Front-end Híbrido e *Back-end* Deformável Aplicados em Sistemas de KWS Robustos ao Ruído

Ênio dos Santos Silva e Rui Seara

***Resumo*— Com o objetivo de desenvolver sistemas de detecção de palavras-chave (*keyword spotting* - KWS) robustos ao ruído, este trabalho de pesquisa discute sobre o uso de *front-end* híbrido e *back-end* deformável, operando em ambientes acústicos com diferentes tipos e níveis de ruído. Nesse contexto, o *front-end* híbrido combina a extração manual (de magnitude e fase de sinais de fala) com a extração automática de atributos realizada através de redes residuais profundas (*deep residual networks* - ResNets). Além do mais, no bloco de *back-end*, a fim de maximizar a probabilidade de compatibilidade dos atributos extraídos com as palavras-chave dos sistemas de KWS, o uso de ResNets deformáveis é também considerado aqui. Assim, sistemas de KWS com *front-end* híbrido e *back-end* deformável são comparados com aqueles que não adotam essas estratégias. Resultados de simulação numérica são mostrados e avaliados com vistas à acurácia de reconhecimento dos sistemas de KWS, confirmando a eficácia dos sistemas propostos neste trabalho.**

***Palavras-Chave*— Detecção de palavras-chave, espectrogramas do sinal de fase, extração de atributos, reconhecimento automático de fala, robustez ao ruído.**

***Abstract*— With the aim of developing noise-robust keyword spotting (KWS) systems, this research work presents an investigation into the use of a hybrid front-end and deformable back-end operating in acoustic environments with different types and noise levels. In this context, a hybrid front-end combines handcrafted speech features (magnitude and phase) with automatic feature extraction performed through deep residual networks (ResNets). Furthermore, in the back-end block, in order to maximize the likelihood of the extracted features with the keywords of the KWS systems, the use of deformable ResNets is also considered here. Thus, KWS systems with a hybrid front-end and deformable back-end are compared with those that do not adopt such strategies. Numerical simulation results are shown and evaluated for keyword recognition performance, confirming the effectiveness of the KWS systems proposed in this work.**

***Keywords*— Keyword spotting, phase spectrograms, feature extraction, automatic speech recognition, noise robustness.**

I. Introdução

O desenvolvimento de sistemas de detecção de palavras-chave (*keyword spotting* - KWS) vem ganhando destaque nos últimos anos, devido especialmente aos avanços na área de aprendizado profundo que têm possibilitado implementações eficientes desses sistemas em uma variedade de aplicações, desde *smartphones* até dispositivos domésticos inteligentes [1].

Sistemas de KWS são estruturas projetadas para reconhecer palavras-chave em sinais de fala, operando tanto de maneira *off-line* em arquivos de áudio quanto de forma *online* em *streaming* de áudio [2], [3]. Essas palavras-chave podem desencadear ações específicas, como ativação de assistentes de voz, controle de dispositivos inteligentes, busca de informações em bancos de dados, dentre outras aplicações [1], [2]. Nesse contexto, os sistemas de KWS são frequentemente implementados para facilitar a comunicação "mãos-livres" com dispositivos eletrônicos [1]. No entanto, esses dispositivos geralmente estão distantes do locutor e operam em ambientes onde o ruído de fundo e a reverberação local podem impactar severamente o desempenho dos sistemas de KWS. Visando o desempenho satisfatório desses sistemas em condições reais de operação, especialmente em aplicações práticas sujeitas a cenários envolvendo sinais com baixa razão sinal-ruído (*signal-to-noise ratio* - SNR), é fundamental levar em conta o desenvolvimento de sistemas robustos ao ruído. Todavia, apesar de a robustez ao ruído ainda ser um problema crítico em aplicações do mundo real, a maioria dos trabalhos de pesquisa do estado-da-arte em KWS não tem ainda abordado de forma satisfatória o impacto do ruído em suas concepções [3]-[5].

