Amplitude Modulated Sinusoidal Models for Audio Modeling and Coding

Mads Græsbøll Christensen*, Søren Vang Andersen, and Søren Holdt Jensen

Department of Communication Technology,
Aalborg University, Denmark
{mgc,sva,shj}@kom.auc.dk

Abstract. In this paper a new perspective on modeling of transient phenomena in the context of sinusoidal audio modeling and coding is presented. In our approach the task of finding time-varying amplitudes for sinusoidal models is viewed as an AM demodulation problem. A general perfect reconstruction framework for amplitude modulated sinusoids is introduced and model reductions lead to a model for audio compression. Demodulation methods are considered for estimation of the time-varying amplitudes, and inherent constraints and limitations are discussed. Finally, some applications are considered and discussed and the concepts are demonstrated to improve sinusoidal modeling of audio and speech.

1 Introduction

In the last couple of decades sinusoidal modeling and coding of both speech and audio in general has received great attention in research. In its most general form, it models a segment of a signal as a finite sum of sinusoidal components each having a time-varying amplitude and a time-varying instantaneous phase. Perhaps the most commonly used derivative of this model is the constant-frequency constant-amplitude model known as the basic sinusoidal model. This model is based on the assumptions that the amplitudes and frequencies remain constant within the segment. It has been used for many years in speech modeling and transformation [1]. The model, however, has problems in modeling transient phenomena such as onsets, which causes so-called pre-echos to occur. This is basically due to the quasi-stationarity assumptions of the model being violated and the fundamental trade-off between time and frequency resolution. Also, the use of overlap-add or interpolative synthesis inevitably smears the time-resolution.

Many different strategies for handling time-varying amplitudes have surfaced in recent years. For example, the use of time-adaptive segmentation [2] improves performance greatly at the cost of increased delay. But even then pre-echos may still occur in overlap regions or if interpolative synthesis [1] is used. Also, the use of exponential dampening of each sinusoid has been extensively studied [3–5], although issues concerning quantization remain unsolved. Other approaches include the use of one common dampening factor for all sinusoids [6], the use of

* This work was conducted within the ARDOR project, EU grant no. IST–2001–34095.

asymmetric windows [7], the use of an envelope estimated by low-pass filtering of the absolute value of the input [8] and the approaches taken in [9, 10]. In [9] lines are fitted to the instantaneous envelope and then used in sinusoidal modeling, and in [10] transient locations are modified in time to reduce pre-echo artifacts. The latter requires the use of dynamic time segmentation. Also, tracking of individual speech formants by means of an energy separation into amplitude modulation and frequency modulation (FM) contributions has been studied in [11–13].

In this paper we propose amplitude modulated sinusoidal models for audio modeling and coding applications. The rest of the paper is organized as follows: In section 2 the mathematical background is presented. A general perfect reconstruction model is derived in section 3, and in section 4 a model which addresses one of the major issues of audio coding regardless of type, namely pre-echo, is presented along with a computationally simple estimation technique. Finally, in section 5 some experimental results are presented and discussed and section 6 concludes on the work.

2 Some Preliminaries

The methods proposed in this paper are all based on the so-called analytic signal, which is derived from the Hilbert transform. First, we introduce the Hilbert transform and define the analytic signal and the instantaneous envelope. Then we briefly state Bedrosian's theorem, which is essential to this paper.

Definition 1 (Discrete Hilbert Transform). *Let $x_r(n)$ be a discrete real signal. The Discrete Hilbert transform, $\mathcal{H}\{\cdot\}$, of this, denoted $x_i(n)$, is then defined as (see e.g. [14])*

$$x_i(n) = \mathcal{H}\{x_r(n)\} = \sum_{m=-\infty}^{\infty} h(m)x_r(n-m) \ . \tag{1}$$

where $h(n)$ is the impulse response of the discrete Hilbert transform given by

$$h(n) = \begin{cases} \frac{2\sin^2(\pi n/2)}{\pi n}, & n \neq 0 \\ 0, & n = 0 \end{cases} \ . \tag{2}$$

