Rede Convolucional Deformável Aplicada a Sistemas de KWS Robusto ao Ruído

Ênio dos Santos Silva e Rui Seara

***Resumo*— Este trabalho apresenta uma discussão sobre o uso de convolução deformável em sistemas de detecção de palavras-chave (*keyword spotting* - KWS) para operação em ambientes de baixa razão sinal-ruído (*signal-to-noise ratio* - SNR). A robustez ao ruído ainda é um problema crítico em aplicações de reconhecimento automático de fala do mundo real. Para mitigar esse problema, visando a obtenção de características discriminativas que descrevam melhor as pistas acústicas em cenários com baixa SNR, o uso de redes neurais convolucionais deformáveis (DefCNN) é considerado no processo de extração de características do sinal de fala. Resultados de simulação numérica são apresentados visando avaliar a acurácia do reconhecimento de palavras-chave, confirmando a eficácia do uso de DefCNN em aplicações de KWS.**

***Palavras-Chave*— Convolução deformável, detecção de palavras-chave, reconhecimento automático de fala, robustez ao ruído.**

***Abstract*— This work presents a discussion on the use of deformable convolution in keyword spotting (KWS) systems to operating in low signal-to-noise ratio (SNR) environments. Noise robustness is still a critical problem in practical real-world automatic speech recognition applications. To mitigate this problem and aiming to obtain discriminative attributes that best describe acoustic cues in low SNR environments, the use of deformable convolutional neural networks (DefCNN) is considered in the speech signal feature extraction process. Numerical simulation results are shown for keyword recognition performance, confirming the effectiveness of the use of DefCNN in KWS applications.**

***Keywords*— Deformable convolution, keyword spotting, automatic speech recognition, noise robustness.**

I. Introdução

Sistemas de detecção de palavras-chave (*keyword spotting* - KWS) vêm ganhando cada vez mais importância no cotidiano das pessoas e se tornando uma prática usual em várias aplicações [1]-[5]. Apesar de os sistemas de KWS atuais funcionarem satisfatoriamente em cenários controlados, seus desempenhos são fortemente degradados em condições de baixa razão sinal-ruído (*signal-to-noise ratio* - SNR) [5]-[8]. Nesse contexto, diversos trabalhos de pesquisa do estado-da-arte em reconhecimento automático de fala (*automatic speech recognition* - ASR) vêm se dedicando a mitigar os efeitos do ruído por meio da investigação de estratégias variadas [1], [6]-[10]. Tais estratégias compreendem abordagens de aprendizado profundo que propõem investigar métricas de avaliação de funções de perda e/ou empregar técnicas de redução de ruído e realce do sinal de fala [6], [7], [11]. Dessa forma, sistemas de KWS robustos ao ruído são realizados usando mascaramento tempo-frequência, seja por meio de um *design* ad hoc do processo de extração de características discriminativas do sinal de fala (*front-end*) e/ou explorando a capacidade das redes neurais convolucionais (*convolutional neural network* - CNN) para extrair mapas de características locais em um modelo de ponta a ponta [12]. Essas técnicas vêm sendo formuladas através do aprendizado supervisionado e do uso de máscaras estimadas a partir de sinais ruidosos, descrevendo as relações tempo-frequência entre fala limpa e com ruído [12]. No entanto, em aplicações do mundo real, quando ruídos desconhecidos são encontrados (diferentes tipos de ruído não presentes nos dados de treinamento), o desempenho dessas técnicas é substancialmente degradado [11].

Por outro lado, o reconhecimento de fala entre humanos é uma tarefa notavelmente robusta em relação ao ruído de fundo, independente do tipo de ruído envolvido [13]. Em particular, considerando que os fonemas (unidade de som de fala) possuem uma configuração espectral-temporal específica (pistas acústicas) necessária para o reconhecimento de palavras, uma razão plausível para um humano reconhecer sons em um ambiente ruidoso está associada à seletividade espectral-temporal, permitindo extrair características acústicas específicas dos sons da fala, tais como formantes e transições consoante-vogal [13]. Essa inferência implica que tanto o cérebro humano quanto seu sistema auditivo levem em consideração alguns mascaramentos psicoacústicos de tal forma que, ao invés de eliminar o ruído, concentre toda a atenção na região de interesse da fala [13], [14]. Nesse contexto, em [15], um mecanismo de atenção é investigado usando uma rede neural para identificar regiões de interesse nos sinais de fala. Em [16], um mecanismo de atenção (do tipo *self-attention*) é considerado através de uma abordagem combinando CNN profunda com memória longa de curto prazo bidirecional (*bidirectional long short-term memory* - BiLSTM). Já em [4], com o objetivo de expandir a abrangência da busca por características discriminativas nas camadas convolucionais, o uso de uma CNN residual (ResNet) com convolução dilatada (CNN dilatada) é considerado.