Tipicamente, sistemas de KWS do estado-da-arte podem ser divididos em dois blocos principais: *front-end* e *back-end* [1]. O primeiro bloco é responsável pela extração de atributos discriminativos dos sinais de fala [4], enquanto o segundo tem o objetivo de identificar (classificar) se os atributos extraídos correspondem a alguma palavra-chave pré-definida no vocabulário do sistema de KWS [5]. Particularmente, o bloco de *front-end* é implementado *ad hoc* através da extração "manual" de atributos do sinal de fala (como, por exemplo, os coeficientes cepstrais em escala-Mel [2], [6]) e/ou automaticamente através de algoritmos de aprendizado de máquina, este último explorando a capacidade das redes neurais profundas (*deep neural networks* - DNNs) em extrair mapas de atributos locais diretamente dos dados de áudio [utilizando uma estratégia de aprendizagem de ponta a ponta (*end-to-end*) [4], [7]].

Atualmente, apesar dos avanços obtidos em *front-ends* automáticos através de aprendizado de atributos [4], [7], as estratégias de extração manual, principalmente aquelas que utilizam atributos relacionados à escala Mel, continuam sendo amplamente empregadas em sistemas de KWS e têm demonstrado ser ainda alternativas bastante interessantes, sobretudo para cenários com sinais de alta SNR [3], [6], [8]. Por outro lado, em cenários com sinais de baixa SNR, a combinação de técnicas de realce (visando a redução de ruído no sinal de fala) associada com sistemas de reconhecimento automático de fala (*automatic speech recognition* - ASR) vem sendo, nas últimas décadas, amplamente investigada [9]-[11]. Nesse contexto, o uso de toda a informação disponível no sinal,

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).

através dos espectros de magnitude e de fase da transformada de Fourier de curto termo (*short-time Fourier transform* - STFT), é considerado [12]-[15]. No entanto, apesar de a inteligibilidade da fala ruidosa poder ser melhorada através de técnicas de realce, tal abordagem não tem sido ainda definitivamente consagrada para ser utilizada como um *front-end* eficaz em sistemas de ASR robustos [1], [11]. A dificuldade em combinar algoritmos de realce com sistemas de ASR pode ser atribuída à distorção da fala inerente ao processo de realce, resultando em um descasamento importante entre as etapas de treinamento e de teste nos sistemas de ASR [11]. Para contornar tais dificuldades, em [9], é mostrado que o realce de atributos é mais adequado para tarefas de ASR do que o realce dos sinais originais de fala. Assim, em sistemas de ASR robustos ao ruído, a estratégia de realce no domínio dos atributos vem sendo um tópico de pesquisa bastante explorado atualmente [1], [9], [16]. Contudo, esse tipo de *front-end* que considera a remoção de ruído através do realce de atributos, em nosso conhecimento, não tem sido ainda suficientemente investigado para ser utilizado em sistemas de KWS.

No bloco de *back-end*, o desempenho satisfatório dos sistemas de KWS do estado-da-arte está associado à eficácia na utilização de DNNs para modelar a verossimilhança dos atributos com as palavras-chave (classes) correspondentes. Nessas circunstâncias, as arquiteturas de rede capazes de capturar padrões tempo-frequência longos são cruciais para melhorar o desempenho dos sistemas de KWS [1]. Esses padrões podem ser capturados usando DNNs com camadas de convoluções temporais e/ou dilatadas, neste último, aumentando o alcance dos campos receptivos das redes nas camadas convolucionais [1], [17], [18]. Nesse contexto, as redes convolucionais deformáveis, discutidas em [5] e [19], possuem filtros convolucionais com campos receptivos adaptativos, oferecendo uma alternativa interessante para capturar padrões tempo-frequência longos, semelhantes aos campos receptivos dilatados. Portanto, visando suplantar as dificuldades encontradas para a implementação dos blocos de *front-end* e de *back-end* em ambientes com sinais de baixa SNR, este trabalho de pesquisa propõe o desenvolvimento de um *front-end* híbrido. Especificamente, o *front-end* proposto realiza a extração manual de atributos baseados nos log-Mel-espectrogramas dos sinais de magnitude e de fase da STFT do sinal de fala, e, em seguida, realiza a extração automática de mapas de atributos por meio de modelos de aprendizado de máquina usando DNNs. Dessa forma, espera-se que, a partir da STFT do sinal de fala (componentes de magnitude e de fase), a estratégia de processamento de atributos (utilizando DNNs) seja capaz de extrair características robustas ao ruído e discriminativas para o processo de classificação. Adicionalmente, para o bloco de *back-end*, propõe-se a utilização de redes deformáveis a fim de capturar os padrões tempo-frequência longos nos atributos extraídos pelo *front-end* proposto.

