Detecção e Suavização de Cliques Naturais em Bancos de Fala Visando Síntese Concatenativa de Alta Qualidade

Monique V. Nicodem, Rui Seara e Fernando S. Pacheco

***Resumo*— Este artigo apresenta uma técnica para detecção e suavização de cliques involuntários, gerados pelo trato vocal humano, que degradam a qualidade de sistemas de síntese de fala. A abordagem proposta pretende melhorar os bancos de fala existentes nos sistemas concatenativos. A técnica proposta para detecção de cliques considera a filtragem passa-faixa nas sub-bandas de freqüência com baixa energia em um segmento de fala, modelagem auto-regressiva, erro de predição e comparação com um limiar adequado. A suavização, por sua vez, realiza uma ponderação do segmento em que o clique é detectado através de uma função janela visando reduzir a percepção do referido clique. Resultados experimentais preliminares comprovam a aplicabilidade da técnica proposta.**

***Palavras-Chave*— Detecção de cliques, melhoria de sinais de fala, síntese concatenativa de fala.**

***Abstract*— This paper presents a technique for detecting and smoothing involuntary clicks generated by the human vocal tract, which degrade the quality of text-to-speech systems. This approach is useful for high quality corpus-based concatenative speech synthesis. The proposed click detection technique is based on bandpass filtering the low energy subbands of a speech signal, autoregressive modeling, prediction error, and a thresholding approach. By using a windowing technique in the smoothing phase one reduces considerably the undesired click effects. Preliminary experimental results ratify the applicability of the proposed approach.**

***Keywords*— Click detection, speech enhancement, concatenative speech synthesis.**

I. INTRODUÇÃO

Em sistemas de síntese de fala concatenativa com alta qualidade, a fala sintetizada é obtida considerando-se o uso de um grande *corpus* (da ordem de dezenas de horas). Tal *corpus* é previamente gravado e são identificados (anotados) os instantes iniciais e finais de cada unidade [1], [2]. Na etapa de síntese, um processo de busca é usado para selecionar os melhores segmentos através da otimização de critérios *ad hoc* que levem a uma fala sintética de alta qualidade. O resultado final do processo de síntese é obtido através da concatenação dos segmentos selecionados [3].

É importante mencionar que, nos dias atuais, a etapa de gravação do *corpus* não constitui qualquer problema para alcançar a qualidade de fala requerida. Isso porque, geralmente, tal gravação é realizada em estúdios profissionais e monitorada por especialistas da área, garantindo um nível baixo de ruído como também mínima distorção [4]. Por outro lado, a fala humana, mesmo a de um locutor profissional, apresenta degradações eventuais produzidas pelo próprio aparelho fonador, as quais se evidenciam em uma aplicação de síntese concatenativa. Tais degradações, causadas tanto por variações de *pitch* e energia (*jitter* e *shimmer* [5]) quanto por cliques, são as principais responsáveis por deteriorar a qualidade perceptual da fala sintetizada. Essa deterioração deve-se em parte à ocorrência de cliques em segmentos contíguos e/ou à associação com possíveis descontinuidades introduzidas pelo processo de concatenação.

É importante aqui mencionar que cliques não representam elementos de degradação para todos os idiomas existentes. Em alguns idiomas africanos, dentre os quais pode-se enumerar *!Xóõ*, *!Xũ*, Nama, Zulu e *Xhosa*, e no idioma australiano *Damin* [6], cliques constituem fonemas consonantais e carregam informações úteis da língua [7]–[10].

Os cliques sob consideração neste trabalho não representam propriamente fonemas. Caracterizam-se por descontinuidades presentes em segmentos de fala. Manifestam-se na forma de pequenos estalos, praticamente imperceptíveis na fala corrente. São produzidos de maneira involuntária, o que nos leva a nomeá-los cliques involuntários. Para reduzir a audibilidade de tais cliques, é necessário primeiro detectá-los para, posteriormente, realizar um tratamento apropriado. Esse tratamento pode consistir na supressão, interpolação ou atenuação das amostras do sinal de fala degradadas por cliques.

