Sociedade de Engenharia de Áudio

Artigo de Convenção

Apresentado na IX Convenção Nacional
11 – 13 de Abril de 2005, São Paulo, SP

Obtenção de Marcas de Pitch em Sinais de Voz para Síntese por Concatenação Temporal

Vagner L. Latsch e Sergio L. Netto
Programa de Engenharia Elétrica, COPPE/UFRJ
Código Postal 68503, Rio de Janeiro, RJ, 21941-972, Brasil
latsch@lps.ufrj.br sergioln@lps.ufrj.br

RESUMO

Neste artigo é proposto um método para obtenção das marcas de pitch em segmentos de sinais de voz a serem concatenados pelo algoritmo TD-PSOLA. O sistema proposto utiliza a captação de um sinal auxiliar, através de um microfone de contato, para obter informações mais intensas a respeito da atividade das cordas vocais. Isto contribuiu para melhorar o desempenho de uma detecção automática de marcas. Os resultados se mostraram promissores, inclusive para casos críticos de detecção, onde se mostrou necessária pouca ou nenhuma correção manual.

INTRODUÇÃO

Os conversores texto-fala, também conhecidos como TTS (text-to-speech), são sistemas que produzem fala sintética correspondente à leitura de um texto. De modo geral, estes sistemas se baseiam na concatenação de unidades sonoras que são devidamente processadas para se incorporar a entoação desejada à voz sintetizada. Um dos algoritmos mais comumente utilizados para este processamento, devido ao seu baixo custo computacional, é o chamado TD-PSOLA (Time Domain - Pitch Synchronous Overlap and Add). Tal algoritmo requer a detecção das chamadas marcas de pitch que indicam o instante de tempo do fechamento da glote em um dado sinal de voz. Este processo como um todo é bastante intenso e consistirá no foco do presente artigo.

Neste contexto, após uma breve introdução ao problema de conversão TTS, faremos uma apresentação do algoritmo TD-PSOLA. Mostraremos, então, a necessidade da obtenção das marcas de pitch, ilustrando claramente este conceito e a dificuldade de detectá-las de forma precisa. Será vista, em seguida, uma metodologia de obtenção das marcas utilizando-se de um sinal auxiliar obtido por um microfone de contato colocado junto à garganta do locutor. Mostraremos por fim como este sinal auxiliar facilita a implementação de todo o processo e torna o resultado obtido muito preciso e confiável. Ao fim, concluímos o artigo apresentando os principais resultados obtidos.

CONVERSÃO TEXTO-FALA IRRESTRITA

Sistemas TTS têm aplicação em diferentes áreas, como por exemplo, na consulta de *emails* por telefone, na leitura de menus e orientações em centros de atendimento automático e até mesmo no auxílio a deficientes visuais em diferentes áreas.

Em alguns casos, como nos serviços de auxilio às listas telefônicas onde se requer a pronúncia automática dos números telefônicos, o "texto" a ser convertido está limitado aos algarismos de 0 a 9. Neste caso, para gerar a fala sintética correspondente à pronúncia do número de telefone desejado, é intuitivo sugerir a concatenação de segmentos de voz contendo a pronúncia dos algarismos.

Já os demais exemplos acima citados de aplicações de sistemas TTS tratam da conversão de texto *irrestrito*, onde o texto a ser convertido não está limitado a um conjunto de palavras ou frases. Em tais casos, a concatenação pura e simples de palavras se torna impraticável se considerarmos a quantidade de palavras existentes e as suas variantes. Baseado em conceitos fonéticos, que tratam de mapear os diversos sons existentes em uma língua (fones), e fonológicos, que observam a organização destes sons construindo significado, tem-se a proposta do uso de unidades menores de concatenação, limitadas pela sílaba, pelos próprios fones, ou unidades intermediárias.

Assim, depois de definido o tipo de unidade a ser usada e quais serão necessárias para gerar as palavras e frases de uma língua em um sistema TTS, é feita a coleta destas unidades a partir de sinais de fala previamente gravados, compondo um banco de unidades [1].

Deste modo, um sistema de conversão de texto irrestrito, inicialmente fará a conversão dos caracteres do texto (grafemas) em unidades fonológicas (fonemas), e em seguida irá obter do banco as unidades necessárias para gerar a seqüência dada e concatená-las gerando o sinal de fala sintético correspondente.

