TR-IT-0230

Voice quality modification using periodic-aperiodic decomposition and spectral processing of the voice source signal

Christophe d'Alessandro

July 24, 1997

ABSTRACT
Voice quality is currently a key issue in speech synthesis. On the one hand, the lack of realistic intra-speaker voice quality variation results in poor naturalness in synthesis methods using small corpora and signal processing (e.g diphone synthesis). On the other hand, voice quality mismatches is one of the main source of concern for methods based on large corpora and labelling (e.g. word or subword units concatenation systems,like CHATR). A new method for voice quality modification is designed. It takes advantage of a spectral theory for voice source signal representation. An algorithm based on periodic-aperiodic decomposition and spectral processing (using the short-term Fourier transform) is described. The use of adaptive inverse filtering in this framework is also discussed. Applications of this algorithm may include: pre-processing of speech corpora, modification of voice quality parameters together with intonation in synthesis, voice transformation. Some experiments were perfomed, showing convincing voice quality modifications for various speakers.

Contents

Chapter 1

Introduction

Dealing with voice quality is currently a key issue in speech analysis and speech synthesis. As speech processing technology progress, more emphasis and more research effort is put on the extra-linguistic aspects of speech.

Although most of the state-of-the art text-to-speech (TTS) synthesis sytems are highly intelligible, they still are felt unatural and machine-produced. Many fundamental problems are still to be solved, in differents areas such as:

1. text analysis (e.g. grapheme to phoneme conversion, text chuncking)
2. text understanding;
3. intonation modelling (fundamental frequency and duration);
4. voice quality modelling (voice quality, vocal effort);
5. modelling interactions between intonation, voice quality and segmental aspects;
6. signal processing for transparent quality parametric modification of speech segments.

In some speech synthesis systems, e.g. voice response systems for ticket reservation, text analysis and meaning is given. Thus grapheme-to-phoneme conversion and prosodic targets are given, and only a pre-recorded limited vocabulary is used. Even in this simple and "ideal" situation, it is not always possible to synthesize perfect human sounding speech. Differences in voice source, for instance differences in vocal effort, are often perceived by listeners as synthesis or concatenation errors, because of the change in quality across segments. Concatenation errors are often noticeable, in such systems, particularly when the vocabulary required by the application is medium-sized, or large (1000 or 10000 segments). The situation is even worse in TTS. Thus, improving signal processing for concatenation synthesis is needed even for systems using massive data bases.

On the one hand, the lack of realistic intra-speaker voice quality variation results in poor naturalness in synthesis methods using small corpora and signal processing (e.g diphone synthesis). On the other hand, voice quality mismatches is one of the main source of failure for methods based on large corpora and labelling (e.g. word or subword units concatenation systems,like CHATR).

Thus, improving signal processing for parametric concatenation synthesis is needed. The signal processing methods should fulfill constraints:

1. transparency: no audible degradation should be introduced, or (weaker but more realistic constraint) no audible degradation should be introduced when parametric variation is small. It must be noted that even well-spread method like PSOLA or MBROLA pitch and duration modification method do not fulfill the weaker form of the transparency constraint.
2. significance: the parameters of the signal processing method should be meaningfull for speech production (and perception). Ideally the method should be able to deal with parameters like "vocal effort", "pitch" (i.e. the perceived correlates of fundamental frequency *and* vocal quality modifications when one speaks higher or lower). A weaker constraint is to deal with meaningfull but lower level parameters (e.g. fundamental frequency, periodic/aperiodic ratio in the voice source, open quotient of the glottal signal etc.).
3. robustness: the method should not be signal or data-base dependant, and should be robust against recording condition and (small amounts of) additive noise. In addition it should be simple and fast (!)

The aim of this study is to design a new parametric modification system for speech signal, which is able to modify low-level aspects of voice quality. The parameters that are envisaged are voice source parameters, and possibly vocal tract parameters. The speech material considered is recorded speech sample, without special requirement on recording conditions and voice types. Thus constraints (1) and (3), and the weak form of constraints (2) are aimed at.

As was stated earlier, it would be better to deal directly with high level parameters for voice quality description (to control the synthesizer with commands such as "speak louder" or "speak softer", "take a happy or a sad voice"). This is not yet possible as there is relatively scarse knowledge on covariation of prosodic (here with the meaning of intonation) and voice quality parameters, and on vocal effort, emotional and speaking style effects on speech signal. However we believe that it will be possible to design higher level rules to drive low-level parameter modifications. For some situations, we shall show that it is quite straightforward (for example global lowering vocal effort for a sentence), although it might be rather intricate in other situations.

Contrary to most of the recent works on source modelling, we prefered spectral processing. It must be emphasized that spectral processing is mathematically equivalent to time-domain processing, only if complex spectra (and signal) are considered. In this case, time and frequency domains are dual throught the Fourier transform. But spectral magnitudes and spectral phases do not play the same role (althought they are merged in time domain). One can take advantage of this separation in spectral processing because:

- spectral processing does not require the use of calibrated measurement equipment for speech recording. For example phase distortion is acceptable for spectral processing, although it is a well-known source of problems for time-domain processing, because it can change a lot the signal waveform. Most of the electro-acoustic equipments (studio microphones, non anechoic recording rooms, tape recorders ...) introduce phase distortion.

- spectral parameters have been more closely linked to the perceptual features of voice quality than time-domain parameters (see below);
- we shall show that one can design simple methods (both conceptually and in terms of processing) for parametric modifications.

The main spectral parameters found in the litterature [10] [18] for synthesizing voices with different qualities are:

1. spectral tilt (a good definition of which is still lacking);
2. amplitude of the first few harmonics;
3. increase in the first formant bandwidth;
4. noise in the voice source.

In contrast, the parameters generally used for glottal signal modelling are defined in time domain:

1. amplitude of voicing;
2. open quotient;
3. fundamental period;
4. return phase;
5. amplitude of additive noise

Linking these two sets of parameters is therefore a key point for studying voice quality as a function of voice source parameters.

In any case, the first step of the analysis/synthesis process must be decomposition of the periodic and aperiodic components of the source. For this we used an algorithm proposed in [7] (see below). For describing voice quality, the spectral parameters of the periodic component are generally measured on the magnitude spectrum. Furthermore, they are mostly used for speech analysis because no exact formulas are available for linking these parameters with time-domain glottal flow models used for synthesis. Therefore a spectral model of the periodic glottal flow spectrum is needed, because the glottal flow models that have been proposed [12] [10] but are defined in time domain. In a previous work, we made an analytic spectral study of these models, to link their parameters with spectral parameters [9]. We also studied estimation of voice source parameters in the spectral domain using this analytic study.

Following this work, it seemed possible to modify voice quality by processing the speech magnitude spectrum, using linear filtering schemes. The scope of my visit at ATR was to work on such methods, to test them, and to envisage their integration within CHATR.

In section 2, the signal model and analytic formulas for the spectrum of glottal flow models are reviewed, and a theory for spectral modification is developped. Section 3 presents the algorithms used for modification of voice quality (adaptive inverse filtering, periodic-aperiodic decomposition, periodic component spectral modification, adaptive filtering). Section 4 gives some experimental results. Section 5 concludes, and discuss application of this work to CHATR.

