VALIDATION OF TEMPORAL & SPECTRAL NOISE PARAMETERS USING (RE)SYNTHESIS.

Peter Pabon° & Guus de Krom
Research Institute for Language and Speech, University of Utrecht, the Netherlands.
°Institute for Sonology, Royal Conservatory, the Hague, the Netherlands.

ABSTRACT

Our main objective in this study was to develop a numerical algorithm or processing scheme that, starting from normal speech samples, produces a natural rough or breathy sound quality. Three methods are presented that illustrate the limited validity of spectral models for period-to-period variation and noise.

INTRODUCTION

Apart from being a playful effect processor, a "roughner" or a "breathinizer" has a serious use in the calibration of algorithms for voice-quality measurement. Although the auditory effects produced by such a processor may seem obvious or evident, the underlying models are often not. To produce a natural rough or breathy quality, the corresponding acoustic characteristics must also be modelled very accurately. Knowing the credentials of the acoustic effect means knowing how to develop the ultimate quality measurement device. Apart from being fun, the option of evaluating the quality by listening to the results makes the research less abstract and more effective. Often, experimenting with test signals quickly helps to define improvements of a model.

Usually, the only criterion for the validity of an acoustic voice-quality parameter is the comparison to perceptual or clinical ratings. The correlations that are found generally indicate only global relations. As a result, no clues, only educated guesses can be given how improvements should be made.

In this paper we try to take a by-way around this approach by using analysis/resynthesis based on the Overlap-Add (OLA) method [1]. With this method, much can be said on forehand about the quality of a model or spectral rating. Modern DSP technology allows a real-time implementation of the OLA method. OLA is a challenging research instrument which allows complex models that use elaborate processing schemes to be judged interactively.

The OLA method

The OLA method is an extension of the general method of spectrum analysis in which a time-varying spectral representation is derived by processing successive time frames. In the OLA method, the inverse process is also formalised. A frequency-to-time-domain transformation followed by an addition of subsequent frames restores the continuous time signal. OLA enables us to perform specific operations on the time-signal by manipulating the related spectral features.

A major limitation of the OLA method is that every spectral modification should represent a valid manipulation of the related time-domain signal. For instance, too rigorous cutting in the amplitude spectrum could disrupt the window-structure in the time domain and thereby produce discontinuities in the resynthesized output signal. Each spectral modification should prevent the complex liaison of amplitude and phase information from being disunited. To guarantee this, each spectral (re)organisation should always represent a realistic time-domain principle like filtering, correlation, shifting, integration, and so on. It is questionable if this restraint is a draw-back. Bizarre spectral or cepstral models could be used to rate voice-quality, but when their corresponding time-domain representation is intangible their relevance will probably be difficult to validate.

Our main approach in the simulation of a rough or breathy quality is a perturbation of the phase-relations of a periodic signal, while keeping the amplitude spectrum unchanged. Although the processing is done in the frequency domain, the main effect will be in the time domain.

Method I, Jittering

The first method uses OLA to randomly vary the linear phase component. The time domain result is a time shift of the entire frame. The frequency domain implementation allows for window shifts in fractions of the sample period. This variable window displacement produces a jittered version of the original (see Fig.1).

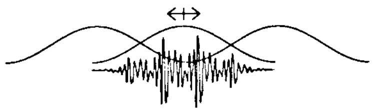

Figure 1, Jittering by random positioning of the frame.

The shape of the window/taper will prevent the emergence of hard discontinuities. The slight phase mismatch is smoothed by the fade-in-fade-out curve of the connected frames. This method is generally used in granular synthesis of single sound samples. If the window is periodically matched to the pitch period, this technique corresponds to a PSOLA method with a slightly perturbed periodicity.

The phase perturbation is proportional for all frequency components and is typically located around the frame boundaries. Seen over the frame, the amplitude spectrum is unaltered. Seen over an analysis period that includes more frames, the harmonic structure is slightly shattered, but the overall spectral envelope remains the same.

