Estratégias de Combinação de Espectrogramas de Magnitude e de Fase Aplicadas em Sistemas Robustos de Detecção de Palavras-Chave

Ênio dos Santos Silva e Rui Seara

***Resumo*— A demanda por sistemas de detecção de palavras-chave (*keyword spotting* - KWS) vem crescendo consideravelmente para as mais diversas aplicações do mundo real. No entanto, o desempenho desses sistemas é fortemente degradado em condições de operação com baixa razão sinal-ruído (*signal-to-noise ratio* - SNR). Visando a obtenção de sistemas de KWS robustos ao ruído, este trabalho de pesquisa investiga o processo de extração de atributos nesses sistemas. Particularmente, o presente trabalho propõe o uso de estratégias de combinação de atributos considerando os espectrogramas de magnitude e de fase dos sinais de fala. Dessa forma, sistemas de KWS utilizando extração de atributos considerando a combinação da magnitude e da fase são contrastados com aqueles que utilizam apenas espectrogramas de magnitude. Resultados de simulação numérica são apresentados e avaliados com vistas à acurácia de reconhecimento de palavras-chave, confirmando a eficácia das estratégias utilizadas neste trabalho.**

***Palavras-Chave*— Comitê de classificadores, detecção de palavras-chave, espectrogramas do sinal de fase, extração de atributos.**

***Abstract*— The demand for keyword spotting (KWS) systems has been increasing considerably for the most diverse applications in the real world. However, the performance of these systems is severely degraded under operating conditions with a low signal-to-noise ratio (SNR). Aiming to obtain KWS systems robust to noise, this research work presents an investigation about the feature extraction process performed in these systems. Particularly, the present work proposes the use of strategies for combining features taking into account the magnitude and phase spectrograms of speech signals. In this way, KWS systems using feature extraction considering combinations of the magnitude and phase are contrasted with those using only magnitude spectrograms. Numerical simulation results are shown and assessed with respect to the accuracy of keyword recognition, confirming the effectiveness of the strategies used in this work.**

***Keywords*— Stacked generalization, keyword spotting, phase spectrograms, feature extraction.**

I. Introdução

Atualmente, o aumento do interesse da indústria por serviços de tecnologias envolvendo o processamento de fala vem motivando a comunidade científica para o desenvolvimento de sistemas robustos aplicados a diferentes cenários do mundo real [1], [2]. Nos últimos anos, a demanda por sistemas que utilizam tecnologias de fala (por exemplo, aplicações de vídeo/áudio conferências, atendimentos automatizados via *call-center*, assistentes pessoais virtuais, dentre outras aplicações) tornou-se um caminho exitoso de aproximar empresas e clientes [1], [2]. Recentemente, a coleta de dados de áudio, provenientes de serviços de tecnologias de fala, vem sendo cada vez mais utilizada por empresas nos mais diferentes segmentos [1]. Esse processo vem gerando uma quantidade importante de dados de áudio não estruturados, mas que contêm informações de grande valia para atender uma relação satisfatória entre empresas-clientes. Nesse caso, sistemas que possibilitem uma interação automatizada homem-máquina, bem como sistemas de buscas por palavras-chave relacionadas a determinados contextos, dentre outras aplicações, vêm se destacando como ferramentas essenciais para obtenção de algumas informações relevantes visando a expansão comercial de indústrias e empresas [1], [2]. Dessa forma, sistemas destinados à detecção de palavras-chave (*keyword spotting* - KWS) vêm se tornando um procedimento rotineiro para extrair certas informações de arquivos de áudio e/ou reconhecer certos comandos de fala em *streaming* de áudio [3], [4], [5].

Os sistemas atuais de reconhecimento automático de fala (*automatic speech recognition* – ASR) têm exibido desempenho satisfatório em cenários acústicos com níveis de ruído controlados, contudo, em ambientes com baixa razão sinal-ruído (*signal-to-noise ratio* - SNR), a operação desses sistemas se torna severamente prejudicada [6]. Nesse contexto, 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 levado em consideração (de forma eficaz) os efeitos do ruído [7], [8].

