Sistemas de Reconhecimento Automático de Fala Baseados em Redes Neurais Profundas Usando Espectrogramas do Sinal de Fase

Ênio dos Santos Silva e Rui Seara

***Resumo*— Este trabalho apresenta uma investigação sobre o uso de espectrogramas de fase aplicados a sistemas de reconhecimento automático de fala (*automatic speech recognition* - ASR) baseados em redes neurais profundas (*deep neural network* - DNN). Particularmente, visando a obtenção de atributos discriminativos robustos ao ruído, a função atraso de grupo modificada é considerada na etapa de extração de log-espectrogramas de fase em escala Mel (log-Mel-espectrogramas) de sinais de fala. Nesse contexto, o uso de log-Mel-espectrogramas de fase em sistemas de ASR associados com DNNs do estado da arte (ASR-DNN) é discutido. Além do mais, o desempenho dos sistemas de ASR-DNN é avaliado em ambientes acústicos com baixa razão sinal-ruído (*signal-to-noise ratio* - SNR). Resultados de simulação são mostrados confirmando a eficácia da utilização de log-Mel-espectrogramas de fase em sistemas de ASR-DNN.**

***Palavras-Chave*— Atraso de grupo, espectrogramas do sinal de fase, extração de atributos, reconhecimento automático de fala, redes neurais profundas.**

***Abstract*— This paper presents an investigation on the use of phase spectrograms in automatic speech recognition (ASR) systems based on deep neural networks (DNN). Particularly, in order to obtain robust discriminative attributes, the modified group delay function is taken into account in the feature extraction stage using log-Mel spectrograms of the phase from speech signals. In this way, the use of phase log-Mel spectrograms in ASR systems associated with state-of-the-art DNNs (ASR-DNN) is discussed. In addition, the performance of ASR systems is assessed in acoustic environments with low signal-to-noise ratio. Simulation results are shown confirming the effectiveness of the use of phase log-Mel spectrograms in ASR-DNN systems.**

***Keywords*— Group delay, phase spectrograms, feature extraction, automatic speech recognition, deep neural network.**

I. Introdução

Até poucos anos atrás, sistemas de reconhecimento automático de fala (*automatic speech recognition* - ASR) vinham sendo implementados usando modelos escondidos de Markov (*hidden Markov model* - HMM) associados com modelos de misturas de gaussianas (*Gaussian mixture models* - GMM) [1]. Atualmente, sistemas de ASR do estado da arte são implementados inteiramente por modelos baseados em redes neurais profundas (*deep neural network* - DNN), denominados ASR-DNN, ou por modelos híbridos de HMMs com DNNs [1]-[4].

Ênio dos Santos Silva e Rui Seara, LINSE–Laboratório de Circuitos e Processamento de Sinais, Departamento de Engenharia Elétrica e Eletrônica, Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil, e-mails: enio@linse.ufsc.br; seara@linse.ufsc.br. Este trabalho foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Em sistemas de ASR usando HMM-GMM, a estratégia de extração de atributos discriminativos de sinais de fala vem sendo implementada há um longo tempo através do uso de coeficientes cepstrais em escala Mel (*mel frequency cepstral coefficients* - MFCC) [4]. No entanto, em contraste com os resultados obtidos em sistemas de ASR baseados em HMM-GMM, as estratégias usando MFCC já não vêm mais apresentando resultados satisfatórios seja em sistemas de ASR híbridos ou mesmo em sistemas de ASR-DNN [4], [5]. Dessa forma, em linha com os avanços de pesquisas atuais em sistemas de ASR usando DNN, as estratégias para a extração de atributos discriminativos de sinais de fala vêm se consolidando como um tópico importante de pesquisa na área de processamento de fala [2], [5], [6]. Geralmente, sistemas de ASR-DNN operam com atributos de entrada provenientes da representação tempo-frequência dada pelo logaritmo do espectrograma de magnitude em escala Mel, denotado aqui log-Mel-espectrograma, da transformada de Fourier de curto termo (*short-time Fourier transform* - STFT) do sinal de fala [2], [3], [7]. Nesse contexto, o uso de espectrogramas de magnitude vem sendo comumente adotado na literatura [1], [2]. No entanto, como discutido em [2], ainda existem vários questionamentos sobre a capacidade dos log-Mel-espectrogramas de magnitude representarem adequadamente as principais características dos sinais de fala, especialmente, quando se opera em ambientes acústicos com baixa razão sinal-ruído (*signal-to-noise ratio* - SNR).

