On the quality of speech produced by impulse driven linear systems
Verhelst, W.D.E.

Published in:

DOI:
10.1109/ICASSP.1991.150386

Published: 01/01/1991

Document Version
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the author’s version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Citation for published version (APA):
ON THE QUALITY OF SPEECH PRODUCED BY IMPULSE DRIVEN LINEAR SYSTEMS

Werner VERHELST

Eindhoven University of Technology
Institute for Perception Research
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

ABSTRACT

It is shown that a recently proposed technique for high quality waveform manipulation can be formulated as a pitch-excited vocoder. This 'waveform vocoder' produces high quality speech over a wide range of prosodic modifications, showing natural sounding speech can be produced using an impulse driven linear synthesis model.

In a pilot experiment, waveform vocoding techniques were applied on the LPC residue to investigate the relative importance of amplitude and phase in the synthesis of male and female voices. It was found that amplitude information contributed more to speech quality than phase information, and that, for male voices, amplitude information alone was sufficient to make the synthetic speech quality almost indistinguishable from that of natural speech.

1. INTRODUCTION

In the basic model of speech production, voiced speech is considered to be the output of an impulse driven linear system, as illustrated in figure 1. Here, the time-varying filter $h(n, m)$ models the combined spectral shaping effects due to the radiated glottal wave and the vocal tract response, while the pseudo-period of the impulse train represents voice pitch. (Energy can be represented by a gain factor, which can be attributed to either the source, or the filter response.)

![Figure 1: Impulse driven linear model for speech production.](image)

In what we will refer to as pitch-excited linear systems, or pitch-excited vocoders, the filter sequence is further assumed to be independent of the source. While this property can represent an advantage for low bitrate coding, it is very important for applications such as speech modification and speech synthesis, since it allows prosodic parameters such as pitch, timing, and energy to be readily manipulated.

From observations of natural speech waves, it is clear that the time-varying impulse response typically lasts longer than a pitch period, such that a deconvolution problem will have to be solved during speech analysis. From the viewpoint of the simplified production model of figure 1, the fact that there is a substantial overlap between successive impulse responses implies that the speech waveform is a time-aliased signal, and does not in itself contain sufficient information to uniquely determine $h(n, m)$. Also, practical speech processing systems contribute an amount of extra information embedded in their analysis strategies. Often part of this information can be found in the form of a model which is proposed for the filter impulse response (such as the all-pole model in the case of LPC).

While such parametric (re-)synthesis systems can produce intelligible speech, their quality lacks the naturalness and richness in acoustical detail of human speech. In the light of the above discussion, a possible cause for these problems could lie in an inappropriate solution of the deconvolution problem, or in the use of an overly stylized model for the combined responses of the glottis and the vocal tract. On the other hand, a vast research effort has been spent worldwide in order to improve on both these points (such as in homomorphic prediction [1], and ARMA modeling [2], respectively). This could raise suspicion that the actual cause of quality problems lies with the pitch-excited linear filtering model itself, in so far as it is probably the most simple production model that is conceivable, and does not take into account the complex interaction between vocal source and vocal tract that is involved in more realistic speech production [3].

2. A PITCH-EXCITED LINEAR SYSTEM PRODUCING NATURAL SOUNDED SPEECH

2.1. High Quality Speech Waveform Manipulation

As mentioned in the introduction, parametric vocoders, such as the well-known LPC vocoder, have problems capturing the acoustical complexity and richness of natural speech. One could therefore consider the alternative of working directly with speech waveforms. For applications such as speech modification or speech synthesis, this poses the important problem of devising strategies which will allow pitch and duration of the speech wave to be manipulated, while leaving timbre unaffected.

As shown in [4], pitch-synchronized complex spectrograms offer interesting possibilities for the manipulation of prosodic features of the speech wave, and different algorithms can be classified according to analysis conditions, modification strategies, and reconstruction schemes. It was reported in [4] that, in a formal experiment on the quality of diphone synthesized speech, different versions of these
so-called PSOLA algorithms obtained comparable scores amongst eachother, and proved to be far superior to standard LPC. The non-parametric vocoder which we introduce in the next subsection can be considered equivalent to what is called a 'simplified overlap-add TD-PSOLA' algorithm in the terminology of [4], and allows important prosodic modifications of speech without affecting voice quality or naturalness.