Em [8] e [9], são discutidas diversas estratégias de redução de ruído e realce do sinal de fala. Recentemente, com o desenvolvimento das técnicas de aprendizado profundo, grandes avanços vêm sendo alcançados nessas áreas de aplicação. Nesse contexto, o uso de toda a informação disponível no sinal de fala através dos espectros de magnitude e de fase da transformada de Fourier de curto termo (*short-time Fourier transform* - STFT) é valorizado. Tais informações (magnitude e fase) são processadas por meio de redes neurais profundas (*deep neural network* – DNN) e vêm apresentando resultados promissores para operar em cenários com baixa SNR.

Em [5], [10], [11] e [12], sistemas de ASR e KWS têm tirado proveito da informação do sinal de fase obtido através da STFT. Em [10] e [11], os sinais de magnitude e fase são considerados de forma independente. Enquanto, em [12], os sinais (magnitude e fase) são considerados conjuntamente; entretanto, em nosso conhecimento, tal abordagem não está sendo ainda aplicada aos sistemas de ASR atuais. Visando aliar as vantagens do uso do sinal de fase em aplicações de realce de fala com os atuais avanços dos sistemas de ASR empregando técnicas de aprendizado profundo, em [5], são apresentados

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).

sistemas de KWS baseados em modelos de DNN do estado-da-arte, utilizando sinais de fase combinados com sinais de magnitude. No entanto, apesar de [5] apresentar resultados relevantes, a investigação sobre estratégias de combinação entre os sinais de magnitude e de fase não foi lá completamente explorada.

Neste trabalho, com o propósito de obter sistemas de KWS robustos orientados a aplicações do mundo real e operando em condições com baixa SNR, três estratégias de combinação usando sinais de magnitude e de fase (obtidos através da STFT) são propostas e discutidas. Resultados de simulação numérica são apresentados e avaliados com vistas à acurácia de reconhecimento de palavras-chave, confirmando a eficácia das estratégias utilizadas neste trabalho.

II. Sistemas de Detecção de Palavras-Chave

Sistemas de KWS têm como objetivo a identificação automática de palavras-chave, operando de maneira *online* em *streaming* de áudio ou de forma *offline* em arquivos de áudio [3], [4]. Preferencialmente, ambos os modos de operação devem proporcionar alta acurácia de reconhecimento, apresentando desempenhos robustos em aplicações práticas sujeitas a cenários acústicos com baixa SNR (para detalhes, veja [7]).

Tipicamente, sistemas de KWS do estado-da-arte podem ser divididos em dois blocos principais: *front-end* e *back-end* [6]. O primeiro bloco é usualmente implementado por redes neurais convolucionais (*convolutional neural network* - CNN) e é projetado para a extração de atributos discriminativos dos sinais de fala [6]; enquanto, o segundo bloco pode ser implementado através de redes de *perceptron* de múltiplas camadas (*multilayer perceptron* - MLP) e tem o objetivo de, a partir dos atributos extraídos no bloco de *front-end*, identificar (classificar) se os atributos examinados correspondem a alguma palavra-chave pré-definida no vocabulário do sistema de KWS [4]. A Fig. 1 ilustra a arquitetura típica de um sistema de KWS do estado-da-arte.

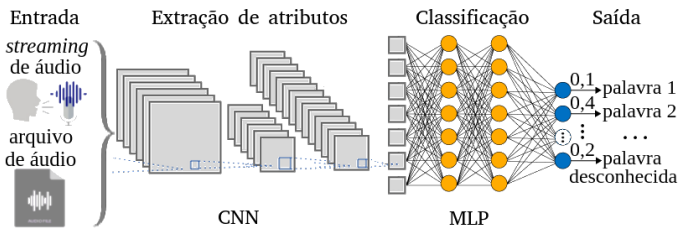

Fig. 1. Ilustração de uma arquitetura típica de sistemas de KWS.

