# Considerações Sobre o Uso do Sinal de Fase em Sistemas de Reconhecimento Automático de Fala

Ênio dos Santos Silva e Rui Seara

***Resumo*— Este trabalho apresenta uma investigação sobre o uso do espectro de fase, oriundo da transformada de Fourier (*Fourier transform phase spectrum* - FTPS), em sistemas de reconhecimento automático de fala (*automatic speech recognition* - ASR). Historicamente, em sistemas de ASR, a utilização do sinal de fase tem sido usualmente negligenciada. No entanto, pesquisas recentes têm mostrado a importância do FTPS em diversas aplicações de processamento de fala. Especificamente, visando o aprimoramento de sistemas de ASR, a função atraso de grupo é considerada na etapa de extração de atributos (*front-end*), bem como na etapa de construção do modelo acústico. Adicionalmente, o desempenho de sistemas de ASR, usando *front-ends* baseados na função atraso de grupo, é avaliado para ambientes acústicos com baixa razão sinal-ruído. Resultados de simulação obtidos aqui permitem inferir acerca do impacto da informação da fase do sinal de fala (melhoria média de 3,32%) no desempenho de sistemas de ASR.**

***Palavras-Chave*— Atraso de grupo, extração de atributos, informação da fase, reconhecimento automático de fala.**

***Abstract*— This paper presents an investigation on the use of the Fourier transform phase spectrum (FTPS) in automatic speech recognition (ASR) systems. Historically, the use of the phase information in ASR systems has commonly been neglected. However, recent research works have shown the importance of the FTPS in some applications of speech processing. Specifically, in order to enhance ASR systems, the group delay function is taken into account in the front-end stage as well as for achieving better performance of acoustic models. In addition, the performance of ASR systems using front-ends based on the group delay function is assessed in acoustic environments with low signal-to-noise ratio. Simulation results allow inferring about the impact of the phase of the speech signal (average improvement of 3,32%) on ASR systems for performance.**

***Keywords*— Group delay, feature extraction, phase information, automatic speech recognition.**

## I. Introdução

Atualmente, sistemas de reconhecimento automático de fala (*automatic speech recognition* - ASR) operam com atributos de entrada provenientes da transformada de Fourier de curto termo (*short-time Fourier transform* - STFT) [1] ou com atributos provenientes diretamente do sinal de fala bruto no domínio do tempo [2]. Independente do modo de operação, a extração de atributos é uma etapa fundamental em qualquer sistema de ASR [2], [3]. Nesse contexto, ambos os modos de operação buscam identificar e reter apenas atributos que mais contribuem para a geração de um modelo acústico (MA) que seja capaz de discriminar satisfatoriamente um determinado conjunto de fonemas [4], [5].

Especificamente, os sistemas de ASR que operam com atributos provenientes da STFT geralmente utilizam apenas informações do espectro de magnitude [1], [6]. Historicamente, em sistemas de processamento da fala, a utilização do espectro de fase oriundo da STFT vem sendo comumente negligenciada [5]-[7]. No entanto, pesquisas recentes têm mostrado a importância da utilização da fase em diversas aplicações de processamento de fala [5]-[10]. Em [9], é investigado o processo de aprendizagem automática de atributos discriminativos em conjunto com a construção de um MA. Particularmente, em [9], a informação de fase é considerada através da utilização da STFT na sua forma de representação retangular. Em [10], cepstros complexos são usados como atributos para a aplicação em conversores de texto para fala. Já em [5] e [8], atributos criados a partir das derivadas do espectro de fase no domínio do tempo e da frequência, representados pela frequência instantânea e pelo atraso de grupo (*group delay* - GD), respectivamente, são utilizados em aplicações de realce do sinal de fala e em sistemas de ASR.

Na literatura da área, trabalhos que adotam o GD como atributos discriminativos relatam resultados satisfatórios em sistemas de ASR [6], [8], [11]-[13]. Em [13], a robustez ao ruído dos atributos de GD é discutida. Em [11], o comprimento da janela de análise da STFT para a extração de atributos de magnitude e de fase é investigado. Já em [12], uma estratégia de combinação de atributos de magnitude e de fase é proposta. Além disso, em [12], é também proposta a utilização de bancos de filtros em escala Mel (*Mel-frequency* - MF) aplicados à combinação dos atributos supracitados.

