SPECTRAL METHODS FOR VOICE SOURCE PARAMETERS ESTIMATION

Boris Doval *Christophe d'Alessandro* *Benoît Diard*

LIMSI-CNRS, BP 133, F91403 Orsay, France.
E-mail: doval@limsi.fr cda@limsi.fr diard@limsi.fr

ABSTRACT

A spectral approach is proposed for voice source parameters representation and estimation. Parameter estimation is based on decomposition of the periodic and the aperiodic components of the speech signal, and on spectral modelling of the periodic component. The paper focusses on parameters estimation for the periodic component of the glottal flow. A new anticausal all-pole model of the glottal flow is derived. Glottal flow is seen as an anticausal 2-pole filter followed by a spectral tilt filter. The anticausal filter has complex poles, instead of the real poles that are usually assumed. Time-domain and frequency domain parameters are linked by analytic formulas. Two spectral domain algorithms are proposed for estimation of open quotient. The first one is based on measurement of the first harmonics, and the second one is based on spectral modelling. Experimental results demonstrate the accuracy of the estimation procedures.

1. INTRODUCTION

In this paper we present our recent work on algorithms for automatic analysis of voice quality parameters. These parameters are studied in the spectral domain. Voice source measurements are needed for high quality speech synthesis, because voice quality is currently a key issue for naturalness. This is particularly true in the situation of speech synthesis using concatenation of natural speech segments (e.g. diphones or non uniforms units). 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. Contrary to most of the recent works on source modelling, we prefered spectral processing, because it has a number of advantages:

- it does not require an accurate inverse filtering of the signal (because the effects of phase and amplitude are better separated in spectral domain);
- it seems more closely linked to the perceptual features of voice quality than time-domain processing;
- we shall show that one can design simpler methods (both conceptually and in terms of processing) for parameter estimations.

The main spectral parameters found in the litterature [7] [6] for synthesizing voices with different qualities are: 1/ spectral tilt; 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 (for instance AV (amplitude of voicing), O_q (open quotient), T_0 (fundamental period), f_t (spectral tilt), and AN (amplitude of noise)). Linking these two sets of parameters is therefore a key point for studying voice quality as a function of voice source parameters.

The first step of the analysis (or analysis/synthesis) process is to perform decomposition of the periodic and aperiodic components of the source. For this we used an algorithm proposed in [1]. 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 (see [2]). This gives a measurement of noise in the voice source, and a periodic component. For describing voice quality, the spectral parameters of the periodic component are generally measured on the amplitude 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 [4] [7] but are defined in time domain. We first made an analytic spectral study of these models, to link their parameters with spectral parameters [3].

Following this work, it seemed possible to process voice quality by processing the speech amplitude spectrum, using simple linear filtering schemes. However, for analysis or modifications it appears necessary to first estimate the values of frequency domain parameters. The scope of the paper is to discuss algorithms for frequency domain estimation of the periodic component parameters. In section 2, analytic formulas for the spectrum of glottal flow models are reviewed. Section 3 presents 2 types of algorithm for es-

timation of the open quotient. Section 4 gives some experimental results. Section 5 concludes.

2. PERIODIC COMPONENT SPECTRAL MODELLING

Using an analytic formulation of the Fourier transform of glottal flow models we show that it is possible to derive simple spectral descriptions of the glottal flow characteristics. We worked on both the LF-model [4] [5] and the KLGLOTT88 model [7]. Firstly, results on the KLGLOTT88 model U_k are reported here. This model is defined by the parameters AV (amplitude of voicing), O_q (open quotient), T_0 (fundamental period) and f_t (spectral tilt). The analytic formula for the spectrum of the KLGLOTT88 model is computed in [3]. When the frequency ν tends to infinity, we can show that this spectrum is equivalent to its first term: $\tilde{U_k}(\nu) \sim \frac{-27AV\exp(-j2\pi\nu O_q T_0)}{4(2\pi\nu)^2}$ Therefore, the spectral slope is -12 dB/oct. Along the same line, the principal term of the series expansion of the KLGLOTT88 spectrum when ν tends to 0 is given by: $\tilde{U_k}(\nu) \rightarrow \frac{9}{16}AV(O_q T_0)^2$ and the spectral slope of $\tilde{U_k}(\nu)$ is null in 0. The magnitude spectrum of $\tilde{U_k}(\nu)$ has a -12 dB/oct slope for high fequencies, and is constant for low frequencies. It can be represented by a second-order low-pass filter. The cutoff frequency of this filter can be computed as the crossing point of the lines defined by the equations above. This cutoff frequency f_g is given by: $f_g = \frac{\sqrt{3}}{\pi}\frac{1}{O_q T_0}$ (1) .