II. Sistema de KWS

Neste artigo, o sistema de KWS considerado analisa o sinal $x(n)$ (denotado como um quadro do sinal de fala no domínio do tempo) em dois canais de processamento distintos constituídos pelo par de blocos: *front-end* manual seguido de *front-end* automático. Nessa proposta, a partir da STFT $X_n(\mathrm{e}^{j\omega})$ de $x(n)$, um canal extrai manualmente atributos relacionados ao log-Mel-espectrograma de fase, enquanto o outro extrai atributos de log-Mel-espectrograma de magnitude. Em seguida, cada canal, utilizando redes residuais profundas (*deep residual networks* - ResNets) [17], aprende sobre a extração automática de novos atributos discriminativos. Dessa forma, os atributos extraídos nos dois canais são então concatenados e enviados para o bloco de *back-end*. Finalmente, visando capturar atributos tempo-frequência longos, o bloco de *back-end* utiliza ResNets com camadas convolucionais deformáveis (*deformable deep residual networks* - DefResNets) conectadas a redes de *perceptron* de múltiplas camadas (*multilayer perceptron* - MLP) para a estimação das palavras-chave associadas ao sinal $x(n)$. O sistema de KWS considerado é ilustrado na Fig. 1.

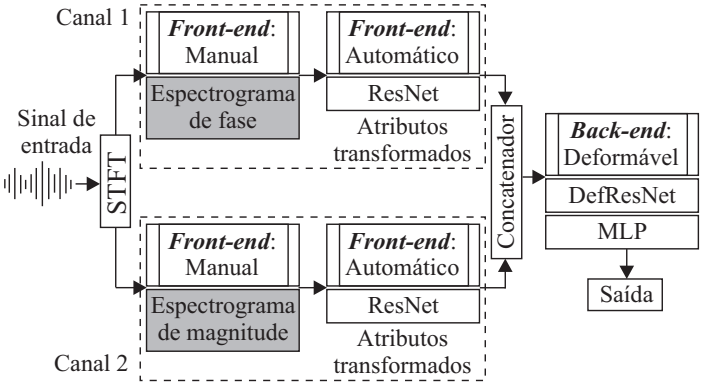

Fig. 1. Diagrama de blocos para a implementação dos sistemas de KWS propostos.

A. Considerações Sobre o Uso do Sinal de Fase

Em ambientes com sinais de alta SNR, log-Mel-espectrogramas de magnitude são capazes de capturar satisfatoriamente certas características tempo-frequência (assinaturas espectrais) que descrevem unidades específicas de sons de fala (fonemas, palavras, dentre outras) [1], [6]. No entanto, em condições de baixa SNR, enquanto o sinal de magnitude pode ser fortemente corrompido pelo ruído, o sinal de fase pode preservar informações valiosas sobre a estrutura tempo-frequência do sinal [12]-[14]. Figurativamente, pode-se interpretar que, no espectro de magnitude, a informação se assemelha a uma modulação em amplitude (AM), enquanto no espectro de fase é como se ocorresse uma modulação em frequência (FM), onde a informação está codificada na variação da fase do sinal [20]. Essas conjecturas presumem que sinais representados na forma de AM ou FM possuem características distintas em relação à robustez ao ruído. Nessa perspectiva, [1] sugere que trabalhos futuros mais promissores sobre sistemas de KWS robustos explorem soluções que possam se beneficiar das informações de fase.

Por sua vez, devido à característica de "empacotamento" da fase (*phase wrapping*) (módulo 2π), o espectro do sinal de fase $\theta_n[X_n(\mathrm{e}^{j\omega})]$, em sua forma original, não representa a "verdadeira" fase dos sinais e, consequentemente, não apresenta diretamente uma estrutura (tempo-frequência) adequada