Neste artigo, visando a identificação de atributos discriminativos do sinal de fala para a obtenção de sistemas de ASR, a investigação de estratégias de extração de atributos (*front-ends*) utilizando a fase da STFT é aqui avaliada. Além do mais, a fim de estimular a comunidade brasileira de processamento da fala para pesquisas destinadas ao idioma português brasileiro, este trabalho adota a língua portuguesa (PT-BR) como *corpus* acústico. Particularmente, devido à alta complexidade computacional necessária para a construção de redes neurais profundas [14] e a escassez de um *corpus* acústico adequado para o PT-BR[1] [15], este trabalho de pesquisa (assim como discutido também em [16] e [17]) utiliza modelos ocultos de Markov (*hidden Markov models* - HMMs) para a construção dos MAs.

Especificamente, neste trabalho, são investigados os *front-*

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).

[1]Pesquisas do estado da arte, para o idioma inglês, trabalham com *corpora* acústicos de 2000 a 20000 horas.

*ends* baseados em atributos de GD, como também atributos provenientes de combinações com a magnitude da STFT e atributos aplicados a bancos de filtros em MF. Adicionalmente, os *front-ends* aqui investigados são avaliados em ambientes acústicos com baixa razão sinal-ruído (*signal-to-noise ratio* - SNR), como também utilizando diferentes comprimentos de janela de análise da STFT. Os resultados de simulação obtidos permitem inferir acerca da influência da fase do sinal de fala no desempenho de sistemas de ASR para o PT-BR, destacando a eficácia dos *front-ends* avaliados.

## II. Função Atraso de Grupo

Denotando $x(n)$ como um quadro do sinal de fala no domínio do tempo, segmentado por janelas de diferentes comprimentos, sua correspondente STFT $X(\mathrm{e}^{j\omega})$ é dada por

$$X(\mathrm{e}^{j\omega}) = |X(\mathrm{e}^{j\omega})|\mathrm{e}^{j\theta(\mathrm{e}^{j\omega})} \tag{1}$$

onde $|X(\mathrm{e}^{j\omega})|$ representa o espectro de magnitude da STFT (*Fourier transform magnitude spectrum* - FTMS) e $\theta(\mathrm{e}^{j\omega})$ representa o espectro de fase da STFT (*Fourier trasnform phase spectrum* - FTPS). Embora ambos os espectros contenham informações importantes do sinal de fala, no FTPS, a identificação dessas informações é prejudicada devido ao problema do "empacotamento" da fase (*phase wrapping*) (módulo $2\pi$) [6], [18], [19]. Particularmente, os valores do $\theta(\mathrm{e}^{j\omega})$ oriundos de (1) estão confinados entre $-\pi$ e $\pi$. Dessa forma, $\theta(\mathrm{e}^{j\omega})$ não representa a "verdadeira" fase do sinal. Assim, ao contrário do FTMS, o FTPS não apresenta diretamente uma estrutura adequada para tarefas de classificação [6].

Uma das possíveis soluções para contornar o problema do empacotamento da fase no FTPS é a utilização da função atraso de grupo (*group delay function* - GDF) $\tau(\mathrm{e}^{j\omega})$, a qual é definida como a derivada da função de fase[2] $\theta(\mathrm{e}^{j\omega})$ [18]. Assim,

$$\tau(\mathrm{e}^{j\omega}) = -\frac{d\theta(\mathrm{e}^{j\omega})}{d\omega}. \tag{2}$$

Nesse caso, para evitar o processo de desempacotamento da fase, GDF $\tau(\mathrm{e}^{j\omega})$ pode ser computada de maneira análoga ao cálculo do cepstro complexo, usando a transformada de Fourier das sequências reais $x(n)$ e $nx(n)$ [19]. A seguir, uma breve descrição desse procedimento é apresentada (para mais detalhes, veja [18] e [19]).

### A. Cálculo da GDF $\tau(e^{j\omega})$

De acordo com [19], considerando $\hat{X}'(\mathrm{e}^{j\omega})$ como a derivada do logaritmo natural de $X(\mathrm{e}^{j\omega})$, tem-se

$$\hat{X}'(\mathrm{e}^{j\omega}) = \frac{d}{d\omega}\ln\{X(\mathrm{e}^{j\omega})\} = \frac{X'(\mathrm{e}^{j\omega})}{X(\mathrm{e}^{j\omega})}. \tag{3}$$

Alternativamente, a partir de (1), obtém-se

$$\hat{X}'(\mathrm{e}^{j\omega}) = \frac{d}{d\omega}\ln|X(\mathrm{e}^{j\omega})| + j\frac{d\theta(\mathrm{e}^{j\omega})}{d\omega} = \frac{X'(\mathrm{e}^{j\omega})}{X(\mathrm{e}^{j\omega})}. \tag{4}$$

