# Disfarce de Voz em Tempo Real com Múltiplas Modulações SSB de Tempo Discreto

D. W. de França Silva e G. Jerônimo da Silva Jr.

***Resumo*— Este trabalho apresenta um método de disfarce de voz eletrônico, que tem por objetivo impedir o reconhecimento de um locutor, utilizando múltiplas modulações SSB (single-sideband) em tempo discreto. Um estudo qualitativo do desempenho do sistema é apresentado.**

***Palavras-Chave*— Disfarce de voz, modulação SSB, transformada de Hilbert, processamento digital de sinais.**

***Abstract*— This work presents a design method to electronic disguised voice, which aims to prevent the speaker recognition, using multiple discrete time SSB (single-sideband) modulations. A qualitative study of system performance is presented.**

***Keywords*— Disguised voice, SSB modulation, Hilbert transform, digital signal processing.**

## I. Introdução

Uma das muitas aplicações em processamento digital de voz é o sistema de reconhecimento automático de locutor (ASR, do inglês *automatic speaker recognition*), que consiste em identificar um locutor a partir de um sinal de voz e encontra aplicações forense e em sistemas de segurança [1], [2].

Um sistema eletrônico de disfarce de voz (EDV) muda aspectos da voz do locutor para dificultar ou impedir seu reconhecimento, utilizando algum processamento digital. Isso encontra aplicações quando o locutor não quer ser reconhecido, como em casos de divulgação de depoimento de testemunhas, popularmente conhecido como *voz de pato*, ou em sistemas que dificultam a gravação de voz não autorizada por meio de escutas escondidas. Embora existam vários aplicativos que fazem diferentes tipos de EDV, bem como o estudo de seus efeitos em sistemas ASR, existe pouca literatura aberta de como esses efeitos são gerados [2]-[5]. Este trabalho propõe um sistema EDV em tempo real, por meio de múltiplas modulações SSB em tempo discreto [6], [7].

Na Seção II é descrito o processo de modulação SSB em tempo discreto, bem como sua implementação. O projeto do efeito múltiplo SSB é descrito na Seção III. Os resultados dos experimentos qualitativos são apresentados na Seção IV e a conclusão do artigo encontra-se na Seção V.

## II. Preliminares

Um sistema de modulação SSB consiste em deslocar o sinal para uma dada frequência, removendo uma das bandas laterais, *upper side band* (USB) ou *lower side band* (LSB), diferentemente do que ocorre em uma modulação por amplitude DSB, em que a banda de passagem dobra [6].

D. W. de França Silva e G. Jerônimo da Silva Jr¸ Grupo de Processamento de Sinais, Departamento de Eletrônica e Sistemas, Universidade Federal de Pernambuco, Recife-PE, Brasil, E-mails: diogeneswallis@outlook.com, gilson.silvajr@ufpe.br.

Um componente importante nessa modulação é a transformada discreta de Hilbert, que pode ser obtida através da convolução do sinal por um filtro com resposta em frequência dada por $H_{\rm h}(e^{j\omega}) = -j\text{sign}(\omega)$ ($\text{sign}(\omega) = \omega/|\omega|, \omega \neq 0, \text{sign}(0) = 0$), para $|\omega| < \pi$. Calculando a transformada de Fourier de tempo discreto (TFTD) inversa, obtêm-se o filtro de Hilbert, dado por

$$h_{\rm h}[n] = \frac{1-(-1)^n}{\pi n},$$

e a transformada de Hilbert discreta do sinal $x[n]$ é dada por

$$\hat{x}[n] = \sum_{m=-\infty}^{\infty} \frac{1-(-1)^m}{\pi m} x[n-m],$$

pode ser implementada de forma aproximada através de um filtro FIR, utilizando uma técnica conhecida como janelamento [8]. Considerando isso, uma modulação SSB pode ser realizada pelo diagrama da Figura 1.