Chapter 2

Theoretical framework

2.1 Signal model

Linear acoustic theory describes speech production in terms of a source/filter model. This model is made of a volume velocity source, which represents the glottal signal, a filter, associated to the vocal tract, and a radiation component, which relates the volume velocity at the lips to the radiated pressure in the far acoustic field . This decomposition is acceptable for phonetics, wich describes speech in analog terms, "phonation" standing for "source", and "articulation" standing for "filter". From the point of view of physics, this model is only an approximation, whose main advantage is simplicity. It is considered valid for frequencies below 4 to 5 kHz, where the assumption of plane wave propagation in the vocal tract is acceptable.

This acoustic model can be written directly in terms of linear systems in the domain of signal processing, as far as source/filter interaction can be neglected. As a matter of fact it is still possible to account for some source/filter interaction effetcs in the source/filter model. For instance the effect of glottal leakage can be simulated by increasing the bandwidth of the 1rst formant in the filter, together with modification of the source parameters. In its simplest form, the source filter model can be written as:

$$
\begin{aligned}
s(t) &= e(t) * v(t) * l(t) \\
\\
S(\omega) &= \mid S(\omega) \mid e^{j\theta(\omega)} \\
&= E(\omega) \times V(\omega) \times L(\omega)
\end{aligned}
$$

where $s(t)$ is the speech signal, $v(t)$ is the vocal tract impulse response, $e(t)$ is the vocal excitation source, $l(t)$ is the impulse response of the sound radiation component, and where $S(\omega)$, $V(\omega)$, $E(\omega)$, $L(\omega)$ are the Fourier transforms of $s(t)$, $v(t)$, $e(t)$, $l(t)$ respectively.

This equation suggests that spectral processing should be easier than time-domain processing. The source component $e(t), E(\omega)$ is a compound signal, which can be represented with the sum of a quasi-periodic component (described by its fundamental frequency and its waveform) and a noise component:

$$\begin{aligned}
s(t) &= [p(t) + r(t)] * v(t) * l(t) \\
&= [\sum_{i=-\infty}^{+\infty} \delta(t - it0) * u_g(t) + r(t)] * v(t) * l(t) \\
S(\omega) &= [P(\omega) + R(\omega)] \times V(\omega) \times L(\omega) \\
&= [(\sum_{i=-\infty}^{+\infty} \delta(\omega - if0)) \mid U_g(\omega) \mid e^{j\theta_{ug}(\omega)} + \mid R(\omega) \mid e^{j\theta_r(\omega)}] \\
&\quad \times \mid V(\omega) \mid e^{j\theta_v(\omega)} \times \mid L(\omega) \mid e^{j\theta_l(\omega)}
\end{aligned}$$

where $p(t)$ is the quasi-periodic component of excitation, $u_g(t)$ is the glottal flow signal, $t0$ is the fundamental period, $r(t)$ is the noise component of glottal excitation, δ is Dirac distribution , $P(\omega)$, $R(\omega)$, $U_g(\omega)$, are the Fourier transforms of $p(t)$, $r(t)$, $u_g(t)$, respectively, and where $f0 = 1/t0$ is the fundamental frequency of voicing.

As far as intra-speaker voice quality is concerned, the most important component is the source component, which is described by r, u_g, and $f0$. This means that modifying this component will change voice quality but not voice personality, although changing the filter component will alter voice personality, but will preserve voice quality. Again, this is only an approximation, and refined methods would likely modify both components for achieving realistic modifications of eiter voice quality or voice personality. In the next section, we shall review time-domain and spectral-domain models of the source signal.

2.2 Frequency domain representation of glottal flow models

In this section we review the work done at LIMSI on spectral representation and modelling of the source component. The LF model [12] and the KLGLOTT88 model [10] are considered. Special attention is paid to the aperiodic component of the source, which is generally underestimated in glottal flow models, but which is crucial for some types of voice quality (soft voices, breathy voice, harsh voices).

2.2.1 Spectrum of the KLGLOTT88 model

This model is used in the Klatt synthesizer, and in the "MIT tradition" of voice source studies. Motivations for this model were pragmatic, with formant synthesis applications in mind. It is a composite model which contain basically three components: a noise component, a periodic glottal waveform U_g^k, which is passed through a spectral tilt filter. The periodic component of this model is characterized by four parameters: the fundamental frequency $f0$, the amplitude of voicing AV, the open quotient O_q, and the frequency of a spectral tilt filter TL. The periodic and aperiodic components are added. The equation of this model are:

$$U_g^k(t) = at^2 - bt^3$$

with

$$\begin{aligned} a &= (27AV)/(4T_0O_q^2) \\ b &= (27AV)/(4T_0^2O_q^3) \end{aligned}$$

After some tedious calculation one can show [9] that the spectrum of $u_g^k(t)$ is (with $\nu = 2\pi/\omega$):

$$\tilde{U}_g^k(\nu) = \frac{27jAV}{2O_q(2\pi\nu)^2}[\frac{j\exp(-j2\pi\nu O_qT_0)}{2} + \frac{1+2\exp(-j2\pi\nu O_qT_0)}{2\pi\nu O_qT_0} + 3j\frac{1-\exp(-j2\pi\nu O_qT_0)}{(2\pi\nu O_qT_0)^2}]$$

2.2.2 Spectrum of the LF model

The LF model was proposed and is used in the "KTH tradition" of voice source analysis (see e.g. [12] [13] [11]). Motivation for this model were more theoretical than pragmatic, although the model is eventually used formant synthesis. The model considers only the periodic part U_g^f of the voice source (which is not realistic in real voices). It is a purely time-domain model, described by several altenative (and equally intricated) set of five parameters. The five parameters commonly used to describe the LF-model are T_0, E_e, R_g, R_k, R_a. According to the analysis in [9], the definition and effect of these parameters are:

1. T_0 is the fundamental period, and will only change the harmonic frequencies.

2. E_e is the maximum flow declination rate, and will only change the overall harmonic amplitudes.

3. R_g is the ratio of T_0 over twice the peak flow time T_e . It behaves much like the open quotient O_q. The spectral effect of an increased R_g is to expand the frequency scale, resulting in shifting energy from low frequency harmonics to medium frequency harmonics.

4. R_k is the inverse of the speed quotient : $R_k = (T_e - T_p)/T_p$. It will change the waveform skewness, and will essentially affect the first harmonics amplitude.

5. R_a mesures the duration of the return phase : $R_a = T_a/T_0$. It will change the spectral tilt adding a $-6dB/oct$ above a frequency which depends on R_a, R_g and R_k, and then will essentially affect high order harmonics amplitude.

The open quotient is related to both R_g and R_k :

$$O_q = (1 + R_k)/(2R_g)$$

A given set of parameters does not guaranty that a plausible speech waveform is produced: the parameters must satisfy their theoretical ranges : $E_e > 0$, $T_0 > 0$, $R_g > 0.5$, $1 > R_k > 0$, $R_a > 0$. But they must also verify the following constraints:

1. $R_k < 2R_g - 1$, which guaranties that the closing time is within the period $t0$;
2. $R_a < 1 - (1 + R_k)/(2R_g)$, which guaranties that the return phase is a decreasing exponential;
3. if $R_k > 0.5$, the negative maximum of the flow derivative is no longer E_e. Thus, to keep the meaning of E_e as the maximum flow declination rate, one must force $R_k < 0.5$.
4. the condition of zero net gain of flow during a fundamental period implies area balance in the flow derivative, and thus implies an implicit equation between the model parameters.