Agora, usando a propriedade da diferenciação da transformada de Fourier, na qual $nx(n) \longleftrightarrow j\frac{dX(\mathrm{e}^{j\omega})}{d\omega}$, e assumindo $y(n) = nx(n)$, tem-se $Y(\mathrm{e}^{j\omega}) = jX'(\mathrm{e}^{j\omega})$. Dessa forma,

$$X'(\mathrm{e}^{j\omega}) = -jY(\mathrm{e}^{j\omega}) = -jY_{\mathrm{R}}(\mathrm{e}^{j\omega}) + Y_{\mathrm{I}}(\mathrm{e}^{j\omega}) \tag{5}$$

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

onde $Y_{\mathrm{R}}(\mathrm{e}^{j\omega})$ e $Y_{\mathrm{I}}(\mathrm{e}^{j\omega})$ denotam, respectivamente, a parte real e a parte imaginária da transformada de Fourier de $y(n)$. Então, substituindo (5) em (4), tem-se

$$\hat{X}'(\mathrm{e}^{j\omega}) = \frac{\left[Y_{\mathrm{I}}(\mathrm{e}^{j\omega}) - jY_{\mathrm{R}}(\mathrm{e}^{j\omega})\right]\left[X_{\mathrm{R}}(\mathrm{e}^{j\omega}) - jX_{\mathrm{I}}(\mathrm{e}^{j\omega})\right]}{|X(\mathrm{e}^{j\omega})|^2} \tag{6}$$

com $X_{\mathrm{R}}(\mathrm{e}^{j\omega})$ e $X_{\mathrm{I}}(\mathrm{e}^{j\omega})$ representando a parte real e a parte imaginária de $X(\mathrm{e}^{j\omega})$, respectivamente.

Agora, efetuando algumas manipulações algébricas em (6), obtém-se

$$\begin{aligned}\hat{X}'(\mathrm{e}^{j\omega}) \;=\; & \frac{Y_{\mathrm{I}}(\mathrm{e}^{j\omega})X_{\mathrm{R}}(\mathrm{e}^{j\omega}) - Y_{\mathrm{R}}(\mathrm{e}^{j\omega})X_{\mathrm{I}}(\mathrm{e}^{j\omega})}{|X(\mathrm{e}^{j\omega})|^2} \\ & -j\frac{\left[Y_{\mathrm{I}}(\mathrm{e}^{j\omega})X_{\mathrm{I}}(\mathrm{e}^{j\omega}) + Y_{\mathrm{R}}(\mathrm{e}^{j\omega})X_{\mathrm{R}}(\mathrm{e}^{j\omega})\right]}{|X(\mathrm{e}^{j\omega})|^2}\end{aligned} \tag{7}$$

onde, de acordo com (2) e (4), a parte imaginária de (7) denota a GDF $\tau(\mathrm{e}^{j\omega})$. Finalmente, substituindo (7) em (2), determina-se uma expressão para obter a GDF $\tau(\mathrm{e}^{j\omega})$. Assim,

$$\tau(\mathrm{e}^{j\omega}) = \frac{Y_{\mathrm{I}}(\mathrm{e}^{j\omega})X_{\mathrm{I}}(\mathrm{e}^{j\omega}) + Y_{\mathrm{R}}(\mathrm{e}^{j\omega})X_{\mathrm{R}}(\mathrm{e}^{j\omega})}{|X(\mathrm{e}^{j\omega})|^2}. \tag{8}$$

Dessa forma, a GDF $\tau(\mathrm{e}^{j\omega})$ pode ser obtida diretamente considerando as STFTs de $x(n)$ e $y(n)$.

A Fig. 1 ilustra exemplos de atributos de um sinal de fala. Especificamente, a Fig. 1(a) mostra o sinal de fala $x(n)$ e as Figs. 1(b), (c) e (d) ilustram, respectivamente, os correspondentes FTMS, a GDF $\tau(\mathrm{e}^{j\omega})$ e o envelope da FTMS.