para ser usada nos sistemas de KWS [12], [14]. Para contornar esse problema, [12] e [14] consideram utilizar a função atraso de grupo $\tau_n(\mathrm{e}^{j\omega})$, que é definida como a derivada negativa de $\theta_n[X_n(\mathrm{e}^{j\omega})]$ (nesse caso, $\theta_n[X_n(\mathrm{e}^{j\omega})]$ representa a função "desempacotada" do espectro de fase). Particularmente, uma versão modificada da função atraso de grupo $\tilde{\tau}_n(\mathrm{e}^{j\omega})$ (*modified group delay* - MOGD) é computada para se obter $\theta_n[X_n(\mathrm{e}^{j\omega})]$. Assim,

$$\tilde{\tau}_n(\mathrm{e}^{j\omega}) = \frac{Y_{\mathrm{I}}(\mathrm{e}^{j\omega})X_{\mathrm{I}}(\mathrm{e}^{j\omega}) + Y_{\mathrm{R}}(\mathrm{e}^{j\omega})X_{\mathrm{R}}(\mathrm{e}^{j\omega})}{|X_n(\mathrm{e}^{j\omega})|^{2\gamma}} \qquad (1)$$

onde γ denota 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 $Y_{\mathrm{I}}(\mathrm{e}^{j\omega})$ representam, respectivamente, a parte real e a parte imaginária da STFT de $\mathbf{x}(n)$ e de $n\mathbf{x}(n)$ (para mais detalhes, veja [12], [13] e [14]).

B. Considerações Sobre o Uso de ResNets

Em sistemas de KWS, a precisão na representação dos sinais de fala é crucial para a correta identificação das palavras-chave. Nesse sentido, para operar satisfatoriamente em ambientes com sinais de baixa SNR, assim como tem sido discutida em [1], a abordagem proposta neste trabalho considera utilizar ResNets para aprender representações discriminativas dos dados (atributos) processados. O uso de conexões residuais permite o treinamento de modelos mais profundos de forma rápida e eficaz, melhorando significativamente o desempenho dos sistemas de KWS. Tais conexões permitem que o modelo aprenda representações mais complexas dos dados, incluindo padrões tempo-frequência longos.

C. Sobre a Utilização de Convoluções Deformáveis

O estado da arte em sistemas de KWS pressupõe que a utilização de ResNets com rede neural convolucional (*convolutuinal neural network* - CNN) dilatada oferece uma solução promissora para o bloco de *back-end*, proporcionando robustez e melhor desempenho na identificação de palavras-chave em ambientes adversos [1], [18]. Particularmente, em contraste com a operação de dilatação [18] dos campos receptivos $\mathcal{F}$ (geometricamente fixos e simétricos) empregada nas CNNs dilatadas, as CNNs deformáveis [19] permitem que os campos receptivos sejam geometricamente adaptáveis (deformáveis) ($\mathcal{F} + \Delta\mathcal{F}$), moldando-se, através da operação de convolução deformável [19], aos formatos das características dos atributos processados. Nesse contexto, considerando que $x(i,j)$ caracteriza o valor do mapa bidimensional de atributos $\mathbf{X}$ localizado nas coordenadas $\{i,j\}$, o cálculo do mapa de atributos de saída $\mathbf{Y}$ nas correspondentes coordenadas pode ser expresso como segue:

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

onde l e k denotam os deslocamentos associados às coordenadas $\{i,j\}$ e ao *kernel* $w(l,k)$. Dessa forma, a operação de convolução deformável computa a soma das multiplicações de $w(l,k)$ pelos valores do mapa de atributos $\mathbf{X}$ obtidos por meio de um campo receptivo deformável $\mathcal{F}+\Delta\mathcal{F}$ nas coordenadas $\{i-l-\Delta l, j-k-\Delta k\}$. Esse procedimento é ilustrado na Fig. 2 (para mais detalhes, veja [5] e [19]).

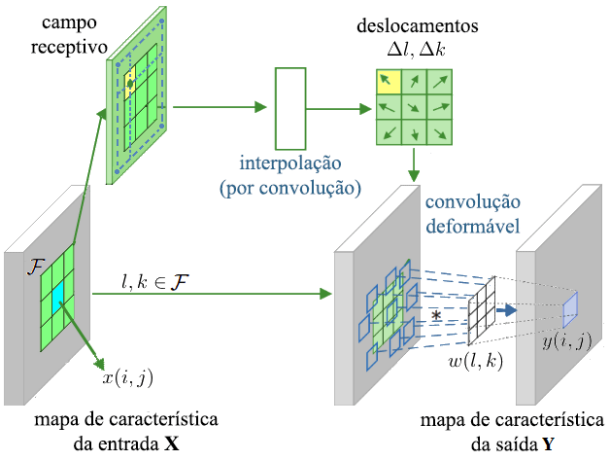

Fig. 2. Esquema ilustrando uma operação de convolução deformável. Figura adaptada de [19].