A. Espectrogramas do Sinal de Fala

No contexto geral de sistemas de ASR, atributos de tempo-frequência (espectrogramas) provenientes dos espectros de magnitude $|X_n(\mathrm{e}^{j\omega})|$, obtidos da STFT[1] do sinal de fala $\mathbf{x}(n)$, têm apresentado resultados satisfatórios em aplicações de ASR operando em ambientes acústicos com alta SNR [6], [10]. Nesses ambientes, os espectrogramas de magnitude são capazes de capturar as características harmônicas e de transição entre diferentes unidades fonéticas (fonemas) [10]. A fim de preservar essas características acústicas, em ambientes com baixa SNR, a combinação de diferentes tipos de espectrogramas vem sendo investigada na literatura aberta [11], [12]. Dessa forma, [5], [10] e [11] propõem a utilização de espectrogramas do sinal de fase compostos por espectros de fase $\theta_n[X_n(\mathrm{e}^{j\omega})]$. Particularmente, devido ao problema do "empacotamento" da fase (*phase wrapping*) (módulo 2π), $\theta_n[X_n(\mathrm{e}^{j\omega})]$ não representa a "verdadeira" fase dos sinais e, consequentemente, não apresenta diretamente uma estrutura adequada para ser usada nos sistemas de KWS. Para contornar esse problema, [5], [10] e [11] usam uma versão modificada da função atraso de grupo $\tilde{\tau}_n(\mathrm{e}^{j\omega})$ (*modified group delay* - MOGD) para representar o espectro de fase da STFT, a qual pode ser definida como uma aproximação da derivada do espectro de fase[2] $\theta_n[X_n(\mathrm{e}^{j\omega})]$ e expressa por

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

onde γ caracteriza 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 [5] e [10]).

B. Arquiteturas de Sistemas de KWS

1) Bloco de Front-end: Aqui, redes residuais profundas (*deep residual networks* - ResNet) [3] são consideradas para realizar a extração de atributos discriminativos dos sinais de fala no bloco de *front-end*. Em resumo, uma ResNet pode ser vista como um conjunto de CNNs empilhadas sequencialmente, em que cada conjunto é constituído por duas camadas convolucionais em série possuindo uma conexão de atalho que liga diretamente a entrada com a saída desses conjuntos. Tais conjuntos são comumente denominados blocos residuais [6], [7]. Particularmente, neste trabalho de pesquisa, cada bloco residual consiste de duas camadas convolucionais com 45 filtros de convolução, de dimensão 3×3 (3×3 conv, 45), seguidas por uma função de ativação de unidade linear retificada (*rectified linear unit* - ReLU) e uma camada de normalização em lote (*batch normalization* - BN). Assim, como descrito em [3], uma ResNet com seis blocos residuais (denominada Res15), seguida por uma operação de subamostragem (do tipo *pooling* médio [6]), também é aqui utilizada.

2) Bloco de Back-end: Conforme ilustrado na Fig. 1, a etapa de classificação é realizada por redes de MLP. Particularmente, é utilizada aqui uma rede de MLP com camadas completamente conectadas (*fully connected* - FC) compostas por 12 neurônios com ativação *Softmax* [6].

Desta forma, os blocos de *front-end* e de *back-end* descritos nesta seção são utilizados como base para o desenvolvimento e avaliação dos sistemas de KWS usando as estratégias de combinação de atributos aqui propostas. A Tabela I apresenta uma descrição detalhada sobre os blocos que compõem os sistemas de KWS investigados neste artigo.

C. Base Acústica

Particularmente, os sistemas de KWS são avaliados aqui usando a segunda versão do banco de dados de comandos de

[1]Representada em sua forma polar $|X_n(\mathrm{e}^{j\omega})|\mathrm{e}^{j\theta_n[X_n(\mathrm{e}^{j\omega})]}$.