Apesar de levar em conta os avanços mencionados anteriormente, em [17], é experimentalmente demonstrado que estas arquiteturas para modelagem de sinais de áudio não apresentam fortes evidências na captura de pistas acústicas de fala em ambientes ruidosos. Essa observação pode estar associada às estruturas geométricas fixas dos campos receptivos considerados nas CNNs tradicionais [9], [10], [17]-[19]. Para investigar esse problema, em [17], um novo sistema de realce

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).

de fala é proposto. Especificamente, é discutida uma nova estratégia de realizar camadas convolucionais (denominada CNN harmônica). Em contraste com a dilatação fixa retangular dos campos receptivos empregada na CNN dilatada [4], a CNN harmônica considera uma dilatação variável de acordo com os múltiplos da frequência fundamental do sinal de fala analisado [17]. No entanto, assim como a CNN tradicional, tanto a CNN dilatada quanto a harmônica ainda impõem restrições na forma geométrica dos campos receptivos. Nesse contexto, em contraste com as CNNs tradicionais, que possuem campos receptivos geometricamente fixos e simétricos, as CNNs deformáveis [18] permitem que campos receptivos se adaptem ao formato das características do sinal de entrada, tornando a rede mais robusta a variações do sinal. Apesar de resultados promissores obtidos em áreas como sistemas mecânicos e processamento de imagens [18]-[20], CNNs deformáveis ainda não foram completamente investigadas em aplicações de ASR [8]-[10]. Logo, a estratégia empregada em CNN deformável surge como uma alternativa importante aos sistemas de ASR robustos ao ruído.

Portanto, neste trabalho de pesquisa, visando a seleção de características discriminativas de sinais de fala e inspirado na capacidade do cérebro humano em discriminar adequadamente fala e não fala por meio de mascaramento psicoacústico (prestar atenção apenas nas regiões de interesse, ou seja, em locais do espaço de características que são especialmente relevantes para reconhecimento de fala), investigamos aqui sistemas de KWS robustos ao ruído usando CNNs deformáveis.

II. Fundamentação Teórica

Nesta seção, é apresentada uma breve revisão da operação de convolução deformável. Em seguida, o uso de campos receptivos deformáveis é considerado para identificar/rastrear pistas acústicas em sinais de fala.

A. Revisitando a Operação de Convolução Deformável

Na filtragem espacial linear, a operação de convolução entre um mapa de características de entrada bidimensional (2D) (representado por uma matriz $\mathbf{X}$) e um filtro de convolução $m \times n$ (*kernel*) $\mathbf{W}$ é dada por

$$y(i,j) = \sum_{l \in \mathcal{F}} \sum_{k \in \mathcal{F}} x(i-l, j-k)\, w(l,k), \qquad (1)$$

onde $y(i,j)$ denota a amostra do mapa de características de saída localizada na coordenada $\{i,j\}$, e l e k indicam os deslocamentos associados à coordenada $\{i,j\}$, visando determinar os locais de amostragem da entrada $\mathbf{X}$ [21][1]. Nessa operação, os deslocamentos l e k são definidos por uma grade regular (campo receptivo) $\mathcal{F} = \{l_h, k_p\}\ \forall\ \{h,p \in \mathbb{Z} \mid h \in [1,m] \text{ e } p \in [1,n]\}$. Por exemplo, considerando um *kernel* de convolução 3×3 associado a um campo receptivo $\mathcal{F} = \{(-1,-1),(-1,0),\ldots,(1,0),(1,1)\}$, apenas as amostras correspondentes aos vizinhos mais próximos de $x(i,j)$ são usadas para o cálculo de $y(i,j)$.

