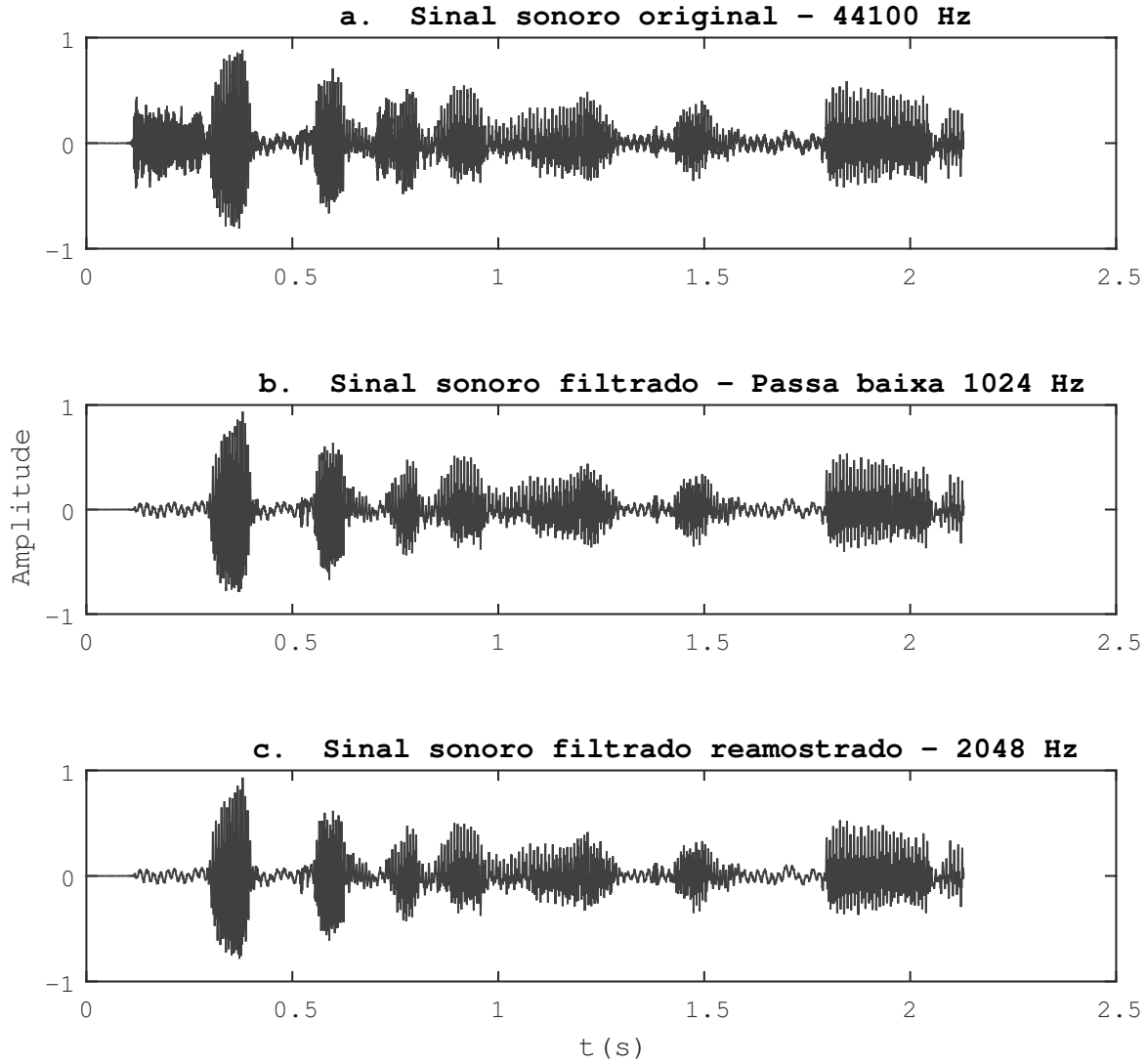
**Tabela 1:** Conjunto de 10 frases gravadas por cada indivíduo. São frases curtas, de 8 sílabas, que procuram explorar a riqueza dos fonemas da Língua Portuguesa. As gravações foram normalizadas com atenuação de 1 Db. As gravações foram editadas para apresentar 2 segundos de duração, cada.