After some tedious calculation, one can show that the derivative of the spectrum of the LF-model is:

$$\begin{aligned}
\tilde{U}_g^{f'}(\nu) = E_0 \frac{1}{(a - j2\pi\nu)^2 + \omega_g^2} \times [\omega_g + \\
\exp((a - j2\pi\nu)T_e)((a - j2\pi\nu)\sin(\omega_g T_e) - \omega_g \cos(\omega_g T_e))] \\
+ E_e \frac{\exp(-j2\pi\nu T_e)}{\epsilon T_a j2\pi\nu(\epsilon + j2\pi\nu)} \times \\
[\epsilon(1 - \epsilon T_a)(1 - \exp(-j2\pi\nu(T_0 - T_e)) - \epsilon T_a j2\pi\nu]
\end{aligned}$$

where the variables E_0, ω_g, T_e, T_a, and ϵ are functions of the model parameters, and where a is obtained solving an implicit equation.

2.2.3 Modelling the spectral correlates of the open quotient

For the LF model as well as for the KLGLOTT88 model, the analytic expressions of the spectrum can be used for several purposes. Both amplitude and phase spectra can be studied, and the assymptotic behaviour of the spectra can be made explicit. These model can be considered fundamentally as impulse responses of low-pass filters, with relatively small bandwidth. Thus they should be represented by a little set of spectral parameters, which can essentially be derived by studying the assymptotic behaviour of the spectrum in the neighborough of 0 and $+\infty$ (for the moment we are dealing only with continuous time systems).

Along this line we showed that the open quotient O_q of the KLGOTT88 model is directly linked with the glottal formant frequency f_k:

$$f_k = \frac{\sqrt{3}}{\pi} \frac{1}{O_q T_0}$$

This result is obtained by studying the asymptotic behavior of the magnitude spectrum:

$$\nu \to 0, \qquad |\tilde{U_g^k}(\nu)| \to \frac{9}{16} av(O_q T_0)^2$$
$$\nu \to \infty, \qquad |\tilde{U_g^k}(\nu)| \sim -\frac{27}{4} \frac{av}{(2\pi\nu)^2}$$

This shows that the behaviour of the spectrum is constant in the neighborough of 0, it has a slope of 0 dB/oct. The spectrum has a - 12 dB/oct slope when ν tends to infinity. The frequency f_k corresponds to the crossing point of the two assymtots. If we consider speech, the glottal flow component must be derived (as the effect of sound radiation can be approached by a derivation), and the (0,-12) dB/oct behaviour of $|\tilde{U_g^k}(\nu)|$ is transformed through derivation into a (+6,-6) dB/oct behaviour. Thus a maximum appears in the spectrum $|\tilde{U_g^{k'}}(\nu)|$, which corresponds to the glottal formant. "Glottal formant" stands for a local maximum in the amplitude spectrum which is located below the first vocal tract formant: a well-known fact in spectrogram reading. Using the glottal formant frequency, the open quotient can be used for computing the amplitudes of the few first harmonics (for instance the amplitude ratio of the 2 first harmonics H1-H2). Unlike other glottal flow models, the open quotient is the only parameter defining H1-H2 in the KLGLOTT88 model.

On the contrary, for the LF-model, H1-H2 depends on two parameters. Using a different technique, namely curve fitting on H1-H2 variation as a fonction of the LF-model parameters, one can show that H1-H2 can be approximated by the following equation:

$$H1 - H2 \simeq 12(\frac{O_q}{0.7})^2(1 - (1 - \frac{R_k}{0.7})^2) - 6$$

with a 1 dB approximation (i.e. the difference between this equation and the actual function computed with the help of analytic formulas is bounded by 1 dB).

2.2.4 Temporal correlates of spectral tilt

The spectral tilt is an important parameter of voice quality. "Spectral tilt" is a rather vague term, standing for the observed spectral attenuation of high frequencies in speech spectra, and thus it is related to the spectrum behaviour when the frequency tends towards $+\infty$. However, the exact shape of this spectral attenuation is not very well understood at this point in time. Therefore, the parameters used for spectral tilt are not very explicit in a number of studies, and probably are not defined at all.

For the KLGLOTT88-model, spectral tilt is a spectral parameter. It is represented by a second-order low-pass digital filter. This results in a -12 dB/oct attenuation for frequencies above the filter cutoff frequency1. The parameter TL gives the cutoff frequency of this filter following the equation:

$$f_t = 3000/\sqrt{10^{TL/10} - 1}$$

For the LF-MODEL, the situation is more intricate, as it is a fully temporal glottal flow model. Using the analytic expression of the LF-model spectrum, one can show that:

1. If the parameter R_a is set to 0, then $\mid \tilde{U}_g^{k'}(\nu) \mid\sim E_e/(2\pi\nu)$ when $\nu \rightarrow +\infty$, which corresponds to a spectral slope of $-6dB/oct$. This is a $3dB$ approximation of the spectrum above frequency $\max(2a, \frac{4\pi R_g}{T0}\cot(\pi(1+R_k)))$, which is given by the second prominent term in the development.

2. If R_a is not equal to zero, then an extra $-6dB/oct$ is added to the spectrum, leading to a $-12dB/oct$ spectral slope, above a cutoff frequency which can be computed as $f_c = F_a + a/2\pi + \frac{R_g}{T0}\cot(\pi(1+R_k))$, where $F_a = 1/2\pi R_a$. In comparison to the predicted cutoff frequency value of F_a given by Fant [1] [3], this analytically calculated value gives a correction term that is not negligible : for instance, with $R_g = 1.3$, $R_k = 0.3$ and $R_a = 0.1$, (which corresponds to a plausible speech waveform) then $F_a = 160Hz$ although the cutoff frequency is equal to $f_c = 290Hz$; in this case, taking F_a instead of f_c leads to a more than $5dB$ error in the determination of the spectral tilt. Notice that the amplitude of the first harmonics is also affected by this parameter.

In conclusion, the spectral tilt depends mostly on the parameter R_a. This parameter is responsible for an extra $-6dB/oct$ attenuation above frequency f_c. However, f_c depends also on R_g and R_k according to the analytic expression $f_c = F_a + a/2\pi + \frac{R_g}{T0}\cot(\pi(1+R_k))$. Thus, contrary to the statement in [12], we show that f_c cannot always be approximated by F_a.

2.2.5 On the phase spectrum

In the preceeding sections, we considered mainly the effect of the time domain parameters on the glottal flow amplitude spectrum. As a matter of fact, all the studies on the perceptual correlates of voice source models are dealing with on ly amplitude spectra (or time-domain parameters). Information on the phase spectrum is scarcely used, excepted for synthesis. In [9], we showed that for the KLGOTT88-model, the phase of $\tilde{U}_g^k(\nu)$ can be split in a linear component and a non-linear component. The linear component is only due to the delay between 0 and the epoch of glottal closure. The non-linear component is linked to the glottal flow waveform. We showed that this component is close to the phase gain of a low pass filter, except that it corresponds to a anticausal impulse response.

