A Frequency Domain Approach to ARX-LF Voiced Speech Parameterization and Synthesis

Alan Ó Cinnéide, David Dorran, Mikel Gainza and Eugene Coyle

Audio Research Group, Dublin Institute of Technology
Kevin Street, Dublin 8, Republic of Ireland
alan.ocinneide@dit.ie, david.dorran@dit.ie, mikel.gainza@dit.ie, eugene.coyle@dit.ie

Abstract

The ARX-LF model interprets voiced speech as the an LF derivative glottal pulse exciting an all-pole vocal tract filter with an additional exogenous residual signal. It fully parameterizes the voice and has been shown to be useful for voice modification. Because time domain methods to determine the ARX-LF parameters from speech are very sensitive to the time placement of the analysis frame and not robust to phase distortion from e.g. recording equipment, a magnitude-only spectral approach to ARX-LF parameterization was recently developed.

This paper describes extensions to this frequency domain approach to obtain continuous robust ARX-LF parameters for voiced speech segments. A listening test of 50 participants comparing synthetic speech produced by this method with a time-domain ARX-LF parameterization approach under real and simulated recording conditions was conducted and it was found that the frequency domain approach was generally preferred.

Index Terms: ARX-LF model, voice coding, speech synthesis

1. Introduction

Physical models of the speech production system offer a straight-forward means of producing natural and realistic voices, and efforts based on these models have found some success. Many recent approaches [1, 2, 3] interpret the voiced speech signal following the ARX-LF model [4], i.e. as the output of an Auto-Regressive (AR) vocal tract filter excited by the sum of an exogenous (X) signal and a pulse of the Liljencrants-Fant (LF) model of derivative glottal flow [5].

In order to obtain the vocal tract filter and glottal flow source parameters, existing ARX-LF approaches invariably involve sophisticated glottal inverse filtering techniques [1, 4]. These techniques operate in the time-domain where they require accurate estimation of the placement in time of each speech pulse. This requirement necessitates the use of phase linear equipment during recording of the speech signal [6]. It has been claimed that most electro-acoustic equipment will introduce phase distortion [7], thus a magnitude-only spectral approach to ARX-LF parameterization would prove more applicable to a wider range of recorded speech.

Recently, a frequency domain glottal inverse filtering algorithm was developed that is robust to the time-placement of the analysis frame and can be used to obtain ARX-LF model parameters from speech [8]. However, this technique required an estimate of the glottal formant to limit a sampled subspace search of LF model parameters. In this work, the technique is extended to a full subspace search, and thus not requiring any prior information about the glottal source. Using this parameterization technique and a dynamic programming algorithm to obtain continuous parameters across multiple frames, a method for the robust determination of the ARX-LF parameters of voiced speech segments is proposed.

The goal of this paper is to describe this new frequency domain ARX-LF based parameterization algorithm and compare its performance with an existing time domain approach. The work is structured as follows. First, a description of the ARX-LF model is given. The new frequency domain ARX-LF parameter determination analysis is outlined and an overlap-add synthesis method described. After briefly outlining a time-domain method of ARX-LF parameterization, experiments comparing the synthetic speech from the frequency domain method of ARX-LF parameterization and the time domain approach are then described. The results are discussed and finally the work is concluded.

2. Models of Voiced Speech

Expressed in the Z-domain, the acoustic theory of speech production [9] views speech \( S(z) \) as the multiplication of glottal flow signal \( G(z) \) with a vocal tract filter \( V(z) \) which is then radiated at the lips \( L(z) \). As lip radiation \( L(z) \) is usually modeled as a differentiating filter and the relationship between the speech chain components assumed linear, it is often combined with the glottal flow \( G(z) \) to form the glottal flow derivative \( G' \)(z) = G(z)L(z). The speech production process can then be expressed:

\[
S(z) = G'(z)V(z)
\]  

(1)

The ARX-LF model of speech is a specific interpretation of the source-filter model as described in Equation 1. In this case, the glottal derivative signal \( G'(z) \) is assumed to be comprised of the LF model and an exogenous residue signal, containing any aspiration noise and all periodic components not captured by the LF model. This model may be expressed in the Z-domain as:

\[
S_{\text{ARX-LF}}(z) = (U_{\text{LF}}(z) + E(z))V(z)
\]  

(2)

where \( U_{\text{LF}}(z) \) represents the LF model contribution and \( E(z) \) represents the exogenous signal.