f_g depends only on the product $O_q T_0$, the open duration. Therefore, one can find a very simple spectral interpretation of the time-domain parameter O_q: it defines the cutoff frequency of a second order low-pass filter. In the KLGLOTT88 model, the waveform $U_k(t)$ is filtered by a first order low-pass filter. This filter is defined by its cutoff frequency f_t, and add an extra -6 dB/oct attenuation. Therefore, the magnitude spectrum of the KLGLOTT88 model can be split in three regions. In a first region, between frequencies 0 and f_g the spectral slope is 0 dB/oct, with a constant value given above. In a second region, between f_g and f_t, the spectral slope is -12 dB/oct. In a third region, above f_t, the spectral slope is -18 dB/oct. It must be pointed out that this representation is very interesting, because there is a one-to-one correspondance between the breakpoints in the magnitude spectrum and the independant time-domain parameters of the model.

Another link can be established between O_q and the glottal flow spectrum, by considering the amlitudes of the two first harmonics (ratio $H_1 - H_2$, where H_1 and H_2 are the amplitudes of the first two harmonics (in dB)). In [3] we show that the amplitude ratio of the two first harmonics depends mostly on the open quotient and the speed quotient (R_g and R_k in the LF-model, see [4]). This ratio increases with the open quotient and its range increases with R_k as shown in the following $1dB$ approximation : $H_1 - H_2 \simeq 12(\frac{O_q}{0.7})^2(1-(1-\frac{R_k}{0.7})^2) - 6$, valid for medium values of O_q.

The analytic formula for the complex spectrum of $U_k(t)$ is also useful for studying the phase spectrum, and thus the time-domain glottal pulse shape. One can show that the phase of $\tilde{U_k}(\nu)$ 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. If this component is removed, only a non-linear component remains. This component is linked to the glottal flow shape. Such a phase spectrum is close to the phase gain of a low pass filter, except that it corresponds to an anticausal impulse response. As a matter of fact, removing the linear phase component is equivalent to shifting the whole waveform in time domain. Then the waveform becomes anticausal.

In summary, we found that, using the analytic formulation of the spectrum, it is possible to design an all-pole filter which is comparable to the KLGLOTT88 model, or the LF-model. This filter is a 3rd order low-pass filter, with an anticausal pair of poles, and a simple real pole. The simple real pole is given directly by the f_t parameter of the KLGLOTT88 model. The other pair of pole is directly linked to the open quotient O_q of the KLGLOTT88 model. It is better to model the low-frequency region using a pair of complex-conjugate poles instead of two real poles. This better represents the "glottal formant" that can be observed in actual speech.

Finally, if one wants to preserve the glottal pulse shape it is necessary to design an anticausal filter. If one wants to preserve the finite duration property of the glottal pulse, it is necessary to truncate the impulse response of the filter.

3. GLOTTAL FLOW SPECTRAL PARAMETERS ESTIMATION

As a result of this linear source model, the estimation of the glottal formant may be seen as a filter estimation procedure. Another solution is to consider the ratio $H_1 - H_2$, which gives O_q, and then the glottal formant. Both solutions are considered here.

3.1. Source estimation

The source/filter digital model of speech production is $S(z) = U_g(z)V(z)L(z)$, where $S(z)$ is the speech signal, $U_g(z)$ is the glottal source, $V(z)$ is the vocal tract filter, and $L(z)$ represents lip radiation. The effect of $L(z)$ is considered to be a derivation, $V(z)$ is an all-pole filter. In spectral domain $U_g(z)$ is usually taken as a 2^{nd} order filter with 2 real poles. Our theoretical study and experiments on natural speech indicated that a filter with resonant poles better fits both the impulse response and the spectral envelope for

$U_g(z)$. Then we used the linear filter described above for the glottal source.

The following inverse filtering procedure has been used to derive the source signal from a given speech signal : 1/ estimation of the vocal tract transfert function : use a preemphasized version of the signal, $S_p(z) = S(z)/\tilde{U}(z)$, where $\tilde{U}(z) \sim U_g(z)L(z)$ is a derivation filter; 2/ suppression of the effect of lip radiation : use a deemphasized version of the signal, $S_d(z) = S(z)/\tilde{L}(z)$, where $\tilde{L}(z) \sim L(z)$ is an integration filter ; 3/ the source signal is obtained as : $\tilde{U_g}(z) = S_d(z)/S_p(z) \sim U_g(z)$. From a practical point of view, $\tilde{U}(z)$ is a single real zero digital filter, $\tilde{L}(z)$ is the inverse of $\tilde{U}(z)$, and $\tilde{U_g}(z)$ is computed as $S_d(z)$ filtered by the LPC inverse filter of $S_p(z)$.