This shed light on the role of phase in the spectrum of glottal flow model: phase is giving the overal shape of the waveform in time-domain, and this shape is close to the (finite) impulse response of a linear anticausal filter. If phase distortion is introduced in the speech recording or storage process, no audible differences will be noticeable, but the time domain waveform will be distorded accordingly. In this case no time-domain model will be able to represent the actual data, although a spectral model will still be effective, if one consider amplitude spectra.

2.3 Aperiodic component

In the preceeding sections we considered the periodic component of excitation. But the glottal flow signal $\{e(t), E(\omega)\}$ encompass also an aperiodic component $\{r(t), R(\omega)\}$. For some types of voice quality, the noise component of the source signal r is a very important feature. It can be described by several parameters:

1. its amplitude relative to u_g, coined here as the periodic-to-aperiodic ratio (PAPR);
2. its power spectrum $|R(\omega)|^2$. The shape of this power spectrum can be described by another spectral tilt parameter for the noise component, which can differ a lot of the spectral tilt defined for the periodic component;
3. its temporal modulation. If this component is considered as a sample of a stochastic process, this is related to the amplitude probability distribution of the noise. In some case the aperiodic component is almost Gaussian (no temporal structure), but in other case it is weakly or strongly modulated by the vocal fold vibration (aspiration noise or voiced frication noise), or by the vocal tract motion (plosive bursts). There is some evidence that noise modulation can be significant for perception. If one consider only this component as a deterministic signal, modulation of the noise is described by the structure of its phase spectrum $\theta_r(\theta_r(\omega)\omega)$. A modulation index can be varied by processing the phases.

Chapter 3

Algorithms

The preceeding equations are the keys for spectral domain modification of the voice source parameters, and thus of voice quality. Direct formulas are available for computing the effect of spectral changes on time-domain parameters.

A summary of the correspondance between spectral parameters and time-domain parameters is reported in Table 1.

For the periodic component of the source signal, the spectral modifications that are envisaged affect only the magnitude spectrum. Modification of voice quality can be achieved in spectral domain by simple linear filtering of the source spectra. For the periodic component, only the magnitude spectrum has to be modified. This is because the phase spectrum is almost not significant for voice quality perception. Therefore zero-phase filters have to be used.

An elaborate time-frequency method for modification of the aperiodic component has been proposed in a previous work [14]. In the present work we considered just simple modifications of the PAPR, and of the phase spectrum.

In this section we shall review the algorithms used for performing voice quality transformation. These algorithms are implemented in C programs, that are available in the department 2 at ATR-ITL.

The algorithms for spectral domain voice quality transformation are chained as follows:

1. **Inverse filtering** The aim of this stage is to derive an approximation to the voice source. One can perform directly modification on the speech signal rather than on the source, but in this case both source and filter will be modified, and the effect of spectral modification will not correspond accurately to the equations listed above.

2. **Periodic-aperiodic decomposition** The aim of this stage is to separate the two components of the source signal.

3. **Periodic component modification** The periodic component U_g can be modified according to the equations cited above. zero-phase filtering implement modification of H1, H2 (...Hn), and of spectral tilt (any aspects of the amplitude spectra can be modified).

4. **Aperiodic component modification** The aperiodic component can be modified also in spectral domain. Only modification of amplitude and of modulation is envisaged here.

	Time Domain	Frequency Domain
filter	impulse response	spectral enveloppe (formants)
periodic component	fundamental period	fundamental frequency
	AV (KLGLOTT88)	overall gain
	E_e (LF)	overall gain
	O_q (KLGLOTT88)	glottal formant
	R_g R_k (LF)	H1-H2
	R_a (LF)	Spectral Tilt
aperiodic component	PAPR	PAPR
	modulation index	phase spectrum
		spectral tilt

Table 3.1: correspondance between time and frequency domain description of glottal flow models

5. **Modified source recomposition** A modified source signal is performed by addition of the modified periodic and the modified aperiodic components.

6. **Vocal tract modification** At this stage it is possible to modify vocal tract parameters by modifying the synthesis filter. This is not implemented yet.

7. **Synthesis filtering** Of course this stage has to be performed only if inverse filtering was performed.

In the remaining parts of this section, we shall review each of this stages of processing and describe the corresponding programs.

3.1 Adaptive inverse filtering

The first step of analysis is inverse filtering. Many methods have been proposed in the litterature. We implemented the IAIF (Iterative Adaptive Inverse Filtering) method [1]. In this method, inverse filtering is done on a frame-by-frame basis (it is adaptive). Consider now discrete time signals. the source filter equation can be written in the Z-domain as:

$$S(z) = P(Z)U_g(z)V(z)L(z)$$

where P is a pulse train is the signal is voiced, or a noise if it is unvoiced, U_g is the glottal flow, V is the vocal tract, and L is lip radiation. The problem is that only $S(z)$ is given, and thus an iterative method must be used for deconvolution of $E(z)$ on the one hand, and $V(z)L(z)$ on the other hand. As a first approximation, one can consider that $L(z)$ is a derivation filter (1rst order, all-zero) , with the following transfert function:

$$L(z) = 1 - az^{-1}$$

another approximation (used only for the first iteration) is that $E(z)$ is a second (or fourth) order all-pole filter:

$$E(z) = \frac{1}{(1 + az^{-1})^2}$$

the vocal tract is assumed to be a all-pole filter (with order p, p=18 for example for a 16 kHz sampling frequency).

$$V(z) = \frac{1}{\sum_{i=0}^{p} a_i z^{-i}}$$

The method is conceptually simple and contains 10 processing blocks.

1. *The effect of the glottal pulseform to the speech spectrum is preliminary estimated by first-order LPC analysis.* This means that the coumpound effect of $G_1(z) \simeq U_g(z)L(z)$ is estimated by LPC. This is a 1rst order filter.

$$S(z) = P(Z)G_1(z)V(z)$$

2. *The estimated glottal contribution is eliminated by filtering the speech signal.* Thus the signal $S_1(z)$ is obtained as:

$$S_1(z) = \frac{S(z)}{G_1(z)} = \frac{P(z)E(z)L(z)V(z)}{G_1(z)} \simeq P(z)V(z)$$

3. *The first estimate for the vocal tract is computed by applying LPC analysis to the output of the previous block.* A filter $VT_1(z)$ is estimated by LPC,, applied to $S_1(z)$.

$$VT_1(z) = \frac{S_1(z)}{P(z)} = \frac{P(z)E(z)L(z)V(z)}{G_1(z)} \simeq P(z)V(Z)$$

4. *The effect of the vocal tract is eliminated from the speech signal by inverse filtering.* A new signal $S_2(z)$ which is obtained:

$$S_2(z) = \frac{S(z)}{VT_1(z)} = \frac{P(z)U_g(z)V(z)L(Z)}{VT_1(z)} \simeq= P(z)U_g(z)L(Z)$$

5. *the first estimate of the glottal excitation is obtained by cancelling the lip radiation effect by integrating.* The first estimate is obtained by filtering the effect of $L(z)$ in $S_2(z)$:

$$E_1(z) = \frac{S_2(z)}{L(z)} \simeq P(z)U_g(z)$$

6. *The second iteration starts by computing a new estimate for the effect of the source to the speech spectrum. This time, LPC analysis of order 2 to 4 is used. The signal from which the glottal estimation is the previous integrated signal.* A new approximation of the source $E_2(z)$ is computed using $E_1(z)$:

$$G_2(z) = \frac{E_1(z)}{P(z)}$$

7. *The effect of the estimated glottal contribution is eliminated.* A signal $S_3(z)$ accounting for vocalt tract, lip rariation and perioic pulses is derived:

$$S_3(z) = \frac{S(z)}{G_2(z)} = \frac{P(z)U_g(z)L(Z)V(z)}{G_2(z)} \simeq P(z)L(Z)V(z)$$