O presente trabalho propõe uma técnica de detecção e tratamento de cliques involuntários. Essa técnica, em nosso conhecimento, não tem sido ainda apresentada na literatura da área. O procedimento em questão é aplicado como um estágio de pré-processamento *offline* ao banco de fala (posterior à segmentação e rotulagem). Tal abordagem não compromete o desempenho computacional do processo de síntese.

A utilização de uma técnica para detectar cliques naturais é motivada pelos procedimentos existentes para localizar ruídos impulsivos em gravações de áudio antigas. A literatura sobre detecção desses ruídos apresenta técnicas baseadas em filtragem passa-altas, análise multirresolução via *wavelets*, abordagem *Bayesiana* e redes neurais [11]–[13]. No presente trabalho, escolhemos utilizar a detecção de cliques baseada na avaliação do erro de predição (obtido via análise preditiva linear) em algumas bandas de freqüência definidas do sinal de

Monique V. Nicodem, Rui Seara e Fernando S. Pacheco, LINSE – Laboratório de Circuitos e Processamento de Sinais, Departamento de Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, SC, E-mails: monique@linse.ufsc.br, seara@linse.ufsc.br, fernando@linse.ufsc.br.

Este trabalho foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Financiadora de Estudos e Projetos (FINEP) e Dígitro Tecnologia Ltda.

fala, considerando-se uma abordagem similar à utilizada em discriminação de ruídos impulsivos em gravações antigas [14]. Considerando o tipo de clique particular (com amplitude muito inferior à de cliques presentes em gravações de áudio antigas), a técnica de detecção utiliza ainda outros procedimentos, tais como a análise de energia em sub-bandas e filtragem passa-faixa.

Técnicas de tratamento de ruídos impulsivos em gravações de áudio antigas usando interpolação e supressão são consideradas na literatura [15]. No presente trabalho, propõe-se o uso de uma técnica de suavização motivada pela sua simplicidade de aplicação. Tal técnica consiste em mascarar o trecho do segmento contendo um clique através de uma janela de ponderação especialmente construída para esse fim.

Este trabalho está organizado como segue. A Seção II apresenta uma modelagem dos cliques gerados pelo aparelho fonador humano. As Seções III e IV apresentam propostas para detecção e suavização de tais cliques involuntários. Por fim, resultados experimentais e conclusões são apresentados, respectivamente, nas Seções V e VI.

II. Modelagem de Cliques Involuntários

Como anteriormente mencionado, os cliques involuntários se manifestam na forma de pequenos estalos, quase imperceptíveis em gravações que compõem bancos de fala visando síntese concatenativa.

O mecanismo de surgimento de tais cliques não é ainda explorado satisfatoriamente na literatura. No entanto, podemos tentar explicar tal mecanismo utilizando como referência a produção de cliques consonantais e emergentes [9], [10], [16], [17]. A produção de cliques consonantais envolve um fluxo de ar ingressivo atravessando uma constrição total ou parcial formada entre a língua e um ponto articulatório. Os cliques emergentes, por sua vez, surgem a partir da formação de uma cavidade de ar entre dois pontos de articulação. Quando essa cavidade sofre expansão de volume, o relaxamento de um dos pontos articulatórios leva à produção do clique. Dessa maneira, acreditamos que os cliques involuntários sejam originados por mecanismos similares.

A Fig. 1 ilustra um segmento de fala apresentando um clique involuntário. O segmento considerado nessa figura, selecionado dentre gravações que compõem um *corpus* de fala, é degradado pela presença de um clique localizado entre as amostras #130 e #145. O clique em questão pode ser considerado como um sinal espúrio (ruído) adicionado ao sinal de fala ideal (isento de cliques). Tal padrão é sempre encontrado quando cliques involuntários estão presentes em gravações.