A useful way of looking at the Hilbert transform, an perhaps a more intuitive definition, is in the frequency domain:

$$X_i(\omega) = H(\omega)X_r(\omega), \quad \text{with} \quad H(\omega) = \begin{cases} j, & \text{for} \quad -\pi < \omega < 0 \\ 0, & \text{for} \quad \omega = \{0, \pi\} \\ -j, & \text{for} \quad 0 < \omega < \pi \end{cases} \ . \tag{3}$$

where $X_i(\omega)$ and $X_r(\omega)$ are the Fourier transforms (denoted $\mathcal{F}\{\cdot\}$) of $x_i(n)$ and $x_r(n)$, respectively, and $H(\omega)$ is the Fourier transform of $h(n)$. The so-called analytic signal and instantaneous envelope are then defined as

$$x_c(n) = x_r(n) + jx_i(n) \quad \text{and} \quad |x_c(n)| = \sqrt{x_r^2(n) + x_i^2(n)} \ , \tag{4}$$

respectively. With these definitions in place, we now state Bedrosian's theorem [15].

Theorem 1 (Bedrosian). *Let $f(n)$ and $g(n)$ denote generally complex functions in $\ell^2(\mathbb{Z})$ of the real, discrete variable n. If*

1. *the Fourier transforms $F(\omega)$ of $f(n)$ is zero for $a < |\omega| \leq \pi$ and the Fourier transform $G(\omega)$ of $g(n)$ is zero for $0 \leq |\omega| < a$, where a is an arbitrary positive constant, or*
2. *$f(n)$ and $g(n)$ are analytic, then*

$$\mathcal{H}\{f(n)g(n)\} = f(n)\mathcal{H}\{g(n)\} \ . \tag{5}$$

For proof of the continuous case see [15]. The theorem holds also for periodic signals in which case the Fourier series should be applied.

3 Sum of Amplitude Modulated Sinusoids

In this section we consider a perfect reconstruction framework based on a model consisting of a sum of amplitude modulated sinusoids:

$$\hat{x}(n) = \sum_{q=1}^{Q} \gamma_q(n) A_q \cos(\omega_q n + \phi_q) \quad \text{for} \quad n = 0, \ldots, N-1 \ , \tag{6}$$

where $\gamma_q(n)$ is the amplitude modulating signal, A_q the amplitude, ω_q the frequency, and ϕ_q the phase of the qth sinusoid. We note in the passing that the aforementioned exponential sinusoidal model [3–5] fall in to this category.

Assume that the signal has been split into a set of subbands by a perfect reconstruction nonuniform Q-band filterbank such as [16] having a set of cut-off frequencies Ω_q for $q = 0, 1, \ldots, Q$ where $\Omega_0 = 0$ and $\Omega_Q = \pi$. Then we express the contents of each individual subband $x_q(n)$ as an amplitude modulated sinusoid placed in the middle of the band, i.e.

$$x_q(n) = \gamma_q(n) A_q \cos(\omega_q n + \phi_q) = \gamma_q(n) s_q(n) \ , \tag{7}$$

where $\omega_q = \frac{\Omega_q + \Omega_{q-1}}{2}$, $\gamma_q(n) \in \mathbb{C}$, i.e. the modulation is complex. We start our demodulation by finding the analytic signal representation of both the left and right side of the previous equation:

$$\gamma_q(n)s_q(n) + j\mathcal{H}\{\gamma_q(n)s_q(n)\} = x_q(n) + j\mathcal{H}\{x_q(n)\} \ , \tag{8}$$

which according to Bedrosian's theorem is equal to

$$\gamma_q(n)s_q(n) + j\mathcal{H}\{\gamma_q(n)s_q(n)\} = \gamma_q(n)\left(s_q(n) + j\mathcal{H}\{s_q(n)\}\right) \tag{9}$$

$$= \gamma_q(n) A_l \exp\left(j(\omega_q n + \phi_q)\right) \ . \tag{10}$$

This means that we can simply perform complex demodulation in each individual subband using a complex sinusoid, i.e.

$$\gamma_q(n) = (x_q(n) + j\mathcal{H}\{x_q(n)\}) \frac{1}{A_q} \exp(-j(\omega_q n + \phi_q)) \ . \tag{11}$$

In this case we have a modulation with a bandwidth equal to the bandwidth of the subband, $\Delta_q = \Omega_q - \Omega_{q-1}$.

It is of interest to relax the constraint on the frequency of the carrier. Here we consider a more general scenario, where the carrier may be placed anywhere in the subband, i.e. $\Omega_{q-1} \leq \omega_q \leq \Omega_q$. In this case, the modulation is asymmetrical around the carrier in the spectrum. An alternative interpretation is that the carrier is both amplitude and phase modulated simultaneously.