Número	Frase
1	Hoje ela mexe muito
2	Feliz aniversário
3	Minha mãe gosta de uva
4	Suco de limão com pera
5	Minha axila é limpa
6	Eu tomo chá de gengibre
7	O treino da rede neural
8	Envelheço na cidade
9	Isso é engenharia
10	Base de frases gravadas

**Tabela 2:** Conjunto de 10 frases gravadas por cada indivíduo. São frases curtas, de 8 sílabas, que procuram explorar a riqueza dos fonemas da Língua Portuguesa. As gravações foram normalizadas com atenuação de 1 Db. As gravações foram editadas para apresentar 2 segundos de duração, cada.

Número	Ruído
1	Trânsito
2	Escritório
3	Construção
4	Chuva
5	Vento

**Tabela 3:** Conjunto de 5 ruídos adicionados às gravações. Duração de 2 segundos, cada. Sinais de ruído foram normalizados com atenuação em 20 Db — cerca de 10 vezes menos intensos do que as gravações.

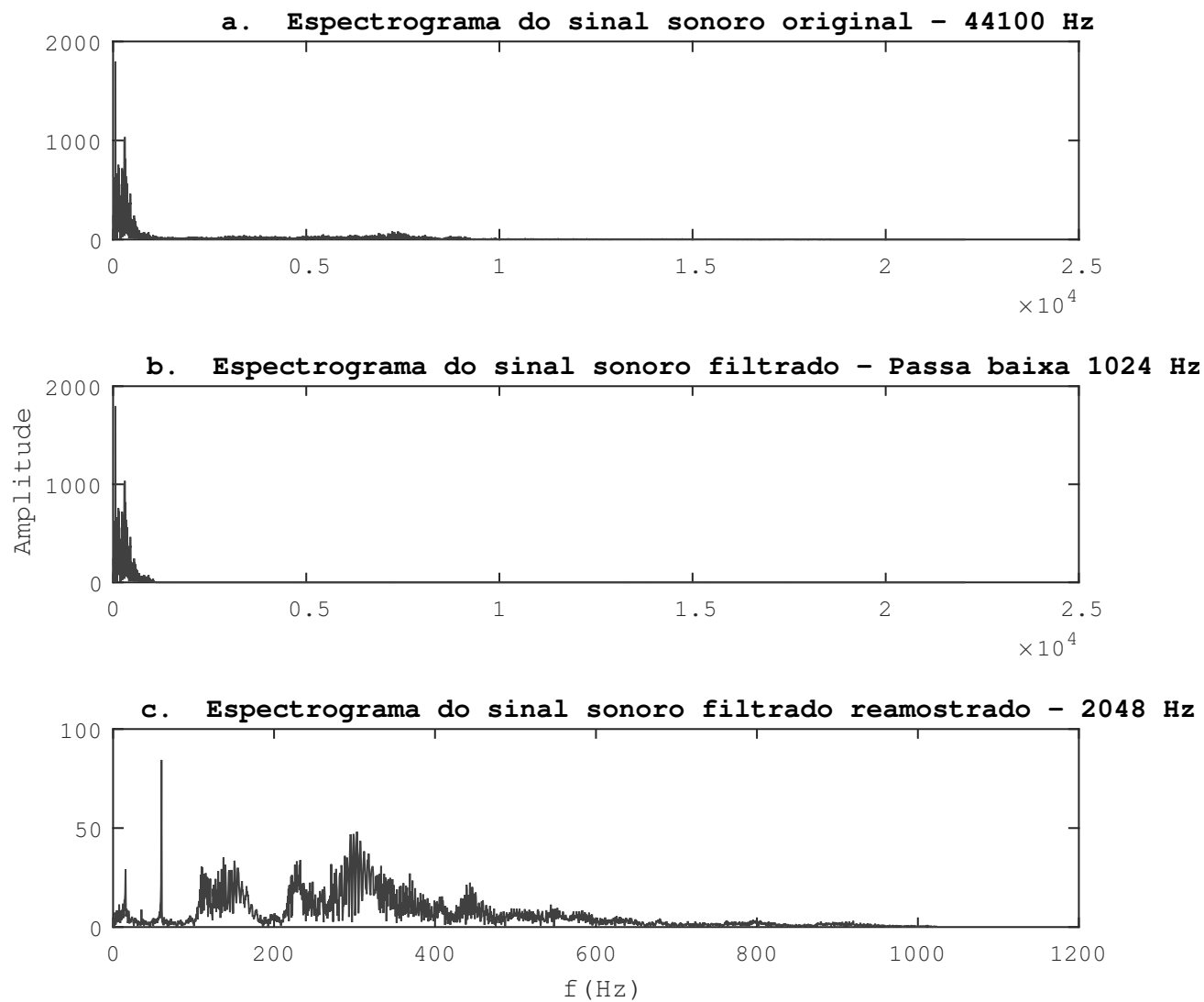


**Figura 1:** Manipulações realizadas em um sinal sonoro.

*a. Sinal original com taxa de amostragem de 44100 Hz.*

*b. Sinal filtrado através de Chebyshev Tipo II passa baixa [1024; 1152] Hz, com atenuação de 1 Db para frequência inferiores a 1024 Hz e atenuação de 60 Db — atenuação por um fator divisor de 1000 — para frequências acima de 1152 Hz.*

*c. Sinal reamostrado para 2048 Hz.*



**Figura 2:** Espectrogramas dos sinais *a*, *b* e *c* da figura 1, respectivamente.

A taxa de amostragem do sinal nos informa quanto ao número de pontos amostrados por segundo. O espectrograma resultante da aplicação da transformada de Fourier contém informação pertinente à metade da taxa de amostragem do sinal — taxa de amostragem de Nyquist [5]. Ao aplicarmos a FFT a uma amostra de áudio amostrado a 44100 Hz (figura 1.a), o espectrograma resultante (figura 2.a) contém dados para frequências de até 22050 Hz, inclusive.