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

TABELA I
ARQUITETURA PADRÃO PARA A REALIZAÇÃO DE SISTEMAS DE KWS

Processos	Camadas	Blocos	Formato de Saída
Espectrograma	STFT	Entrada	49 x 49 x 1
Res15	3 x 3 conv, 45 + ReLu	*Front-end*	49 x 49 x 45
	3 x 3 conv, 45 ReLu + BN 3 x 3 conv, 45 ReLu + BN (x 6)		49 x 49 x 45 49 x 49 x 45 49 x 49 x 45 49 x 49 x 45 49 x 49 x 45
	3 x 3 conv, 45 ReLu + BN		49 x 49 x 45 49 x 49 x 45
Subamostragem	Avg. Pooling	Transição	1 x 45
Classificador	MLP FC, 12 + Softmax	*Back-end*	1 x 12

fala do Google (*Google speech command database* - GSCD-v2) [13]. A base acústica GSCD-v2 consiste de 105.829 segmentos de áudio (de um segundo de duração cada) contendo 35 palavras distintas. Além disso, a base GSCD-v2 também fornece arquivos de áudio contendo ruídos ambientes do mundo real (não artificial). A partir dos diferentes tipos de ruído não artificiais disponíveis na GSCD-v2, novos arquivos de comandos de fala são gerados artificialmente e apresentam SNRs de 5, 10 e 20 dB. Visando a avaliação dos sistemas de KWS em ambientes acústicos com baixa SNR, esses arquivos de áudio ruidosos são adicionados ao conjunto de teste, seguindo os procedimentos discutidos em [4].

III. ESTRATÉGIAS DE COMBINAÇÃO DE ATRIBUTOS

Agora, visando investigar diferentes estratégias de combinação dos atributos provenientes do espectro de magnitude e/ou do espectro de fase da STFT, um bloco combinador é adicionado à arquitetura padrão dos sistemas de KWS discutida na Seção II. A Fig. 2 apresenta as estratégias de implementação dos sistemas de KWS consideradas neste artigo. Para fins de comparação, a estratégia usando apenas espectrogramas de magnitude [tipicamente adotada na literatura, veja Fig. 2(a)] é utilizada como referência para a avaliação de desempenho das demais estratégias. A seguir, as estratégias para implementar os sistemas de KWS são apresentadas.

A. Estratégia Usando Espectrogramas Individuais

As estratégias ilustradas pelas Figs. 2(a) e 2(b) correspondem ao uso exclusivo de espectrogramas de magnitude e de fase, respectivamente. Essas estratégias têm sido investigadas também em [3], [10] e [11].

B. Estratégia Usando Combinação Linear de Espectrogramas

Conforme mostrado na Fig. 2(c), esta estratégia combina diretamente os espectrogramas de magnitude e de fase provenientes da STFT do sinal de fala. A estratégia em tela corresponde a uma combinação linear realizada através de uma rede de MLP usando um neurônio com função de ativação linear. Tal combinação é realizada na entrada do bloco de *front-end*. Assim, o espectro linearmente combinado é fornecido ao bloco de *front-end* para ser processado por uma Res15 a fim de extrair atributos discriminativos para a etapa seguinte de classificação.

C. Estratégia Usando Combinação via Concatenação

Conforme ilustrado pela Fig. 2(d), esta estratégia implementa a combinação de atributos (de magnitude e de fase) após o processo de extração efetuado pelo bloco de *front-end*, isto é, a combinação é realizada no domínio transformado (dos correspondentes espectrogramas) obtido através de suas ResNets associadas. Nessa etapa, os atributos transformados são combinados via concatenação. Na sequência, esses atributos são fornecidos ao bloco de *back-end* para serem classificados pela rede de MLP descrita na Seção II-B.2.