Por outro lado, em uma rede convolucional deformável, o campo receptivo $\mathcal{F}$ é modificado por um deslocamento adicional $\Delta\mathcal{F} = \{\Delta l, \Delta k\}$. Assim, para o cálculo $y(i,j)$, são consideradas diferentes amostras de entrada localizadas nas coordenadas $\{i-l-\Delta l, j-k-\Delta k\}$. Em resumo, a operação de convolução deformável consiste em duas etapas: i) amostragem usando uma grade deformável sobre o mapa de características de entrada $\mathbf{X}$; e ii) soma dos valores amostrados pelo campo receptivo deformável $\mathcal{F}+\Delta\mathcal{F}$ e ponderados pelo *kernel* de convolução $w(l,k)$ [18]. Essa operação é ilustrada na Fig. 1 e pode ser expressa como

$$y(i,j) = \sum_{l,\Delta l} \sum_{k,\Delta k} x(i-l-\Delta l, j-k-\Delta k)\, w(l,k). \qquad (2)$$

A ideia subjacente da CNN deformável é permitir o uso de campos receptivos adaptativos. Assim, ao contrário da convolução tradicional, uma convolução deformável tem maior flexibilidade nos locais de amostragem [18], [20]. Nesse contexto, como ilustrado na Fig. 1, a partir do mapa de características de entrada $\mathbf{X}$, o deslocamento $\Delta\mathcal{F}$ é obtido por meio de interpolação utilizando uma camada convolucional adicional. Assim, considerando os locais de amostragem de deslocamento $(\mathcal{F}+\Delta\mathcal{F})$, o *kernel* de convolução $\mathbf{W}$ é obtido usando (2) (para mais detalhes, veja [18]).

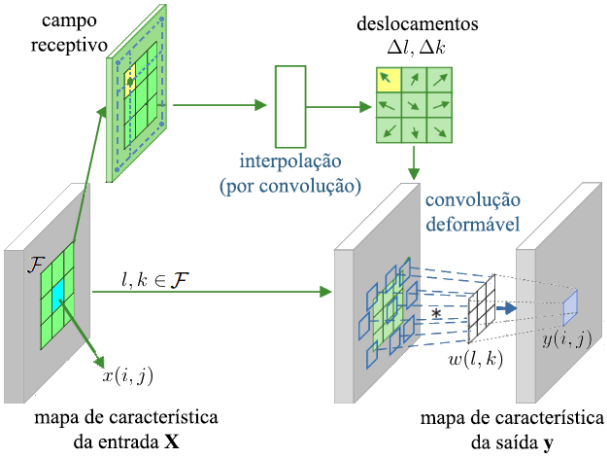


Fig. 1. Ilustração da convolução deformável. Figura adaptada de [18].

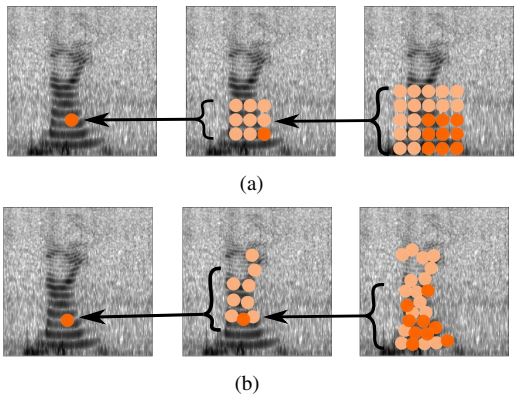
B. Pistas Acústicas e Campos Receptivos Deformáveis

Devido aos locais de amostragem fixos apontados pelos campos receptivos em um modelo de CNN tradicional, as entradas das camadas convolucionais são limitadas a serem analisadas apenas na forma geométrica fixa do campo receptivo em cada ponto da entrada [22]. Assim, esse processo dificulta a identificação de pistas acústicas espalhadas para além dos locais de amostragem apontados por esses campos receptivos (geometricamente fixos) e, consequentemente, faz com que as características extraídas frequentemente sofram o efeito do ruído de fundo [20].