III. Simulações Numéricas

Neste trabalho, todos os experimentos são realizados usando uma GPU Nvidia RTX 3090 de 24 GB em um sistema Debian Linux com 24 CPUs Intel Core i7 de 5,20 GHz e 64 GB de RAM. Todos os códigos são implementados na linguagem de programação Python, com o auxílio da biblioteca PyTorch. Para treinamento, validação e teste dos sistemas de KWS, tem sido utilizada a segunda versão do conjunto de dados de comandos de fala do Google (*Google speech command database* - GSCD-v2). Esse conjunto de dados consiste de 105829 arquivos de áudio (amostrados em 16 kHz) correspondentes a 35 comandos de fala. A obtenção dos conjuntos de treinamento, validação e teste segue os procedimentos descritos em [5], [12], que utiliza listas pré-definidas de arquivos de áudio do GSCD-v2 para a separação desses conjuntos. Além disso, o GSCD-v2 inclui arquivos de áudio com ruídos artificiais (ruído branco e ruído rosa) e ruídos reais. Para avaliar os sistemas de KWS em ambientes com sinais de baixa SNR, são gerados outros arquivos de áudio com SNR de $-5, 0, 5, 10$ e 20 dB, utilizando os diferentes tipos de ruído disponíveis no GSCD-v2. Particularmente, assim como discutido em [12], devido ao processo de aquisição dos áudios da base de dados considerada não seguir um padrão de qualidade controlado, 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, a SNR verdadeira será menor do que a especificada.

Os sistemas de KWS propostos na Seção II (ilustrado na Fig. 1), que utilizam dois canais de processamento paralelos no bloco de *front-end* (onde cada canal processa exclusivamente atributos provenientes dos espectrogramas de magnitude ou de fase) e a DefResNet no bloco de *back-end*, são avaliados quanto ao seus desempenhos de acurácia (porcentagem de palavras classificadas corretamente) no conjunto de teste. Esses desempenhos são comparados com o desempenho dos demais sistemas mencionados na Tabela I. Nesse contexto, os Sistemas de KWS S1 e S2 correspondem a, respectivamente, um sistema que utiliza no bloco de *front-end* apenas atributos de magnitude e a um sistema que utiliza apenas atributos de fase. Os sistemas de KWS propostos combinam ambos os tipos de atributos, sendo que o Sistema Proposto #1 não leva em

consideração o uso de convoluções deformáveis (discutidas na Seção II-C) no bloco de *back-end*, enquanto o Sistema Proposto #2 considera o uso de convoluções deformáveis. Dessa forma, visando obter modelos com a mesma complexidade computacional, todos os sistemas de KWS são implementados utilizando o bloco de *front-end* com processamento em dois canais. A Tabela II mostra a estruturação da arquitetura dos sistemas de KWS propostos que são utilizadas como referência para implementar os Sistemas S1 e S2.

TABELA I

SISTEMAS DE KWS CONSIDERADOS

Sistemas de KWS	Atributos no *Front-end*		Processamento no *Back-end*
	Canal 1	Canal 2	
S1	Magnitude	Magnitude	ResNet
S2	Fase	Fase	ResNet
Proposto #1	**Magnitude**	**Fase**	**ResNet**
Proposto #2	**Magnitude**	**Fase**	**DefResNet**

TABELA II

ESTRUTURAÇÃO DA ARQUITETURA DO SISTEMA DE KWS PROPOSTO #2

Blocos	Camadas de processamento		Formato de saída
Front-end Manual	Espectrogramas de magnitude e de fase		$40 \times 66 \times 2$
Front-end Automático	Canal 1: ResNet	64 *kernels* 7×7 + ReLu $4\times$ [$2\times$ [N_c *kernels* 3×3 BN + ReLu]] Subamostragem	$5 \times 9 \times 512$
	Canal 2: ResNet	64 *kernels* 7×7 + ReLU $4\times$ [$2\times$ [N_c *kernels* 3×3 BN + ReLu]] Subamostragem	$5 \times 9 \times 512$
Back-end	Concatenador		$5 \times 9 \times 1024$
	Def ResNet	64 *kernels* deformáveis 7×7 + ReLu $4\times$ [$2\times$ [N_c *kernels* 3×3 BN + ReLu]] Subamostragem	1×512
	MLP totalmente conectada, 35		1×35

