REDUCING THE NATURAL CLICK EFFECT WITHIN DATABASE FOR HIGH QUALITY CORPUS-BASED SPEECH SYNTHESIS

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

LINSE – Circuits and Signal Processing Laboratory
Department of Electrical Engineering
Federal University of Santa Catarina
88040-900 – Florianópolis – SC – Brazil
{monique, seara, fernando}@linse.ufsc.br

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 verify the applicability of the proposed approach.

1. INTRODUCTION

In high quality concatenative speech synthesis systems, synthetic speech is obtained considering a large speech corpus (in the order of tens of hours). Such a corpus is previously recorded and each unit is identified (annotated) with its initial and final instant [1], [2]. At synthesis time, a search process is carried out to select the best segments through an ad hoc optimization criterion, aiming at a high quality synthetic speech. Such a speech is obtained through concatenation of the selected segments.

Nowadays, the corpus-recording phase is no longer an issue in attaining the desired speech quality. This fact is due that the recording is usually carried out in professional studios, monitored by qualified technicians, ensuring a low noise level as well as minimum distortion [3]. On the other hand, human speech – even the one produced by professional speakers – presents accidental degradations generated naturally by the human vocal tract which are more discerning in concatenative speech synthesis applications. Such degradations caused by pitch and energy changes (jitter and shimmer) as well as clicks are the main causes for harming the perceptual quality of the synthetic speech. The harms caused by clicks are due to their interaction with possible discontinuities inserted in the concatenation process and/or their association in adjacent segments.

Before proceeding, it is important to consider that clicks do not mean degradation elements for every existing language. In some African languages, such as *!Xóõ*, *!Xũ*, Nama, Zulu and Xhosa, and in an Australian language named *Damin* [4], clicks represent consonantal phonemes and carry useful language information [5]–[8].

In our case, clicks under investigation do not represent phonemes themselves. They are characterized as discontinuity-like effects existing in some speech segments. Such clicks noticed as small cracks (almost imperceptible in ordinary speech) are produced in an involuntary way, which leads one to name these natural degradations as involuntary clicks. To reduce their audibility one should process them by using one of the following techniques: suppression and interpolation [9], concealment, or attenuation.

In this paper we propose involuntary click detection and smoothing. To our knowledge this procedure has never been described in the open literature. Such an approach is applied to speech databases as an offline preprocessing stage, not increasing the computational complexity of the synthesis process per se.

Existing approaches to detect impulsive noise (clicks) in old disc recordings (caused by dust, scratches, granularity, among others) have motivated us to use similar approaches also for detecting natural clicks. Click detection in old audio recordings is based on techniques as highpass filtering, wavelet analysis, Bayesian approach, and artificial neural networks [9], [10]. In our approach, we opt for using click detection based on prediction error assessment (obtained through linear predictive analysis) over some selected frequency bands of the speech signal, technique similar to the one used in [9] for old recording click discrimination. Allowing for the particular click type (much smaller amplitude than the one of old audio recordings), the proposed technique also uses other procedures such as subband energy analysis and bandpass filtering.

This work was partially supported by the Brazilian National Research Council (CNPq), Research and Projects Financing (FINEP), and Dígitro Technology Ltd.

The current research work also proposes a click smoothing approach, which consists of masking part of the segment containing a click by using a weighting window specially made for this purpose.

2. INVOLUNTARY CLICK MODELING

As previously mentioned, involuntary clicks manifest themselves as small cracks, almost imperceptible in the recordings of speech databases used for concatenative synthesis.

The production mechanism of involuntary clicks has not yet been exploited in the literature. On the other hand, we can try to explain such a mechanism by observing the production of consonantal and emergent clicks [7], [8], [11]. The former is created when an ingressive airflow crosses a partial or total constriction produced between the tongue and an articulation place. In turn, emergent clicks are due to the volume expansion of an air cavity, which is formed between two articulatory constrictions associated with the speech coarticulation process. Then, by relaxing one articulatory constriction a click sound may be produced. In this way, we suppose that involuntary clicks may also be generated by similar mechanisms.