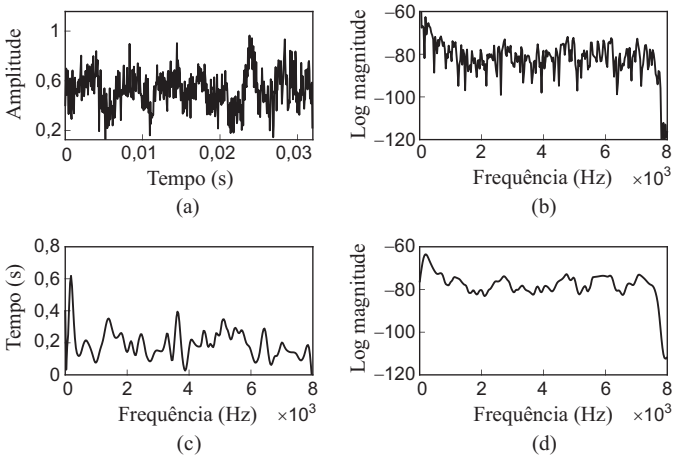



Fig. 1. Exemplos de atributos de um sinal da fala. (a) Sinal de fala $x(n)$. (b) Espectro de magnitude $|X(\mathrm{e}^{j\omega})|$. (c) Função atraso de grupo $\tau(\mathrm{e}^{j\omega})$. (d) Envelope espectral de $|X(\mathrm{e}^{j\omega})|$.

### B. Considerações Sobre a GDF

Uma dificuldade inerente às GDFs vem da localização dos zeros de $X(z)$ à medida que esses se aproximam da circunferência de raio unitário no plano $z$. Note em (8) que o denominador de $\tau(\mathrm{e}^{j\omega})$ contém o termo $|X(\mathrm{e}^{j\omega})|^2$, dessa forma, os zeros de $X(z)$, próximos da circunferência de raio unitário, podem causar instabilidade na GDF. Para superar tal dificuldade, em [12], tem sido sugerido usar a função produto espectral (*product spectrum function* - PSF) $Q(\mathrm{e}^{j\omega})$ no lugar

de $\tau(\mathrm{e}^{j\omega})$, a qual é denotada como segue:

$$\begin{aligned} Q(\mathrm{e}^{j\omega}) &= \tau(\mathrm{e}^{j\omega})|X(\mathrm{e}^{j\omega})|^2 \\ &= X_{\mathrm{R}}(\mathrm{e}^{j\omega})Y_{\mathrm{R}}(\mathrm{e}^{j\omega}) + X_{\mathrm{I}}(\mathrm{e}^{j\omega})Y_{\mathrm{I}}(\mathrm{e}^{j\omega}). \end{aligned} \quad (9)$$

Além disso, pode-se ainda utilizar as estratégias discutidas em [8], [11] e [12] para o tratamento das raízes próximas à circunferência de raio unitário. Dentre elas, destacam-se as que utilizam bancos de filtros em MF. Estas últimas são baseadas nos coeficientes cepstrais em escala Mel (*Mel frequency cepstral coefficients* - MFCC [1]), cujo processo de extração de atributos é sumarizado a seguir:

1) Cálculo do FTMS $|X(\mathrm{e}^{j\omega})|$ de $x(n)$.
2) Cálculo da potência espectral $|X(\mathrm{e}^{j\omega})|^2$.
3) Aplicação de bancos de filtro MF em $|X(\mathrm{e}^{j\omega})|^2$.
4) Obtenção das energias de cada banco de filtro (*filter-bank energies* - FBEs).
5) Cálculo da transformada discreta do cosseno do logaritmo natural das FBEs.
6) Seleção dos atributos MFCCs[3].

Além disso, para a obtenção de GDFs confiáveis, em [7] e [11], é destacada a importância do comprimento da janela de análise da STFT. Nesse contexto, com uma escolha adequada da função de janelamento, a GDF, como ilustrada na Fig. 1(c), pode realçar apropriadamente as frequências ressonantes e antirressonantes do correspondente sinal da fala.

## III. Sistemas de ASR

Neste artigo, os atributos discutidos na seção anterior são utilizados como *front-ends* dos sistemas de ASR aqui implementados. Além da etapa de *front-end*, um sistema típico de ASR é composto por mais quatro blocos principais, a saber: dicionário fonético (DF), MA, modelo de linguagem (ML) e decodificador, conforme ilustrado na Fig. 2 [17].

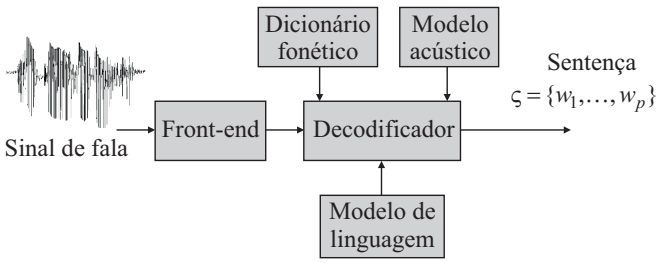



Fig. 2. Diagrama de blocos de um sistema de ASR típico.