É preciso levar em conta que as unidades de concatenação, sejam elas quais forem, estão sujeitas à variação, de acordo com a posição ocupada dentro de uma frase ou com a entoação aplicada. Por exemplo, no caso dos números de telefone, considerando cada algarismo uma unidade, a pronúncia do algarismo 2 na seqüência numérica **2**555-5555 será diferente na seqüência 5555-555**2**. Assim, para obter uma entoação correta e natural, seria necessário armazenar todas as variantes de uma unidade ou então usar um método capaz de modificá-las principalmente em intensidade, duração e freqüência fundamental, que são os principais fatores para caracterizar a entoação ou prosódia.

Um dos métodos mais populares, que tem sido usado em diversos sistemas TTS atuais, devido à sua simplicidade de implementação e ao reduzido custo computacional, é algoritmo TD-PSOLA descrito brevemente a seguir.

ALGORITMO TD-PSOLA

O algoritmo TD-PSOLA [2] é baseado na técnica de *overlap and add* (OLA), na qual um sinal periódico com diferente escala temporal e/ou pitch é reconstruído através da aplicação janelas síncronas com o pitch. Estas janelas são aplicadas ao sinal, centradas em *marcas de pitch* com largura típica de dois períodos, e são alongadas ou encurtadas, removidas ou repetidas para obter o sinal modificado [3].

A qualidade oferecida pelo TD-PSOLA no contexto de síntese por cópia é perto da perfeição [3]. Porém, quando o algoritmo é utilizado na modificação de segmentos concatenados, provenientes de outros contextos, se as marcas não são posicionadas de forma consistente, o resultado são erros de fase na superposição das janelas, principalmente na vizinhança de concatenação [3].

Para a determinação destas marcas, podem ser usados algoritmos de detecção de pitch chamados de PDAs (*Pitch Detection Algoritms*) [4]. Estes algoritmos podem ser divididos em duas categorias: PDAs no domínio do tempo e PDAs por análise em termo curto. Os PDAs no domínio do tempo oferecem a estimativa do pitch período a período, mas são sensíveis as degradações do sinal na janela de análise. PDAs de termo curto, por outro lado, são mais robustos mas oferecem uma estimativa do pitch médio ao longo de um número de períodos, isto porque o método conta com a similaridade do sinal de voz entre períodos de pitch adjacentes. Assim, se vários períodos de pitch estão contidos em um segmento de análise, o valor do pitch estimado é um valor médio para todo o segmento [5].

Idealmente as marcas de pitch deveriam ser introduzidas nos trechos sonoros na posição de um evento específico no ciclo de pitch e nos trechos surdos regularmente espaçadas [3]. Um evento no ciclo de pitch muito utilizado é o instante de fechamento glotal (GCI, g*lotal closure instant*), ponto onde ocorre a maior excitação do trato vocal. Porém, a detecção precisa destes instantes diretamente do sinal de voz, seja de forma automática ou mesmo de forma visual, apresenta uma enorme dificuldade significando um grande consumo de tempo [3].

Uma solução possível para a detecção precisa do instante de fechamento glotal é o uso de um equipamento chamado *eletroglotógrafo* (EGG) que mede a atividade das cordas vocais [6]. Este equipamento, porém, apresenta um custo elevado, da ordem de alguns milhares de dólares. Uma alternativa de baixo custo é vista a seguir.

MICROFONE DE CONTATO

Alguns sistemas de aquisição de voz em ambientes extremamente ruidosos têm usado microfones em contato com o pescoço, chamado de *throat microphone*, por apresentar reduzida captação de ruído ambiente. Alguns autores [7] têm proposto a utilização deste tipo de microfone para melhorar o desempenho de sistemas de reconhecimento em ambientes ruidosos. Em [8] os autores utilizaram um acelerômetro em contato com a pele, na altura da glote, e observaram que o sinal captado, quando comparado com o sinal de um EGG, representa o som gerado pela vibração das cordas vocais.