Tendo em vista a identificação de atributos discriminativos do sinal de fala, em [6], uma atenção especial é dada para a utilização do sinal de fase em aplicações de processamento de fala. Nesse contexto, em [8] e [9], atributos criados a partir das derivadas do sinal de fase (tanto no domínio do tempo quanto no domínio da frequência), representados pelo atraso de grupo (*group delay* - GD) e pela frequência instantânea, são utilizados com sucesso em aplicações de ASR e em aplicações de realce do sinal de fala. Já em [10], visando explorar os benefícios da utilização de ambos os atributos (magnitude e fase), uma estratégia de combinação de atributos vem sendo discutida e avaliada para ser considerada em sistemas de ASR usando HMM-GMM.

Particularmente, em [9], modelos de DNN vêm sendo utilizados e vêm confirmando a importância do sinal de fase em aplicações de realce de sinais de fala corrompidos por ruídos de diferentes níveis de SNR. Já em [8], uma versão modificada do GD (*modified group delay* - MOGD) e os MFCCs são utilizados para representar, respectivamente, as

informações de fase e de magnitude da STFT do sinal de fala. Além do mais, também em [8], sistemas de ASR baseados em HMM-GMM e sistemas híbridos HMM-DNN vêm sendo investigados. Embora alguns resultados promissores tenham sido obtidos em [8], a literatura atual aponta que a utilização de MFCC, assim como o uso de sistemas híbridos HMM-DNN, não são capazes de alcançar resultados tão bons quando comparados com os possíveis resultados obtidos em sistemas de ASR-DNN [1], [2], [4], [5], [11].

Neste trabalho, visando a obtenção de atributos discriminativos capazes de representar estruturas harmônicas e pistas fonético-acústicas de sinais de fala, operando em ambientes acústicos com baixa SNR, o uso de log-Mel-espectrogramas de fase de sinais de fala em sistemas de ASR-DNN é discutido. Especificamente, neste trabalho de pesquisa (assim como abordado também em [8]), os atributos de MOGD são considerados na etapa de extração de log-Mel-espectrogramas de fase. Em contraste com a estratégia proposta em [8], os sistemas de ASR investigados aqui são compostos integralmente por DNNs. Além disso, os resultados obtidos em [10] motivaram uma estratégia de combinação de atributos (magnitude e fase), a qual é aqui discutida e avaliada em ambientes acústicos com baixa SNR. Resultados de simulação confirmam a eficácia da utilização de log-Mel-espectrogramas de fase em sistemas de ASR-DNN.

II. Espectrogramas do Sinal de Fala

Vamos considerar agora $\mathbf{x}(n)$ caracterizando um quadro do sinal de fala no domínio do tempo; sua correspondente STFT $X_n(\mathrm{e}^{j\omega})$ é dada por

$$X_n(\mathrm{e}^{j\omega}) = |X_n(\mathrm{e}^{j\omega})|\mathrm{e}^{j\theta_n[X_n(\mathrm{e}^{j\omega})]} \tag{1}$$

onde $|X_n(\mathrm{e}^{j\omega})|$ denota o espectro de magnitude da STFT e $\theta_n[X_n(\mathrm{e}^{j\omega})]$ representa o seu correspondente espectro de fase [10]. Nesse contexto, o espectrograma $X(N, \mathrm{e}^{j\omega})$ é a representação tempo-frequência de N quadros do sinal $\mathbf{x}(n)$ dado pela sequência temporal dos espectros de magnitude e de fase, isto é,

$$X(N, \mathrm{e}^{j\omega}) = \begin{cases} |X_1(\mathrm{e}^{j\omega})|, |X_2(\mathrm{e}^{j\omega})|, \ldots, |X_N(\mathrm{e}^{j\omega})| \\ \theta_1[X_1(\mathrm{e}^{j\omega})], \theta_2[X_2(\mathrm{e}^{j\omega})], \ldots, \theta_N[X_N(\mathrm{e}^{j\omega})]. \end{cases} \tag{2}$$

Particularmente, em sistemas de ASR-DNN, o aprendizado de máquina está relacionado à extração de atributos relevantes dos espectrogramas e à classificação fonética e/ou linguística desses atributos [4], [9]. Nesse contexto, é importante que as informações relevantes dos espectrogramas sejam mantidas. Então, visando considerar todas as informações contidas no sinal de fala, a seguir, são discutidos os procedimentos para a obtenção dos espectrogramas de magnitude e de fase do correspondente sinal.