D. Estratégia Usando Combinação via Comitê de Classificadores

Conforme ilustrado na Fig. 2(e), esta estratégia de combinação opera sobre a saída individual (saída 1 e saída 2) dos blocos de classificação (*back-end*) apresentados na Seção II-B.2. Nesse contexto, a classificação final é obtida usando um comitê de classificadores (*stacked generalization ensemble*) [14] através de uma rede de MLP FC de 24 neurônios com função de ativação ReLu, seguida por uma rede de MLP FC de 12 neurônios com função de ativação *Softmax*.

Os blocos combinadores de atributos apresentados nesta seção são ilustrados nas Fig. 2(c), 2(d) e 2(e) (destacados nas cores amarela, roxa e verde) e mostrados na Tabela II.

TABELA II
ARQUITETURA ESPECÍFICA DOS BLOCOS COMBINADORES PROPOSTOS

Combinadores	Camadas	Formato de Entrada	Formato de Saída
Combinação Linear	MLP FC, 1 + Ativ. Linear	49 x 49 x 2	49 x 49 x 1
Concatenação	Concatenador	1 x 45 x 2	1 x 90
Comitê de Classificadores	MLP FC, 24 + Ativ. ReLu	1 x 12 x 2	1 x 24
	MLP FC, 12 + Ativ. Softmax	1 x 24	1 x 12

IV. SIMULAÇÕES NUMÉRICAS

O método de geração dos conjuntos de treinamento, validação e teste segue os mesmos procedimentos discutidos em [4], no qual listas de arquivos de áudio disponíveis na GSCD-v2 são utilizadas para a separação dos conjuntos. A partir desses conjuntos, as arquiteturas de KWS são treinadas, validadas e avaliadas através da tarefa de classificação de 12 comandos de fala, sendo dez comandos principais, além de segmentos contendo apenas ruído ambiente (para serem classificados como silêncio) e segmentos de fala contendo outras palavras, que devem ser classificados como palavras desconhecidas (palavras fora do vocabulário da aplicação de KWS). Logo, os modelos são treinados visando a maximização da acurácia de reconhecimento de 12 comandos de fala individuais. Aqui, os modelos são treinados por 35 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 [4], 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[3] de 5, 10, 20 e $+\infty$ dB.

[3]Tipicamente, em sistemas de ASR, ambientes acústicos com SNR $\leq$ 10dB são considerados ambientes de baixo nível de SNR [7].