3.2. Open quotient estimation

Once the source signal has been estimated, the next step is to derive the glottal formant frequency. This is done by matching a second order filter on the signal, for instance by an LPC procedure, or by an envelope fitting procedure of a digital or analog filter. If LPC is used, care must be taken to deduce the glottal formant frequency f_g from the pole values given by the LPC. In fact, in the previous theoretical development, f_g is the cutoff frequency of an analog filter whereas the LPC gives a digital filter. The procedure we used to compute the estimated f_g from the location of the poles is as follows : compute the resonance frequency f_r of the digital estimated filter from the complex poles $p = \rho e^{\pm j\theta}$ (this is always possible because the filter is always resonant) :

$$f_r(\rho,\theta) = \frac{1}{2\pi}\arccos(\frac{1}{2}(\rho+\frac{1}{\rho})\cos(\theta))$$

then fit an analog filter with same resonance frequency and amplitude, which gives the cutoff frequency of the analog filter :

$$f_g(\rho,\theta) = f_r(\rho,\theta)/\sqrt[4]{1-(\frac{\sin(\theta)(1-\rho^2)}{1-2\rho\cos(\theta)+\rho^2})^2}$$

The estimated open quotient is then deduced from equation (1). It must be noticed that estimation of the open quotient requires the knowledge of the fundamental frequency f_0, whereas the glottal formant is independant of f_0.

A second approach has been tried, in which the open quotient O_q is directly derived from measurement of the amplitudes H_1 and H_2 of the 2 first harmonics of the source signal. The difference $H_1 - H_2$ is related to laryngealisation (see [7] and [6]), and we established (see [3]) the relationship between $H_1 - H_2$ and O_q. This relation is not inversible, so an iterative procedure such as Newton's must be used. Estimation of H_1 and H_2 is achieved by estimation of the raw frequency of the harmonics, given the fundamental frequency. Then a precise value of both their frequency and amplitude is computed by an interpolation procedure.

4. EXPERIMENTS

Some experiments have been carried out on synthetic and natural speech with both approaches. The first experiment validates the open quotient estimation procedure with a known source : a KLGLOTT88 waveform with varying open quotient is generated as a source signal, and the estimated open quotient is compared to the true one (see figure 1). The same procedure is tested on synthetic vowels, again with a known source : the KLGLOTT88 waveform is passed through a vowel filter which simulates the vocal tract, for 5 French vowels. On figure 1 are shown the evolution against time of the "true" open quotient and the estimated open quotient with both approaches. One may notice that the estimation is less robust when the first formant frequency is near the glottal formant frequency (for instance in the vowel [i] or [u]), and when the fundamental frequency is high.

The estimation scheme has also been tested on natural speech. An example is given in figure 2 where the evolution of the open quotient is plotted against time. Other voice source parameters are also plotted for the same sentence in figure 2 : $f0$, short-term energy of the periodic component, and short-term energy of the aperiodic component. This allows for estimation of AV, and of AN. Thus all the time-domain voice source parameters (and the spectral-domain voice quality parameters) are estimated, spectral tilt excepted. Covariation of source parameters is noticeable in figure 2. for instance at the end of the sentence, the periodic component energy decreases, the aperiodic component energy increases, the fundamental frequency increases, the open quotient increases. The variations of all these parameters indicate a lower vocal effort.

5. CONCLUSION

In this paper we present algorithms for voice quality analysis. First a decomposition of the periodic and aperiodic components of the signal is performed. The periodic/aperiodic ratio and the amplitude of voicing are a first set of voice quality parameters obtained with this procedure. Other parameters are derived in the spectral domain, with the help of an analytic study of glottal flow models spectra. We focussed on estimation of the open quotient parameter. It is performed by using all pole modelling of the speech signal and of the estimated source, or by deducing its value from the amplitude difference of the 2 first harmonics. Some experiments have been carried out to show the accuracy of the estimations. Future work will be devoted to spectral tilt estimation, and to modification of voice quality through

modification of voice source parameters in natural speech segments.

REFERENCES

[1] C. d'Alessandro, B. Yegnanarayana, and V. Darsinos. "Decomposition of speech signals into deterministic and stochastic components." Proc. ICASSP'95, 760–763.

[2] V. Darsinos, C. d'Alessandro, and B. Yegnanarayana. "Evaluation of a periodic/aperiodic speech decomposition algorithm." Proc EUROSPEECH'95, 393–396.

[3] B. Doval and C. d'Alessandro. "Spectral correlates of glottal waveform models: an analytic study." Proc. ICASSP'97, 1295–1298.

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

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

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

[7] 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.