Neste trabalho propõe-se o uso de um microfone de contato para a captação da vibração da glote com o objetivo de obter os GCIs de forma barata e precisa. O "microfone" utilizado trata-se de um disco piezoelétrico cerâmico, ilustrado na Fig. 1, geralmente utilizado como captador em instrumentos musicais acústicos, como violão, violino etc.

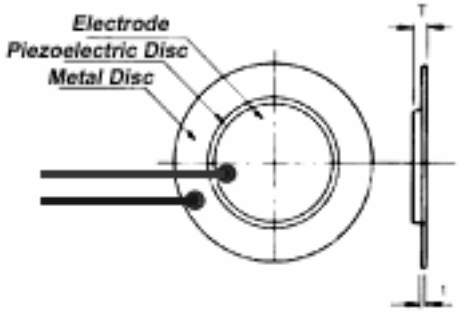

Fig. 1: Formato de um disco piezoelétrico.

O disco de metal, mostrado na Fig. 1, é colado a uma base plástica somente pelas bordas, de maneira que o centro fique livre. Esta base plástica é então fixada a uma fita de velcro, formando um colar, mostrado na Fig. 2. Este colar é colocado ao redor do pescoço de maneira que o disco piezoelétrico

fique localizado na região frontal do pescoço, o mais baixo possível, conforme mostrado na Fig. 2. O ajuste de pressão do colar (apertado ou frouxo), esta diretamente associado à qualidade do sinal, portanto a pressão ideal é aquela em que o colar fique o mais justo possível, sem causar grande desconforto.

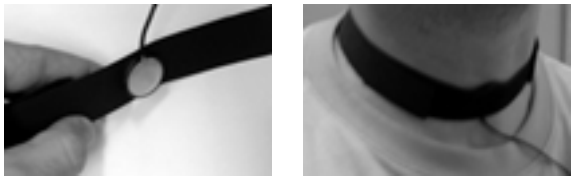

Fig. 2: Disco piezoelétrico fixado ao colar de velcro e colocação do colar na base do pescoço.

Além do sinal do microfone de contato, o sinal de voz precisa ser captado em simultâneo, no entanto, a maioria das placas de som não possui entrada para dois microfones em simultâneo (estéreo). Uma solução é utilizar a entrada *line-in* da placa de som, que pode ser utilizada em modo estéreo, porém é necessário um pré-amplificador para os microfones. Deste modo, foram montados dois pré-amplificadores, conforme a nota referenciada em [9]. Os sinais obtidos em simultâneo pelos dois microfones se mostram defasados de acordo com a distância entre os microfones, deste modo para manter esta distância fixa ao longo da gravação, foi usado um microfone acoplado aos fones de ouvido e próximo a boca. Deste modo, o atraso dependerá principalmente das características físicas do usuário. Em média este atraso equivale ao tempo de propagação para o som percorrer uma distância típica de 20 cm variando em +/- 5 cm. Deste modo, temos um atraso no sinal do microfone convencional da ordem de (0,6 +/- 0,15) ms. Os sinais obtidos foram amostrados na freqüência de 22050 Hz o que resulta em um atraso em torno de 12 amostras. Na Fig. 3 é mostrado no gráfico superior o sinal obtido pelo microfone convencional, onde foi compensado o atraso de 12 amostras, e no gráfico abaixo o sinal do microfone de contato.