Dessa maneira, para modelar cliques involuntários, é adotada uma representação similar ao de um ruído aditivo intermitente incorporado ao sinal de fala ideal. Assim, um sinal $y(n)$ contendo cliques involuntários é modelado por

$$y(n) = x(n) + i(n)r(n), \tag{1}$$

onde $x(n)$ é o sinal de fala ideal, $i(n)$ denota um processo de chaveamento assumindo valores $\{0, 1\}$, o qual indica a ausência ou presença de clique, e $r(n)$ representa os cliques que degradam o sinal de fala ideal. A modelagem aqui proposta é inspirada na representação de ruídos impulsivos presentes em gravações de áudio antigas [11]. O objetivo de realizar tal modelagem consiste em possibilitar a geração de cliques artificiais semelhantes aos naturais (involuntários). A geração artificial, por sua vez, facilita a avaliação das ferramentas de detecção de cliques à medida que se conhece *a priori* suas localizações.

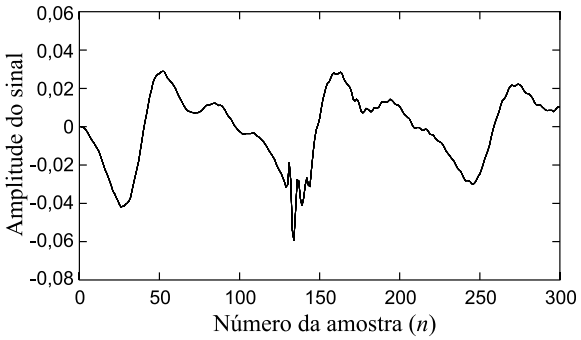

Fig. 1. Segmento de fala vozeado contendo um clique involuntário.

III. Detecção de Cliques

A detecção consiste em localizar as regiões degradadas por cliques involuntários em um sinal de fala. A técnica aqui considerada é motivada pelo conceito de que um clique é evidenciado quando, em uma sub-banda específica, sua energia supera a de suas amostras contíguas. A Fig. 2 ilustra tal fenômeno. Nessa figura, é mostrado o espectrograma de um segmento com 3200 amostras (freqüência de amostragem de 16 kHz) do fone [ã] contendo um clique involuntário indicado pela seta entre as amostras #2608 e #2623. Na figura, é possível identificar o clique com relativa facilidade na sub-banda de freqüências que se estende aproximadamente de 2 kHz a 5 kHz. A identificação é facilitada porque, na referida sub-banda, a energia do clique se destaca em relação à energia do sinal analisado nas regiões temporalmente próximas ao clique. O que implica em se dizer que, relativamente ao sinal de clique, a região sob análise, considerando o sinal ideal, é de baixa energia.

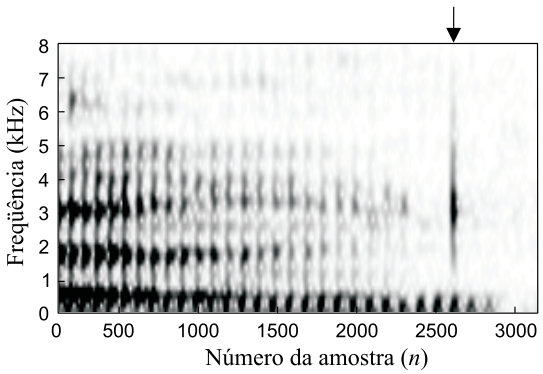

Fig. 2. Espectrograma de um segmento de fala contendo um clique involuntário.