A estrutura mostrada na Tabela II considera blocos de ResNets que são organizados em quatro conjuntos ($4\times$) de acordo com os seus correspondentes N_c números de *kernels* ($N_c \in \{64, 128, 256, 512\}$). Especificamente, cada bloco é composto inicialmente por uma camada convolucional com 64 *kernels* 7×7 e ativação com unidade linear retificada (*rectified linear unit* - ReLU). Em seguida, dois conjuntos ($2\times$) compostos por uma camada convolucional com N_c *kernels* 3×3, normalização (*batch normalization* - BN) e ReLu são utilizados. Após cada conjunto de blocos residuais, uma operação de subamostragem (do tipo *average-pooling* [3]) é realizada. No bloco de *back-end*, além da DefResNet, uma camada de MLP totalmente conectada com 35 neurônios é implementada para a classificação final dos comandos de fala. Particularmente, seguindo as configurações de hiperparâmetros consideradas em [5], aqui o tamanho do lote (*batch size*) usado é de 128 e o algoritmo de otimização adotado é o Adam [1], com taxa de aprendizagem inicial de 10^{-3} e fator de decaimento (no platô) de $0,1$. Dessa forma, os sistemas de KWS são treinados e validados visando maximizar a acurácia de reconhecimento dos 35 comandos de fala disponíveis em GSCD-v2. Todos os modelos dos sistemas de KWS são treinados por 160 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 que apresente 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 discutido em [5] e [11], tais modelos são treinados (e validados) usando a parte do conjunto de dados isenta de ruído (SNR = $+\infty$ dB) e são testados em ambientes acústicos com sinais com SNRs de $-5, 0, 5, 10, 20$ e $+\infty$ dB.

IV. RESULTADOS E ANÁLISE DE DESEMPENHO

A Tabela III mostra os resultados de acurácia dos sistemas de KWS mencionados na Tabela I, avaliados no conjunto de teste com intervalo de confiança (IC) de 95%, obtidos através de simulação de Monte Carlo (MC), usando cinco repetições independentes para cada sistema. Especificamente, a Tabela III apresenta os resultados da acurácia média de cada sistema de KWS operando em diferentes tipos de ruídos com SNR $= [-5, 0, 5, 10, 20]$ dB e também a acurácia dos sistemas de KWS avaliados no conjunto de teste correspondente apenas à parte do GSCD-v2 isenta de ruído (SNR $= +\infty$ dB).

TABELA III

ACURÁCIA (%) DE RECONHECIMENTO COM IC DE 95%.

Sistemas de KWS	SNR (dB)						
	-5	0	5	10	20	$+\infty$	Média
S1	16,95	44,32	65,46	78,81	89,76	94,87	67,45
S2	17,36	45,38	68,54	80,56	90,76	94,26	69,03
Proposto #1	**19,33**	**47,04**	**69,06**	**80,51**	**91,01**	**95,18**	**69,93**
Proposto #2	**20,01**	**48,64**	**69,97**	**80,57**	**91,02**	**95,25**	**70,09**

Na Fig. 3, as variações na acurácia de teste dos sistemas são mostradas por meio de diagrama de caixa (obtidos a partir de simulação de MC) para diferentes níveis de SNRs.