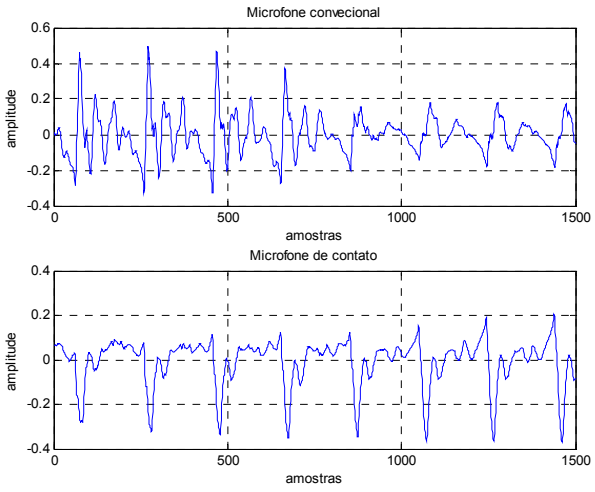

Fig. 3: sinal de voz e do microfone de contato

Observando os sinais obtidos pelo microfone de contato e pelo microfone convencional, nota-se que durante os trechos sonoros, ou seja onde ocorre atividade glotal, o sinal demonstra características mais regulares do que o sinal de voz. Esta característica se justifica pelo fato de que as vibrações captadas pelo contato provêm principalmente da laringe (considerada um tubo com dimensões constantes) que produz harmônicos quase invariantes ao longo do tempo. Por outro lado, para o sinal de voz, as diferentes configurações do trato vocal para a produção de diferentes sons, produzem diferentes harmônicos (formantes), que se mantém regulares por curtos períodos de tempo.

Na Fig 3, observa-se que a ocorrência de picos no sinal do microfone de contato podem ser bons indicativos para os GCIs. Porém, em alguns casos, estes picos têm sua amplitude reduzida não sendo possível detectá-los como máximos locais. Isto, de modo geral, inviabiliza a obtenção dos GCIs diretamente do sinal do microfone de contato, sendo necessário então um procedimento de detecção um pouco mais elaborado, como visto a seguir.

OBTENÇÂO DOS GCIs

De modo geral, destacamos neste trabalho duas técnicas para determinação dos GCIs a partir do sinal de voz: as técnicas baseadas no resíduo da predição linear e as técnicas baseadas no conceito de máxima verossimilhança. Estas famílias de algoritmos são discutidas em seqüência.

Métodos Baseados no Resíduo da Predição Linear

Muitos sistemas de análise da voz são baseados no modelo linear fonte-filtro, constituído por um filtro digital linear autoregressivo, que modela o trato vocal e uma fonte de excitação periódica considerada como um sinal representativo da atividade glotal.

Vários algoritmos precursores da detecção automática de eventos no sinal de voz, como por exemplo, os descritos em [10], [11], [12], [13] e [14], baseiam-se na idéia de que em segmentos curtos (menores do que um período de pitch) que não contêm uma excitação, o modelo de predição linear é mais adequado e conseqüentemente o erro de predição é menor. Por outro lado, quando um instante de excitação, ou o instante de fechamento glotal, está incluído no segmento de análise o erro de predição linear é maior. Deste modo, o ponto onde ocorre um grande erro de predição pode ser usado para indicar o instante do fechamento glotal [13].

Em [14], a partir da suposição de que o resíduo de predição linear exibe picos correspondentes aos GCIs, os autores observam que devidos a alguns fatores, como por exemplo, a estimação não acurada das formantes e da largura de bandas na etapa de análise, múltiplos picos podem ocorrer no sinal de resíduo, tornando difícil a estimação precisa [14] dos GCIs. Deste modo, o método proposto em [14] procura reduzir estas ambigüidades. Para isto, inicialmente o sinal de resíduo é processado no domínio da freqüência, aplicando-se uma janela de Hanning à sua FFT, para reduzir as componentes de baixas e altas freqüências. Em seguida, é obtido o contorno da Transformada de Hilbert de modo a atenuar os efeitos de fase introduzidos na obtenção do resíduo.

Na aplicação do método ao sinal do microfone de contato, inicialmente foi aplicado um filtro de pré-ênfase enfatizando as altas freqüências, para tornar seu decaimento espectral similar ao sinal de voz. Em seguida a aplicação do método de detecção pode ser observada na Fig. 4.

Observa-se que os máximos, correspondentes aos pontos de excitação do trato vocal, ocorrem com maior amplitude do que no sinal de voz, sendo mais fácil detectá-los em meio ao ruído. Porém, os picos intermediários aos supostos GCIs também ocorrem com maior amplitude.