A Fig. 3 mostra o diagrama em blocos da técnica de detecção proposta. Verifica-se que a primeira etapa da detecção consiste em segmentar o sinal de fala utilizando uma janela

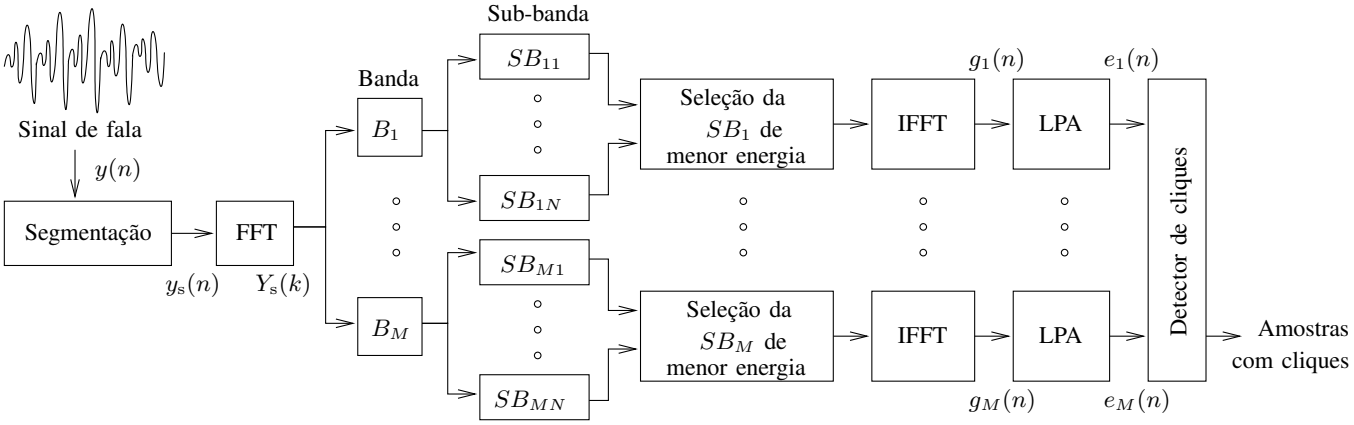

Fig. 3. Procedimento proposto para detecção de cliques involutários em segmentos de fala.

de Hanning, com subseqüente classificação vozeado/não-vozeado. Cada segmento de fala é, por sua vez, transformado para o domínio da freqüência com o auxílio da transformada de Fourier discreta (DFT). A DFT é aqui obtida através de um algoritmo de transformada rápida de Fourier (FFT). O segmento de fala, agora no domínio da freqüência, é dividido em M bandas B_1,B_2,...,B_M. Por sua vez, cada uma das M bandas é ainda subdividida em N sub-bandas SB_{ij} para $i = 1,\ldots,M$ e $j = 1,\ldots,N$. A próxima etapa consiste em selecionar para cada banda a sub-banda de menor energia, desde que essa energia não ultrapasse um limiar estipulado. Dessa forma, são atribuídos valores nulos aos coeficientes da FFT não correspondentes à sub-banda selecionada. Se a energia dessa sub-banda ultrapassar o limiar, considera-se uma região de alta energia em que um possível clique estaria mascarado. Nesse caso, o processo de detecção para a banda considerada é interrompido. Após a seleção das sub-bandas de menor energia de cada banda, obtém-se a transformada de Fourier discreta inversa (IDFT) via um algoritmo FFT inverso (IFFT), resultando em até M sinais de baixa energia. Para cada sinal, é realizada uma análise preditiva linear (LPA), técnica usualmente considerada para restauração de sinais de áudio [11]. A análise preditiva engloba uma estimação dos parâmetros de um modelo auto-regressivo (AR), com subseqüente cálculo do erro de predição normalizado $e_i(n)$ obtido comparando-se o sinal real $g_i(n)$ com a sua estimativa $\widehat{g_i}(n)$, para $i = 1,\ldots,M$. Quando o valor absoluto do erro de predição para um dos M sinais superar um estipulado limiar, conclui-se pela existência de um clique. Tal função é desempenhada pelo detector de cliques mostrado na Fig. 3. Visando a melhoria de desempenho da detecção, o mesmo procedimento é considerado para o sinal de fala reverso.