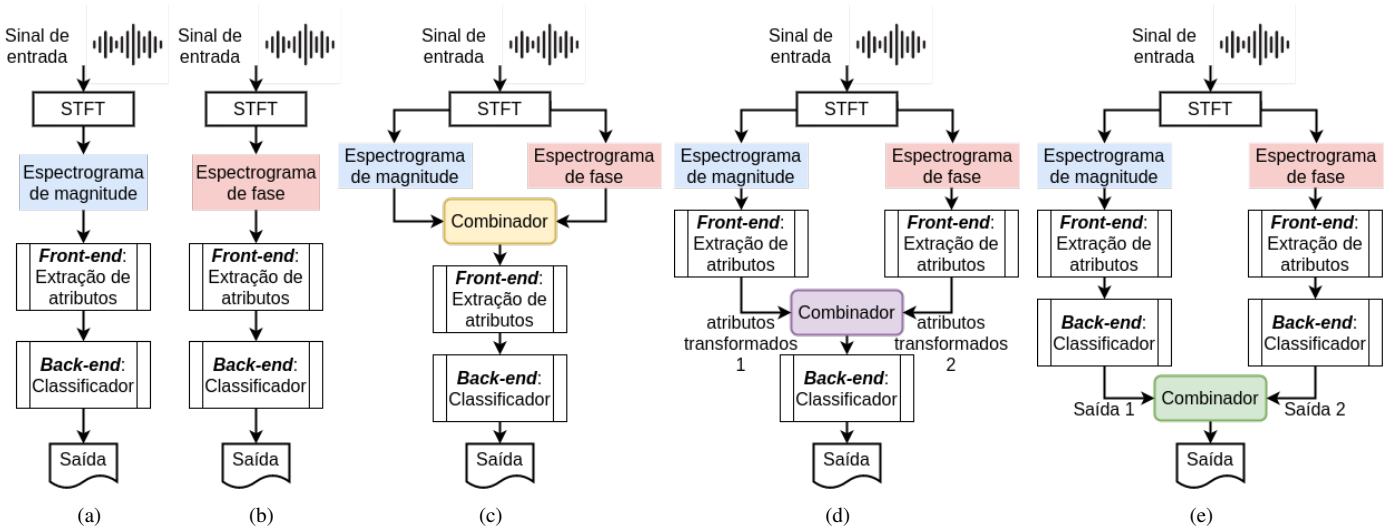

Fig. 2. Estratégias de extração de atributos. (a) Espectrograma de magnitude [3]. (b) Espectrograma de fase [10], [11]. (c) Combinação linear (proposta). (d) Combinação usando concatenação (proposta). (e) Combinação usando comitê de classificadores (proposta).

Para fins de comparação dos sistemas de KWS, os hiperparâmetros correspondentes aos modelos utilizados neste trabalho são mantidos inalterados para todos os experimentos e seguem as configurações estabelecidas em [3]. Dessa forma, os sistemas de KWS discutidos nas seções anteriores são treinados com minilotes (*minibatches*) de 64 segmentos de sinais de fala, utilizando um otimizador baseado no algoritmo do gradiente estocástico descendente (*stochastic gradient descent*) com momento de 0,9 e taxa de aprendizagem de $0,1$. Além disso, os parâmetros treináveis das redes usam inicialização normal de *He* e regularização l_2 de 10^{-5} (para mais detalhes, veja [3]).

V. Resultados e Análise de Desempenho

A Tabela III apresenta a acurácia média de reconhecimento com intervalo de confiança (IC) de 95% computado a partir de simulações de Monte Carlo (MC) usando três ensaios experimentais independentes. Especificamente, os resultados apresentados na Tabela III são obtidos a partir da avaliação do conjunto de teste correspondente à parte da GSCD-v2 isenta de ruído (SNR$= +\infty$ dB). Dessa forma, os resultados dos sistemas de KWS (baseados em Res8, Res15 e Res26) publicados em [3] também são apresentados nesta tabela para fins de comparação. Adicionalmente, com o objetivo de indicar a complexidade computacional de cada sistema de KWS, a Tabela III também mostra o número de parâmetros treináveis desses modelos.

Para a avaliação de desempenho dos sistemas de KWS em cenários acústicos com baixa SNR, a parte do conjunto de teste corrompida por ruídos não artificiais (apresentada na Seção II-C) é agora considerada. Assim, a Fig. 3 apresenta, através de diagramas de caixa, as variações da acurácia de reconhecimento (obtidas por simulações de MC) para diferentes níveis de SNR.

Particularmente, da Tabela III, o sistema de KWS intitulado "Res15 usando Magnitude" corresponde a uma reimplementação similar à do sistema Res15 dado em [3], diferenciado apenas pelo sinal de entrada utilizado. Em [3], o sinal de entrada considerado é baseado em coeficientes cepstrais em escala Mel (*Mel-frequency cepstral coefficients* - MFCC). Enquanto os sistemas de KWS propostos neste trabalho levam em consideração log-Mel espectrogramas [6] (de magnitude e/ou de fase). Dessa forma, os resultados de [3] indicam que a arquitetura Res15 obteve desempenho satisfatório em cenários isentos de ruídos, corroborando os resultados obtidos pelos sistemas desenvolvidos neste trabalho.