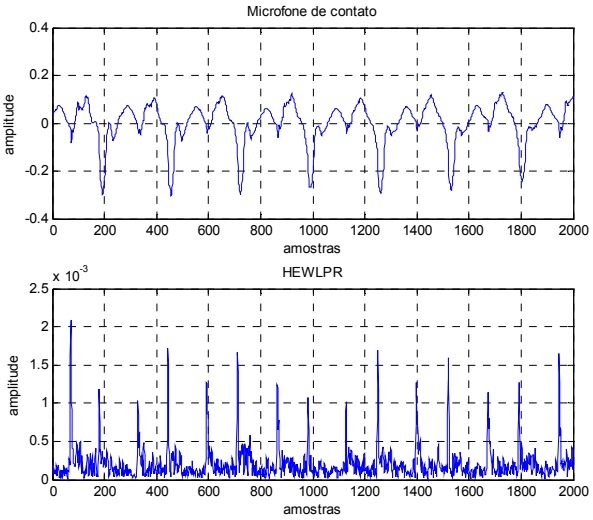

Fig. 4: Obtenção dos GCIs por resíduo da predição linear.

Supõe-se que o sinal do microfone de contato forneça informações mais intensas não somente sobre o instante de fechamento da glote, mas também em outros instantes de batimento das cordas vogais, como por exemplo, o instante de abertura da glote. Por um lado, o método confirma a suposição de que um microfone de contato forneceria informações mais nítidas sobre o movimento da glote, por outro lado, a dificuldade em separar o GCIs torna o método ineficiente para esta proposta.

Métodos Baseados na Máxima Verossimilhança

Esta metodologia foi proposta em [5] para estimar os GCIs, adaptada da teoria de detecção de épocas (ou eventos) por máxima verossimilhança em aplicações para radar. Assim como na seção anterior, este método assume que o sinal de voz dentro de um período de pitch é induzido por um pulso em uma época, geralmente definida como a representação de um GCI.

Assumindo que a produção da voz pode ser modelada por um sistema linear autoregressivo, o sinal modelo devido a uma época pode ser expresso como:

$$\hat{s}(n) = \begin{cases} \sum_{i=1}^{p} a_i \hat{s}(n-i) & 0 < n \le \infty \\ G & n = 0 \\ 0 & n < 0 \end{cases} \qquad (1)$$

onde G é uma constante positiva arbitrária e p a ordem do polinômio.

Em seguida, é suposto que a diferença entre o sinal observado, $s(n+n_0)$ $n \in [0, N-1]$ (onde n_0 é uma seqüência de atrasos de alinhamento) e o sinal modelo é um processo gaussiano e que as N observações constroem um processo gaussiano com N dimensões independentes e variância uniforme σ.

Assim, dado $x(n) = s(n+n_0) - \hat{s}(n)$, a densidade de probabilidade condicional, ou função de verossimilhança, será descrita por:

$$p(X \mid \theta) = \frac{1}{(2\pi\sigma^2)^{N/2}} . \exp\left\{ \frac{-\sum_{n=0}^{N-1} \left[s(n+n_0) - s(n) \right]^2}{2\sigma^2} \right\} \qquad (2)$$

onde θ é o espaço de parâmetros $\theta = \{\sigma, a_1, a_2, a_3, ..., n_0\}$.

Deste modo, quando o valor dos parâmetros maximizarem a função de verossimilhança, significa que uma época ocorreu. Maximizar a função de verossimilhança pode ser substituído por maximizar o logaritmo da verossimilhança e portanto a função a ser maximizada torna-se:

$$\ln\left[p(X \mid \theta) \right] = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{\sum_{n=0}^{N-1} \left[s(n+n_0) - s(n) \right]^2}{2\sigma^2} \qquad (3)$$

Assim conclui-se que não é possível encontrar uma expressão explícita para um valor ótimo de n_0. Porém, resolvendo algebricamente a potência interna ao somatório, e observando as possibilidades de máximos na função de $\ln[p(X \mid \theta)]$ em função de n_0, tem-se que o termo $\sum_{n=0}^{N-1} \left[s(n+n_0) \hat{s}(n) \right]$ é dominante. Este termo é chamado de sinal MLED (m*aximum-likelihood epoch determination*) e trata-se da correlação cruzada entre o sinal de voz e o sinal modelo. Portanto, em função de n_0, os máximos no sinal MLED equivalem aos máximos na função de verossimilhança. Em seguida, os coeficientes do sinal modelo, que produzem um máximo na função de verossimilhança, são deduzidos como os coeficientes de predição linear obtidos pelo método da autocorrelação. Neste sentido, o sinal modelo é considerado como os coeficientes de um filtro casado [5].