Em resumo, o ML de um sistema de ASR fornece a probabilidade $p(\varsigma)$ de ocorrer uma sentença $\varsigma = \{w_1, \ldots, w_p\}$ de $p$ palavras. O MA determina a verossimilhança acústica entre os atributos do sinal de fala e as palavras da sentença $\varsigma$. Geralmente, os sistemas de ASR dividem sentenças em palavras e palavras em unidades básicas, tais como fones ou trifones [16], [17]. Particularmente, o DF efetua o mapeamento das palavras para essas unidades básicas e vice-versa. Na etapa do decodificador, com base no ML, no MA e após o *front-end* converter o sinal da fala em atributos discriminativos, a sentença $\varsigma$ mais provável ao contexto linguístico e ao sinal da fala é estimada.

[3]Usualmente, são selecionados os 13 primeiros atributos.

### A. Métrica de Avaliação

Na maioria das aplicações de ASR, a figura de mérito utilizada para avaliar tais sistemas é a taxa de erro de palavra (*word error rate* - WER), definida como

$$\mathrm{WER} = \frac{D+R}{W} \times 100\% \quad (10)$$

onde $W$ é o número de palavras na sequência de entrada, e $R$ e $D$ são, respectivamente, o número de erros de substituição e o de deleção na sequência de palavras reconhecidas, quando comparados com a sequência correta.

## IV. *Corpus* Acústico

Para que seja possível a obtenção de WERs satisfatórias, os sistemas de ASR necessitam de um *corpus* com grande variabilidade acústica. Dessa forma, a disponibilidade de um *corpus* adequado é um fator primordial para o desenvolvimento de MAs precisos. Um *corpus* típico possui arquivos de fala com as suas transcrições associadas, grande quantidade de textos, como também um DF. Para atender tais requisitos, no desenvolvimento do sistema de ASR, nós utilizamos o DF e os *corpora* de fala e de texto disponibilizados em [20], [21] e [22]. A Tabela I apresenta as características do *corpus* acústico utilizado aqui para dados amostrados em 16 kHz, considerando tipo de fala de leitura.

TABELA I
Características do *corpus* acústico

| Etapa | Duração | Vocabulário | Sentenças | Locutores |
|---|---|---|---|---|
| **Treinamento** | 4,70 h | 2933 | 3959 | 100 |
| **Teste** | 0,90 h | 2731 | 700 | 35 |

### A. Adição de Ruído

Na prática, os sistemas de ASR sofrem degradações importantes em seu desempenho quando operam em ambientes com baixas SNRs. Neste trabalho, a fim de investigar o desempenho dos sistemas de ASR em ambientes ruidosos, o *corpus* acústico original [21] é corrompido com um ruído do tipo *babble* (ruído caracterizado por sons de sussurro) com os seguintes níveis de SNRs: 0, 5, 10, 15 e 20 dB. Nesse contexto, dado o sinal de fala original $x(n)$ e o sinal de ruído $r(n)$, o sinal $x_r(n)$ corrompido é obtido como

$$x_r(n) = \alpha x(n) + \beta r(n) \quad (11)$$

onde $\alpha$ e $\beta$ são determinados de acordo com o nível desejado de SNR.

## V. Estratégias de ASR usando *front-ends* GDFs

Neste trabalho de pesquisa, com base nos procedimentos descritos em [16] e [17], a ferramenta HTK tem sido adotada para o treinamento dos MAs. Particularmente, os MAs são obtidos aqui usando HMMs contínuos com transições de estado, assumindo a topologia esquerda-direita. Dessa forma, a função densidade de probabilidade (da observação dos fonemas em cada estado de uma HMM) é estimada através de modelos de misturas de gaussianas (*Gaussian mixture model* - GMM). Assim, cada HMM-GMM constitui modelos dependentes do contexto (trifones *cross-word*) computados a partir de 38 monofones. Para a concepção dos MLs, a ferramenta MITLM [17] tem sido utilizada considerando modelos

de trigramas estimados através da técnica de suavização de *Kneser-Ney* [17].

Visando investigar o desempenho dos sistemas de ASR que consideram os atributos baseados na GDF apresentados na Seção II, a etapa de *front-end* mostrada na Fig. 2 é discutida aqui. Particularmente, conforme ilustrado na Fig. 3, os atributos $|X(\mathrm{e}^{jw})|$, $\tau(\mathrm{e}^{jw})$ e $Q(\mathrm{e}^{jw})$ são processados segundo os procedimentos (passos de 2 a 6) referentes à extração de MFCCs. Dessa forma, além dos MFCCs convencionais, tal processo resulta na extração de coeficientes cepstrais da GDF e da PSF em escala Mel (*Mel frequency group delay cepstral coefficient* - MFGDCC e *Mel frequency product spectrum cepstral coefficient* - MFPSCC, respectivamente). Adicionalmente, valores de energia e de variação temporal dos coeficientes são adicionados aos atributos resultantes do processo supracitado.