Fig. 1 illustrates a speech segment containing an involuntary click. Such a segment, selected among recordings that compose a speech corpus, has a click located between samples #130 and #145. This click can be visualized as a spurious signal (noise) added to an ideal speech signal (without clicks). Such pattern is always found when involuntary clicks are present in recordings.

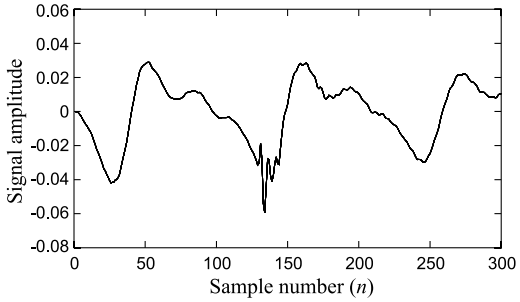

Fig. 1. Voiced speech segment containing a natural click.

Then, let us model an involuntary click as an intermittent additive noise incorporated into an ideal speech signal. Thus, a signal $y(n)$ containing involuntary clicks is expressed as

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

where $x(n)$ denotes the ideal speech signal, $i(n)$ characterizes a switching function assuming $\{0, 1\}$, which indicates absence or presence of a click, and $r(n)$ represents the click signal itself. The model proposed here is inspired in the representation of impulsive noise present in old audio recordings [9]. The objective of creating such a model consists of generating artificial clicks similar to natural ones (involuntary). These artificial clicks in turn facilitate the assessment of detection tools since we have *a priori* knowledge about the considered clicks as well as their locations.

3. CLICK DETECTION

Click detection consists of determining regions within the speech signal where involuntary clicks occur. The used approach considers that a click is evidenced in a certain subband when its energy is much larger than the one of the speech signal in the corresponding subband. Fig. 2 illustrates such a phenomenon. This figure shows the spectrogram of a segment with 3200 samples (sampling rate of 16 kHz) from the phone [ã] containing an involuntary click (indicated by an arrow in the figure) between samples #2608 and #2623. Note that we can easily identify this click in the frequency subband which extends from 2 kHz to 5 kHz. In this case its identification becomes easier, since within the referred subband the click energy surpasses the speech signal energy in the click neighborhood. Then, taking into account the ideal signal, the region under analysis has a lower energy than the one of the click.

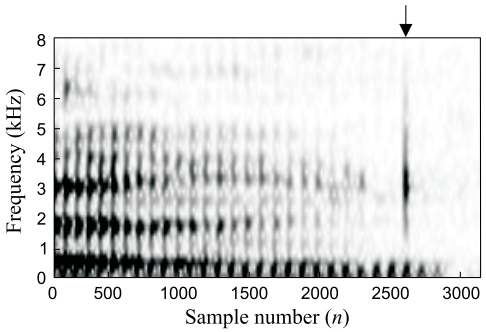

Fig. 2. Spectrogram of a speech signal with a natural click.

Fig. 3 shows the block diagram of the proposed detection technique. Note that the first stage consists of segmenting the speech signal by using a Hanning window, and classifying them as voiced/unvoiced. In the following phase, each speech segment is transformed into the frequency domain using the discrete Fourier transform (DFT). Here the DFT is obtained through a fast Fourier transform algorithm (FFT). The speech segment in the frequency domain is divided into M bands $B_1, B_2, \ldots, B_M$. Then, each of these M bands is now subdivided into N subbands SB_{ij} for $i = 1, \ldots, M$ and $j = 1, \ldots, N$. The next stage selects for each band the subband of lowest energy, providing that such energy value does not exceed a prescribed threshold. Such a selection is carried out imposing zero-values on the FFT coefficients of the unselected subbands. In addition, if the selected subband energy exceeds the fixed threshold value, a high-energy region is reached. In this case, a possible click would be masked and the detection process for the considered band is interrupted. In the following, for each selected