Method II, phase perturbation

In the first method, the entire frame was shifted as a block. In the second method, the frame positioning is randomised per frequency band. This effect is realised by convolving the signal with white noise. For band-limited random modulation of the phase, this is the best approach as it will guarantee continuity in both domains. Another way to look at this process, is to consider it as an all-pass filtering technique, where the phase delay of the all-pass filter is updated dynamically with each frame. Basically, the process is a cross-correlation or phase-vocoding technique [2] using white noise as a modulator (see Figure 2).

As can be seen in Figure 2, the amplitude spectrum is wobbly, but still remains maximally flat given the chosen phase curve. The overall spectral envelope is preserved and the speech output is still intelligible, although the quality is severely rough. The random time shift per frequency band breaks up all time-synchronisation that used to lead to sharply defined periodic excitation moments. If the low frequency part of the spectrum is processed in this way, our perception of a clear pitch is largely gone.

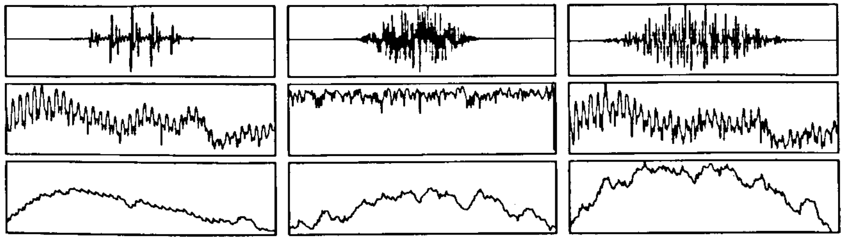

Figure 2, Perturbation by multiplying the signal spectrum with a white noise spectrum. (top) time signals, (middle) log spectra, range 80 dB, (bottom) phase spectra, range 16π, (left) speech signal, (middle) white noise and (right) noise modulated speech signal.

Phase versus amplitude information

Both methods were based on the principle that the periodicity information held by the phase spectrum was altered while the periodicity information in the amplitude spectrum was largely preserved. In Figure 2, the shape of the harmonics in the amplitude spectrum is deformed, but their average spacing remains the same. The auto-correlation-function, the cepstrum, even the complex cepstrum that also includes phase information, could still show a peak that indicates a clear fundamental periodicity, while the remains of this periodicity in the time-domain are completely lost by the randomisation of the phase information. Of course, the situation that the phase information is largely scrambled while the amplitude information is not, is not likely to appear to that extent in natural signals.

Still, the fact that large perturbations can occur without a corresponding large modification of the harmonic structure in the amplitude spectrum shows that every perturbation or noise model that is solely based on this amplitude representation is incomplete.

Additional information can of course be found in the regularity of the phase spectrum. However, it is questionable if the FFT phase spectrum is a good candidate for a separate model of (a)periodicity. Even if an elegant phase-unwrapping method is available, any correlational measure based on FFT phase information will suffer from the fact that a large part of the phase spectrum is non-deterministic. Overall, the curve may seem smooth, but, when inspected on an enlarged scale, the curve is jagged. For the FFT phase spectrum, a strange duality exist. If there are spectral components that have a clearly defined phase, they are more likely to be overshadowed by neighbouring spectral components that have not. The more concentrated the information due to periodicity, the less defined (and thus the more jumpy) the phase values of the other undefined components in between. A comparable principle is found with the cepstrum; the sharper the harmonics, the sharper and more jumpy the dips in the log-spectrum and the more noisy the base line of the cepstrum.

Phase/frequency stability is the base for the concentration of spectral energy in harmonics. The Fourier transform translates stability to a distinct spectral amplitude, thereby leaving only a superficial footprint on the mostly irregular terrain of the phase spectrum. On this terrain, the harmonics form stepping stones at which FFT phase information makes sense.