Em um período do sinal MLED, aparecem não só os máximos locais, onde a correlação cruzada é máxima, que correspondem aos GCIs, mas também a outros falsos candidatos. A razão em amplitude entre o pulso principal e os outros pulsos varia substancialmente e depende das propriedades do sinal, criando ambigüidade na decisão de escolha [5]. Para contornar este problema, os autores propõem o uso de um "sinal de seleção", similar à aplicação de uma janela, para enfatizar o contraste entre o pulso principal e os pulsos secundários. Os autores demonstram que o contorno da transformada de Hilbert (ou módulo do sinal analítico) do sinal MLED pode ser utilizado como sinal de seleção. A média pode ainda ser subtraída para tornar o sinal de seleção mais parecido com um pulso, sendo possível anular o sinal entre pulsos adjacentes.

Aplicando o método ao sinal de voz, os autores observaram experimentalmente que o indicativo para o GCI é melhor definido a 50% da amplitude do máximo (de zero até o ponto máximo do pulso, à esquerda) e este critério é empírico [5]. Esta imprecisão no posicionamento do GCI relatada pelos autores foi verificada em vários sinais, porém o critério de correção sugerido pelos autores nem sempre é eficiente.

Aplicando o método ao mesmo sinal do experimento anterior, passado igualmente por um filtro de pré-ênfase, o resultado da detecção é mostrado na Fig 5.

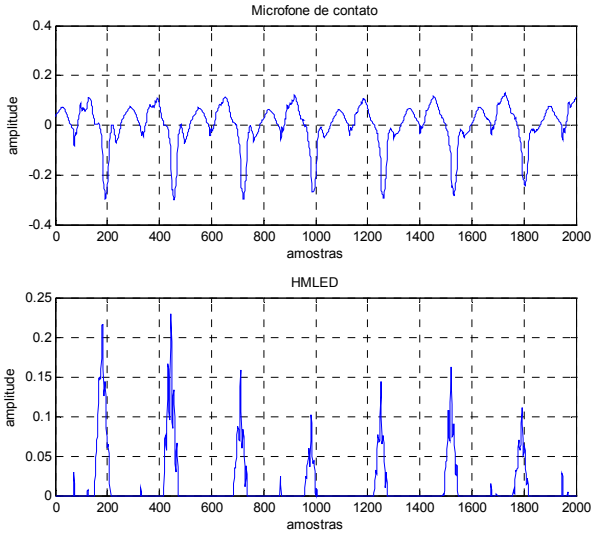

Fig. 5: Obtenção dos GCIs por máxima verossimilhança.

Conforme dito anteriormente, o método quando aplicado ao sinal de voz, apresenta uma imprecisão na detecção dos GCIs, no qual era necessário um método empírico de correção. Surpreendentemente, o método quando aplicado ao sinal do microfone de contato, apresentou esta imprecisão somente em alguns poucos casos e da ordem de 0,3 ms. Além disso, foi constatado que o módulo da transformada de Hilbert do sinal MLED é mais preciso do que próprio sinal MLED ou da multiplicação dos dois sinais conforme proposto em [5]. De fato, somente a subtração da média global do módulo da transformada de Hilbert permitiu a diferenciação entre os trechos sonoros e surdos.

SISTEMA PARA OBTENÇÃO DOS CGIS

Para o auxílio na construção de um banco de unidades para síntese por concatenação, foi implementado um aplicativo na linguagem C++, para o sistema operacional Windows, que permite ao usuário gravar os sinais em modo estéreo exibindo-os em janelas paralelas; detectar automaticamente o atraso entre os sinais; obter as marcas de GCIs automaticamente e editar estas marcas para efeito de sintonia fina.

A detecção do atraso é feita baseada no resíduo de predição linear. Apesar do sinal de resíduo do microfone de contato exibir outros picos de excitação além daqueles contidos no sinal de voz, observa-se que a correlação entre os resíduos é capaz de determinar o atraso presente no sinal de voz. Porém, diferenças de polaridade nos picos em amplitude nos resíduos, influenciam o resultado da correlação. Deste modo, foi utilizado o método [14] para reduzir as ambigüidades nos sinais de resíduo e em seguida calcular a correlação entre os dois num intervalo característico para o atraso.