2.2. A High Quality Waveform Vocoder

Analysis strategy

From the input utterance $s(n)$, we construct a sequence of analysis pitchmarks $p_a(k)$, which contains the sample indices of the zero-crossings at the beginning of consecutive pitch periods (figure 2).

$$p_a(k)$$

---

Using this sequence of pitchmarks, we can derive an analysis procedure as illustrated in figure 3.

$$h(n,m)$$

---

The input sequence $i(n)$ is constructed as an impulse train with a pseudo-period given by the length of the input pitch period

$$i(n) = \sum_{k=-\infty}^{\infty} \delta(n - p_a(k)).$$

The time-varying filter impulse response at time instants $p_a(k)$ is obtained by a simple windowing procedure, applied to the input speech wave

$$h(n,p_a(k)) = s(n).w(n,p_a(k)),$$

where $w(n,p_a(k))$ represents an asymmetric hanning-type window, centered at $p_a(k)$

$$w(n,p_a(k)) = 0.5(1 - \cos(\pi \frac{n - p_a(k - 1)}{p_a(k) - p_a(k - 1)}))$$

$$p_a(k - 1) < n \leq p_a(k)$$

$$= 0.5(1 + \cos(\pi \frac{n - p_a(k)}{p_a(k + 1) - p_a(k)}))$$

$$p_a(k) < n < p_a(k + 1)$$

Synthesis strategy

Using the procedure described above, a speech wave can be analysed to obtain pitch information $p_a(k)$, and a sequence of synthesis filters $h(n,p_a(k))$, just like in any other pitch-excited vocoding scheme. Therefore, the synthesis strategy does not have to be different from standard approaches (the only particularity is that synthesis filters are non-parametric FIR filters, and are specified at non-uniform sampling intervals [as in variable frame-rate vocoders]). Also, in accordance with the general pitch-excited synthesis model of figs. 1 and 3, the synthesis equation is

$$\hat{s}(n) = \sum_{m=-\infty}^{\infty} i(m).h(n,m),$$

where $i(n)$ and $h(n,m)$ represent the source and filter parametertracks, which correspond to the desired synthesis operation, and which are derived in the usual vocoder fashion. For the case of pitch modification, for instance,

$$i(n) = \sum_{k=-\infty}^{\infty} \delta(n - p_a(k))$$

consists of a train of impulses spaced according to the desired pitch. The synthetic speech wave is then simply obtained as

$$\hat{s}(n) = \sum_{k=-\infty}^{\infty} h(n,p_a(k)).$$

Again as with standard vocoders, impulse responses $h(n,p_a(k))$ are obtained by interpolating between the samples $h(n,p_{a}(k))$ of the time-varying impulse response that are available from the analysis $h(n,m) = h(n,p_a(\text{argmin}(m-p_a(k))))$ in the case of zero-order interpolation.

In a similar way, speaking rate modification can be achieved by appropriate time-scaling of analysis parametertracks. Figures 4 and 5 show examples of pitch modification and speaking rate conversion, respectively.

2.3. Discussion

With the waveform vocoder described above, analysis-resynthesis trivially leads to an identity operation since analysis windows are constructed such that they add to one. Therefore, $\hat{s}(n) = \sum_{k=-\infty}^{\infty} h(n,p_{a}(k)) = s(n)$ $\sum_{k=-\infty}^{\infty} w(n,p_{a}(k)) = s[n]$. Nevertheless, the system must be rightfully called a high-quality non-parametric vocoder, as it allows modifications of prosodic parameters in the usual vocoding fashion, and maintains high quality and naturalness.

---

1Other window types can also be used, provided that they smoothly taper down to zero. (For the experiment in section 3, an asymmetric trapezoidal window was used.)
By specifying the synthesis filter as a sequence of finite duration impulse responses, the deconvolution problem was circumvented (rather than solved), and the attractiveness of vocoding systems was reconciled with the high quality of waveform processing. We therefore reach the conceptually important conclusion that natural sounding speech can be produced using an impulse driven linear synthesis model (i.e., the most simple production model that is conceivable).

Figure 4. Illustration of pitch modification.

Figure 5. Illustration of speech rate conversion.

3. A STUDY OF ACOUSTIC FEATURES FOR HIGH QUALITY VOCODERS

In practice, the waveform vocoder can be applied in speech modification and text-to-speech systems to obtain better speech quality than with current LPC-based systems. Waveform vocoding strategies, including prosodic manipulations, can also be applied on the LPC residue. As illustrated by figure 6, this property can be used to systematically investigate which acoustic features of LPC vocoders require most improvement to produce high quality speech.