Ainda da Tabela III, nota-se que todas as estratégias de combinação apresentam resultados equivalentes ao resultado do sistema de KWS usando apenas magnitude. Nesse sentido, é dado um destaque maior para o sistema de KWS utilizando a estratégia de combinação linear (veja Seção III-B), por esta apresentar maior acurácia média ($97,2\%$) e por manter a mesma quantidade de parâmetros treináveis (complexidade computacional) da arquitetura Res15 [3]. Nesse contexto, também é importante ressaltar que o sistema de KWS usando a estratégia de combinação via comitê de classificadores exibe um resultado ligeiramente superior ao resultado obtido pelo sistema utilizando apenas magnitude, porém apresenta complexidade computacional maior.

De forma similar ao sistema que emprega a estratégia de combinação via comitê de classificadores, a arquitetura Res26 dada em [3] (contendo uma ResNet com 26 camadas convolucionais) também apresenta complexidade computacional maior quando comparada com o sistema Res15; no entanto, esse aumento de complexidade não chega a ser compensado por um aumento na acurácia de reconhecimento. Tal observação pode indicar que o aumento do número de camadas convolucionais não tem sido diretamente proporcional ao ganho de desempenho em acurácia obtido pela Res26. Contudo, para o sistema de KWS, utilizando a estratégia de combinação via comitê de classificadores proposta aqui, o aumento de complexidade do sistema (dado pelo aumento do número de camadas convolucionais) tem proporcionado ganhos relevantes de acurácia. Assim, pode-se inferir que (considerando as arquiteturas baseadas em ResNet e o tamanho fixo da GSCD-v2) a adição de camadas convolucionais resulta em ganhos

de desempenho, sobretudo, quando essas camadas processam atributos provenientes de entradas distintas (sinais de magnitude e de fase).

A partir da Fig. 3, avaliando o desempenho dos sistemas de KWS em cenários acústicos com baixa SNR, verifica-se que o uso apenas do espectrograma de fase proporciona ganhos significativos na acurácia de reconhecimento do sistema KWS considerado. Além disso, a complexidade computacional desse sistema (veja Tabela III) é a mesma daqueles que utilizam apenas o sinal de magnitude. Em contrapartida, considerando um cenário acústico isento de ruído, o desempenho dos sistemas de KWS que usam apenas o espectrograma de fase é inferior aos dos demais sistemas.

Finalmente, a partir da Tabela III e da Fig. 3, nota-se que os sistemas de KWS que usam as estratégias de combinação por concatenação e por comitês de classificadores apresentam desempenhos satisfatórios tanto em cenários acústicos isentos de ruídos quanto em cenários acústicos ruidosos, mesmo considerando os tipos/níveis de ruído não presentes no conjunto de treinamento. Com vistas à complexidade computacional, esses sistemas apresentam uma quantidade maior de parâmetros do que os sistemas que usam apenas a magnitude ou a fase; no entanto, assim como nos sistemas que empregam apenas a fase, pode-se inferir (através do ganho de acurácia de reconhecimento) que suas camadas convolucionais são capazes de extrair atributos mais significativos para a etapa final de classificação.

TABELA III

ACURÁCIA DE RECONHECIMENTO COM IC DE 95% E COMPLEXIDADE COMPUTACIONAL EM TERMOS DO NÚMERO DE PARÂMETROS TREINÁVEIS.

Arquiteturas	Acurácia (%)	Parâm. Treináveis
Res8 (Implementado em [3])	$94,1 \pm 0,351$	110k
Res15 (Implementado em [3])	$\mathbf{95,8 \pm 0,484}$	**238k**
Res26 (Implementado em [3])	$95,2 \pm 0,184$	438k
Res15 usando Magnitude	$\mathbf{97,1 \pm 0,245}$	**238k**
Res15 usando Fase	$96,7 \pm 0,258$	238k
Res15 usando Comb. Linear	$\mathbf{97,2 \pm 0,129}$	**238k**
Res15 usando Concatenação	$97,0 \pm 0.183$	478k
Res15 usando Comitê de Class.	$97,1 \pm 0.122$	481k