A obtenção automática dos GCIs foi feita utilizando o módulo da transformada de Hilbert do sinal MLED, onde os máximos locais são detectados em segmentos de curta duração.

Para verificação do método, foram observados casos onde a detecção dos GCIs a partir do sinal de voz é extremamente difícil, principalmente para consoantes vozeadas.

Nas Fig. 6, 7, 8 e 9 são mostrados dois gráficos onde são mostrados o sinal de voz e os GCIs obtidos diretamente do sinal de voz e do sinal do microfone de contato, respectivamente. Para notação dos segmentos e na transcrição fonética foram utilizados os símbolos da Associação Internacional de Fonética (IPA) e a nomenclatura utilizada para as consoantes segue a definida em [15].

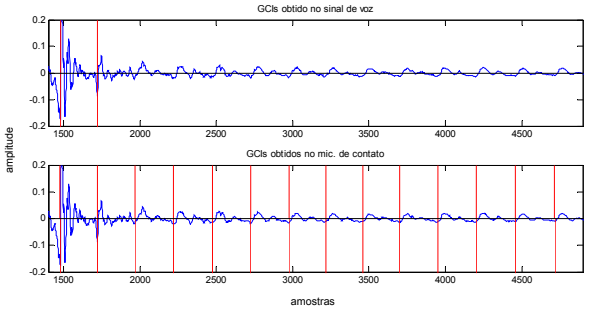

Fig. 6: Segmento do sinal de voz contendo a consoante Oclusiva Bilabial Vozeada /b/, recortada da palavra "abril" - /aˈbriu/, e as marcas dos GCIs obtidos do sinal de voz e do microfone de contato.

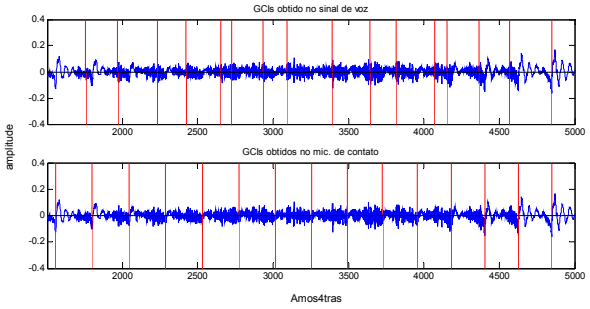

Fig. 7: Segmento do sinal de voz contendo a consoante Fricativa Alveolar Vozeada /z/, recortada da palavra "casa" - /ˈkaza/, e as marcas dos GCIs obtidos do sinal de voz e do microfone de contato.

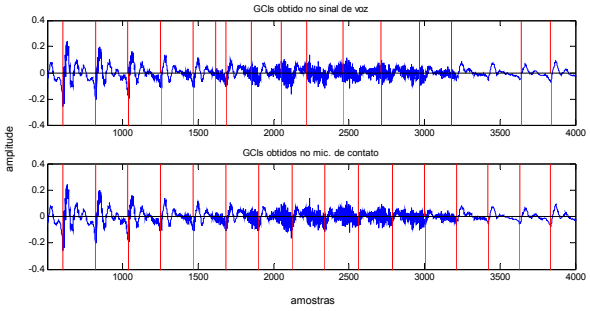

Fig. 8: Segmento do sinal de voz contendo a consoante Fricativa Alveopalatal Vozeada /ʒ/, recortada da palavra "mesmo" - /ˈmeʒmu/ e as marcas dos GCIs obtidos do sinal de voz e do microfone de contato.

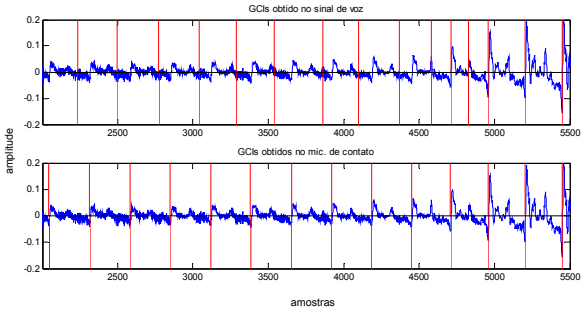

Fig. 9: Segmento do sinal de voz contendo a consoante Fricativa Alveolar Vozeada /v/, recortada da palavra "avante" - /aˈvãtʃi/, e as marcas dos GCIs obtidos do sinal de voz e do microfone de contato.