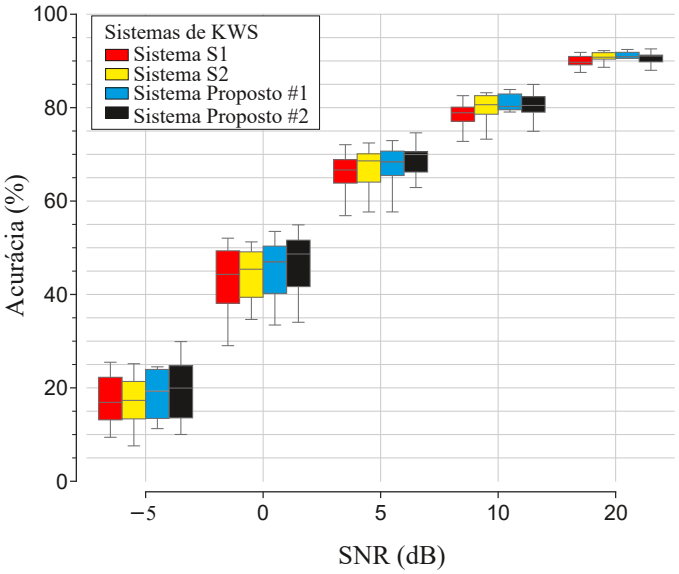

Fig. 3. Diagrama de caixa ilustrando os resultados de acurácia dos sistemas de KWS operando em ambientes ruidosos.

A partir dos resultados obtidos (por simulação de MC), comparando o desempenho do Sistema Proposto #1 com o dos Sistemas S1 e S2, pode-se inferir que a estratégia de utilização do bloco de *front-end* híbrido proposto neste artigo (implementada no Sistema Proposto #1, combinando extração manual e automática de atributos) é mais eficaz para operar sistemas de KWS sujeitos a diferentes tipos e níveis de

ruído, quando comparada com os outros sistemas discutidos aqui. Essa estratégia de processamento, em contraste com os Sistemas S1 e S2, possibilita a extração de atributos mais discriminativos que melhor descrevem as assinaturas espectrais dos sinais de fala, mesmo considerando tipos e níveis de ruído que não estão presentes no conjunto de treinamento. Esses experimentos confirmam a hipótese de que o desempenho dos sistemas de KWS utilizando atributos obtidos via magnitude ou fase sofrem variações na presença de diferentes tipos e níveis de ruído. Nesse sentido, o Sistema S2 que opera com atributos oriundos da fase proporciona maior acurácia em cenários ruidosos do que o Sistema S1 que opera com atributos obtidos da magnitude. No entanto, em cenários isentos de ruído, o Sistema S1 apresenta maior acurácia do que o Sistema S2. Nesse contexto, o Sistema Proposto #1 combina as vantagens dos Sistemas S1 e S2, apresentando maior acurácia tanto em cenários ruidosos quanto em cenários isentos de ruído.

Finalmente, ao comparar o Sistema de KWS Proposto #2 com o Sistema Proposto #1, baseados nos resultados apresentados na Tabela III e na Fig. 3, pode-se observar que o uso de DefResNet no bloco de *back-end* leva a um melhor desempenho de classificação em comparação com o modelo que não utiliza DefResNet. Isso corrobora a hipótese de que redes residuais deformáveis são mais eficazes para capturar padrões tempo-frequência longos (refinando a modelagem da verossimilhança dos atributos com as palavras-chave dos sistemas de KWS) em comparação com redes residuais sem *kernels* deformáveis.

V. Conclusões e Considerações Finais

Neste trabalho de pesquisa, foram investigados sistemas de KWS implementados através de arquiteturas de DNNs residuais utilizando *front-end* híbrido (manual e automático) e *back-end* deformável, visando robustez para operação em ambientes contaminados por ruído. Nessa proposta, o bloco de *front-end* levou em conta a combinação da expertise humana na seleção de características manuais (por meio da extração de espectrogramas de magnitude e de fase) com a eficiência do aprendizado de máquina (através do uso de ResNets) na extração automática de atributos. Além disso, no bloco de *back-end*, redes residuais deformáveis (DefResNet) foram utilizadas visando maximizar a verossimilhança dos atributos obtidos com as palavras-chave dos sistemas de KWS. Tais sistemas foram avaliados de acordo com a acurácia de reconhecimento em condições de baixa SNR. Nesse contexto, os sistemas de KWS propostos neste artigo apresentaram melhores desempenhos em comparação com os Sistemas de KWS S1 e S2 que levam em consideração unicamente atributos provenientes dos espectrogramas de magnitude ou de fase, ou que não utilizam convoluções deformáveis. Dessa forma, pode-se inferir que os sistemas de KWS propostos possibilitaram representações robustas dos dados acústicos e a obtenção de atributos mais discriminativos ao processo de classificação/detecção das palavras-chave. Os resultados de acurácia obtidos através de simulações numéricas corroboraram a eficácia dos sistemas de KWS propostos, demonstrando seus potenciais para aplicações práticas demandantes de reconhecimento de palavras-chave.