VI. CONCLUSÕES E CONSIDERAÇÕES FINAIS

Neste trabalho de pesquisa, o processo de extração de atributos discriminativos para sistemas de KWS foi investigado. Esse processo levou em conta estratégias de combinação de espectrogramas de magnitude e de fase dos sinais de fala, sendo avaliado de acordo com a acurácia de reconhecimento para operação em ambientes com baixa SNR. Nesse contexto, sistemas de KWS treinados usando estratégias de combinação via concatenação ou comitê de classificadores proporcionaram desempenhos melhores quando comparados aos sistemas que fazem uso apenas de espectrogramas de magnitude. Os resultados de acurácia obtidos corroboraram a eficácia do uso das estratégias de combinação de atributos (de magnitude e de fase) investigadas neste artigo. Para trabalhos futuros, visando aprimorar os resultados das estratégias aqui investigadas, algumas modificações podem ser realizadas nas arquiteturas das redes. Por exemplo, o bloco combinador linear poderia ser implementado com um número maior de neurônios na rede MLP (responsável pela combinação dos espectrogramas de magnitude e de fase). Além disso, outra possibilidade seria

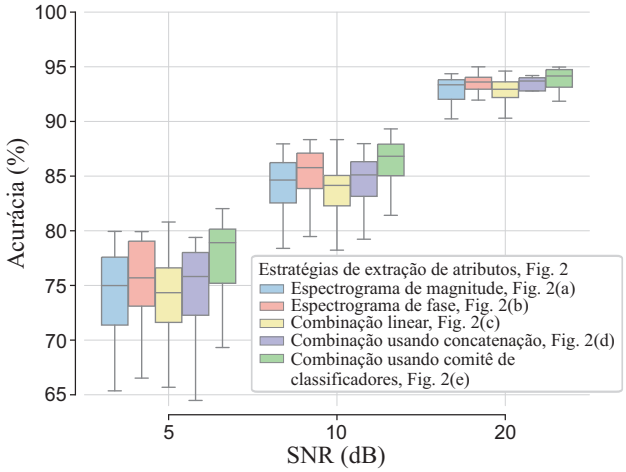

Fig. 3. Diagrama de caixa de acurácia dos sistemas de KWS operando em ambientes ruidosos.

a inclusão de novas etapas de extração de atributos após os blocos combinadores, seja operando atributos transformados concatenados (considerados na combinação via concatenação) ou vetores de classificação (utilizados na combinação via comitê de classificadores).

REFERÊNCIAS

[1] D. Top, "State of automatic speech recognition," *Opus Research*, Mar. 2021.

[2] A. Fong and M. Usman, "Machine learning for end consumers," *IEEE Trans. Consum. Electron.*, vol. 9, no. 5, pp. 77–78, Aug. 2020.

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

[4] M. Zeng and N. Xiao, "Effective combination of DenseNet and BiLSTM for keyword spotting," *IEEE Access*, vol. 7, no. 1, pp. 10 767–10 775, Jan. 2019.

[5] 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.

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

[7] A. Pervaiz, F. Hussain, H. Israr, M. A. Tahir, F. R. Raja, N. K. Baloch, F. Ishmanov, and Y. B. Zikria, "Incorporating noise robustness in speech command recognition by noise augmentation of training data," *Sensors*, vol. 2326, no. 8, pp. 1–19, Apr. 2020.

[8] 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. Audio, Speech and Language Process.*, vol. 28, no. 1, pp. 2756–2769, Oct. 2020.

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

[10] A. Dutta, G. Ashishkumar, and C. R. Rao, "Phase Based Spectro-Temporal Features for Building a Robust ASR System," in *Proc. Int. Speech Communication Association (INTERSPEECH)*, Shanghai, China, Oct. 2020, pp. 1668–1672.

[11] 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.

[12] D. Zhu and K. K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Montreal, Canada, 2004, pp. 125–128.

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

[14] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. USA: Wiley-Interscience, 2004.