Um ponto essencial no processo de detecção consiste em obter os limiares adequados para o detector de cliques. Assim, o seguinte procedimento é aplicado sobre um segmento sob análise:

i) em cada segmento, 1% das amostras de maior amplitude do valor absoluto do erro de predição normalizado é descartada;

ii) a amostra de maior amplitude do sinal resultante após (*i*) é armazenada. Tal valor é multiplicado por $1{,}5$ para se definir o requerido limiar.

A Fig. 4 ilustra um limiar de detecção obtido considerando-se o procedimento anterior, como também o valor absoluto do erro de predição normalizado para um trecho de sinal de fala que apresenta um clique.

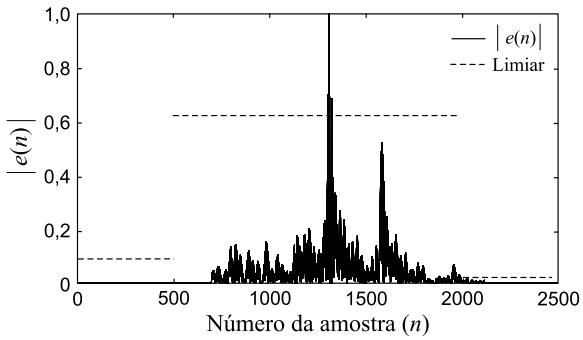

Fig. 4. Valor absoluto do erro de predição normalizado e limiar de detecção.

É importante mencionar a existência de outros procedimentos para a obtenção de tal limiar. Em detecção de distúrbios impulsivos provenientes do meio de gravação, considera-se a estimativa do desvio padrão do sinal de excitação [11] ou, ainda, a mediana do valor absoluto do erro de predição [14], [18]. Entretanto, tais procedimentos se mostraram menos eficazes do que o apresentado anteriormente para o tipo de aplicação em questão.

Destacamos ainda que a técnica proposta se restringe ao tratamento de cliques involuntários existentes em segmentos de fala vozeados, como também não-vozeados de baixa energia. Essa restrição consiste em uma solução aceitável, visto que a percepção de cliques em segmentos não-vozeados de alta energia é atenuada através do efeito de mascaramento existente em tais segmentos.

IV. Suavização dos Cliques

A suavização de cliques consiste em um estágio de processamento que objetiva reduzir a audibilidade dos cliques

involuntários. O presente trabalho propõe, para redução de audibilidade, o uso de uma função janela da seguinte forma:

$$h(n) = 1 - \alpha_1 w_{\rm h}(n) + \alpha_2 w'_{\rm h}(n), \qquad 0 \leq n \leq 4P, \qquad (2)$$

onde $w_{\rm h}(n)$ é uma janela de Hanning com $4P+1$ coeficientes, $w'_{\rm h}(n)$, um sinal com P amostras iniciais e finais nulas, tendo $2P+1$ amostras centrais correspondendo a uma outra função janela de Hanning, $0 \leq \alpha_1 \leq 1$ e $0 \leq \alpha_2 \leq 1$ são parâmetros que ajustam o peso de cada uma das janelas envolvidas em $h(n)$, respectivamente. É ainda importante considerar, em nosso caso, que $\alpha_1 > \alpha_2$.

A Fig. 5 ilustra uma janela de suavização para $\alpha_1 = 1$, $\alpha_2 = 0{,}1$ e $P = 180$.

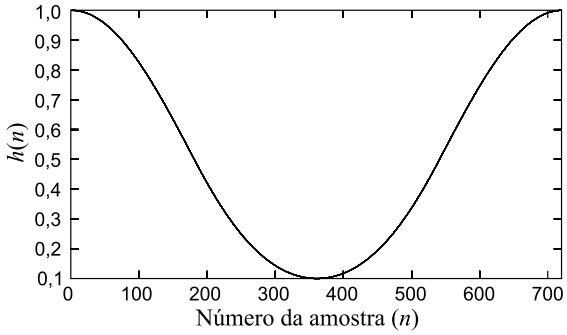

Fig. 5. Janela de suavização com $\alpha_1 = 1$, $\alpha_2 = 0{,}1$ e $P = 180$.