Referências

[1] I. López-Espejo, Z. Tan, J. H. L. Hansen, and J. H. Jensen, "Deep spoken keyword spotting: An overview," *IEEE Access*, vol. 10, pp. 4169–4199, 2022.

[2] I. Syafalni, C. Amadeus, N. Sutisna, and T. Adiono, "Efficient real-time smart keyword spotting using spectrogram-based hybrid CNN-LSTM for edge system," *IEEE Access*, vol. 12, pp. 43 109–43 125, 2024.

[3] P. H. Pereira, W. Beccaro, and M. A. Ramírez, "Evaluating robustness to noise and compression of deep neural networks for keyword spotting," *IEEE Access*, vol. 11, pp. 53 224–53 236, 2023.

[4] I. López-Espejo, R. C. C. Shekar, Z. Tan, J. H. Jensen, and J. H. L. Hansen, "Filterbank learning for noise-robust small-footprint keyword spotting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.

[5] E. S. Silva and R. Seara, "Rede convolucional deformável aplicada a sistemas de KWS robustos ao ruído," in *Anais do Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT)*, São José dos Campos, SP, Out. 2023, pp. 1–5.

[6] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122 136–122 158, 2022.

[7] P. Vitolo, R. Liguori, L. D. Benedetto, A. Rubino, and G. D. Licciardo, "Automatic audio feature extraction for keyword spotting," *IEEE Signal Processing Letters*, vol. 31, pp. 161–165, 2024.

[8] I. López-Espejo, Z. Tan, and J. H. Jensen, "An experimental study on light speech features for small-footprint keyword spotting," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Granada, Spain, Nov. 2022, pp. 131–135.

[9] P. Wang and D. Wang, "Enhanced spectral features for distortion-independent acoustic modeling," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Graz, Austria, Nov. 2019, pp. 476–480.

[10] C. Yang, Y. M. Saidutta, R. S. Srinivasa, C. H. Lee, Y. Shen, and H. Jin, "Robust keyword spotting for noisy environments by leveraging speech enhancement and speech presence probability," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Dublin, Ireland, Aug. 2023, pp. 4913–4917.

[11] Y. Yang, A. Pandey, and D. Wang, "Time-domain speech enhancement for robust automatic speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc (INTERSPEECH)*, Dublin, Ireland, Aug. 2023, pp. 4913–4917.

[12] E. S. Silva and R. Seara, "Sistemas de reconhecimento automático de fala baseados em redes neurais profundas usando espectrogramas do sinal de fase," in *Anais do Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT)*, Florianópolis, SC, Nov. 2020, pp. 1–5.

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

[14] A. Dutta, A. P. Gudmalwar, and C. V. R. Rao, "Phase based spectro-temporal features for building a robust ASR system," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Shanghai, China, Oct. 2020, pp. 1668–1672.

[15] G. A. Prabhakar, B. Basel, A. Dutta, and C. V. R. Rao, "Multichannel CNN-BLSTM architecture for speech emotion recognition system by fusion of magnitude and phase spectral features using DCCA for consumer applications," *IEEE Trans. Consum. Electron.*, vol. 69, no. 2, pp. 226–235, May 2023.

[16] G. Wei, Z. Duan, S. Li, X. Yu, and G. Yang, "LFEformer: Local feature enhancement using sliding window with deformability for automatic speech recognition," *IEEE Signal Processing Letters*, vol. 30, pp. 180–184, 2023.

[17] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 11 966–11 976.

[18] A. Coucke, M. Chlieh, T. Gisselbrecht, D. Leroy, M. Poumeyrol, and T. Lavril, "Efficient keyword spotting using dilated convolutions and gating," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, UK, May 2019, pp. 6351–6355.

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

[20] E. Loweimi, J. Barker, O. S. Torralba, and T. Hain, "Robust source-filter separation of speech signal in the phase domain," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Stockholm, Sweden, Aug. 2017, pp. 414–418.