CONCLUSÕES

Foi estabelecido um método para obtenção precisa e automática das marcas de pitch em segmentos de sinais de voz, a serem utilizadas pelo algoritmo TD-PSOLA na concatenação destes segmentos. As marcas coincidem com os instantes de fechamento glotal, reduzindo a possibilidade de erros de fase na concatenação.

O método utiliza a gravação simultânea do sinal de voz e o sinal obtido por um disco piezoelétrico em contato com a pele, localizado na base do pescoço, no qual se mostrou conter mais informações sobre a atividade glotal do que o próprio sinal de voz. O disco piezoelétrico usado, assim como o circuito pré-amplificador teve caráter experimental, deste modo sugere-se que as características deste dispositivo sejam melhor exploradas assim como circuitos adequados à instrumentação. Observa-se ainda que o sinal obtido pelo microfone de contato é menos sujeito a ruído ambiente; que em trechos sonoros o sinal apresenta características mais estacionárias do que o sinal de voz; e que as consoantes sonoras são enfatizadas.

Foi descrito assim todo um método (semi-)automático para detecção automática do GCIs. Os resultados se mostraram bastante promissores, especialmente quando o método é aplicado a casos críticos onde a obtenção precisa dos GCIs a partir apenas do sinal de voz apresenta grande dificuldade quando realizada pelos métodos tradicionais.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Black, A.W., Lenzo, K.L., "Building Synthetic Voices", *FestVox 2.0*, 2003.

[2] Charpentier, F., Moulines, E. "Pitch-Synchronous wave form processing techniques for text-to-speech synthesis using diphones", *Proceedings of Eurospeech 89,* v. 2, pp. 13-19, 1989.

[3] Dutoit, T., *An introduction to text-to-speech synthesis*, Kluwer Academic Publishers, London, 1997.

[4] Hess, W., *Pitch Determination of speech Signals*, Springer-Verlag, Berlin, 1983.

[5] Cheng, Y. M., O'Shaughnessy, D., "Automatic and Reliable Estimation of Glottal Closuse Instants and Period", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 37, n. 12, pp. 1805-1815, 1989.

[6] Krishnamurthy, A. K., Childers, D. G., "Two-Channel Speech Analysis", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 34, No. 4, pp.730-743, 1986.

[7] Graciarena, M., Franco, H., Sonmez, K., et al., "Combining Standard and Throat Microphones for Robust Speech Recognition", *IEEE Signal Processing Letters*, v. 10, n. 3, 2003.

[8] Askenfelt, A., Gauffin, J., Sundberg, J., et al., "A comparison of contact microphone and electroglottograph for the measure of fundamental frequency", *Journal of Speech and Hearing Research*, v. 23, n. 2, pp. 258-273, 1980.

[9] Cittadinni, R., Poulan, F., "TS971 based Electret Condenser Microphone amplifier", *AN1534 Application Note*, STMicroeletronics, 2002.

[10] Wong, D.Y., Markel, J.D., Gray, A.H., "Least squares glottal inverse filtering from the acoustic speech waveform", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 27, pp. 350-355, 1979.

[11] Smits R. , Yegnanarayana, B., "Determination of Instants of Significant Excitation in Speech Using Group Delay Function", *IEEE Transactions on Speech Audio Processing*, v. 3, pp. 325-333, 1995.

[12] Strube, H. W., "Determination of the instants of glottal closure from the speech wave", *J. Acoust. Soc. Amer.,* v. 56, n. 5, pp. 1625-1629, 1974.

[13] ChangXue MA, Kamp Y. K., Willems L. F., "Frobenius Norm Approach to Glottal Closure Detection from the Speech Signal", *IEEE Transactions on Speech and Audio Processing*, v.2, pp. 258-265, 1994.

[14] Ananthapadmanabha T.V., Yegnanarayana, B., "Epoch Extration from linear Prediction Residual for Identification of Closed Glottis Interval", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 27, n. 4, pp. 309-319, 1979.

[15] Thaís C. S., *Fonética e Fonologia do Português*, Editora Contexto, São Paulo, 2003.