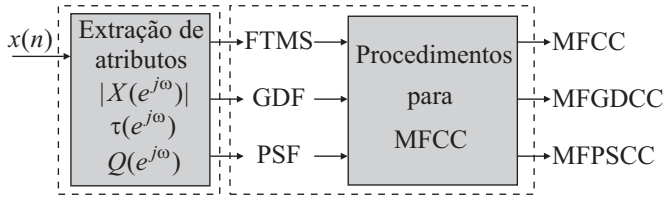



Fig. 3. *Front-end* usando o processo de extração de atributos em MF.

Levando em conta o Teorema de Parserval [19], o coeficiente cepstral $c_0$ é usado para estimar a energia de um quadro do sinal de fala. Já os coeficientes de variação temporal são computados como segue:

$$\mathbf{\Delta}_t = \frac{\sum_{\phi=1}^{\Theta}(\mathbf{c}_{t+\phi} - \mathbf{c}_{t-\phi})\phi}{2\sum_{\phi=1}^{\Theta}\phi^2} \tag{12}$$

onde $\mathbf{\Delta}_t$ é o vetor de coeficiente de variação temporal de um quadro $t$ obtido usando os correspondentes coeficientes estáticos $\mathbf{c}_{t-\Theta}$ a $\mathbf{c}_{t+\Theta}$. Especificamente aqui, é usado o valor $\Theta = 2$. Dessa forma, (12) também é aplicada aos coeficientes $\mathbf{\Delta}_t$ para a obtenção dos coeficientes $\mathbf{\Delta\Delta}_t$.

## VI. Resultados e Análise de Desempenho

Para a análise de desempenho, o *corpus LapsBenchmark* [21] é adotado. Aqui, as contribuições das informações do FTPS, representadas pelos *front-ends* MFPSCC e MFGDCC (discutidos nas Seções II e V), são medidas através da WER, definida em (10).

Especificamente, sistemas de ASR com *front-ends* utilizando atributos MFCC, MFPSCC e MFGDCC são avaliados em ambientes acústicos com SNRs de 0, 5, 10, 15 e 20 dB, bem como em ambientes isentos de ruído ($+\infty$ dB). Ressalta-se que os resultados obtidos com o sistema utilizando atributos MFCC consideram apenas a informação do FTMS. Tais resultados são adotados como referência para a avaliação de ganhos na WER provenientes da exploração do FTPS. Além disso, é investigado o uso de janelas de análise de *Hamming* de comprimentos entre 32 e 256 ms com sobreposição de 10 ms.

A Tabela II apresenta o desempenho dos sistemas de ASR usando os *front-ends* supracitados. Nessa tabela, são mostrados apenas os melhores resultados de cada experimento. Nesse contexto, a fim de encontrar os pontos ótimos de operação dos sistemas de ASR, o número de estados da HMM (NE) e o número de gaussianas da GMM (NG) dos MAs são avaliados conforme os resultados da WER. Na Fig 4, são mostrados os resultados obtidos para uma SNR $= 15$ dB. Em particular, as Figs. 4(a), (b), (c) e (d) apresentam, respectivamente, as WERs (em função do NE e do NG) para os comprimentos da janela de análise (janela de *Hamming*) de 256, 128, 64 e 32 ms.

Da Tabela II e da Fig. 4, nota-se que o desempenho do *front-end* MFGDCC [que utiliza apenas a informação da fase $\tau(\mathrm{e}^{jw})$] não apresenta contribuições relevantes para os resultados das WERs quando comparado ao desempenho do MFCC, especialmente para janelas de análise com comprimentos menores do que 128 ms. Em contrapartida, o desempenho do *front-end* MFPSCC que, como definido em (9), utiliza informações do FTPS $\tau(\mathrm{e}^{jw})$ e do FTMS $|X(\mathrm{e}^{jw})|$ geralmente apresenta desempenho superior ao do MFCC. Particularmente, o *front-end* MFPSCC apresenta ganhos absolutos de 6,05; 3,31; 1,33; 0,16 e 1,19% (em ambientes de 5, 10, 15, 20 e $+\infty$ dB, respectivamente) quando comparado com o *front-end* MFCC.

Em contraste com [12], o presente trabalho investiga sistemas de ASR que utilizam sinais de fala com frequência de amostragem de 16 kHz. Nesse contexto, devido à utilização de uma maior faixa espectral, tais sistemas apresentam em média desempenho 5% superior a sistemas que consideram sinais amostrados em 8 kHz [17].

