EVALUATION OF A GLOTTAL ARMA MODELLING SCHEME

A.P. Lobo and W.A. Ainsworth

Dept. of Communication and Neuroscience,
University of Keele, Keele,
Staffordshire ST5 5BG U.K.

It is well known that glottal pulse shapes differ from person to person and also for the same person for utterances in different contexts. Since our objective was to determine the effect of the glottal waveshape on synthetic speech we have implemented an analysis by synthesis system whose characteristics can be controlled through a set of parameters to realize any desired voice source characteristics. A natural continuous voiced utterance was analyzed and it is shown that the source and the vocal tract parameters are well estimated by combining a 6 parameter source model with ARMA analysis. The better performance of this system over two other methods i.e. closed phase LPC covariance analysis [5] and robust LPC analysis [2] (for the case of non gaussian excitations) is demonstrated in terms of formant tracking ability and efficiency of resynthesis.

INTRODUCTION

A major inconsistency in conventional LP speech analysis methods is that while white noise is assumed in the analysis, voiced speech is periodic. Consequently impulse or pulse like excitation has to be used in synthesis. This in itself is insufficient to correctly estimate the formant frequencies and bandwidths of the vocal tract since the estimated spectral envelope will describe not only the vocal tract transfer function but will also contain spectral characteristics of the glottal pulse. To overcome this limitation a glottal ARMA analysis scheme is proposed in which we have explicitly modelled a voice source in combination with pole-zero modelling of the vocal tract.

DESCRIPTION OF THE ANALYSIS-SYNTHESIS SYSTEM

1. THEORY AND GLOTTAL MODEL PROPOSAL

The production model, fig.1, used assumes that speech is the output of a time varying vocal tract system excited by a time varying glottal excitation.

In order to separate the essential features of the glottal source directly from the speech waveform inverse filtering has to be employed in one form or the other. The adjustment of the inverse filter is however often based on subjective qualitative judgements regarding the shape of the glottal wave. In order to establish objective quantitative criteria, modelling of the glottal wave has to be introduced. Thus a parametric representation capable of describing the most essential characteristics of the glottal waveform are of interest for the general study and classification of various modes of glottal excitation and for high quality speech synthesis.

A model, fig.2, is proposed for the glottal flow pulse whose derivative is composed of sinusoidal as well as polynomial segments and is described by the following 6 parameters viz. opening duration(R), closing duration(F), interval from closure to max. negative flow(D), peak positive amplitude(EI), slope at closure(EE) and slope immediately after closure(C) [1, 4]. The equations describing the model are written below: (refer fig.2)
where \( b = \frac{CD}{D-3(T-W)} \)

2. ESTIMATION OF THE VOICE SOURCE AND VOCAL TRACT PARAMETERS

The AR and MA parameters are simultaneously estimated in a linear procedure, using the described parameters of the source waveform as input [1].

The overall minimization procedure is described as follows.

1) generation of the voice source signal,
2) estimation of the vocal tract transfer function using the generated source signal and the speech signal,
3) evaluation of the prediction error.

The sequence (1)-(3) is iterated as the glottal parameter space is searched for a combination that gives the best description of the input signal in terms of the minimum prediction error. A schematic of the estimation procedure is shown in fig. 3.

![Diagram](image)

The nonlinear optimization is performed by a Levenberg-Marquardt algorithm [6, 7] (Levenberg 1944, Marquardt 1963) which essentially examines three dimensions at a time (3 time and 3 amplitude dimensions). In the solution of linear system of equations singular value decomposition is used to avoid the occurrence of singular matrices in the calculations. The analysis is carried out pitch synchronously with the frame length equal to a period. Both the speech and voice source signals are pre-emphasized to achieve pre-whitening of the inputs to the linear estimation procedure. The pitch contour is obtained automatically from a synchronously recorded and appropriately delayed electroglottogram to account for the propagation delay from the glottis to the vocal tract. The glottogram also serves to identify initial estimates of the events in a glottal cycle. The speech signal is also processed through a noise canceleing stage [3] to eliminate low frequency breath and ambient noise and followed by a phase compensation stage (to eliminate phase distortion introduced by the recording equipment) by filtering the time reversed speech signal through an all pass filter.