Com o objetivo de explorar a flexibilidade nos locais de amostragem proporcionados pelos campos receptivos deformáveis, as CNNs deformáveis são utilizadas para identificar

[1]Em uma rede convolucional tradicional, o *kernel* geralmente é simétrico e, portanto, a correlação cruzada pode ser usada no lugar da operação de convolução [21], [22]. No caso da rede convolucional deformável, o *kernel* usualmente não é simétrico; dessa forma, temos de usar a operação de convolução.

e rastrear pistas acústicas. Como mostrado na Fig. 2, em contraste com a localização fixa usada na convolução tradicional [ver Fig. 2(a)], na convolução deformável, os locais de amostragem (indicados pelo campo receptivo) podem ser ajustados adaptativamente de acordo com as pistas acústicas (configuração espectral-temporal específica) [veja Fig. 2(b)].



(a)

(b)

Fig. 2. Exemplos de locais de amostragem possíveis de serem realizados pelos campos receptivos considerados. (a) Convolução tradicional. (b) Convolução deformável.

III. Arquiteturas para Sistemas de KWS

Sistemas de KWS são responsáveis por detectar palavras-chave em áudios de fala, visando apresentar alta precisão de reconhecimento [3]. Atualmente, as redes neurais de ponta a ponta vêm se tornando a arquitetura padrão para sistemas de KWS [2]. Tipicamente, os sistemas de KWS do estado-da-arte podem ser divididos em dois blocos principais. O primeiro bloco é construído a partir de CNNs e é projetado para a extração de características discriminativas de sinais de fala, enquanto o segundo é construído a partir de camadas totalmente conectadas (*fully connected* - FC) e tem o propósito de classificar determinadas expressões em um dado sinal de fala.

Neste trabalho, a convolução deformável é considerada para ser usada no bloco de extração de características. Nesse contexto, duas arquiteturas importantes de sistemas de KWS do estado-da-arte (Deep ResNet [4] e DenseNet-BiLSTM com mecanismo de atenção [16]) são usadas para avaliar o desempenho dos sistemas de KWS propostos em termos de precisão de reconhecimento para diferentes níveis de ruído de fundo. As arquiteturas consideradas aqui (veja Fig. 3) são descritas de forma resumida nas próximas seções.

A. Rede Residual Profunda

A rede residual profunda (*deep residual network* - Deep ResNet) [4], [22] pode ser vista como um conjunto de blocos de CNNs (blocos residuais) empilhados (sequencialmente), em que cada bloco consiste em duas camadas convolucionais conectadas em série, incluindo uma conexão de atalho adicional que conecta diretamente a entrada com a saída de cada bloco. Como em [4], neste trabalho de pesquisa, uma arquitetura Deep ResNet com seis blocos de CNN residual (denominada ResNet15) foi utilizada. Em particular, cada bloco residual consiste em duas camadas convolucionais com 45 filtros de convolução de dimensão 3×3 (3×3 conv, 45), seguidas por uma função ReLU (*rectified linear unit*) e uma camada de normalização em lote (*batch normalization* - BN), os quais realizam o processo de extração de características. Finalmente, uma operação de subamostragem (do tipo *pooling* médio [22]) também é utilizada e uma camada FC [com ativação *softmax* de 35 neurônios (unidades)] é implementada no processo de classificação. A Fig. 3(a) ilustra a sequência de operações realizadas por esse modelo de arquitetura. Aqui, como também em [4], o uso de CNNs dilatadas é considerado, e, portanto, uma arquitetura ResNet15-Dilatada também é utilizada para fins de avaliação de desempenho.

B. Rede Densa BiLSTM com Mecanismo de Atenção

Nesta arquitetura, como discutido em [16], o bloco de extração de características é realizado por meio de CNNs densamente conectadas (DenseNet), enquanto o bloco de classificação é realizado usando camadas BiLSTM com um mecanismo de atenção do tipo *self-attention*, seguido de camadas FC. Dessa forma, a saída do bloco de extração de características é utilizada como entrada para a primeira camada BiLSTM, composta por três camadas com 64 neurônios (unidades) em cada célula LSTM. Particularmente, essas camadas extraem características de séries temporais e usam um mecanismo de atenção (*self-attention*) para selecionar apenas os termos temporais mais importantes. Aqui, o cálculo de BiLSTM com atenção permanece como descrito em [16].