## VII. Conclusões e Considerações Finais

Neste trabalho de pesquisa, a utilização da informação do FTPS em sistemas de ASR foi investigada. Tais informações foram consideradas nos *front-ends* MFPSCC e MFGDCC, sendo avaliadas segundo as WERs dos correspondentes sistemas de ASR. Através dos resultados obtidos, pôde-se inferir que o *front-end* MFPSCC proporcionou maior riqueza espectral ao treinamento de MAs, resultando em atributos mais significativos ao ASR, especialmente, para ambientes com baixas SNRs. Nesses casos, o MA do sistema de ASR treinado com os atributos MFPSCC apresentou melhor desempenho, quando comparado ao treinamento usando uma versão que leva em consideração apenas as informações de magnitude (MFCC). Os resultados obtidos da WER confirmam a eficácia da utilização do FTPS.

Espera-se que este trabalho estimule novas pesquisas destinadas ao PT-BR, assim como a criação de um *corpus* acústico apropriado para a realização de investigações adicionais sobre o impacto da fase do sinal de fala, utilizando outras estratégias de aprendizado, tais como as baseadas em redes neurais profundas.

## Referências

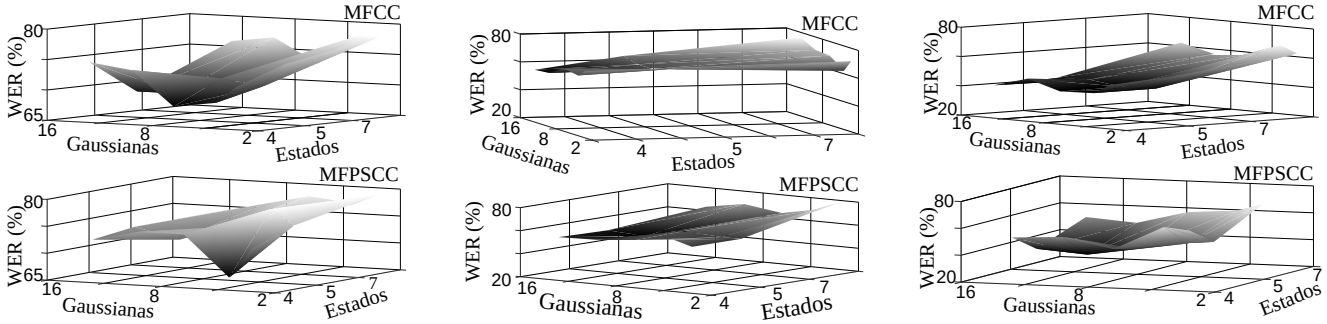
[1] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, no. 1, pp. 19 143–19 165, Feb. 2019.

[2] V. Passricha and R. K. Aggarwal, *From Natural to Artificial Intelligence: Convolutional Neural Networks for Raw Speech Recognition*, R. Lopez-Ruiz, Ed. Rijeka, Croatia: IntechOpen, 2018.

TABELA II

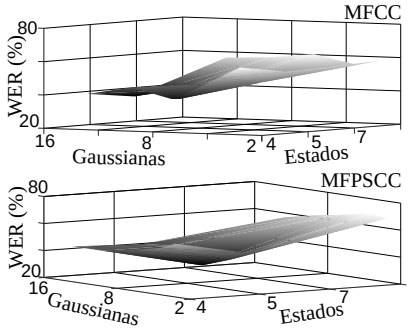
DESEMPENHO DOS SISTEMAS DE ASR AVALIADOS SEGUNDO A WER (%) EM AMBIENTES COM A PRESENÇA DE RUÍDO DO TIPO *BABBLE*

| SNR (dB) | HMM-GMM usando *front-ends* juntamente com $\mathbf{c}_0$, $\mathbf{\Delta}_t$ e $\mathbf{\Delta\Delta}_t$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MFCC | | | | MFPSCC | | | | MFGDCC | | | |
| | 32 ms | 64 ms | 128 ms | 256 ms | 32 ms | 64 ms | 128 ms | 256 ms | 32 ms | 64 ms | 128 ms | 256 ms |
| 0 | 100 | 100 | 88, 91 | 95, 74 | 92, 97 | 100 | 89, 02 | 75, 34 | 100 | 95, 74 | 100 | **72, 26** |
| 5 | 69, 91 | 84, 85 | 69, 06 | 87, 97 | **63, 01** | 84, 25 | 70, 07 | 70, 58 | 100 | 85, 04 | 94, 32 | 66, 08 |
| 10 | 41, 53 | 51, 14 | 54, 71 | 82, 04 | **38, 22** | 47, 32 | 54, 93 | 65, 03 | 75, 74 | 66, 46 | 72, 56 | 69, 17 |
| 15 | 33, 22 | 34, 09 | 49, 33 | 78, 09 | **31, 89** | 39, 24 | 42, 61 | 65, 82 | 52, 53 | 48, 87 | 67, 37 | 54, 56 |
| 20 | 30, 11 | 32, 96 | 44, 85 | 78, 09 | **29, 95** | 33, 47 | 43, 56 | 61, 86 | 35, 43 | 38, 12 | 54, 93 | 54, 52 |
| $+\infty$ | 29, 08 | 31, 83 | 44, 85 | 76, 11 | **27, 89** | 31, 16 | 41, 67 | 64, 24 | 39, 09 | 41, 05 | 52, 86 | 55, 99 |