Alternatively, we can split the modulation into an upper (usb) and a lower sideband (lsb). These can be obtained by calculating the analytic signal of $\gamma_q(n)$ and $\gamma_q^*(n)$, which is similar to zeroing out the negative frequencies:

$$\gamma_{q,usb}(n) = \frac{1}{2} \left(\gamma_q(n) + j\mathcal{H}\{\gamma_q(n)\}\right) \tag{12}$$

$$\gamma_{q,lsb}(n) = \frac{1}{2} \left(\gamma_q^*(n) + j\mathcal{H}\{\gamma_q^*(n)\}\right) \ . \tag{13}$$

The complex modulating signal can be reconstructed as

$$\gamma_q(n) = \gamma_{q,usb}(n) + \gamma_{q,lsb}^*(n) \ . \tag{14}$$

The modulating signal can be written as $\gamma_q(n) = C + b(n)$, where $b(n)$ is zero mean. For $C \neq 0$, this is the case where the sinusoidal carrier is present in the spectrum in the form of a discrete frequency component. For the special case that $C = 0$, we have what is known as suppressed carrier AM, i.e. the carrier will not be present in the spectrum. In the context of speech modeling this representation may be useful in modeling non-tonal parts, e.g. unvoiced speech, whereas the non-suppressed AM ($C \neq 0$) case may be well-suited for voiced speech.

In the particular case that the modulating signal is both non-negative and real, i.e. $\gamma_q(n) \in \mathbb{R}$ and $\gamma_q(n) \geq 0$, the demodulation simply reduces to

$$\gamma_q(n) = \frac{1}{A_q} |x_q(n) + j\mathcal{H}\{x_q(n)\}| \ , \tag{15}$$

as the instantaneous envelope of the carrier is equal to 1. This last estimation is lossy as opposed to the previous demodulations. Notice that in the perfect reconstruction scenario, the filtering of the signal into subbands and subsequent demodulation can be implemented efficiently using an FFT.

An alternative to the filterbank-based sum of amplitude modulated sinusoids scheme, which requires that the sinusoidal components are well spaced in frequency is the use of periodic algebraic separation [17, 18]. This allows for demodulation of closely spaced periodic components provided that the periods are known.

4 Amplitude Modulated Sum of Sinusoids

In this section a model for audio compression is introduced. This model addresses one of the major problems of audio coding regardless of type, namely pre-echo control. The perfect reconstruction model of the previous section has an amplitude modulating signal of each individual sinusoid. Here, we explore the notion of having more sinusoids in each subband and that modulating signal being identical for all sinusoids in the subband. This is especially useful in the context of modeling onsets and may even be used in the one-band case for low bit-rate or single source applications. The model of the qth subband is:

$$\hat{x}_q(n) = \gamma_q(n) \sum_{l=1}^{L_q} A_{q,l} \cos(\omega_{q,l} n + \phi_{q,l}) = \gamma_q(n) \hat{s}_q(n) \ , \tag{16}$$

where $\hat{s}_q(n)$ is the constant-amplitude part. In the one-band case where $x_q(n) = x(n)$ the models in [6–9] all fall into this category. These, however, do not reflect human sound perception very well as pre-echos may occur in the individual critical bands (see e.g. [19]). Neither do they take the presence of multiple temporally overlapping sources into account. The sum of amplitude modulated sinusoids, however, does take multiple sources into account.

The basic principle in the estimation of the modulating signal $\gamma_q(n)$ is that it can be separated from the constant-amplitude part of our model $\hat{x}_q(n)$ under certain conditions. First we write the instantaneous envelope of equation 16, i.e.

$$|\hat{x}_q(n) + j\mathcal{H}\{\hat{x}_q(n)\}| = |\gamma_q(n)\hat{s}_q(n) + j\mathcal{H}\{\gamma_q(n)\hat{s}_q(n)\}| \ . \tag{17}$$

Since we are concerned here with sinusoidal modeling, we constrain the modulation to the case of non-suppressed carrier and the physically meaningful non-negative and real modulating signal. Equation (17) can then be rewritten using Bedrosian's theorem:

$$|\gamma_q(n)\hat{s}_q(n) + j\mathcal{H}\{\gamma_q(n)\hat{s}_q(n)\}| = \gamma_q(n)|\hat{s}_q(n) + j\mathcal{H}\{\hat{s}_q(n)\}| \ . \tag{18}$$