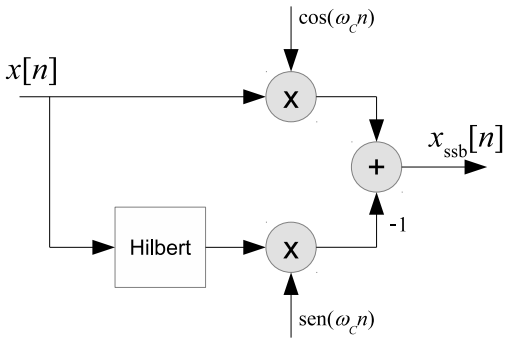


Fig. 1. Modulação SSB em tempo discreto.

Seja $X(e^{j\omega})$ a TFTD de $x[n] \in \mathbb{R}$, com

$$X(e^{j\omega}) = X_e(e^{j\omega}) + X_d(e^{j\omega}),$$

em que $X_e(e^{j\omega}) = 0$ para $0 < \omega < \pi$, essa é a parte à esquerda da TFTD de $x[n]$, e $X_d(e^{j\omega}) = 0$ para $-\pi < \omega < 0$ é a parte à direita.

Utilizando a propriedade da modulação em tempo discreto [7],

$$X_{\rm SSB}(e^{j\omega}) = X_e(e^{j(\omega+\omega_c)}) + X_d(e^{j(\omega-\omega_c)}).$$

Se $\omega_c > 0$, tem-se a modulação USB, que afasta o espectro do sinal $x[n]$ da origem, deslocando o *pitch* de $\omega_c$ a direita [1]. No caso em que $\omega_c < 0$, tem-se a modulação LSB, que aproxima o espectro do sinal $x[n]$ da origem, deslocando o *pitch* de $\omega_c$ a esquerda.

As bandas laterais podem se sobrepor, causando distorções que podem comprometer a inteligibilidade do sinal. Para evitar esse problema, um filtro passa banda pode ser colocado opcionalmente antes do modulador SSB.

## III. Projeto do efeito múltiplo SSB

Considerando um disfarce de voz utilizando único modulador SSB em tempo discreto (a maioria dos sistemas EDV descritos em aplicativos ou blogs implementam o deslocamento do pitch via uma única modulação SSB), é possível recuperar a voz original utilizando uma demodulação SSB. A ideia da proposta é utilizar múltiplas modulações em paralelo e somá-las, de forma que a sobreposição dos sons cause um efeito na voz que não seja possível de recuperar. A Figura 2 mostra um esquema do *efeito múltiplo SSB* (EMSSB) proposto, utilizando 4 moduladores.

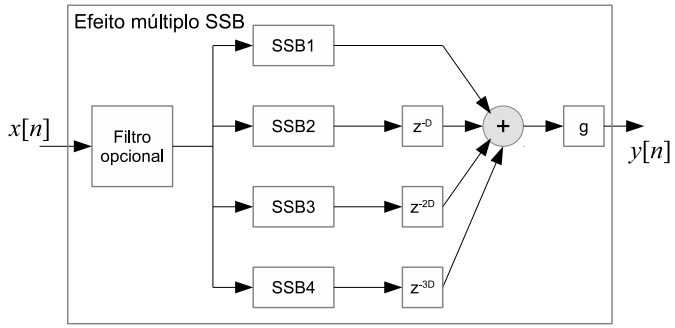


Fig. 2. Diagrama do efeito proposto, chamado de múltiplo SSB, com 4 moduladores.

Cada bloco SSB possui frequência de portadora escolhida aleatoriamente, para que os sinais se sobreponham na frequência. Note que as frequências podem ser positivas ou negativas, equivalendo a modulação USB ou LSB respectivamente. Além disso, cada SSB é submetido a retardos diferentes, espaçados de $D$ amostras, isto é, a saída do $(i+1)$-ésimo SSB está atrasado de $t_d = T_s D$ da saída do $i$-ésimo SSB. Ao fim do processo, um ganho (ou atenuação), $g$, é aplicado ao sinal para normalizar a amplitude e evitar saturações nos alto-falantes.