EXPERIMENTS ON COMPARISON OF DIFFERENT ANALYSIS METHODS

The above method was used along with two other methods viz. closed phase LPC covariance analysis and robust LPC analysis hereafter referred to as CLPM and RBLPM respectively, to analyse a continuous voiced utterance by the author. The utterance was of 2 seconds duration and consisted of the sentence "may we all learn a yellow lion roar". It was recorded using a REVOX 3400 dynamic microphone and sampled directly at 50 kHz into a MASSCOMP 5520 computer system equipped with an EF-12M A/D converter with 12 bit resolution. The utterance was then low pass filtered to 5 kHz by a linear phase FIR filter and then downsampled to 10 kHz.

A quantity called the resynthesis signal to error ratio was calculated for each period of the resynthesized speech signal as a measure of the efficiency of resynthesis. It is the ratio in decibels of the energy in the original speech to the resynthesis error resulting from resynthesizing the speech wave. Thus, higher the value of this ratio, better the quality of the resynthesized speech.

In the case of GARMA, two cases were studied:

1) The pole and zero orders were fixed at 12 and 8 respectively.
2) The pole and zero orders were fixed initially at 12 and 8 respectively. If the resynthesis signal to error ratio fell below a threshold value of 15 dB then the pole and zero orders were allowed to vary between 12 and 5 and between 8 and 0 respectively with the zero order range being used up first. If the threshold was not crossed during this procedure then speech was resynthesized using the pole-zero order combination which yielded the max. value of the resynthesis signal to error ratio for that period.

For each of the above cases the frame length was set equal to the pitch period. The minimization procedure was iterated 7 times. The frame position was shifted +/- 3 points from the position identified by the events of a glottogram cycle and at each instant the analysis was repeated to find the optimum voice source pulse placement.

In the case of CLPM, for each frame of analysis the optimum framemlength between 20 and 28 and filter order between 8 and 12.
was determined depending on the combination which gave the minimum value of the normalized total squared prediction error. The gain term in each period was calculated as the r.m.s. value of the speech segment covered by that period. The analysis was repeated +/- 10 points from the position identified by the glottogram events in a cycle to find the position where the normalized total squared error was minimum.

In the case of RBLPM the frame length was set at 128 sample points and an efficiency tuning constant of 1.5 was used to evaluate Huber's psi function. Initial estimates of the AR parameters were supplied by the CLPM method.

Fig. 4(a,b,c,d) below show the formant/bandwidth contours obtained in each of the analysis schemes. A cross on a vertical cross section indicates formant frequency values obtained for a glottal period. Vertical lines centered on the crosses represent estimated values of formant bandwidths. Isolated crosses represent formants estimated with bandwidths exceeding 700 Hz.

It can be seen that the GARMA analysis with a fixed pole-zero order performs best in terms of the formant/bandwidth tracking capability and smoothness of the contours, fig.4(a), followed by CLPM, fig.4(c) and RBLPM, fig.4(d) methods. Although the formant structure is smoother in the case of the GARMA scheme with a variable pole-zero order, fig.4(b), than the CLPM or the RBLPM schemes, there are gaps in the formant tracks which is probably due to the fact that high values of the resynthesis signal to error ratio were obtained at lower pole-zero orders thus reducing the number of formants estimated.

Fig. 5 shows the tracks of the resynthesis signal to error ratio calculated for each of the above cases. The tracks obtained for the GARMA (fixed pole-zero orders), GARMA
CONCLUSION

The present study has shown that a GARMA analysis by synthesis scheme can be employed to obtain an improved estimation of the vocal tract transfer function with a parametric representation of the voice source signal. Analysis experiments on natural speech confirmed the validity of the method and indicated improved performance over both CLPM and RBLPM in terms of better formant tracking ability and contour smoothness and smaller resynthesis error. A GARMA analysis scheme with a fixed pole-zero order is preferable to one with a variable pole-zero order.

REFERENCES