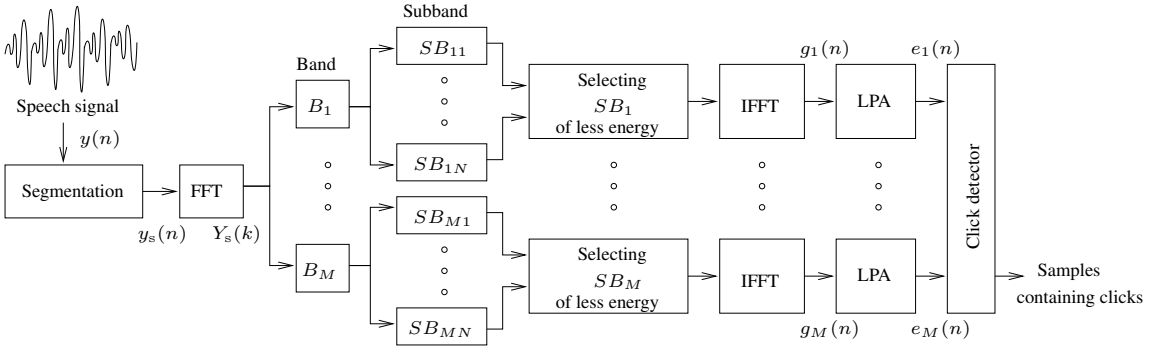

Fig. 3. Proposed scheme for detecting clicks within speech segments.

subband, an inverse discrete Fourier transform (IDFT) is carried out by means of an inverse FFT algorithm (IFFT), resulting in at most M signals. For each signal a linear predictive analysis (LPA) is achieved. Predictive analysis includes an autoregressive estimation (AR); computation of the normalized prediction error $e_i(n)$, obtained comparing the real signal $g_i(n)$ with its estimate $\widehat{g}_i(n)$, for $i = 1, \ldots, M$. If at least one of the prediction error absolute values exceeds a stipulated threshold, we conclude that a click within the analyzed segment exists. Such a task is performed by the click detector block shown in Fig. 3. To increase performance the same process is also carried out for the reversed signal.

An essential point in the detection process is the computation of the appropriated threshold values for the click detector. Then, the following procedure is applied to a segment under analysis: *i*) In each segment 1% of the largest prediction error samples is discarded. *ii*) The largest sample after (*i*) is retained. This value is multiplied by 1.5 to define the required threshold.

Fig. 4 illustrates the obtained detection threshold considering the previous procedure as well as the absolute value of the normalized prediction error for a speech signal containing a click.

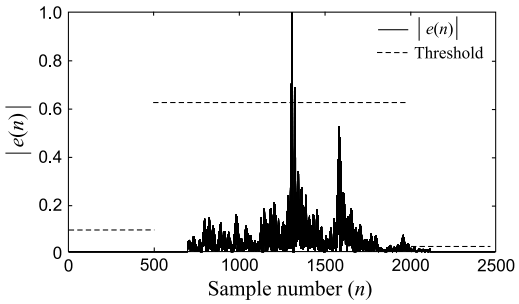

Fig. 4. Absolute value of the normalized prediction error and detection threshold.

Here, it is important to mention the existence of other procedures to obtain such a threshold value. For example in old recordings, impulsive disturbances are detected considering both an estimate of the excitation signal standard deviation [9] and the prediction error absolute value median [12]. However, such procedures have shown to be less effective than the previous approach to determine the threshold value in the present application.

We must also remark that the proposed technique is restricted to deal with involuntary clicks existing in both voiced and low energy unvoiced segments. Such a restriction does not make our approach less effective, since the perception of clicks present in high-energy unvoiced segments is attenuated by the masking effect occurring in such signals.

4. CLICK EFFECT SMOOTHING

Click smoothing is a processing stage aiming at reducing involuntary click audibility. In this paper we consider for such a task the use of a weighting window (applied over the region where a click is detected) described as follows:

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

where $w_{\rm h}(n)$ is a Hanning window with $4P + 1$ coefficients, $w'_{\rm h}(n)$, a signal with P initial and final zero-samples, having $2P + 1$ central samples corresponding to another Hanning window, $0 \leq \alpha_1 \leq 1$ and $0 \leq \alpha_2 \leq 1$ are parameters that control the weight of each $h(n)$ window. For our case, it is still considered that $\alpha_1 > \alpha_2$.