A. Espectrogramas do Sinal de Fase

Embora ambos os espectrogramas (magnitude e fase) contenham informações relevantes do sinal de fala, no espectrograma de fase, a identificação dessas informações é prejudicada devido ao problema do "empacotamento" da fase (*phase wrapping*) (módulo 2π) [6], [10]. Especificamente, os valores de $\theta_n[X_n(\mathrm{e}^{j\omega})]$ oriundos de (1) estão confinados entre $-\pi$ e π, e não representam a "verdadeira" fase do sinal [10].

1) Revisitando a Função Atraso de Grupo Modificada: Uma das possíveis abordagens para contornar o problema do empacotamento do sinal de fase é a utilização da função MOGD $\tilde{\tau}_n(\mathrm{e}^{j\omega})$ [8], [10], a qual pode ser definida como uma aproximação da derivada da função de fase[1] $\theta_n[X_n(\mathrm{e}^{j\omega})]$ e expressa por

$$\tilde{\tau}_n(\mathrm{e}^{j\omega}) = \frac{Y_{\mathrm{I}}(\mathrm{e}^{j\omega})X_{\mathrm{I}}(\mathrm{e}^{j\omega}) + Y_{\mathrm{R}}(\mathrm{e}^{j\omega})X_{\mathrm{R}}(\mathrm{e}^{j\omega})}{|X_n(\mathrm{e}^{j\omega})|^{2\gamma}.} \tag{3}$$

onde γ representa um coeficiente de suavização e $X_{\mathrm{R}}(\mathrm{e}^{j\omega})$, $X_{\mathrm{I}}(\mathrm{e}^{j\omega})$, $Y_{\mathrm{R}}(\mathrm{e}^{j\omega})$ e $Y_{\mathrm{I}}(\mathrm{e}^{j\omega})$ denotam, respectivamente, a parte real e a parte imaginária da STFT de $\mathbf{x}(n)$ e de $n\mathbf{x}(n)$ (para mais detalhes veja [10]).

2) Log-Mel-espectrogramas: Visando representar a característica não linear da percepção auditiva humana, a escala logarítmica Mel vem sendo comumente adotada em processamento de fala [2], [4], [9]. Assim como em [12], aqui também é usada a relação logarítmica entre as escalas de frequências em Hertz (Hz) e em Mel, dada por

$$\mathrm{mel}(f_{\mathrm{Hz}}) = \frac{1000}{\log(2)}\log\left(1 + \frac{f_{\mathrm{Hz}}}{1000}\right) \tag{4}$$

onde mel(.) denota as frequências resultantes na escala Mel e f_{Hz} representa os componentes de frequência em Hz. Além disso, n_{Mel} bancos de filtros triangulares com variação logarítmica de amplitude também são considerados no processo de obtenção dos espectros de frequência em escala Mel. O procedimento para a obtenção desses espectros consiste em computar os valores do logaritmo da energia das n_{Mel} bandas de frequências distribuídas através dos bancos de filtros triangulares com frequências centrais igualmente espaçadas na escala Mel. A Fig. 1 ilustra o processo de distribuição dos espectros de magnitude e de fase (aqui representados por $|X_n(\mathrm{e}^{j\omega})|$ e $\tilde{\tau}_n(\mathrm{e}^{j\omega})$, respectivamente) nos filtros triangulares correspondentes às n_{Mel} bandas de frequência. Dessa forma, aqui os espectrogramas do sinal de fase são representados pelos espectrogramas logarítmicos em escala Mel $X_{\tilde{\tau}}(N, \mathrm{e}^{j\omega})$ oriundos da sequência temporal dos atributos da função MOGD $\tilde{\tau}_n(\mathrm{e}^{j\omega})$.

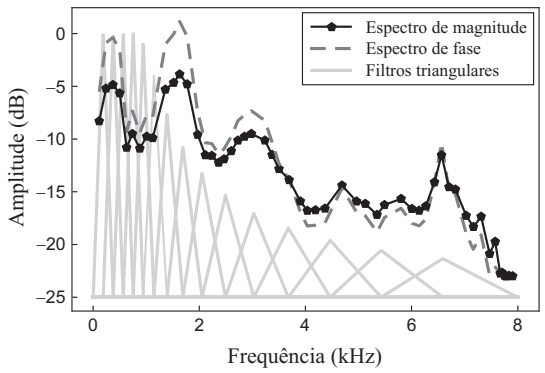


Fig. 1. Distribuição dos espectros de magnitude e de fase nos n_{Mel} bancos de filtros triangulares.