No bloco de extração de características, como ilustrado na Fig. 3(b), uma camada BN, uma função de ativação ReLU e um filtro de convolução de tamanho 5×1 seguido por uma subamostragem média de 2×2 são usados antes dos blocos DenseNet e de transição. Na arquitetura DenseNet, ao contrário de uma arquitetura ResNet padrão, todas as saídas das camadas convolucionais são conectadas às próximas camadas de entrada. Além disso, blocos de transição são usados para ajustar as saídas dos blocos densos. Como em [16], em cada bloco denso, o número de camadas é definido como sendo o mesmo. Assim, um conjunto de sequências [BN+ReLU+ conv 1×1, 40] e [BN+ReLU+ conv 3×3, 10] é repetido 6 vezes para cada bloco denso. A Fig. 3(b) ilustra todas as operações consideradas nesse modelo de arquitetura.

C. Arquiteturas usando Convolução Deformável

Visando então extrair pistas acústicas que sejam mais discriminativas para o processo de classificação, CNNs deformáveis são aplicadas nos blocos de extração de características das arquiteturas de sistemas de KWS apresentadas nas seções anteriores. Conforme ilustrado na Fig. 1, todas as camadas convolucionais com filtro de convolução de dimensão 3×3 (3×3 conv) são modificadas com o objetivo de incluir agora a convolução deformável. Neste trabalho de pesquisa, novas arquiteturas, denominadas Deep ResNet15 Deformável (DefResNet15) e DenseNet Deformável BiLSTM com Mecanismo de Atenção (DefDenseBiLSTM), são discutidas em termos de desempenho de reconhecimento considerando diferentes níveis de ruído de fundo.

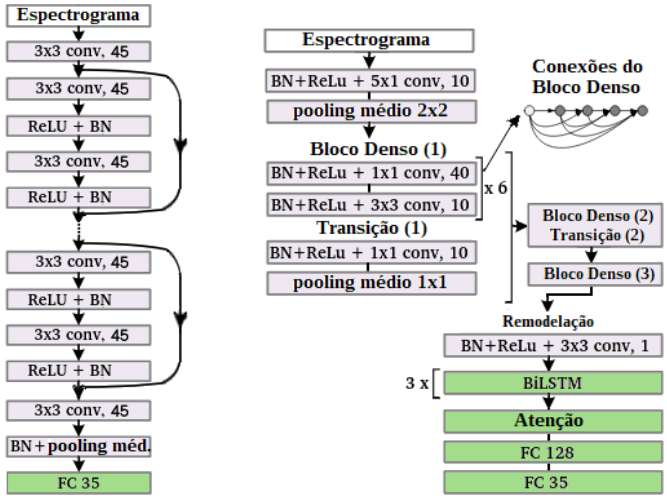


Fig. 3. Modelos de arquitetura de sistemas de KWS do estado-da-arte. (a) Deep ResNet. (b) DenseNet-BiLSTM com mecanismo de atenção.

IV. Simulações Numéricas

Neste trabalho, todos os experimentos foram realizados na linguagem de programação Python usando as bibliotecas TensorFlow e Keras, e o serviço de computação remota "Google Colaboratory", que oferece uma máquina virtual Debian Linux com 4 CPUs Intel Xeon de 2,20 GHz, 25 GB de RAM e uma unidade de processamento gráfico Nvidia Tesla P100 com 16 GB de RAM.

Os experimentos são realizados com base no conjunto de dados GSCD-v2 [23], consistindo de 105.829 sinais de fala correspondentes a 35 palavras (comandos de fala). O método de geração dos conjuntos de treinamento, validação e teste segue os mesmos procedimentos discutidos em [16], no qual listas de arquivos de áudio disponíveis em GSCD-v2 são utilizadas para a separação dos conjuntos. Além disso, esse conjunto de dados também fornece arquivos de áudio com ruído de fundo gerados artificialmente (ruído branco e rosa) e alguns ruídos reais do dia a dia. Assim, neste trabalho de pesquisa, visando avaliar as arquiteturas de sistemas de KWS em ambientes acústicos com baixa SNR, foram gerados novos arquivos de áudio com SNR de 5, 10 e 20 dB, usando os diferentes tipos de ruído disponíveis em GSCD-v2.