(a) (b) (c)

Fig. 4. Desempenho dos sistemas de ASR (usando MFCC e MFPSCC) para sinais da fala com SNR = 15 dB, considerando janelas de análise de (a) 256 ms, (b) 128 ms, (c) 64 ms e (d) 32 ms.




(d)

(continuação Fig. 4)


[3] P. Ghahremani, H. Hadian, H. Lv, D. Povey, and S. Khudanpur, "Acoustic modeling from frequency domain representations of speech," in *Proc. Int, Speech Communication Association (INTERSPEECH)*, Hyderabad, India, Sep. 2018, pp. 1596–1600.

[4] C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 4774–4778.

[5] 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.

[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] K. K. Paliwal and L. D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 1, pp. 153–170, Aug. 2005.

[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] E. Variani, T. N. Sainath, I. Shafran, and M. Bacchiani, "Complex Linear Projection (CLP): A discriminative approach to joint feature extraction and acoustic modeling," in *Proc. Int. Speech Communication Association (INTERSPEECH)*, San Francisco, USA, Sep. 2016, pp. 808–812.

[10] R. Maia and R. Seara, "Speech synthesis based on deep neural networks with direct modeling of amplitude spectra," in *Anais do Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT)*, Campina Grande, PB, Set. 2018, pp. 687–691.

[11] K. Yamamoto, E. Sueyoshi, and S. Nakagawa, "Speech recognition using long-term phase information," in *Proc. Int. Speech Communication Association (INTERSPEECH)*, Makuhari, Japan, Sep. 2010, pp. 1189–1192.

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

[13] S. H. K. Parthasarathi, R. Padmanabhan, and H. A. Murthy, "Robustness of group delay representations for noisy speech signals," *Int. J. Speech Technology*, vol. 14, pp. 361–368, Sep. 2011.

[14] T. N. Sainath, A. Narayanan, R. J. Weiss, E. Variani, and I. Shafran, "Reducing the computational complexity of multimicrophone acoustic models with integrated feature extraction," in *Proc. Int, Speech Communication Association (INTERSPEECH)*, San Francisco, USA, Sep. 2016, pp. 1971–1975.

[15] I. M. Quintanilha, L. W. P. Biscainho, and S. L. Netto, "Towards an end-to-end speech recognizer for portuguese using deep neural networks," in *Anais do Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT)*, São Pedro, SP, Set. 2017, pp. 408–412.

[16] E. Techini, Z. Sakka, and M. Bouhlel, "Robust front-end based on MVA and HEQ post-processing for arabic speech recognition using Hidden Markov Model Toolkit (HTK)," in *Proc. Int. Conf. on Comput. Syst. and Applications (AICCSA)*, Hammamet, Tunisia, Oct. 2017, pp. 815–820.

[17] E. S. Silva and R. Seara, "Extensão artificial de largura de banda aplicada em reconhecimento automático de fala," in *Anais do Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT)*, Santarém, PA, Set. 2016, pp. 80–84.

[18] H. A. Murthy and B. Yegnanarayana, "Group delay functions and its applications in speech technology," *Sadhana*, vol. 36, no. 5, pp. 745–782, Sep. 2011.

[19] A. V. Oppenheim and R. W. Schafer, *Discrete Time Signal Processing*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2010.

[20] C. A. Ynoguti and F. Violaro, "A brazilian portuguese speech database," in *Anais do Simpósio Brasileiro de Telecomunicações (SBrT)*, Rio de Janeiro, RJ, Set. 2008, pp. 15–20.

[21] LaPS-UFPa, "Grupo Fala Brasil," https://github.com/falabrasil/corpora, Accessed: May 2019.

[22] "VoxForge," http://www.voxforge.org, Accessed: May 2019.