8. *The final model for the vocal tract is obtained by applying LPC analysis to the output of the previous block.* The filyter $VT_2(z)$ is estimated using LPC:

$$S_3(z) = VT_2(z)P(z)L(Z)$$

9. *The effect of the vocal tract is eliminated from the speech signal by inverse filtering.* A new signal $E_3(z)$ which is computed:

$$E_3(z) = \frac{S(z)}{VT_2(z)} \simeq P(z)L(Z)U_g(z)$$

10. *The final estimate of the glottal excitation is obtained by cancelling the lip radiation effect by integrating the output of the previous block.* The approximation $G_3(z)$ of the source signal $P(z)U_g(z)$ is:

$$G_3(z) = \frac{E_(z)}{L(z)} \simeq P(z)U_g(z)$$

3.2 Periodic-aperiodic decomposition

The algorithm used for periodic-aperiodic (PAP) decomposition is described in some details in [7] [5] [6] [8]. We shall not repeat all the details here.

We performed a series of tests using synthetic signal, and demonstrated that the algorithm is able to decompose varying mixtures of periodic and aperiodic components, like the noise bursts produced at the glottal closure and the deterministic glottal pulses. This gives a measurement of noise in the voice source, and a periodic component.

However, the structure of the algorithm must be presented, just because modifications of both the periodic and the aperiodic components are integrated inside the PAP algorithm.

The PAP and further modification algorithms can be used with or without prior inverse filtering. If inverse filtering is used, glottal source parameters and glottal source parameter modifications can be computed directly from the equation of Section 2. If inverse filtering is not used, the exact effect of parametric modification will be hardly known, as initial glottal flow component is not explicit. However, for magnitude only zero-phase modification, the effect of the modifications are exactly the same with or without inverse (and synthesis) filtering. As for phase modifications, of course a prior deconvolution will give better results. The following description is borrowed to [6]

The complex addition in the source/filter spectral equation suggests the importance of both the magnitude and phase of each of the components in the signal. The proposed decomposition algorithm contains the following steps, assuming that one works on an excitation signal obtained by inverse filtering :

(a) **Short-term Fourier analysis** The signal is decomposed into short overlapping analysis frames, using a data window (e.g. a 40 ms Hamming window). The frame rate is relatively important, e.g. 200 frames per second for the short-term signal decomposition. A short-time spectrum $E_l(k)$ is computed for each frame, using the Discrete Fourier Transform (DFT) (e.g. N=512 points).

$$E_l(k) = \sum_{n=0}^{N-1} w(n)\ e(n + lH) \exp(-\frac{j2\pi}{N} nk)$$

where s is the discrete-time excitation signal, H is the hop size in number of samples (the spacing between analysis frames), l is the frame index, N is the FFT size, and w is the analysis window.

(b) **Identification of frequency regions of the aperiodic component.** Both periodic and aperiodic components contribute to the DFT coefficients. In the first stage of processing, we identify a subset of the DFT coefficients to form an approximation to the aperiodic component. For this purpose, we determine approximately the frequency regions contributing to the harmonic part and the frequency region contributing to the noise part. This is accomplished by using a prior pitch detection.

(c) **Reconstruction of the aperiodic signal using an iterative procedure.** The knowledge of the location of the harmonics does not enable us to separate the two components by subtraction, because at each frequency point there is contribution due to both the periodic and aperiodic components. According to the spectral source/filter model, it is necessary to use both amplitude and phase, at each frequency, for separating these components in the frequency domain. Obviously, the magnitude and phase are not directly available. We developed an iterative procedure to reconstruct the aperiodic components.

From the frequency distribution of the harmonic regions in the log magnitude spectrum, we hypothesize, to a first approximation, that the valley regions between two harmonics are mostly due to the aperiodic component. To obtain an approximate aperiodic component, $r_l(n)$, of the residual, we can sum only those DFT coefficients ($k \in F_r$) for which the noise component dominate. That is

$$r_l(n) = \sum_{k \in F_r} E_l(k) \exp(\frac{j2\pi}{N} nk)$$

where N is equal to the size of the DFT. Here F_r is the set of frequency points in the valley regions between two harmonics. The width of the harmonics must be fixed (for instance as a function of f0).

Thus, the aperiodic component is set to zero in the harmonic regions, and to the measured DFT values in the regions between the harmonics, i.e., in the noise regions. It is clear that such a comb-filtered noise component cannot represent the aperiodic component in the speech signal. It is necessary to estimate the values of the noise component in the harmonic regions. The contribution of the aperiodic component in the harmonic regions is estimated using an iterative algorithm similar to the Papoulis-Gerchberg extrapolation algorithm. An estimate of the aperiodic component is obtained by iteratively moving from the frequency domain to the time domain and vice versa, through the IDFT and DFT operations.

<u>First iteration:</u>

Suppose we obtained a set of DFT coefficients that form a first approximation $R_l^0(k)$ to the aperiodic component:

$$R_l^0(k) = \begin{cases} E_l(k), & \text{for } k \in F_r \text{ (noise regions)} \\ 0, & \text{otherwise} \end{cases}$$

An IDFT is applied to this first approximation, and the corresponding time-domain signal $r_l^0(n)$ is obtained (for instance 512 samples in our case). A finite duration constraint is imposed in the time domain, i.e. the signal samples of the aperiodic component in the time domain beyond the analysis frame size are set to zero (as we used a $N/2-1=255$ samples analysis windows, the samples from 255 to 511 are set to zero, numbering the samples from 0 to 511). That is, form a signal

$$\hat{r}_l^0(n) = \begin{cases} r_l^0(n), & \text{for } n \ < N/2-1 \\ 0, & \text{otherwise} \end{cases}$$

m^{th} iteration:

Starting with $m=1$, we compute the DFT $\hat{R}_l^{m-1}(k)$ of $\hat{r}_l^{m-1}(n)$, and form the function

$$R_l^m(k) = \begin{cases} E_l(k), & \text{for } k \in F_r \\ \hat{R}_l^{m-1}(k), & \text{otherwise} \end{cases}$$

and compute its IDFT $r_l^m(n)$. The time samples beyond $N/2-1$ are set to zero. That is

$$\hat{r}_l^m(n) = \begin{cases} r_l^m(n), & \text{for } n < N/2-1 \\ 0, & \text{otherwise} \end{cases}$$

The iterative algorithm is continued until the difference (in terms of magnitude of the noise samples) between two successive steps becomes less than a given threshold value, or after a fixed number of iterations. In our experiments, we used $m=10$ iterations. The periodic component is obtained by subtracting the reconstructed aperiodic component noise samples from the residual signal samples in the time domain. The convergence of the proposed algorithm can be proved.