As arquiteturas de sistemas de KWS que usam CNNs deformáveis (veja Seção III-C) são avaliadas comparando os seus desempenhos de acurácia no conjunto de teste em contraste com o desempenho das arquiteturas que não utilizam convoluções deformáveis, descritas nas Seções III-A e B. Especificamente, para avaliar o desempenho dos sistemas de KWS que usam convolução deformável, as arquiteturas DefResNet15 e DefDenseNet-BiLSTM (discutidas na Seção III-C) são aqui implementadas. Em seguida, para fins de comparação, tanto a arquitetura ResNet15 quanto a ResNet15-Dilatada são usadas como referências para avaliar o desempenho da arquitetura DefResNet15, enquanto a arquitetura DenseNet-BiLSTM com atenção é utilizada como referência para avaliar o desempenho da arquitetura DefDenseNet-BiLSTM.

A. Treinamento dos Modelos e Configuração dos Parâmetros

Os modelos de arquitetura de sistemas de KWS são treinados e validados (através de tarefa de classificação) visando a maximização da acurácia de reconhecimento dos 35 comandos de fala disponíveis em GSCD-v2. Aqui, todos os modelos de arquitetura são treinados por 30 épocas, a acurácia da validação é examinada a cada época e um ponto de verificação (*checkpoint*) indicando a melhor acurácia é considerado para selecionar o modelo de melhor desempenho. Em seguida, esse modelo é usado no conjunto de teste para avaliar o desempenho final dos sistemas de KWS. Além disso, assim como também discutido em [16], tais modelos são treinados (e validados) usando a parte da base acústica isenta de ruído (SNR = $+\infty$ dB) e são testados em ambientes acústicos com SNRs[2] de 5, 10, 20 e $+\infty$ dB.

Seguindo as configurações de hiperparâmetros consideradas em [4] e [16], a Tabela I apresenta as configurações utilizadas para as arquiteturas com base em ResNet15 e DenseNet-BiLSTM.

TABELA I
Configuração de Hiperparâmetros

Hiperparâmetros	Modelos de arquitetura	
	ResNet15	DenseNet-BiLSTM
Algoritmo de otimização	Grad. desc. estocástico	Adam estocástico
Taxa de aprendizado inicial	$0,1$ (momento $= 0,9$)	$0,001$
Fator de mult. (no platô)	$0,1$	$0,5$
Tam. do lote (*batch size*)	64	100
Regularização l_2	10^{-5}	—

V. Resultados e Análise de Desempenho

Para verificar o desempenho dos sistemas de KWS considerados, os modelos de arquitetura são avaliados levando em consideração (como entrada) espectrogramas log-Mel de dimensão 49×49 (para modelos baseados em ResNet15) e 126×126 (para DenseNet-BiLSTM). Aqui é importante levar em conta que não estamos comparando o desempenho de modelos de arquiteturas diferentes, mas sim avaliando o benefício do uso da operação de convolução deformável em tais modelos.

A Tabela II apresenta a acurácia de teste de cada modelo para níveis de SNR de 10 e $+\infty$ dB, com intervalo de confiança (IC) de 95%, obtido a partir de simulações de Monte Carlo (MC) considerando cinco testes independentes. Além disso, com o objetivo de destacar a complexidade computacional de cada modelo de arquitetura KWS, a Tabela II mostra o número de parâmetros treináveis nesses modelos. Ainda, na Fig. 4, as variações na acurácia de teste dos modelos são mostradas por meio de diagramas de caixa (obtidos a partir das simulações de MC) para diferentes níveis de SNR.