V. Resultados Experimentais

Para se avaliar o procedimento de detecção apresentado, propõe-se como figura de mérito o índice de detecção correta (IDC) definido por

$$\text{IDC} = \frac{n_{\rm c}}{n_{\rm t}} \times 100, \qquad (3)$$

onde $n_{\rm c}$ representa o número de cliques detectados corretamente e $n_{\rm t}$, o número total de cliques.

Um conjunto de 105 cliques gerado artificialmente, com distribuição estatística (amplitude, duração e fator de amortecimento) semelhante à dos cliques naturais, é adicionado a um sinal de fala isento de cliques. Tal sinal, amostrado à taxa de 16 kHz, apresenta um minuto de duração. Esse sinal é segmentado utilizando uma janela de Hanning. Cada quadro tem duração de 100 ms com 50 ms de recobrimento (*overlap*). Uma FFT de 4096 pontos é aplicada para cada segmento. Obtém-se então uma divisão em M bandas e N sub-bandas. A escolha dos valores de M, N e limiar de energia é feita usando-se uma busca exaustiva, visando-se obter o maior número possível de cliques detectado corretamente.

O desempenho da abordagem proposta é avaliado separadamente para segmentos vozeados e não-vozeados. Para os segmentos vozeados, os valores de M, N e limiar de energia que proporcionam maior IDC são, respectivamente, iguais a 4, 3 e $0{,}45$. O IDC nesse caso é de $91{,}80\%$. Para os não-vozeados de baixa energia, os valores de M, N e limiar de energia que levam ao maior IDC são, respectivamente, 4, 3 e $0{,}25$. Nesse caso, o IDC obtido é $81{,}82\%$.

Outro sinal de fala, com um minuto de duração, amostrado à taxa de 16 kHz e contendo cliques naturais (que para o locutor considerado ocorrem a uma taxa típica de 44 por minuto), é também usado para avaliar a técnica de detecção proposta. Para tal, utilizam-se os mesmos valores de M, N e limiar de energia previamente obtidos. Para tal sinal, o IDC total (vozeado/não-vozeado) é agora $78{,}57\%$.

É importante mencionar que o número de detecções falsas não é considerado como parâmetro para avaliar o desempenho da técnica de detecção proposta. Isso porque o tratamento de suavização aplicado em cliques detectados erroneamente (falsos positivos) não prejudica a qualidade da fala (considerando a freqüência de ocorrência típica dos cliques), visto que a atenuação é suave e se aplica somente a uma região do fonema contendo um clique.

Após a etapa de detecção, é realizada a suavização dos cliques encontrados. Para se avaliar a técnica de suavização, propõe-se o uso de uma medida perceptual definida por

$$\text{IS} = \frac{n_{\rm i}}{n_{\rm c}} \times 100, \qquad (4)$$

onde IS representa o índice de suavização, $n_{\rm i}$ é o número de cliques detectados que se tornaram inaudíveis e $n_{\rm c}$, o número total de cliques detectados.

Para segmentos vozeados, o procedimento de suavização é realizado utilizando $\alpha_1 = 1$, $\alpha_2 = 0{,}1$ e $P = 60$. Para fala não-vozeada, considera-se $\alpha_1 = 1$, $\alpha_2 = 0$ e $P = 180$. Após a suavização, o sinal de fala (contendo cliques naturais) é avaliado por dois ouvintes especialistas em processamento de fala utilizando um fone de ouvido de alta qualidade. Cada ouvinte apontou os instantes onde um som de clique era audível. Posteriormente, tais instantes foram comparados com os instantes obtidos pela técnica automática de detecção de cliques. O IS conjunto obtido na avaliação é igual a $93{,}75\%$, indicando um desempenho satisfatório.