The LF model represents the general flow shape of the glottal flow derivative over one glottal cycle and whose shape can be uniquely described with four parameters. This prevalent model has previously been shown to be useful for speech synthesis and modification [1, 2, 10], not least because of its physiologically significant parameter set which corresponds to the vibratory behavior of the vocal folds.
3. Frequency Domain ARX-LF Model Parameterization

The ARX-LF parameterization method presented in [8] was shown to be robust to the position of the analysis frame and robust to phase distortion of a simulated recording system. The method is briefly summarized in this section, in addition to the extensions necessary to parameterize speech segments.

Using an estimate of the fundamental frequency, pitch synchronous analysis points along a voiced speech segment are calculated. Extracted frames are centered on these points and $2T_0 + 1$ samples in length, where $T_0$ is the local pitch period. In [8], spectral information was determined using the discrete Fourier transform; the present algorithm obtains more accurate spectral samples using a least-squares harmonic analysis [11]. This technique also permits smaller frame sizes. The estimated harmonic magnitudes are then used to estimate an all-pole vocal tract filter for each LF model pulse in a codebook, using Discrete All-Pole Modeling [12]. The codebook has been expanded from the one used in [8] to contain a sampled subset of all possible LF model parameter configurations, rather than only the return phase parameter. A spectral representation of the glottal inverse filtering method is given in Figure 1.

In order to obtain smooth parameterizations through an entire speech segment, a dynamic programming method is utilized. Following the initial analysis step described in the previous paragraph, each frame has associated with it, for each LF model parameter configuration within the codebook:

- an all-pole filter filter representing the vocal tract,
- an estimate of the scale factor of the given LF model parameter configuration, and
- an Itakura-Saito error measure, quantifying the goodness-of-fit.

Using distance functions similar to [1] and [2], a dynamic programming algorithm is used to determine the sequence of LF model pulses and vocal tract filters yielding the lowest overall error. The LF pulse parameter are refined using a magnitude spectrum LF model fitting technique. Finally, both the estimated vocal tract filter and LF model parameters are smoothed here.

One analysis point is placed in frames $T_0$ in length such that the glottal inverse filtering method is given in Figure 1.

5. Experiment

Within this section is the experiment which compares a state-of-the-art ARX-LF time domain parameterization technique to the previously described frequency domain approach. First, the time domain method is briefly outlined, followed by the experimental setup and results.

5.1. Time Domain ARX-LF Model Parameterization

The time domain ARX-LF parameterization method used for comparison was the one described in [1] and extended by [13] and [14]. This method has found application in synthesis of singing voice with quality control [1], speaker conversion [13] and analysis/synthesis of vowel segments [14].

The method uses a convex optimization approach to jointly estimate the all-pole vocal tract filter and parameters of the KLGLOTT88 model [15] by minimizing the squared error of the residual signal in the time domain for a codebook of KL-GLOTT88 parameters. Adaptive pre-emphasis was used to increase robustness [13, 14]. A dynamic programming algorithm is applied to the determined vocal tract and KLGLOTT88 parameters to find the lowest overall error [1]. Once the vocal
tract is estimated, each pulse of the estimated glottal derivative flow signal is parameterized by the LF model using initial parameters mapped from the KLGLOTT88 parameters and a constrained optimization algorithm [14].

The glottal closing instant information required for the time domain algorithm was obtained from the DYPSA algorithm [16]. As this information is particularly crucial for the time-domain algorithm, the glottal closing instants are refined following a first pass of the algorithm by choosing the glottal derivative flow signal minimum close the initial estimation. Pitch was estimated by the SWIPE® algorithm [17].

5.2. Preference Test

Three experiments were performed to test the frequency domain ARX-LF analysis method against the time domain based approach described above. The first experiment compared the performance of both parameterization algorithms using signals recorded using phase linear equipment. The second experiment focused on the phase robustness of each technique and accordingly the test signals were convolved with the impulse response of a non-linear phase recording device, taken from the description given in [18]. These two experiments used the same test data, the first five sentences from the CMU-Arctic database for two speakers, one male (bdl) and one female (slt). A third experiment was performed on speech obtained using an inexpensive headset microphone and a laptop computer which does not exhibit linear phase characteristics. Five sentences from a male and female speaker were recorded.