In a first step, a sequence of LPC filters is obtained from a standard LPC analysis procedure, and used to compute the LPC residue \( r(n) \). Subsequently, the waveform vocoder analysis strategy can be applied on \( r(n) \), using a time-varying pitch-synchronized windowing function. The resulting sequence of FIR synthesis filter responses \( f_{cor}(n, p_k(k)) \) can then be considered to represent a time-varying correction filter: the cascade connection of \( f_{cor}(n, m) \) with the LPC synthesis filter \( f_{LPC}(n, m) \) constitutes an optimized synthesizer that will produce high quality speech over a wide range of prosodic manipulations (figure 7).

After transforming the correction filters responses \( f_{cor}(n, p_k(k)) \) to the frequency domain, using a zero padded FFT, the spectral characteristics \( F_{cor}(\omega, p_k(k)) \) can be modified in selected frequency regions. The modified spectra \( F_{cor}(\omega, p_k(k)) \) are then transformed back to the time domain to obtain a sequence of modified filter impulse responses \( f_{cor}(n, p_k(k)) \), which are used in the waveform vocoder synthesis scheme to obtain a modified residue \( r'(n) \). This modified residue is then used to drive the sequence of LPC synthesis filters, yielding the correspondingly modified speech \( s'(n) \).

Figure 6. Overview of speech analysis, spectral modification, and synthesis.

Figure 7. An optimized speech synthesizer constructed by cascade connecting a waveform synthesis filter to the LPC synthesizer.
versions was judged by twelve subjects in a paired comparison experiment. This resulted in 2880 scores \( y_{ijkm} \), which express the degree of preference for a second version of an ordered pair \((i, j)\), as stated on a 5-point scale \((-2 \ldots +2)\) by subject \(k\) for sentence \(l\) read by speaker \(m\).

\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
\text{Male} & \text{Female} \\
\hline
M1 & M2 & M3 & M4 & \text{F1} & \text{F2} & \text{F3} & \text{F4} \\
\hline
B & MB1 & MB3 & FB1 & KB1 & FK1 & FK3 & FB4 \\
K & MK4 & FK1 & FK3 & FR2 & FR4 \\
R & MR1 & MR2 & MR3 & FR2 & FR4 \\
V & MV2 & MV4 & FY1 & FY1 & FY4 \\
L & MZ1 & MZ4 & FZ2 & FZ5 & FZ5 \\
\hline
\end{array}
\]

Table 1. Utterances used in the experiment (test material codes \(B \ldots L\) are explained in table 5; shown boldfaced is the largest subset that contains balanced male/female utterances.)

<table>
<thead>
<tr>
<th>Code</th>
<th>Sentence</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Original</td>
<td>-0.66</td>
</tr>
<tr>
<td>2</td>
<td>Original A</td>
<td>-0.29</td>
</tr>
<tr>
<td>3</td>
<td>Original A</td>
<td>0.20</td>
</tr>
<tr>
<td>4</td>
<td>Original A</td>
<td>-0.66</td>
</tr>
<tr>
<td>1</td>
<td>Constant A</td>
<td>0.42</td>
</tr>
<tr>
<td>2</td>
<td>Constant A</td>
<td>-0.10</td>
</tr>
<tr>
<td>3</td>
<td>Constant A</td>
<td>0.29</td>
</tr>
<tr>
<td>4</td>
<td>Constant A</td>
<td>-0.61</td>
</tr>
</tbody>
</table>

Table 2. Version numbering as a function of amplitude \((A)\) and phase \((B)\) conditions of the modified correction filter.

In order to investigate possible differences between male and female voices, we analysed two data subsets separately: one subset pertains to 7 male utterances, and the other to the corresponding 7 female utterances as shown in table 2.

The results in figure 9 show that the overall quality loss between natural speech and its LPC counterpart was more important in the case of female voices, and that correction of amplitude information alone was sufficient to give the synthetic speech a quality which was almost indistinguishable from that of natural speech in the case of male voices.

CONCLUSION

It was shown that there exists an equivalence between a high quality waveform manipulation technique, and a pitch-excited vocoder. As this waveform vocoder generates natural sounding speech over a wide range of prosodic modifications, it was concluded that an impulse driven linear synthesis model can produce high quality speech.

It was shown that waveform vocoding techniques can be used to systematically investigate which acoustic features are perceptually important for high quality speech synthesis. In a pilot experiment, it was found that the LPC residuals amplitude information contributes more to overall speech quality than its phase information, and that, for male voices, amplitude information alone is sufficient to give the synthetic speech a quality which is almost indistinguishable from that of natural speech.

ACKNOWLEDGEMENT

I would like to thank J.H. Eggen for contributions to the pilot experiment and some of the illustrations.

REFERENCES