(d) **Synthesis.** The aperiodic component of the excitation signal is obtained for each of the overlapping analysis frames. The periodic component is obtained on each frame by simple time-domain substraction of the aperodic component from the excitation signal:

$$p_l(n) = e_l(n) - r_l^m(n) \qquad \text{for } n < N$$

The perodic and aperiodic component signal of the excitation for the entire utterance are derived from these short-time signals, using an overlapp-add procedure:

$$\begin{aligned} r(n) &= \sum_{l=0}^{L-1} r_l^m(n - lH) \\ p(n) &= \sum_{l=0}^{L-1} p_l(n - lH) \end{aligned}$$

where L is the number of frame in the utterance.

3.3 Spectral modification of voice quality

Spectral modifications of the voice source are easily integrated to the periodic-aperiodic decomposition algorithm, by inserting a spectral modification after iterative reconstruction of the aperiodic component (step (c) of the algorithm). The modification of the PAP algorithm is as follows.

(d') **periodic component modification.** The aperiodic component of the excitation signal is obtained for each of the overlapping analysis frames. The periodic component is obtained on each frame by simple time-domain substraction of the aperiodic component from the excitation signal:

$$p_l(n) = e_l(n) - r_l^m(n) \qquad \text{for } n < N$$

We compute then the DFT $P_l(k)$ of the periodic signal on this frame. Modifications are achieved by simple multiplication of the DFT samples by the frequency samples of the modification filter $M(k)$. One can notice that the PAP decomposition method makes use of the so-called block interpretation of the short term Fourier transform. In this frameowork, multiplication of the DFT samples results in the well-known method for linear time-varying filtering using the DFT with overlap-add.

In some situations, magnitude only modifications are desired. This is equivalent to zero-phase filtering, because the phase spectrum is not altered. When zero-phase filter is desired, the samples of $M(k)$ must be real coefficients.

For filtering real signals by multiplication of the DFT samples, one must take care to modify the sample according to the hermitian symetry, i.e:

$$\begin{array}{lll} P_l(k) & = P_l(k)M(k) & 0 \leq k \leq N/2 \\ P_l(k) & = P_l(k)M(k - \frac{N}{2}) & N/2 < k < N \end{array}$$

Examples of desired modifications are modification of H1-H2, modification of spectral tilt, and modification of the PAPR.

Modification of H1-H2 is achieved by selection of the frequency samples corresponding to these components of the short-term spectrum $(k_{H1}), (k_{H2})$, and by selective multiplication of these samples, by modification factors M_1, M_2:

$$\begin{aligned} P_l(k) &= P_l(k)M_1(k) \quad k \in k_{H1} \\ P_l(k) &= P_l(k)M_2(k) \quad k \in k_{H2} \end{aligned}$$

Modification of spectral tilt involves also selection of the samples to be modified (k_{ST}), for example the samples above a given frequency f_{tl}, and multiplication by a frequency dependant factor. For instance if one wants a A dB/oct spectral tilt above f_{tl}, the multiplicative factor at frequency f will be :

$$A(f) = 10\frac{A\log_2(\frac{f}{f_{tl}})}{20}$$

(e') **aperiodic component modification.** As the aperiodic component of the excitation signal is obtained for each of the overlapping analysis frames, the same type of modification can be applied

. The periodic component is obtained on each frame by simple time-domain substraction of the aperiodic component from the excitation signal:

$$p_l(n) = e_l(n) - r_l^m(n) \qquad \text{for } n < N$$

We compute then the DFT $P_l(k)$ of the periodic signal on this frame. Modifications are achieved by simple multiplication of the DFT samples by the frequency samples of the modification filter $M(k)$. One can notice that the PAP decomposition method makes use of the so-called block interpretation of the short term Fourier transform. In this frameowork, multiplication of the DFT samples results in the well-known method for linear time-varying filtering using the DFT with overlap-add.

In some situations, magnitude only modifications are desired. This is equivalent to zero-phase filtering, because the phase spectrum is not altered. When zero-phase filter is desired, the samples of $M(k)$ must be real coefficients.

For filtering real signals by multiplication of the DFT samples, one must take care to modify the sample according to the hermitian symetry, i.e:

$$\begin{aligned} P_l(k) &= P_l(k)M(k) \quad 0 \leq k \leq N/2 \\ P_l(k) &= P_l(k)M(k - \tfrac{N}{2}) \quad N/2 < k < N \end{aligned}$$

(f') **Synthesis.** The periodic and aperiodic components of the modified excitation signal are obtained for each of the overlapping analysis frames.

The modified perodic and aperiodic component signal of the excitation for the entire utterance are derived from these short-time signals, using an overlapp-add procedure:

$$\begin{aligned} r(n) &= \sum_{l=0}^{L-1} r_l^m(n - lH) \\ p(n) &= \sum_{l=0}^{L-1} p_l(n - lH) \end{aligned}$$

where L is the number of frame in the utterance.

3.4 Adaptive synthesis filtering

If inverse filtering has been performed, the modificed speech signal is reconstructed by synthesis filtering of the modified excitation signal. The synthesis filter is a compound filter:

$$SY(z) = VT_2(z)L(z)$$

Chapter 4

Experiments

4.1 Spectral parameter modification and voice quality

Sone types of voice qualities often found in the litterature are:

whispery phonation the glottis is open, and there is no periodic vibration of the vocal folds.

breathy phonation the glottal closure is incomplete, there is a lot of additive aperiodicity in the source.

creaky phonation there is a lot of structural aperiodicity in the source (jitter, shimmer), and/or f0 is very low..

soft phonation the vocal fold are vibrating, the vocal effort is weak.

pressed phonation the vocal effort is high, but the signal is not necessarily efficient.

falsetto phonation F0 is high, vocal folds vibrate in a different mode.

loud phonation the vocal effort and the signal energy are high.

Some examples of relation between voice quality and spectral parameters of the source are summarized in Table 2.

phonation	H1-H2	Spectral Tilt	PAPR	f0	gain
whispery	not relevant	*	0	not relevant	low
breathy	*	high	low	*	low
creaky	*	high	low	*	very low
soft	high	high	low	*	low
pressed	low	*	*	*	*
falsetto phonation	high	*	*	high	
loud	low	low	high	high	large

Table 4.1: Spectral parameters and voice quality

4.2 Vocal effort modification

Some experiments have been conducted for modification of vocal effort. According to the preceeding analysis, modification of vocal effort requires joint modification of several acoustic parameters. In the examples we tried only very simple rules were used:

lowering vocal effort Vocal effort is lowered by joint modifications of :

1. H1-H2: amplitude of the first harmonic is raised compared to amplitude of the second harmonic (few dB).
2. spectral tilt is increased, above a fixed frequency. This frequency is typically 1500-2500 Hz, and the additional spectral attenuation is 6 to 12 dB/oct.
3. Periodic-APeriodic Ratio is decreased. Typically the amplitude balance between both components is shifted to 30 % periodic 70 % aperiodic.

Increasing vocal effort Vocal effort is increased by joint modifications of:

1. H1-H2: amplitude of the second harmonic is raised compared to amplitude of the second harmonic (few dB).
2. spectral tilt is decreased, above a fixed frequency. This frequency is typically 1500-2500 Hz, and the additional spectral boost is 6 to 12 dB/oct.
3. Periodic-APeriodic Ratio is increased. Typically the amplitude balance between both components is shifted to 70 % periodic 30 % aperiodic.