5. EXPERIMENTAL RESULTS

To evaluate the presented detection procedure we propose as a figure of merit the correct detection index (CDI) defined as

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

where $n_{\rm c}$ denotes the number of correctly detected clicks and $n_{\rm t}$, the total number of clicks.

A set of artificially generated clicks (105), with statistical distribution (amplitude, duration, and damping factor) similar to the one of natural clicks, is added to a speech signal with no click. Such a signal, sampled at 16 kHz, is one minute long. This signal is segmented using a Hanning window. Each frame is 100 ms long and with 50 ms overlap. A FFT of 4096 points is applied to each segment. A division into M bands and N subbands is then obtained. The choice of M, N, and energy threshold values is made using an exhaustive search procedure, aiming at achieving a maximum CDI value.

The performance of the proposed approach is independently assessed at voiced and unvoiced segments. For voiced signals, the values of M, N, and energy threshold which provide a maximum CDI are, respectively, equal to 4, 3, and 0.45. In this case, the CDI value obtained is 91.80%. For low energy unvoiced signals, M, N, and energy threshold which lead to a maximum CDI are, respectively, 4, 3, and 0.25. In this case, the CDI value is equal to 81.82%.

Another one minute long signal, sampled at 16 kHz, and containing natural clicks (which occur at a typical rate of 44 per minute for the considered speaker) has been used to assess the proposed detection technique allowing for the same values of M, N, and energy threshold as previously obtained. For such a signal, the total CDI value (voiced/unvoiced) is now 78.57%.

We must emphasize that the number of false detections is not considered as a parameter to assess the performance of the approach in question. This is due to the fact that the smoothing procedure applied to clicks incorrectly detected (false positives) does not harm the speech quality (considering the typical frequency of click occurrence), since the attenuation is soft and it is only applied to a reduced number of samples of the phoneme.

After the detection phase, a click smoothing stage is applied over the set of detected clicks. To evaluate the smoothing procedure we propose the use of a perceptual measure defined as

$$\text{SI} = \frac{n_\text{i}}{n_\text{c}} \times 100, \tag{4}$$

where SI denotes the smoothing index, n_i is the number of detected clicks which are transformed into inaudible, and n_c, the total number of detected clicks.

For voiced segments the smoothing procedure is carried out using $\alpha_1 = 1$, $\alpha_2 = 0.1$, and $P = 60$. For low energy unvoiced speech we consider $\alpha_1 = 1$, $\alpha_2 = 0$, and $P = 180$. After smoothing, the speech signal (with natural clicks) is assessed by two qualified listeners using a high quality headphone. Each listener has pointed out the instants where a click sound could be listened. After this, such instants are compared with the instants of click detection. The joint SI obtained in the evaluation is 93.75%, indicating a satisfactory performance.

6. CONCLUSIONS AND REMARKS

An approach for involuntary click detection and smoothing in speech signals has been presented. Experimental results have shown the applicability of the proposed approach for detecting and smoothing clicks existing in both voiced and low energy unvoiced speech segments. The smoothing procedure used for click attenuation is useful in large corpora for concatenative speech synthesis systems. Additional studies are being carried out to improve smoothing phase parameters and assess the performance of other approaches for reducing involuntary clicks. As an alternative, interpolation techniques and pruning of database phonemes containing involuntary clicks are being tested. For such, we intend to accomplish a formal perceptual evaluation (*mean opinion score* – MOS) and compare synthetic speech quality for different approaches before and after reducing clicks.

7. REFERENCES

[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*, 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 Int. Conf. Acoust., Speech, Signal Processing*, Salt Lake City, USA, May 2001, pp. 785–788.

[3] 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*, Santa Monica, USA, Sept. 2002, pp. 199–202.

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

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

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

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

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

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

[10] 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*, New York, Oct. 1995, pp. 139–142.

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

[12] P. A. A. Esquef, M. Karjalainen, and V. Välimäki, "Detection of clicks in audio signals using warped linear prediction," in *Proc. IEEE Int. Conf. on Digital Signal Processing*, vol. 2, Santorini, Greece, July 2002, pp. 1085–1088.