The experiments were carried out as follows. Both ARX-LF parameterization methods were used to analyze the voiced speech segments of the signals. The parameters obtained were then used to synthesize speech segments. As the methods described in [1, 13, 14] do not attempt to model the speech residual, the first method of synthesis described in Section 4 is implemented. The ARX-LF model parameterizes only the voiced parts of speech; consequently, unvoiced speech segments are simply added into the output signal.

The signals from the time-domain and frequency domain approaches were then compared with each other using a listening test. The participants, of which there were 50, were asked to listen to both versions of the sentence and to give a score according to their general preference. The preference scores ranged from −3 to +3 corresponding to a strong preference for either the time-domain or frequency domain ARX-LF method. A 0 score denoted no preference for either technique.

The ARX-LF parameterization methods are fully automatic and no further processing was performed on the signals other than described. Due to some errors in the deconvolution procedure, disagreeable discontinuity-type artifacts were generated from the time-domain ARX-LF parameters of 4 sentences of the male speaker bdl, two from both the linear phase and nonlinear phase experiments. These utterances were removed from the dataset. The mean preferences of the remaining signals and their 95% confidence intervals are presented within Figure 3.

6. Discussion

The data from the listening tests shows a tendency of the participants to prefer the speech synthesized with parameters of the frequency domain ARX-LF model technique over the time domain ARX-LF model method. The preference is consistent in all recording conditions scenarios and both sexes.

The only case where there is no clear preference for either technique is that of phase linearly recorded male speech. On the other hand, the data shows that the synthetic speech of female speakers analyzed by the frequency domain approach was particularly preferred over the time domain method under phase linear conditions. This may be due to the difficulty in obtaining accurate glottal closing instant for these voices, which is a critical for the accuracy of time domain ARX-LF parameterization approaches.

Unsurprisingly, the frequency domain ARX-LF model approach was superior under nonlinear phase conditions, i.e. the scenarios where the signals were convolved with the impulse response from a nonlinear phase recording system and recorded with inexpensive equipment. By ignoring phase information, the frequency domain approach to ARX-LF parameterization has been shown to be robust in non-ideal phase conditions. On the other hand, time domain approaches are not very robust to the time placement of the analysis frame. While efforts were made to mitigate this error in this experiment, this was almost certainly a source of some audible artifacts.

As the time domain shape of the signal can become significantly distorted depending on the phase response of the recording system, the concept of the glottal closure instant and time domain processing of the signal generally becomes less clear. The phase response described in [18] is most distorted at very low frequencies (< 45Hz) and subsequently approaches linearity as frequency increases (see Figure 4). In this case, the system would introduce more time domain changes to low frequency signal components. Lower pitch male voices are therefore more likely to be affected by this kind of distortion, and so it is then unsurprising that the frequency domain ARX-LF parameterization method preference for male voices is significantly higher in the nonlinear case than the linear one. Accordingly, the increased preference for the female voice in this case is less.

It is not the intention of this experiment to draw conclusions about the quality of the speech produced by the ARX-LF parameters, only to compare the general preferences of listeners to speech synthesized produced by the model. However, it is useful to mention the perceived differences with the original signals. While both methods retain much of the speakers quality, it was noted for many signal that the frequency domain approach...
to ARX-LF modeling produced speech that was perceived to be “brighter” than the time domain method, and contained more high frequency energy. This may be due to the least-square energy criterion of the time-domain methods tends to determine the higher energy, low frequency formants of the vocal tract at the expense of the low energy, high frequency ones.

Additionally, it is also important to note that the application of the ARX-LF model to speech varies from speaker to speaker. While it was felt not to be the case with the signals used in this experiment, the speech signals synthesized using ARX-LF parameters alone may differ significantly from the original signal.

7. Conclusion

This paper described extensions to an existing frequency domain algorithm for obtaining continuous robust parameters for the ARX-LF model of speech. An experiment was undertaken to compare the synthetic speech generated using the parameters obtained by this method with speech utterances synthesized using parameters obtained by a time domain ARX-LF parameterization technique. The experiment worked with signals that were both recorded using linear phase equipment, convolved with a impulse response of a non-linear phase system, and also with signals recorded using generic non-linear phase audio recording equipment.

The experiment found that the frequency domain approach to ARX-LF parameterization is preferred over other techniques in all recording scenarios for all voices, though the preference was slight for the case with male speakers under phase linear conditions.

Future work continues toward the development of a robust system of ARX-LF model based voiced speech modification.

8. Acknowledgments

The authors extend their thanks to the anonymous reviewers for their helpful comments which improved this work.

9. References