[1]Nessa definição, $\theta_n[X_n(\mathrm{e}^{j\omega})]$ é considerada uma função contínua, uma vez que ela está na forma desempacotada.

III. Sistemas de ASR-DNN

Considerando o estado da arte em aprendizado profundo, tanto a extração de atributos discriminativos quanto o projeto de classificadores podem ser otimizados quando ambas as tarefas são realizadas conjuntamente em uma única rede conexionista [1], [2], [4], [11]. Tipicamente, sistemas de ASR-DNN são baseados em arquiteturas compostas por redes neurais convolucionais (*convolutional neural network* - CNN) conectadas a redes de *perceptron* de múltiplas camadas (*multilayer perceptron* - MLP) [4]. Nessas arquiteturas, as primeiras camadas (correspondentes às CNNs) são projetadas para aprenderem sobre a extração de atributos discriminativos, enquanto as últimas camadas (correspondentes às MLPs) têm a finalidade específica de classificação [2]. Dessa forma, a estrutura supracitada engloba todas as etapas de aprendizado em uma única rede conexionista profunda. A Fig. 2 ilustra a arquitetura típica de um sistema de ASR-DNN.

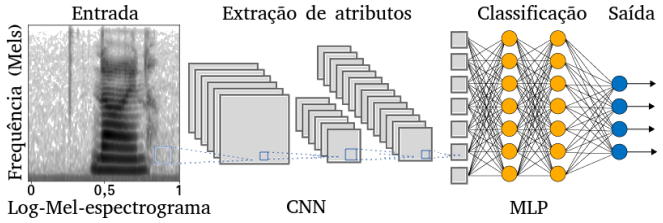


Fig. 2. Ilustração de uma arquitetura típica de sistemas de ASR-DNN.

Na literatura, as arquiteturas propostas em [7] e [13] vêm sendo adotadas em sistemas de ASR-DNN. Tais arquiteturas são brevemente discutidas a seguir.

A. Arquitetura usando CNN Tradicional

A arquitetura proposta em [7] vem sendo avaliada em diversos sistemas de ASR-DNN [1], [2]. Essa arquitetura, denominada cnn-trad-fpool3, recebe espectrogramas do sinal de fala como atributos de entrada e é composta por uma CNN de duas camadas convolucionais (conectadas em série) com normalização em lote (*batch normalization* - BN) e ativação linear retificada (*rectified linear unit* - ReLU) seguida por operações de subamostragem (do tipo *max-pooling* [11]) apenas na dimensão correspondente ao domínio da frequência. A saída da CNN é conectada a uma rede MLP contendo três camadas completamente conectadas (*fully connected* - FC), com BN e ativação ReLU apenas na segunda camada. A camada final da rede MLP implementa uma ativação do tipo *softmax* e é destinada à estimação das probabilidades das classes fonéticas e/ou linguísticas correspondentes aos espectrogramas fornecidos na entrada da rede [11] (veja [7] para mais detalhes). A Tabela I mostra a configuração da rede cnn-trad-fpool3, onde m e r denotam as dimensões dos filtros convolucionais (*kernels* $m \times r$) correspondentes aos domínios do tempo e da frequência, e q representa a taxa de subamostragem usada na operação de *max-pooling*. Além disso, n denota o número de *kernels* de cada camada da CNN e u caracteriza o número de unidades neurais de cada camada FC da rede MLP. Na última camada da rede (FC + *softmax*), u indica o número de unidades de saída (classes fonéticas e/ou linguísticas).

TABELA I
Configuração da rede cnn-trad-fpool3

Arquitetura	**Tipo da camada**	m	r	q	n	u
CNN	Conv. + BN + ReLU	20	8	3	64	-
	Conv. + BN + ReLU	10	4	1	64	-
MLP	FC	-	-	-	-	32
	FC + BN + ReLU	-	-	-	-	128
	FC + *Softmax*	-	-	-	-	35

B. Arquitetura usando Rede Residual

No treinamento de uma rede neural convencional, dado o sinal de entrada $\mathbf{x}$, o objetivo da rede é estimar uma função $h(\mathbf{x})$ capaz de associar $\mathbf{x}$ com determinadas saídas pré-estabelecidas. Hipoteticamente, caso $\mathbf{x}$ seja adicionado à saída da rede, então, ao invés de estimar $h(\mathbf{x})$, a rede será levada a estimar $h(\mathbf{x}) - \mathbf{x}$. Tal procedimento é denominado aprendizado residual [11]. Nesse contexto, a rede residual profunda (*residual networks* - ResNet) [13] pode ser vista como um conjunto de blocos residuais empilhados sequencialmente, em que cada bloco é composto por uma pequena CNN com uma conexão adicional que liga diretamente a entrada com a saída desses blocos (*shortcut connection*).