For this to be true, our constant-amplitude model and the amplitude modulation may not overlap in frequency, i.e. we have that the lowest frequency must be above the bandwidth, BW, of the modulating signal

$$BW < \min_l \omega_{q,l} \ . \tag{19}$$

Using this constraint, we now proceed in the estimation of the amplitude modulating signal $\gamma_q(n)$ by finding the analytic signal of the sinusoidal model

$$\hat{x}_{q,c}(n) = \sum_{l=1}^{L_q} A_l \gamma_q(n) \exp(j\phi_{q,l}) \exp(j\omega_{q,l} n) \ , \tag{20}$$

with subscript c denoting the analytic signal. We then find the squared instantaneous envelope of the model:

$$|\hat{x}_{q,c}(n)|^2 = \sum_{l=1}^{L_q} \sum_{k=1}^{L_q} \gamma_q^2(n) A_{q,l} A_{q,k} \exp(j(\phi_{q,k} - \phi_{q,l})) \exp(j(\omega_{q,k} - \omega_{q,l})n). \quad (21)$$

The squared instantaneous envelope is thus composed of a set of auto-terms ($l = k$) which identifies the amplitude modulating signal and a set of interfering cross-terms ($l \neq k$). From this it can be seen that the frequencies of these cross-terms in the instantaneous envelope is given by the distances between the sinusoidal components. Thus, the lowest frequency in the squared instantaneous envelope caused by the interaction of the constant-amplitude sinusoids is given by the minimum distance between two adjacent sinusoids.

A computationally simple approach is to reduce the cross-terms by constraining the minimum distance between sinusoids and then simply lowpass filter the squared instantaneous envelope of the input signal, i.e.

$$\gamma_q^2(n) = \alpha e_q^2(n) * h_{LP}(n) \ , \quad (22)$$

where $e_q^2(n) = x_q^2(n) + \mathcal{H}\{x_q(n)\}^2$, α is a positive scaling factor and $h_{LP}(n)$ is the impulse response of an appropriate lowpass filter with a stopband frequency below half the minimum distance between two sinusoids, i.e.

$$2BW < \min_{l \neq k} |\omega_{q,l} - \omega_{q,k}| \ . \quad (23)$$

This estimate allows us to find a amplitude modulating signal without knowing the parameters of the sinusoidal model a priori. This is especially attractive in the context of matching pursuit [20]. Note that the constraint in equation (23) is more restrictive than those of theorem 1.

The design of the lowpass filter is subject to conflicting criteria. On one hand, we want to have sufficient bandwidth for modeling transients well. On the other, we want to attenuate the cross-terms while having arbitrarily small spacing in frequency between adjacent sinusoids. Also these criteria have a time-varying nature. A suitable filter which can easily be altered to fit the requirements is described in [8]. Generally, the consequences of setting the cutoff frequency of the lowpass filter too low are more severe than setting it too high. Setting the cutoff frequency too high causes cross-terms to occur in $\gamma_q(n)$, which may result in degradation in some cases, whereas setting the cutoff frequency too low reduces the models ability to handle transients.

An alternative approach in finding $\gamma_q(n)$ would be to estimate the amplitude modulating signals of the individual sinusoids and then combine these according to frequency bands or sources.

5 Results and Discussion

The framework in section 3 has been verified in simulations to attain perfect reconstruction. The choice of model, whether it is some derivative of the sum

of amplitude modulated sinusoids or the amplitude modulated sum of sinusoids, should reflect signal characteristics. Types of sinusoidal signals that can be efficiently modeled using a one-band amplitude modulated sum of sinusoids are single sources that have a quasi-harmonic structure, i.e. pitched sounds. For example, voiced speech can be modeled well using such a model. In figure 1 two examples of onsets of voiced speech are shown (sampled at 8kHz) with the originals at the top, modeled without AM in the middle, and with at the bottom. The fundamental frequency was found using a correlation-based algorithm and the amplitudes and phases were then estimated using weighted least-squares. Segments of size 20ms and overlap-add with 50% overlap was used.