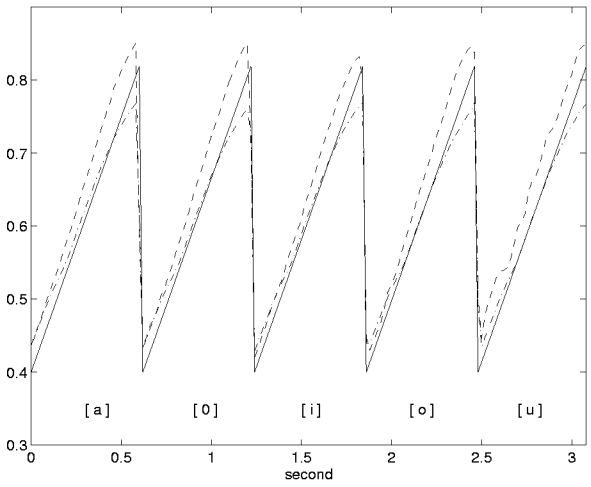

Figure 1. Estimation of the open quotient O_q for a synthetic KLGLOTT88 model source. True O_q (solid line), estimated O_q on the source component only (dotted line), and estimated O_q on the source component passed throught vowel filters (dashed line). $f_0 = 100Hz$. The signal is made of 5 french vowels. During each vowel, the open quotient is varied from 0.4 to 0.83. The source signal is estimated using a 18 order LPC, and the open quotient is estimated through the measurement of $H_1 - H_2$.

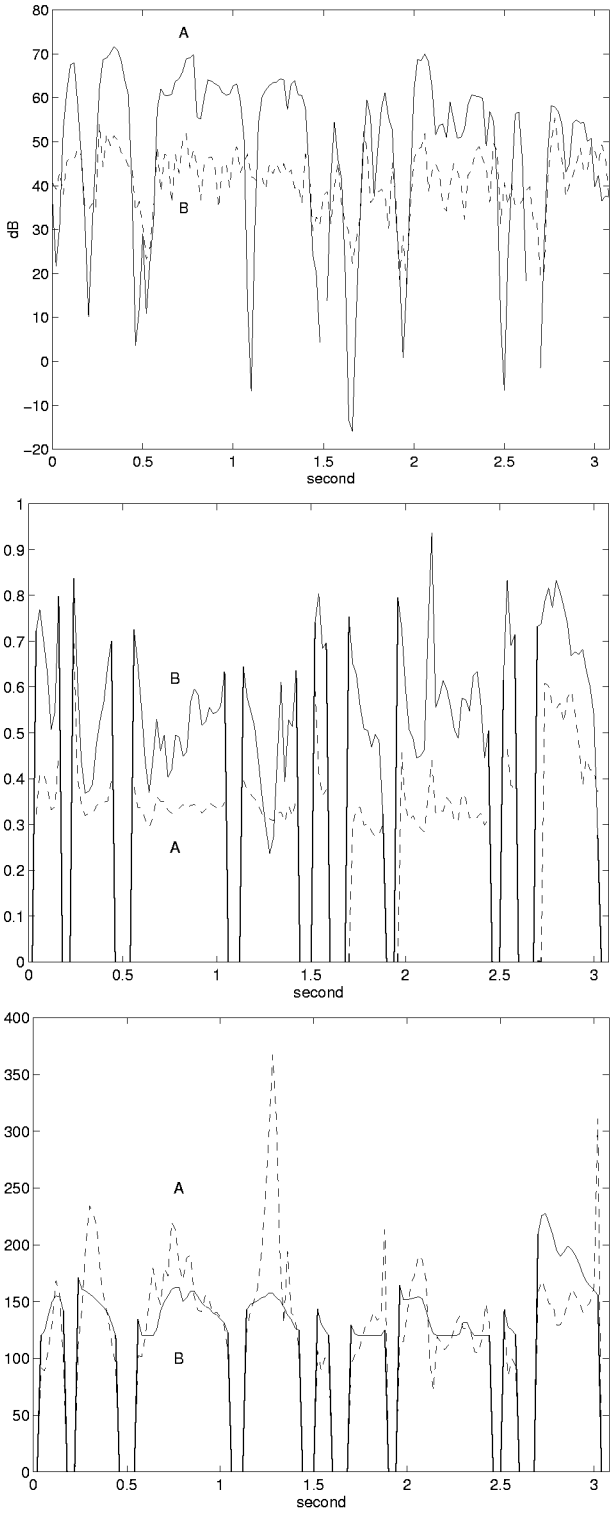

Figure 2. Estimation of voice source parameters on the French sentence : "Je pense que Marie et Jean n'accepteront pas de dire des choses pareilles". Top : energy in the periodic (A) and the aperiodic (B) components. Middle: estimation open quotient (A) 1st method (B) second method. A difference in normalization is noticeable. Bottom : fundamental frequency (B) and glottal formant frequency (A).