As arquiteturas discutidas em [13] implementam CNNs extremamente profundas e, assim como na seção anterior, elas também vêm sendo investigadas em diversos sistemas de ASR-DNN [1], [2]. Particularmente, neste trabalho de pesquisa, uma ResNet com 20 camadas convolucionais (ResNet-20) tem sido utilizada. A Fig. 3 ilustra a sequência de operações realizadas pela ResNet-20. Tipicamente, os blocos residuais são organizados em três conjuntos de acordo com os seus correspondentes números de filtros convolucionais (16, 32 e 64). Especificamente, cada bloco é composto por duas camadas convolucionais com *kernels* 3×3, BN e ativação ReLU, sendo que a última operação de ativação ReLU é realizada após a *shortcut connection*. Após os conjuntos de blocos residuais, uma operação de subamostragem (do tipo *average-pooling* [11]) é realizada e, na saída da rede, uma camada FC com ativação *softmax* é implementada para a estimação das classes fonéticas e/ou linguísticas.

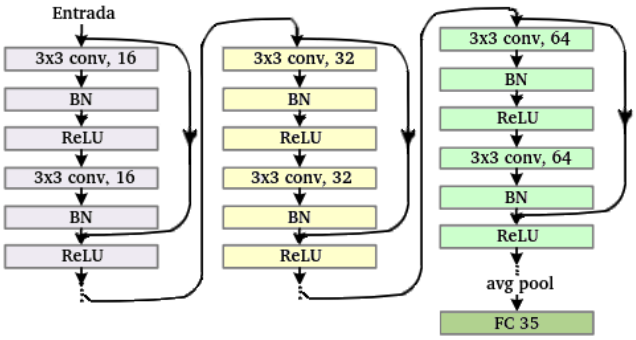


Fig. 3. Ilustração da arquitetura ResNet-20.

IV. Resultados e Análise de Desempenho

A. Base Acústica

Para o desenvolvimento dos sistemas de ASR-DNN, tem sido utilizada a base de dados de comandos de fala disponi-

bilizada em [14] pela empresa *Google*. Essa base consiste de 65000 arquivos de áudio (amostrado em 16 kHz) de duração de um segundo (cada arquivo), sendo muito utilizada em diversas competições de aprendizado de máquina promovidas pelos grupos *TensorFlow* e *AIY* [3], [14]. Nessa base acústica, os arquivos são classificados em 35 comandos de fala, divididos entre 24 comandos principais e 11 subcomandos. Adicionalmente, a base de dados também fornece arquivos de áudio com ruído de fundo, a saber: ruídos gerados artificialmente do tipo branco e rosa, e alguns ruídos reais naturais (não artificiais) do cotidiano. A Tabela II apresenta os comandos considerados na base acústica.

TABELA II
COMANDOS DE FALA PRESENTES NA BASE ACÚSTICA

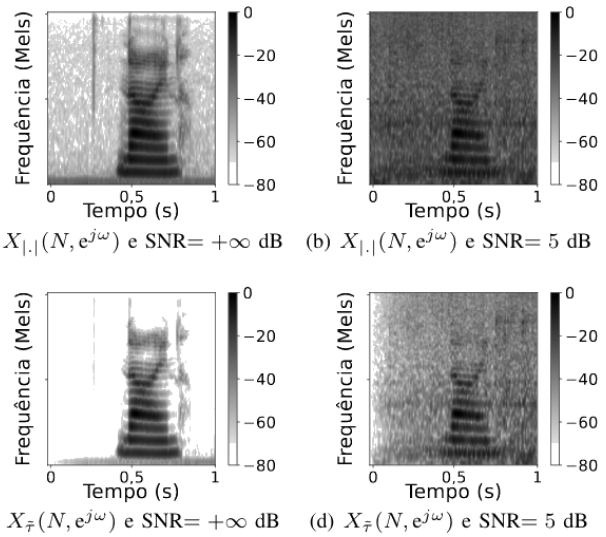
Tipos	**Comandos**
Comandos principais	*down, eight, five, four, go, left, nine, no, off, on, right, seven, six, stop, three, two, up, backward, zero, forward, follow, learn, one, yes.*
Subcomandos	*bed, bird, cat, dog, happy, house, marvin, sheila, tree, wow, visual.*
Ruídos de fundo	*dude miaowing, doing the dishes, white noise, pink noise, exercise bike, running tap.*