A partir da Tabela II e da Fig. 4, verifica-se que o uso da convolução deformável promove ganhos significativos de acurácia (principalmente em ambientes ruidosos) para os sistemas de KWS considerados. Em relação aos modelos baseados em DenseNet-BiLSTM, obteve-se uma mediana de ganho de acurácia de 2,85%, 1,93% e 0,75% em ambientes ruidosos para SNRs de 5, 10 e 20 dB, respectivamente. Além disso, a

[2]Tipicamente, em sistemas de ASR, ambientes acústicos com SNR $\leq$ 10dB são considerados ambientes de baixo nível de SNR [1], [8].

diferença de complexidade computacional desses modelos, em termos do número de parâmetros treináveis (veja Tabela II), não é significativa. Por outro lado, em relação aos modelos baseados em ResNet15, obtiveram-se ganhos substanciais de acurácia tanto para ambientes livres de ruído quanto para ambientes ruidosos. No entanto, nesse contexto, a complexidade da arquitetura DefResNet15 é maior em comparação com a ResNet15 e ResNet15-Dilatada. Esse aumento na complexidade é esperado devido ao uso de uma camada convolucional extra resultante da convolução deformável.

Por meio dos resultados obtidos por simulações de MC, pode-se inferir que as arquiteturas DefResNet15 e DefDenseNet-BiLSTM promovem locais de amostragem mais discriminativos (devido a uma maior flexibilidade no campo receptivo), resultando em atributos discriminativos que melhor descrevem as pistas acústicas de sinais de fala, mesmo considerando tipos/níveis de ruído que não estão presentes no conjunto de treinamento.

TABELA II

TESTE DE ACURÁCIA COM INTERVALO DE CONFIANÇA DE 95% E CARGA COMPUTACIONAL EM TERMOS DO NÚMERO DE PARÂMETROS TREINÁVEIS

Modelos de arquitetura	Acurácia (%)		Parâmetros treináveis
	10 dB	$+\infty$ dB	
ResNet15	79,7±1,693	94,0±0,631	240k
ResNet15-Dilatada	81,6±2,865	95,5±0,422	240k
DefResNet15	83,2±2,435	96,1±0,150	714k
DenseNet-BiLSTM	75,9±5,455	95,9±0,155	406k
DefDenseNet-BiLSTM	76,9±5,615	95,8±0,210	439k

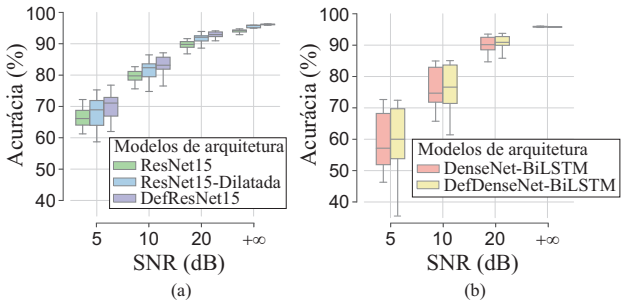


Fig. 4. Avaliação dos modelos de arquitetura de KWS utilizando diagramas de caixa no conjunto de teste. (a) Modelos baseados em ResNet15. (b) Modelos baseados em DenseNet-BiLSTM.

VI. CONCLUSÕES E CONSIDERAÇÕES FINAIS

Neste artigo, foram investigados sistemas de KWS considerando o uso de convolução deformável no processo de extração de características de sinais de fala. Tais sistemas foram implementados usando arquiteturas baseadas em ResNet15 (DefResNet15) e DenseNet-BiLSTM (DefDenseNet-BiLSTM) e avaliados de acordo com a acurácia de reconhecimento dos sistemas de KWS em condições de baixa SNR. Nesse contexto, tanto o DefResNet15 quanto o DefDenseNet-BiLSTM apresentaram melhor desempenho em comparação com as redes que levam em consideração a convolução tradicional. Os resultados de acurácia de reconhecimento dos sistemas de KWS obtidos aqui corroboram a eficácia do uso da convolução deformável para o desenvolvimento desses sistemas.

REFERÊNCIAS

[1] C. Cioflan, L. Cavigelli, M. Rusci, M. de Prado, and L. Benini, "Towards on-device domain adaptation for noise-robust keyword spotting," in *Proc. IEEE 4th Int. Conf. on Artificial Intelligence Circuits and Systems (AICAS)*, Incheon, Republic of Korea, Sep. 2022, pp. 82–85.

[2] L. Liu, M. Yang, X. Gao, Q. Liu, Z. Yuan, and J. Zhou, "Keyword spotting techniques to improve the recognition accuracy of user-defined keywords," *Neural Networks*, vol. 139, pp. 237–245, Jul. 2021.