Apart from the models shown above, the amplitude spectrum can misrepresent what we consider noise or periodic on more occasions. For instance, the amplitude spectrum of a periodic noise burst can show a nice harmonic structure, while the noise within the bursts is not correlated, only the envelope (see Fig. 3).

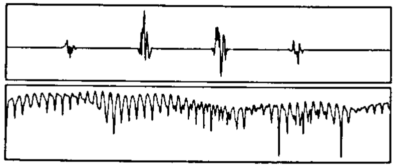

Figure 3, Windowed periodic bursts of white noise (top) and the corresponding logarithmic amplitude spectrum (bottom), vertical scale range 40 dB.

For a signal as in Figure 3, a repetition pitch can be perceived, and a harmonic series can be expected. However, it does illuminate an important question in the design of spectral models for voice (a)periodicity and noise: *In what way is the information on period-to-period correlation represented in the harmonic series?* This question is of relevance for the modelling of a breathy voice quality for which a period-linked noise burst seems an important attribute. It even

seems that the time synchronisation is not the only prerequisite, more complex correlations within the ensemble [3] are likely to play a role. A related idea is that any source (a)periodicity is linked to the period while any (a)periodicity linked to the tract/resonance could be judged on a different time scale. The above question is also of importance for the definition of harmonic-to-noise measures. Any reference to levels in between harmonics is a reference to a noise that is likely to be uncorrelated to the periodicity, e.g. to be randomly distributed over the analysis frame. Depending on the number of periods in the frame, e.g. the harmonic density, this noise can originate from many sources, which makes it an unreliable reference. This does not mean that such a parameter is insensitive, to the contrary, but it is questionable what it discriminates. Again, the overlap-add method allows a check of the above questions by resynthesizing both groups of information.

Method III, killing periodicity

The Fourier analysis principle is based on a phase stability criterion. If a frequency component is stable, it will match to a center-frequency of a band-filter and thus lead to a cumulative result over a given amount of time yielding a spectral peak. To see how effective the tuning/adding is, we have two options: (A) compare amplitudes between two adjoining spectral frames [4], or (B) check the stability of the phase curve as a function of band-filter frequency. Phase stability is weighted in the group delay phase (GDP) function, the differentiated phase curve. The GDP function is a good candidate to mark and thus to remove harmonics from a spectrum. The spectral amplitude is high only due to band-filter phase matching/stability, but the absolute level has no influence on the stabilising principle. A comb-filter design need not be based on a strictly regular pattern in the harmonic series (the pitch).

In our implementation, each harmonic is attenuated using a non-linear mapping function. The resulting output signal shows different degrees of periodicity killing, depending on the averaging time and stability threshold used. In general, the non-harmonic residue illustrates the inaptness of the spectral model to separate period-to-period aperiodicity from noise.

CONCLUSIONS

The first two methods demonstrate the additional role of phase information in the description of voice aperiodicity. The fact that the produced "perturbation" is often perceived as being synthetic, makes us question the completeness and even the validity of common spectral models for voice aperiodicity. The way phase stability is linked to spectral amplitude information, and the way this complex is condensed in the harmonics is vital in our description of period-to-period variability, and the definition of noise as being a non-harmonic residue. Spectral parameters that rate voice (a)periodicity should therefore also include phase (stability) information.

REFERENCES

[1] Allen, J.B., and Rabiner, L.R. (1977). A Unified approach to short-time Fourier analysis and synthesis., *Proc. of the IEEE*, (65), No. 11, pp 1558-1564 .

[2] Moorer, J.A. (1987). The use of phase vocoder in computer music applications, *J. of the Audio Engineering Society*, (26), No. 1/2, pp 42-45.

[3] Pabon, P (1994). A real-time singing voice synthesizer (Alto*). SMAC Proceedings.*. Royal Swedish Academy of Music, Issue No. 79, pp 288-293.

[4] Serra, X. and Smith, J. (1990) Spectral Modeling Synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition, *Computer Music Journal*, (14,) No. 4, pp 12-24.