Esse espectrograma (figura 2.a) nos permite avaliar que a maior parte da informação vocal se concentra em baixas frequências. Para simplificar o modelo contemplado pela rede projetada e remover ruído do sinal de áudio de entrada, desenvolvemos um filtro passa baixa Chebyshev Tipo II, com frequência de corte de 1024 Hz e atenuação de 60 Db — o sinal é atenuado por um fator divisor de 1000. O sinal de áudio filtrado e espectrograma resultante constam nas figuras 1.b e 2.b, respectivamente.

Uma vez que o sinal filtrado apresenta pouca ou nenhuma intensidade para frequências acima de 1024 Hz e preservou parte considerável da informação vocal, ele foi reamostrado para 2048 Hz. O sinal de áudio filtrado reamostrado e espectrograma resultante constam nas figuras 1.c e figura 2.c, respectivamente.

O valor da frequência de corte para filtro passa baixa não foi, de todo, escolhido arbitrariamente; tampouco a duração do áudio: a aplicação da FFT apresenta eficiente custo computacional para sequência de dados cujo tamanho total seja o resultado de uma potência de 2 [6] e [7].

O resultado da FFT aplicada ao sinal filtrado e reamostrado é utilizado como entrada da rede apresentada. Como ele apresenta números complexos, optou-se por utilizar o valor absoluto de cada ponto, correspondente ao módulo ou magnitude do mesmo:

$$|a + bi| = \sqrt{a^2 + b^2}. \quad (2)$$

O tamanho final da sequência de entrada para cada amostra da base de gravações é dado

por:

$$n = 2048 \times t, \quad (3)$$

onde  $t = 2$  segundos é duração de cada amostra, 2048 é a taxa de amostragem do sinal de áudio, e  $n = 4096$  é o tamanho do vetor de entrada.

A rede proposta é uma rede neural artificial de classificação com camada de entrada de 4096 pontos, uma camada escondida com 64 neurônios, e camada de saída com 10 neurônios. Possui função de ativação não linear na camada escondida e utiliza função de ativação *softmax*. A função de treinamento é a função do gradiente conjugado escalado.

A matriz de entrada  $X_{m \times n}$  possui 600 amostras, cada qual com 4096 pontos, das quais 360 amostras foram utilizadas para treinamento, 120 para validação e 120 para testes. A partição da amostra é aleatória.