[3] R. A. Solovyev, M. Vakhrushev, A. Radionov, I. I. Romanova, A. A. Amerikanov, V. Aliev, and A. A. Shvets, "Deep learning approaches for understanding simple speech commands," in *Proc. IEEE 40th Int. Conf. on Electronics and Nanotechnology (ELNANO)*, Kyiv, Ukraine, May 2020, pp. 688–693.

[4] R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Sept. 2018, pp. 8604–8608.

[5] 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.

[6] I. López-Espejo, Z.-H. Tan, and J. Jensen, "A novel loss function and training strategy for noise-robust keyword spotting," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 29, no. 1, pp. 2254–2266, Jul. 2021.

[7] D. Kim, K. Ko, D. K. Han, and H. Ko, "Discriminatory and orthogonal feature learning for noise robust keyword spotting," *IEEE Signal Processing Letters*, vol. 29, pp. 1913–1917, Aug. 2022.

[8] Y. Wang, Y. S. Chong, W. L. Goh, and A. T. Do, "Noise-aware and lightweight lstm for keyword spotting applications," in *Proc. 19th Int. SoC Design Conf. (ISOCC)*, Gangneung-si, Republic of Korea, Oct. 2022, pp. 135–136.

[9] H. B. Nguyen, V. H. Duong, A. X. T. Thi, and Q. C. Nguyen, "Efficient keyword spotting system using deformable convolutional network," *IETE Journal of Research*, pp. 1–9, Jul. 2021.

[10] K. An, Y. Zhang, and Z. Ou, "Deformable tdnn with adaptive receptive fields for speech recognition," in *Proc. Int. Speech Communication Assoc. (INTERSPEECH)*, Brno, Czechia, Sep. 2021, pp. 2067–2071.

[11] C. Yu, R. E. Zezario, S. S. Wang, J. Sherman, Y. Y. Hsieh, X. Lu, H. M. Wang, and Y. Tsao, "Speech enhancement based on denoising autoencoder with multi-branched encoders," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 28, no. 1, pp. 2756–2769, Oct. 2020.

[12] M. Soni, I. Sheikh, and S. K. Kopparapu, "Label-driven time-frequency masking for robust speech command recognition," in *Proc. Int. Conf. on Text, Speech, Dialogue*, Ljubljana, Slovenia, Sep. 2019, pp. 341–351.

[13] F. Khatami and M. A. Escabí, "Spiking network optimized for word recognition in noise predicts auditory system hierarchy," *PLOS Computational Biology*, vol. 16, no. 6, pp. 1–27, Jun. 2020.

[14] N. R. Clark, G. J. Brown, T. Jürgens, and R. Meddis, "A frequency-selective feedback model of auditory efferent suppression and its implications for the recognition of speech in noise." *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1535–41, Sep. 2012.

[15] J. Hou, Y. Shi, M. Ostendorf, M.-Y. Hwang, and L. Xie, "Region proposal network based small-footprint keyword spotting," *IEEE Signal Processing Letters*, vol. 26, no. 10, pp. 1471–1475, Oct. 2019.

[16] M. Zeng and N. Xiao, "Effective combination of DenseNet and BiLSTM for keyword spotting," *IEEE Access*, vol. 7, no. 1, pp. 67–75, Jan. 2019.

[17] Z. Zhang, Y. Wang, C. Gan, J. Wu, J. B. Tenenbaum, A. Torralba, and W. T. Freeman, "Deep audio priors emerge from harmonic convolutional networks," in *Proc. Int. Conf. on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020, pp. 1–12.

[18] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, Venice, Italy, Oct. 2017, pp. 764–773.

[19] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets V2: More deformable, better results," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 9300–9308.

[20] S. Li, W. Yang, A. Zhang, H. Liu, J. Huang, C. Li, and J. Hu, "A novel method of bearing fault diagnosis in time-frequency graphs using InceptionResnet and deformable convolution networks," *IEEE Access*, vol. 8, no. 1, pp. 92 743–92 753, May 2020.

[21] R. E. Gonzalez, Rafael C.; Woods, *Digital Image Processing*, 4th ed. New York, USA: Pearson, 2018.

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

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