VI. Conclusões

No presente trabalho, uma técnica para detecção e tratamento de cliques naturais em sinais de fala tem sido apresentada. Resultados experimentais comprovam a aplicabilidade da abordagem proposta para detectar e suavizar cliques em segmentos de fala vozeados como também em não-vozeados de baixa energia. Estudos adicionais estão sendo realizados para avaliar a influência de falsas detecções sobre a qualidade da fala sintética. Estão sendo também estudadas estratégias de otimização dos parâmetros envolvidos no procedimento de suavização. Além da suavização, outras técnicas de tratamento de cliques involuntários estão sendo consideradas. Como exemplo, estamos avaliando ferramentas de interpolação bem como a eliminação (*pruning*) de fonemas contendo cliques involuntários. Para comparar os procedimentos de tratamento considerados, pretende-se realizar um experimento de avaliação perceptual formal (*Mean Opinion Score* - MOS) a fim de comparar a qualidade da fala sintética produzida por um sistema de síntese concatenativa antes e após o tratamento de cliques.

REFERÊNCIAS

[1] R. Prudon and C. d'Alessandro, "A selection/concatenation text-to-speech synthesis system: Databases development, system design, comparative evaluation," in *Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW4'01)*, Pitlochry, Scotland, Aug./Sept. 2001, pp. 137–142.

[2] M. Chu, H. Peng, H.-Y. Yang, and E. Chang, "Selecting non-uniform units from a very large corpus for concatenative speech synthesizer," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, Salt Lake City, USA, May 2001, pp. 785–788.

[3] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96)*, Atlanta, May 1996, pp. 373–376.

[4] W. Zhu, W. Zhang, Q. Shi, F. Chen, H. Li, X. Ma, and L. Shen, "Corpus building for data-driven TTS systems," in *Proc. IEEE Workshop on Speech Synthesis (TTS'02)*, Santa Monica, USA, Sept. 2002, pp. 199–202.

[5] P. J. Murphy, "Spectral characterization of jitter, shimmer, and additive noise in synthetically generated voice signals," *Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 978–988, Feb. 2000.

[6] E. Pennisi, "The first language?" *Science*, vol. 303, pp. 1319–1320, Feb. 2004.

[7] W. J. Hardcastle and J. Laver, *The Handbook of Phonetic Sciences*. Cambridge, USA: Blackwell, 1997.

[8] P. Ladefoged and A. Traill, "Clicks and their accompaniments," *Journal of Phonetics*, vol. 22, pp. 33–64, 1994.

[9] K. N. Stevens, *Acoustic Phonetics*. Cambridge, USA: MIT Press, 1998.

[10] J. Clark and C. Yallop, *An Introduction to Phonetics and Phonology*, 2nd ed. Cambridge, USA: Blackwell, 1995.

[11] S. J. Godsill and P. J. Rayner, *Digital Audio Restoration - A Statistical Model Based Approach*. London, UK: Springer-Verlag, 1998.

[12] A. Czyzewski, "Some methods for detection and interpolation of impulsive distortions in old audio recordings," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (ASSP'95)*, New York, USA, Oct. 1995, pp. 139–142.

[13] S. J. Godsill, "A Bayesian approach to the restoration of degraded audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 267–278, July 1995.

[14] P. A. A. Esquef, "Restauração de sinais de áudio degradados por ruído impulsivo," Tese de mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, 1999.

[15] S. J. Godsill, "A Bayesian approach to the detection and correction of error bursts in audio signals," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'92)*, vol. 2, San Francisco, USA, Mar. 1992, pp. 261–264.

[16] J. J. Ohala. Emergent stops. Unpublished. [Online]. Available: http://trill.berkeley.edu/users/ohala/papers

[17] ——, "A probable case of clicks influencing the sound patterns of some european languages," *Phonetica*, vol. 52, no. 3, pp. 160–170, 1995.

[18] P. A. A. Esquef, M. Karjalainen, and V. Välimäki, "Detection of clicks in audio signals using warped linear prediction," in *Proc. IEEE International Conference on Digital Signal Processing (DSP'02)*, vol. 2, Santorini, Greece, July 2002, pp. 1085–1088.