Cada amostra é associada à sua classificação correspondente: um vetor binário com 10 posições, referentes aos 10 indivíduos — considerados classes — onde o valor 1 na  $i$ -ésima posição significa que a amostra pertence à  $i$ -ésima classe; o valor 0, de forma análoga, significa que a amostra não pertence à  $i$ -ésima classe. Cada amostra pertence a uma única classe, naturalmente. Esses vetores associados às amostras formam a matriz de alvos  $T_{m \times c}$ , onde  $c = 10$  é o número de classes.

O vetor de saída da rede apresenta, idealmente, as mesmas características de uma classificação associado a uma amostra.

Na seção 3 são expostos os resultados obtidos pela rede projetada.

### III. RESULTADOS

O treinamento da rede de classificação se deu em 10 segundos. O treinamento foi interrompido por se atingir limite inferior do gradiente (múltiplas execuções).

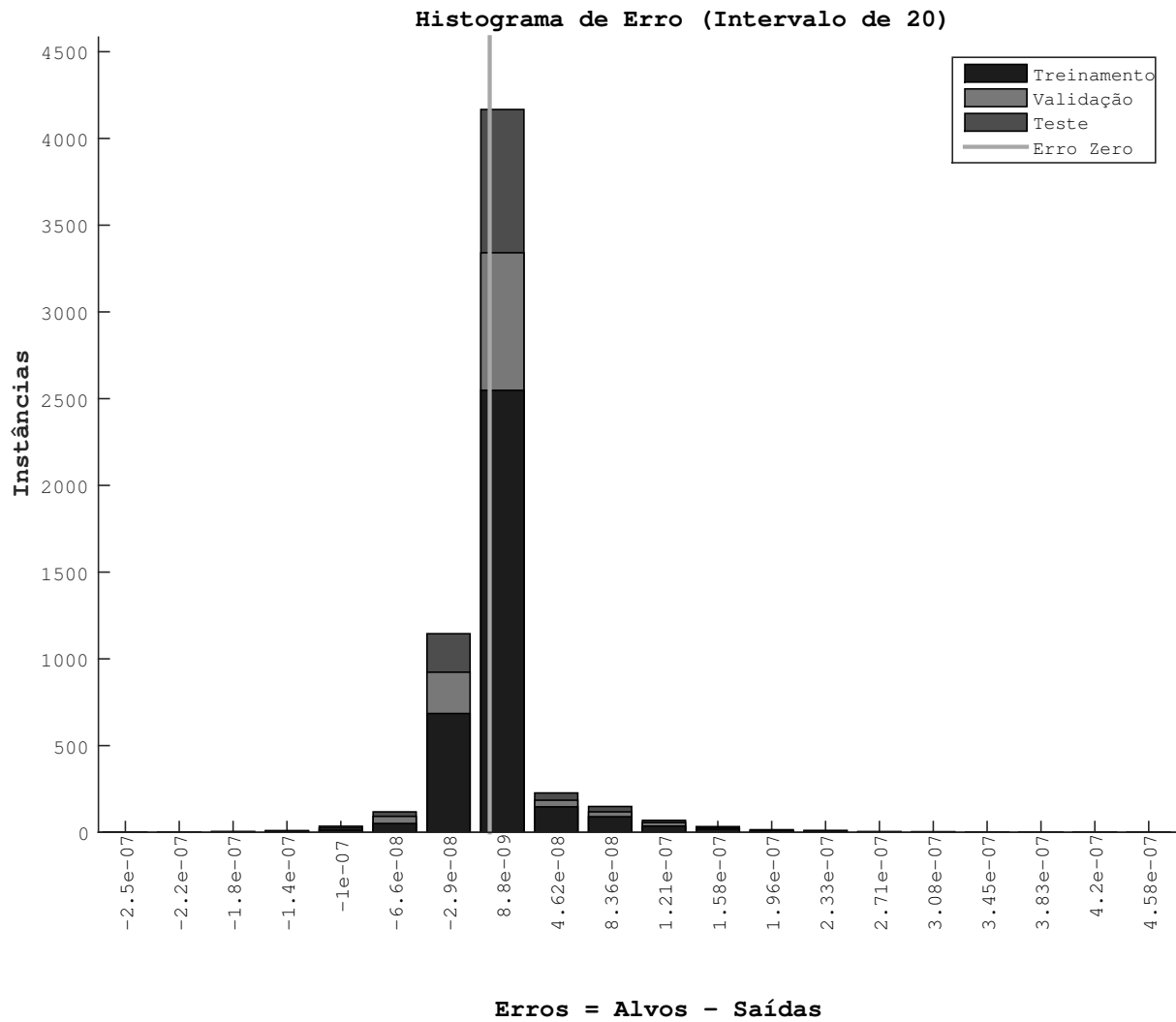
A rede neural classificou corretamente todas as amostras (figura 3) em intervalo de tempo aproximado de 1 segundo. Apesar de a rede classificar as amostras, não existe associação