FO and durations were not modified. Amplitude of the signal was not modified explicitely, but, as amplitude of the signal components was changed, this resulted in some global amplitude change. Global amplitude was decreased when lowering vocal effort, and increased when increasing vocal effort.

Chapter 5

Conclusion

In this report we summarized previous work on periodic-aperiodic decomposition and spectral representation of glottal flow signals. In a another part we presented the designa and implementation of a new algorithm for spectral domain modification of voice quality.

The work presented here has a number of possible applications to speech synthesis:

1. Future work could be devoted to pre-processing data-bases, in the context of synthesis by concatenation systems. This could be useful for reducing the quality difference between speech segments.

2. The method developped could also be used in parametric synthesis of speech, for implementing rules dealing with the vocal effort. This is to be done in relation with stress and accent. Also improved rules for f0/duration/voice quality interaction could be implemented with our method.

3. experiments and research on voice quality perception could also benefit of the transparent quality obtained for voice quality parameters.

Bibliography

[1] Alku P. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11:109–118, 1992.

[2] Campbell, N., and Beckman, M. Stress, prominence, and Spectral Tilt. *ESCA Workshop on Prosody*,Rhodes, Sept. 97, to appear.

[3] Ding, W. and Campbell, N.. Detection of sentence prominence using voice source parameters. *Proc. of the 3rd joint meeting of ASA and ASJ*, Honolulu, Hawaii, 2-6 December 1996, 843-848..

[4] Ding, W., Campbell, N., Higuchi, N.,. and Kasuya, H. Fast and robust joint estimation of vocal tract and voice source parameters. In *International Conference on Acoustics, Speech and Signal Processing*, pages 1291–1294, Munich, avril 1997. Institute of Electronics and Electrical Engineers.

[5] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos. An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE transaction on Speech and Audio*, 1997. in press

[6] C. d'Alessandro, V. Darsinos, and B.Yegnanarayana. Significance of periodic and aperiodic decomposition for analysis of voice sources. *IEEE transaction on Speech and Audio*, 1997. in press.

[7] C. d'Alessandro, B. Yegnanarayana, and V. Darsinos. Decomposition of speech signals into deterministic and stochastic components. In *International Conference on Acoustics, Speech and Signal Processing*, pages 760–763, Detroit, avril 1995. Institute of Electronics and Electrical Engineers.

[8] V. Darsinos, C. d'Alessandro, and B. Yegnanarayana. Evaluation of a periodic/aperiodic speech decomposition algorithm. In *European Conference on Speech Communication and Technology*, pages 393–396, Madrid, Septembre 1995. European Speech Communication Association.

[9] B. Doval and C. d'Alessandro. Spectral correlates of glottal waveform models: an analytic study. In *International Conference on Acoustics, Speech and Signal Processing*, pages 1295–1299, Munich, avril 1997. Institute of Electronics and Electrical Engineers.

[10] Klatt D. and Klatt L. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.*, 87(2):820–857, 1990.

[11] Fant G., Kruckenberg A., Liljencrants J., and Bavegard M. Voice source parameters in continuous speech. Transformation of LF-parameters. *ICSLP*, 1451–1454, Yokohama, 1994.

[12] Fant G., Liljencrants J., and Lin Q. A four-parameter model of glottal flow. *STL-QPSR*, 85(2):1–13, 1985.

[13] Fant G. and Lin Q. Frequency domain interpretation and derivation of glottal flow parameters. *STL-QPSR*, 88(2-3):1–21, 1988.

[14] Richard G. and d'Alessandro C. Analysis/synthesis and modification of the speech aperiodic component. *Speech Communication*, 19:221–244, 1996.

[15] Swerts, M. and Veldhuis, R. Interaction between intonation and glottal-pulse characteristics. *ESCA Workshop on Prosody*,Rhodes, Sept. 97, to appear.

[16] Sluijter, A. Phonetic correlates of stress and accent. Holland Ac. Graphics, The Hague.

[17] Gauffin J. and Sundberg J. Spectral correlates of glottal voice source waveform characteristics. *Journal of Speech and Hearing Research*, 32:556–565, 1989.

[18] Hanson H. M. Glottal characteristics of female speakers. In *PhD Thesis*. Harvard Univ., 1995.

Appendix A

Programs

All the programs are in the directory cda/src, and the sound files in cda/data.

These are all C-language programs, compiled with cc -traditional -lm.

A.1 Iterative Adaptive Inverse Filtering

A.1.1 Shell command

The command is a C-shell (csh) unix command of the form:

```
iaif glume
```

where the speech signal is in the file "glume.s1". The inverse filtered signal is in "glume.ug", and the LPC coefficients (that will be useful for adaptive synthesis filtering) are in the file "glume.vt2".

The unix command file "iaif" is as follows:

```
# Iterative Adaptive Inverse Filtering
# as described in Alku P.
# "Glottal wave analysis with pitch
# synchronous iterative adaptive inverse filtering."
#  Speech Communication, 11:109--118, 1992.
 #
 # Author: Christophe d'Alessandro, ATR-ITL (07/1997)
 #
 #

 # The effect of the glottal pulseform to the speech
 #spectrum is preliminary estimated by first-order LPC analysis.

 # The estimated glottal contribution is eliminated by filtering
# the speech signal.

resid -i$1.s1 -o$1.g1 -c$1lpc.g1 -l1 -w160 -v80

# The first estimate for the vocal tract is computed by applying LPC
# analysis to the output of the previous block.

resid -i$1.g1 -oglume  -c$1lpc.vt1 -l18 -w160 -v80

# The effect of the vocal tract is eliminated from the speech signal
# by inverse filtering

invfil -i$1.s1 -o$1.vt1 -c$1lpc.vt1 -l18  -w80

# the first estimate of the glottal excitation is obtained by
# cancelling the lip radiation effect by integrating.

integ -i$1.vt1 -o$1.vt1i -c0.95

# The second iteration starts by computing a new estimate for the
# effect of the source to the speech spectrum. This time, LPC analysis
# of order 2 to 4 is used. The signal from which the glottal estimation
# is the previous integrated signal.

resid -i$1.vt1i -oglume -c$1lpc.g2 -l2 -w160 -v80

# The effect of the estimated glottal contribution is eliminated.

invfil -i$1.s1 -o$1.g2 -c$1lpc.g2 -l2  -w80

# The final model for the vocal tract is obtained by applying LPC
# analysis to the output of the previous block.

resid -i$1.g2 -oglume -c$1lpc.vt2 -l18 -w160 -v80

# The effect of the vocal tract is eliminated from the speech signal
# by inverse filtering

invfil -i$1.s1 -o$1.vt2 -c$1lpc.vt2 -l18 -w80

# The final estimate  of the glottal excitation is obtained by
# cancelling the lip radiation effect by integrating the output
# of the previous block.

integ -i$1.vt2 -o$1.ug -c0.95
```

A.1.2 LPC autocorelation analysis

```
 LPC autocorelation + residual

  USAGE: resid -i<input file(bin)>
               -o<output file(bin)>
               -c<file with lpc coef.
               -l<numb. of lpc coef.>
               -z<anticausal signal>
               -w<window_length>
```

A.1.3 LPC inverse filtering

```
Inverse filtering using LPC

USAGE: invfil -i<input file(bin)>
              -o<output file(bin)>
              -c<file with lpc coef.>
              -l<number of lpc coef.>
              -z<anticausal>
              -w<step length>
              -a<amplify(float)>
```

A.1.4 LPC signal integration

```
signal integrator

USAGE: integ -i<input file(bin)>
              -o<output file(bin)>
              -c<integration coef>
              -a<amplify(float)>
```

A.2 Periodic-Aperiodic decomposition and spectral processing.