Neste trabalho, a fim de investigar o desempenho dos sistemas de ASR-DNN em ambientes acústicos com baixas SNR, de acordo com os tipos de ruídos especificados na Tabela II, foram gerados áudios corrompidos artificialmente por esses ruídos, com os seguintes níveis de SNRs: 5, 10 e 20 dB. Particularmente, assim como discutido em [3], devido ao processo de aquisição dos áudios da base acústica supracitada não seguir um controle de qualidade adequado, deve-se assumir que os arquivos de áudio originais dessa base já contenham algum tipo de ruído de gravação e que, ao se adicionar artificialmente algum ruído de fundo, o nível da SNR real será menor do que o especificado.

B. Treinamento, Validação e Teste

Para avaliar o desempenho dos sistemas de ASR-DNN considerando o uso de espectrogramas do sinal de fase, as arquiteturas cnn-trad-fpool3 e ResNet-20, discutidas nas Seções III-A e III-B, são implementadas. Então, para efeito de comparação, os modelos de ASR-DNN cnn-trad-fpool3 e ResNet-20 são avaliados considerando os seguintes atributos de entrada: 1) usando apenas log-Mel-espectrogramas de magnitude $X_{|.|}(N, \mathrm{e}^{j\omega})$; 2) usando apenas log-Mel-espectrogramas de fase $X_{\tilde{\tau}}(N, \mathrm{e}^{j\omega})$; e 3) usando em conjunto log-Mel-espectrogramas de magnitude e de fase $\left[X_{|.|}(N, \mathrm{e}^{j\omega}), X_{\tilde{\tau}}(N, \mathrm{e}^{j\omega})\right]$.

A Fig. 4 ilustra os atributos de entrada [log-Mel-espectrogramas de magnitude $X_{|.|}(N, \mathrm{e}^{j\omega})$ e de fase $X_{\tilde{\tau}}(N, \mathrm{e}^{j\omega})$] dos sistemas de ASR-DNN em ambientes com níveis de SNR iguais a $+\infty$ e 5 dB. Nota-se que, assim também como observado na Fig. 1, os espectrogramas de fase $X_{\tilde{\tau}}(N, \mathrm{e}^{j\omega})$ apresentam maiores destaques nas estruturas harmônicas e nas pistas fonético-acústicas do sinal de fala, quando comparados com os correspondentes espectrogramas de magnitude $X_{|.|}(N, \mathrm{e}^{j\omega})$, seja em ambientes isentos de ruído [Fig. 4(a), 4(c)] ou em ambientes ruidosos [Fig. 4(b), 4(d)].



(a) $X_{|.|}(N, \mathrm{e}^{j\omega})$ e SNR$= +\infty$ dB (b) $X_{|.|}(N, \mathrm{e}^{j\omega})$ e SNR$= 5$ dB

(c) $X_{\tilde{\tau}}(N, \mathrm{e}^{j\omega})$ e SNR$= +\infty$ dB (d) $X_{\tilde{\tau}}(N, \mathrm{e}^{j\omega})$ e SNR$= 5$ dB

Fig. 4. Log-Mel-espectrogramas de magnitude e de fase do sinal da fala em ambientes acústicos com SNR igual a $+\infty$ e 5 dB.

Os sistemas de ASR investigados aqui operam com sinais de fala (comandos) de duração de um segundo, segmentados em quadros de 25 ms (usando janelas de $Hanning$) com sobreposição de 10 ms. Dessa forma, seguindo os procedimentos discutidos na Seção II, os log-Mel-espectragramas $X_{|.|}(N, \mathrm{e}^{j\omega})$ e $X_{\tilde{\tau}}(N, \mathrm{e}^{j\omega})$ são computados usando 40 bancos de filtros triangulares ($n_{\text{Mel}} = 40$). Especificamente, os atributos $X_{\tilde{\tau}}(N, \mathrm{e}^{j\omega})$ são computados a partir de $\tilde{\tau}_n(\mathrm{e}^{j\omega})$ com o coeficiente de suavização $\gamma = 1/4$.

A fim de garantir a reprodução e a comparação dos trabalhos de pesquisa que usam a base acústica discutida na seção anterior, em [14], os conjuntos (mutuamente exclusivos) de treinamento, validação e teste têm sido previamente definidos. Então, neste artigo, os mesmos conjuntos pré-definidos em [14] são utilizados.