**Matriz de Confusão**

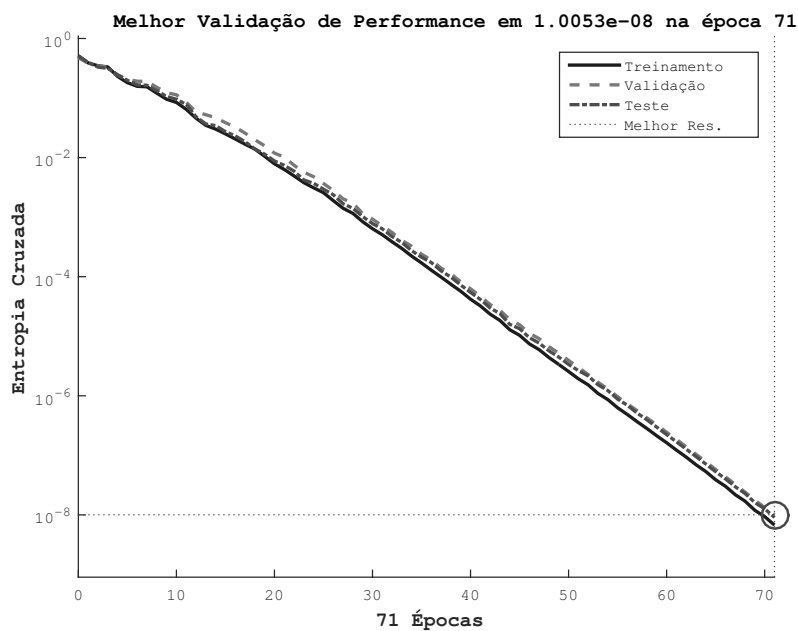
1	60 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
2	0 0.0%	60 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	60 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
4	0 0.0%	0 0.0%	0 0.0%	60 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	60 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	60 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	60 10.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
8	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	60 10.0%	0 0.0%	0 0.0%	100% 0.0%
9	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	60 10.0%	0 0.0%	100% 0.0%
10	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	60 10.0%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%
	1	2	3	4	5	6	7	8	9	10	
	Classe Alvo										

Classe de Saída

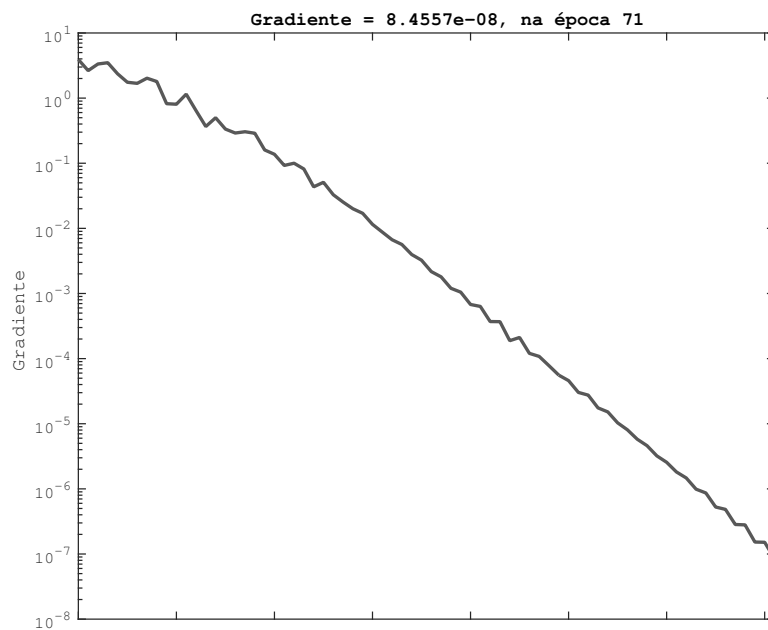
**Figura 3:** Matriz de confusão. A rede classificou corretamente todas as amostras. Cada classe tem 60 amostras correspondentes.