A.2.1 Shell command

The following rules for naming each file (e.g. "soundexample") are used:

soundexample.s1 is the original signal file. There are raw format signal files, 16 kHz, 16 bits, 2'complement, signed integer.

soundexample.pit contains pitch analysis, for each frame (frame rate is defined when running the analysis/synthesis program). Pitch analyses where obtained using a separate program, and any pitch analysis program will be convenient. The ".pit" file is a binary file containing C-language "float" numbers. When the signal is unvoiced, pitch is 0.

soundexample.p the periodic component of the signal.

soundexample.ap the aperiodic component of the signal.

soundexample.mod* the signal with modified voice quality.

```
# Examples of periodic-aperiodic analysis and spectral modifications
# INPUT
#        glume.s1 : signal file
#        glume.pit : pitch file
# OUTPUT
#        glume.p : periodic signal file
#        glume.ap : aperiodic signal file
#        glume.mod : modified signal file
# type papsp for help.
#
#  References :
#    C.d'Alessandro, B.Yegananarayana & V.Darsinos Icassp 95
#    B. Yegnanarayana C.d'Alessandro & V.Darsinos IEEE-trans SAP (in press)
#    C.d'Alessandro B.Yegananarayana & V.Darsinos  IEEE-trans SAP (in press)
#    B. Doval & C. d'Alessandro  Icassp 97
#    B. Doval,  C. d'Alessandro & B. Diard, Eurospeech 97
#
#
# Author: Christophe d'Alessandro, ATR-ITL (07/1997)
#
#




papsp -p$1.pit -l$1.ap -f16000 -e9 -w499 -r80 -b1 -q5 -c3 -h1 -a$1.papr  \
 -m$1.mod1   <$1.s1 >$1.p

papsp -p$1.pit  -l$1.ap -f16000 -e9 -w499 -r80 -b1 -q5 -c3 -h1  -a$1.papr \
 -m$1.mod2  -x2 - -u1300 -v-12 -z0.35   <$1.s1 >$1.p

papsp -p$1.pit  -l$1.ap -f16000 -e9 -w499 -r80 -b1 -q5 -c3 -h1  -a$1.papr \
-m$1.mod3  -y-2 x6 -u2000 -v-6 -z0.4   <$1.s1 >$1.p

papsp -p$1.pit  -l$1.ap -f16000 -e9 -w499 -r80 -b1 -q5 -c3 -h1  -a$1.papr \
-m$1.mod4  -y8 -x-2 -u2000 -v6 -z0.7   <$1.s1 >$1.p
```

A.2.2 Decomposition and spectral modifications

```
 Periodic/Aperiodic Decomposition of speech signals
 and spectral modification

USAGE: papsp [options] <Input_speech_file>Output_periodic_file

       options:
              -p<Pitch_file_name>
              -a<PAP_ratio_file_name>
              -m<modified_file_name>
              -l<Aperiodic_output_file_name>
              -exxx :fft order
              -wxxx :data window length
              -zxxx : PAP ratio in mod synt sig
              -xxxx : amp. fact. H1 (dB)
              -yxxx : amp. fact. H2 (dB)
              -uxxx : spectral tilt frequency (Hz)
              -vxxx : spectral tilt (dB/oct)
              -rxxx :overlap between windows(samples)
              -dxxx :starting sample
              -nxxx :number of samples for analysis
              -bxxx :number of frame for which save fft,cepstrum etc.
              -qxxx :number of iterations in pap-algorithm
              -k :if, then use cepstrum extracted pitch
              -t :if, then use all the cepstrum-rahmonics
                      for the decomposition
              -cxxx :bandwidth/2 of a rahmonic in cepstrum(samples)
              -hxxx :0 use Hanning window
                     1 use Hamming window
                     2 use Blackman-Harris window
```

A.3 Adaptive Synthesis Filter.

A.3.1 Shell command

The command is a C-shell (csh) unix command of the form:

```
asy glume
```

where the modified excitation signal is in the file "glume.ugm". The synthesis filtered signal is in "glume.s1m", and the LPC coefficients (the same as for inverse filtering) are in the file "glume.vt2".

The unix command file "asy" is as follows:

```
# Adaptive Synthesis Filtering
#
# Author: Christophe d'Alessandro, ATR-ITL (07/1997)
#
#


# The modified signal is filtered withthe vocal tract filter

synres -i$1.ugm -o$1.vtm -c$1lpc.g1 -l1 -w80

# The preceeding signal is derived, for simulating lip radiation

deriv  -i$1.vtm -o$1.s1m -c0.95
```

A.3.2 Synthesis filtering

```
Resynthesis of speech from the lpc residual

 USAGE: synres -i<input file(bin)>
               -o<output file(bin)>
               -c<file with lpc coef.>
               -l<number of lpc coef.>
               -w<step length>
               -a<amplify(float)>
```

A.3.3 Signal derivation

```
 signal derivation

 USAGE: deriv -i<input file(bin)>
              -o<output file(bin)>
              -c<integration coef>
              -a<amplify(float)>
```

Appendix B

Sound Demonstration

All the sound examples are in the directory /home/as65/cda/demos.

B.1 Lowering vocal effort

In these examples, vocal effort is lowered by joint modifications of H1-H2 (higher), spectral tilt (higher), Periodic-APeriodic Ratio (lower). FO and durations are not modified. Amplitude of the signal is not modified (but affected by parametric changes).

- Female voice, French speaker.

    1. Original signal `ala131.s1`
    2. Modified signal `ala131.mod4`

- Female voice, Japanese speaker.

    1. Original signal `f1.s1`
    2. Modified signal `f1.mod4`

- Male voice, British speaker.

    1. Original signal `n1.s1`
    2. Modified signal `n1.mod1`

- Male voice, french speaker.

    1. Original signal `fba138.s1`
    2. Modified signal `fba138h16t12000-12p20.s1`

- Male voice, Japanese speaker.

    1. Original signal `MHN_503_B_41.d.s1`
    2. Modified signal `MHN_503_B_41.d.mod2`

B.2 Increasing vocal effort

In these examples, vocal effort is increased by joint modifications of H1-H2 (lower), spectral tilt (lower), Periodic-APeriodic Ratio (higher). FO and durations are not modified. Amplitude of the signal is not modified (but affected by parametric changes).

- Female voice, French speaker.

  1. Original signal `ala131.s1`
  2. Modified signal `ala131.mod6`

- Female voice, Japanese speaker.

  1. Original signal `f1.s1`
  2. Modified signal `f1.mod3`

- Male voice, British speaker.

  1. Original signal `n1.s1`
  2. Modified signal `n1.mod2`

- Male voice, french speaker.

  1. Original signal `fba138.s1`
  2. Modified signal `fba138h2X2.s1`

- Male voice, Japanese speaker.

  1. Original signal `MHN_503_B_41.d.s1`
  2. Modified signal `MHN_503_B_41.d.mod5`