Com o intuito de analisar a eficácia do uso de log-Mel-espectrogramas contendo informações de fase dos sinais de fala, os hiperparâmetros, correspondentes aos modelos utilizados neste trabalho, são mantidos inalterados para todos os experimentos. Dessa forma, os modelos de ASR-DNN discutidos na Seção III são treinados com minilotes (*minibatches*) contendo 32 exemplos de sinais de fala, otimizador *Adam* usando uma taxa de aprendizado de 10^{-3}, inicialização normal de *He* e regularização l_2 de 10^{-4} (para mais detalhes, veja [11]). Particularmente, os modelos são treinados visando a maximização da acurácia na classificação individual de cada comando de fala. Durante a etapa de treinamento, a cada *época*, o desempenho dos modelos é avaliado sobre o conjunto de validação, e aquele que apresentar o melhor desempenho será selecionado. Assumindo uma possibilidade de parada antecipada [11], o treinamento dos modelos pode ser realizado por no máximo 80 *épocas* ou pode ser interrompido caso não apresente qualquer melhora na acurácia, considerando a comparação com os resultados das 40 *épocas* mais recentes. Além disso, tais modelos são treinados (e validados) usando a base acústica isenta de ruído (SNR = $+\infty$ dB) e são testados em ambientes acústicos com SNRs de 5, 10, 20 e $+\infty$ dB.

A Tabela III apresenta a acurácia mediana, resultante

dos sistemas de ASR usando as arquiteturas cnn-trad-fpool3 e ResNet-20, para diferentes níveis de SNR. Tais resultados foram obtidos a partir de simulações de Monte Carlo (MC) (considerando 10 realizações independentes), usando os atributos de entrada $X_{|.|}(N, \mathrm{e}^{j\omega})$, $X_{\tilde{\tau}}(N, \mathrm{e}^{j\omega})$ e $\left[X_{|.|}(N, \mathrm{e}^{j\omega}), X_{\tilde{\tau}}(N, \mathrm{e}^{j\omega})\right]$, representados, respectivamente, pelos símbolos (M), (F) e (M+F). Adicionalmente, na Fig. 5, as variações na acurácia dos modelos, resultantes das simulações de MC, são mostradas por diagramas de caixa.

TABELA III
ACURÁCIA MEDIANA (%) DOS SISTEMAS DE ASR-DNN

Nível SNR (dB)	Arquitetura					
	cnn-trad-fpool3			ResNet-20		
	Atributos de entrada					
	(M)	(F)	(M+F)	(M)	(F)	(M+F)
5	40,26	**40, 90**	36,81	65,02	66,69	**68, 11**
10	59,99	**60, 15**	58,09	79,16	79,83	**80, 62**
20	78,01	**79, 67**	78,23	90,62	90,25	**90, 72**
$+\infty$	**88, 04**	86,35	87,63	95,01	94,82	**95, 13**

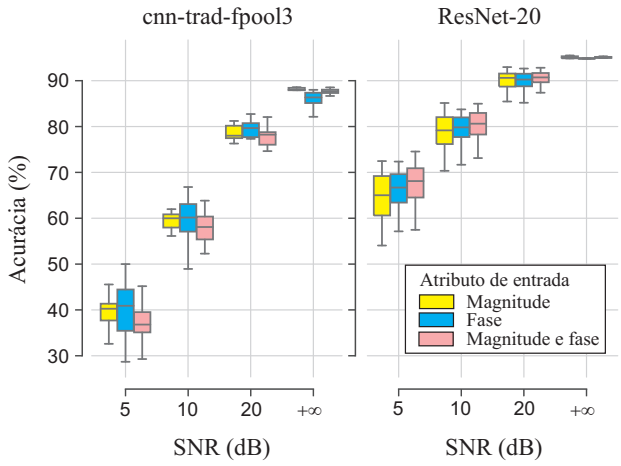


Fig. 5. Diagrama de caixa de acurácia dos sistemas de ASR-DNN.