**Figura 4:** Histograma de erro (intervalo de largura 20).



**Figura 5:** Evolução da entropia cruzada por épocas de treinamento. Apesar do decrescimento aproximadamente linear, o treinamento foi interrompido com finalidade de se evitar overfitting.



**Figura 6:** Evolução do gradiente por épocas de treinamento. Apesar do decrescimento logarítmico, o treinamento foi interrompido com finalidade de se evitar overfitting.



útil entre o número da classe e o indivíduo correspondente.

A performance e o gradiente de treinamento decaem rapidamente (figuras 5 e 6), mas, com o limiar inferior do gradiente atingido, o treinamento é interrompido com finalidade de se evitar *overfitting*. O histograma de erro (figura 4) e a correta classificação das amostras (figura 3) indicam que a aplicação de uma rede de classificação se mostrou adequada para resolver o problema de reconhecimento de locutor.

#### IV. AMEAÇAS À VALIDADE E LIMITAÇÕES

O presente trabalho foi realizado com escopo de tempo reduzido. Como resultado, alguns fatores ameaçam a universalidade dos resultados obtidos.

Os locutores não representam extensa cobertura das possibilidades vocais e o número de classes é relativamente pequeno; para certas aplicações.

A atenuação de 20 Db nos sinais de ruído foi arbitrária e não se encontra amparada por estudos de quaisquer natureza. A redução da taxa de amostragem de 44100 Hz para 2048 Hz possui fraqueza semelhante. Uma taxa de amostragem de 8000 Hz se mostra suficiente, no entanto.

Para que se faça possível a classificação de uma entrada, é necessário que se faça a reamostragem para 2048 Hz da mesma. Este trabalho não avaliou o impacto computacional que isso acarreta.

Na próxima seção são apresentadas as considerações finais.

#### V. CONCLUSÃO

É frequente o uso de redes neurais artificiais nos casos em que se tem pouca informação acerca de um determinado problema e não se tem meios de estabelecer uma solução analítica para o mesmo.

A rede neural artificial proposta apresentou tempo de treinamento factível para aplicações

reais (10 segundos) e apresentou desempenho igualmente factível (600 amostras em aproximadamente 1 segundo), passível de ser implantada em sistemas embarcados e de fazer classificações em tempo real. Todas as amostras foram classificadas corretamente pela rede proposta.

#### REFERÊNCIAS

- [1] Jesus A. O. Neto; Marco A. A. Castro; Leonardo B. Felix. “Reconhecimento de comandos de voz para o acionamento de cadeira de rodas”. *Anais do XVIII Congresso Brasileiro de Automática* (2010), pp. 3819–3824.
- [2] Thiang. “Implementation of speech recognition on MCS51 microcontroller for controlling wheelchair”. *Intelligent and Advanced Systems 2007*. Ed. por IEEE. 2007, pp. 1193–1198.
- [3] Sérgio Eduardo Cardoso. “A inteligência artificial no judiciário”. Dissertação. Universidade Federal de Santa Catarina, 2001.
- [4] Marília A. Amaral; Rodolfo Barriviera; Eduardo Cotrin Teixeira. “Reconhecimento de voz para automação residencial baseado em agentes inteligentes”. *Revista Eletrônica de Sistemas de Informação 3.1* (2004).
- [5] E. Oran Brigham. *The Fast Fourier Transform and Its Applications*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988, pp. 83–87. ISBN: 0133075052.
- [6] E. Oran Brigham. *The Fast Fourier Transform and Its Applications*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988, pp. 131–136. ISBN: 0133075052.
- [7] E. Oran Brigham. *The Fast Fourier Transform and Its Applications*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988, pp. 148–156. ISBN: 0133075052.