Generalizando, o EMSSB possui $N$ moduladores. Com isso, considerando um sinal de voz amostrado, $x[n]$, com um pitch, o sinal de saída, $y[n]$, possui $N$ pitchs. Esse efeito de múltiplos pitchs pode ser multiplicado cascateando vários estágios de EMSSB. Considerando $M$ estágios, cada um com $N_m$ moduladores, o sinal após todo o cascateamento, $y[n]$, terá $\prod_{m=1}^{M} N_m$ pitchs e corresponde a um único EMSSB com esse valor de moduladores. Entretanto, o projetista do sistema deve adequar o número de pitchs, que dificulta o reconhecimento da voz, com a inteligibilidade do sinal, que vai diminuindo com esse número.

## IV. Resultados de experimentos

A simulação foi realizada na ferramenta computacional Scilab utilizando frequência de amostragem do sinal de voz de $f_s = 8$ kHz. Sugere-se que as frequências de portadoras sejam sorteadas entre $-\pi/10 < \omega_c < \pi/10$ rad/amostra, com distribuição de probabilidade uniforme, isso equivale a frequências entre $-400 < f < 400$ Hz. O filtro ante sobreposição opcional passa banda usado é um FLG (fase linear generalizada) com 33 pontos utilizando janela retangular com frequências de cortes $0{,}1\pi$ e $0{,}4\pi$ rad/amostras.

Utilizou-se 4 EMSSB em cascata, com 2, 3 e 4 moduladores SSB em cada. Os experimentos foram realizados utilizando o filtro opcional e sem o filtro opcional. A escolha adequada de $D$ depende da velocidade da conversa do locutor, mas foi utilizado $D = 112$, que implica em $t_d = T_s D \approx 14$ ms. Os resultados foram avaliados de forma qualitativa com 3 níveis: bom, regular e fraco, avaliando a inteligibilidade e o mascaramento (abreviado com letras). Um bom mascaramento significa que a voz no sinal não se parece com a voz da pessoa que originou o som. Foi utilizado um áudio com duração de 31 s, cujo locutor é do sexo masculino e de 36 anos. Os resultados estão na Tabela I. Não foi percebido diferença qualitativa significativa ao utilizar o filtro opcional.

TABELA I

Avaliação da inteligibilidade (I) e mascaramento (M) em função do número de moduladores por EMSSB e do estágio.

| # mod | est 1 | est 2 | est 3 | est 4 |
|---|---|---|---|---|
| 2 mods | I: bom<br>M: reg | I: bom<br>M: bom | I: bom<br>M: bom | I: bom<br>M: bom |
| 3 mods | I: bom<br>M: bom | I: bom<br>M: bom | I: bom<br>M: bom | I: reg<br>M: bom |
| 4 mods | I: bom<br>M: bom | I: bom<br>M: bom | I: reg<br>M: bom | I: fra<br>M: bom |

## V. Conclusões

Um sistema eletrônico de disfarce de voz utilizando múltiplos moduladores SSB em tempo discreto foi proposto. Esse sistema é simples de implementar em tempo real, pois utiliza apenas filtros FIR e multiplicadores. A simulação desse sistema foi realizada e os resultados apresentam as características desejadas, avaliando de forma qualitativa. Como trabalho futuro, deseja-se avaliar o sistema utilizando técnicas conhecidas de reconhecimento automático de locutor.

## Referências

[1] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*. Pearson, 2011.

[2] H. Wu, Y. Wang, and J. Huang, "Identification of electronic disguised voices," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pp. 489–500, March 2014.

[3] T. Tan, "The effect of voice disguise on automatic speaker recognition," in *2010 3rd International Congress on Image and Signal Processing*, vol. 8, Oct 2010, pp. 3538–3541.

[4] G. S. Didla and H. Hollien, "Voice disguise and speaker identification," *Proceedings of Meetings on Acoustics*, vol. 25, no. 1, p. 8, 2015. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/2.0000239

[5] S. Kurian and N. G. Kurup, "Recognition of electronic disguised voices by the means of MFCC," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 5, no. 6, pp. 4603–4609, June 2016.

[6] B. P. Lathi, *Sistemas de Comunicações Analógicos e Digitais Modernos*, 4th ed. LTC, 2012.

[7] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Sinais e Sistemas*, 2nd ed. Pearson Prentice Hall, 2010.

[8] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 3rd ed. New Jersey: Prentice Hall, 2010.