A partir da Tabela III, nota-se o ganho de acurácia promovido pelo uso dos atributos de fase, tanto de maneira direta [usando $X_{\tilde{\tau}}(N, \mathrm{e}^{j\omega})$] nas arquiteturas com cnn-trad-fpool3, quanto de maneira combinada {usando $\left[X_{|.|}(N, \mathrm{e}^{j\omega}), X_{\tilde{\tau}}(N, \mathrm{e}^{j\omega})\right]$} nas arquiteturas que utilizam ResNet-20. Particularmente, considerando a avaliação dos modelos operando em ambientes ruidosos, constata-se que os ganhos de acurácia provocados pelo uso de atributos $X_{\tilde{\tau}}(N, \mathrm{e}^{j\omega})$ são mais evidentes em ambientes com baixos níveis de SNR. Especificamente, o uso de atributos combinados $\left[X_{|.|}(N, \mathrm{e}^{j\omega}), X_{\tilde{\tau}}(N, \mathrm{e}^{j\omega})\right]$ apresenta ganhos de 3,09; 1,46; 0,10 e 0,12% (em ambientes de 5, 10, 20 e $+\infty$ dB, respectivamente), quando comparado aos sistemas que utilizam apenas espectrogramas de magnitude $[X_{|.|}(N, \mathrm{e}^{j\omega})]$. Além disso, considerando apenas os atributos $X_{\tilde{\tau}}(N, \mathrm{e}^{j\omega})$, com base na Fig. 5, observa-se uma maior variância da acurácia nas arquiteturas usando CNNs menos profundas (cnn-trad-fpool3) em relação à variação apresentada nas arquiteturas que usam mais camadas destinadas à extração de atributos discriminativos (ResNet-20). Nesse contexto, acredita-se que a utilização de redes com camadas mais profundas e a otimização dos hiperparâmetros (para cada combinação entre a arquitetura da rede e o atributo de entrada) possam aprimorar ainda mais o desempenho dos sistemas de ASR-DNN usando log-Mel-espectrogramas de fase.

V. CONCLUSÕES E CONSIDERAÇÕES FINAIS

Neste trabalho de pesquisa, o uso de log-Mel-espectrogramas da fase do sinal de fala em sistemas de ASR foi investigado. Tais sistemas foram implementados considerando DNNs consagradas na literatura, CNN tradicional e ResNet, operando com sinais de comando de fala. Através dos resultados obtidos via simulações de MC, pôde-se inferir que o uso de log-Mel-espectrogramas de fase proporcionou maior riqueza espectral ao treinamento das redes neurais, possibilitando a extração de atributos mais significativos ao ASR, notadamente, para ambientes com baixas SNRs. Nesses casos, os sistemas de ASR treinados com os espectrogramas de fase apresentaram melhores desempenhos quando comparados aos sistemas que consideram apenas os espectrogramas de magnitude. Os resultados de acurácia obtidos confirmam a eficácia da utilização de log-Mel-espectrogramas do sinal de fase em sistemas de ASR-DNN.

REFERÊNCIAS

[1] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, no. 1, pp. 19 143–19 165, Feb. 2019.

[2] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, May 2019.

[3] J. Bae and D.-S. Kim, "End-to-end speech command recognition with capsule network," in *Proc. Int. Speech Communication Association (INTERSPEECH)*, Hyderabad, India, Sep. 2018, pp. 776–780.

[4] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach (Signals and Communication Technology)*, 1st ed. New York, USA: Springer, 2015.

[5] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. L. Seltzer, G. Zweig, X. He, J. D. Williams, Y. Gong, and A. Acero, "Recent advances in deep learning for speech research at Microsoft," in *Proc. Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Vancouver, Canada, May 2013, pp. 8604–8608.

[6] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, no. 1, pp. 1–29, Jul. 2016.

[7] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Proc. Int. Speech Communication Assn. (INTERSPEECH)*, Dresden, Germany, Sep. 2015, pp. 1478–1482.

[8] J. Fahringer, T. Schrank, J. Stahl, P. Mowlaee, and F. Pernkopf, "Phase-aware signal processing for automatic speech recognition," in *Proc. Int. Speech Communication Association (INTERSPEECH)*, San Francisco, USA, Sep. 2016, pp. 3374–3378.

[9] N. Zheng and X.-L. Zhang, "Phase-aware speech enhancement based on deep neural networks," *IEEE/ACM Trans. on Audio, Speech and Language Process.*, vol. 27, no. 1, pp. 63–76, Sep. 2019.

[10] E. S. Silva and R. Seara, "Considerações sobre o uso do sinal de fase em sistemas de reconhecimento automático de fala," in *Anais do Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT)*, Petrópolis, RJ, Set. 2019, pp. 1–5.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, 1st ed. London, UK: MIT Press, 2016.

[12] B. McFee, C. Raffel, D. Liang, D. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in python," in *Proc. 14th Python in Sci. Conf.*, Austin, USA, Jul. 2015, pp. 18–25.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, Jun. 2016, pp. 770–778.

[14] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv:1804.03209 [cs.CL]*, vol. 1, pp. 1–11, Apr. 2018.