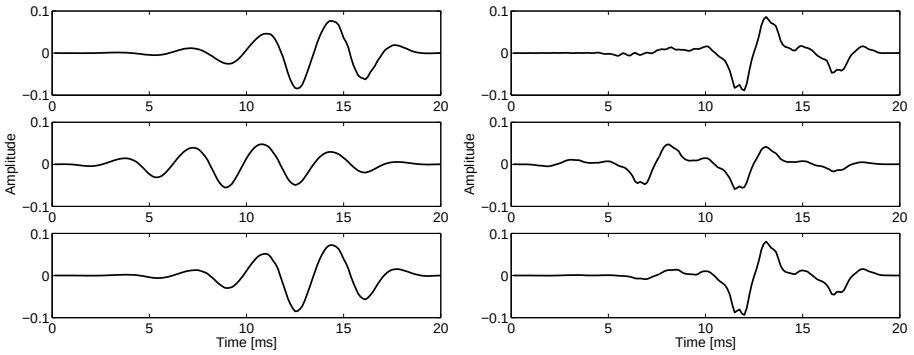

Fig. 1. Signal examples: voiced speech. Original (*top*), *modeled without AM (middle) and with AM (bottom)*

It can be seen that the pre-echo artifacts present in the constant-amplitude model are clearly reduced by the use of the AM scheme. The proposed model and estimation technique was found to consistently improve performance of the harmonic sinusoidal model in transient speech segments with pre-echo artifacts clearly being reduced.

More complex signals composed of multiple temporally overlapping sources, however, require more sophisticated approaches for handling non-stationarities. The glockenspiel of SQAM [21] is such a signal. At first glance this signal seems well suited for modeling using a sinusoidal model. The onsets are, however, extremely difficult to model accurately using a sinusoidal model. This is illustrated in figure 2, again with the original at the top, modeled using constant amplitude sinusoids in the middle and using AM at the bottom. The signal on the left is the entire signal and the signal on the right is a magnification of a transition region between notes. In this case amplitude modulation is applied per equivalent rectangular bandwidth (ERB) (see [19]) and a simple matching pursuit-like algorithm was used for finding sinusoidal model parameters, i.e. no harmonic constraints on the frequencies. Again overlap-add using segments of 20ms and 50% overlap was employed. In this example the sampling frequency was 44.1kHz.

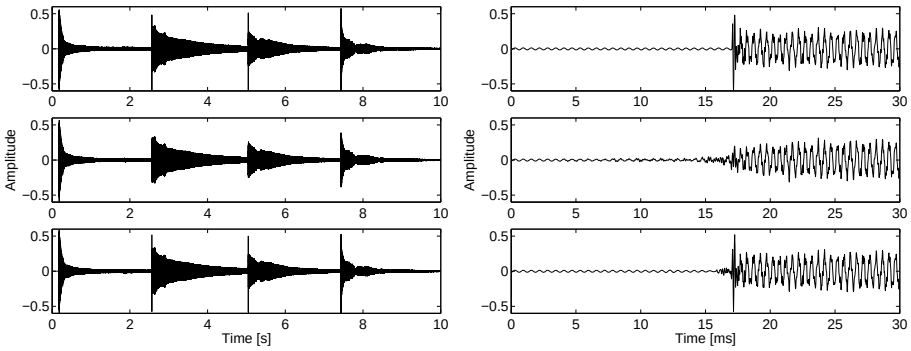

Fig. 2. Signal examples: glockenspiel. Original *(top), modeled without AM (middle) and with AM (bottom)*

It can be seen that the onsets are smeared when employing constant amplitude and that there is a significant improvement when AM is applied, although some smearing of the transition still occurs due to the filtering.

6 Conclusion

In this paper we have explored the notion of amplitude modulated sinusoidal models. First, a general perfect reconstruction framework based on a filterbank was introduced, and different options with respect to modulation and their physical interpretations were presented. Here, one sinusoid per subband is used and everything else in the subband is then modeled as modulation of that sinusoid. This model is generally applicable and can be used for modeling not only tonal signals but also noise-like signals such as unvoiced speech. Then a physically meaningful, compact representation for sinusoidal audio coding and modeling and a demodulation scheme with low computational complexity was presented. In this model, each subband is represented using a sum of sinusoids having one common real, non-negative modulating signal, which is estimated by lowpass filtering the squared instantaneous envelope. The model and the proposed estimation technique was found to be suitable for modeling of onsets of pitched sounds and was verified to generally improve modeling performance of sinusoidal models.

References

1. McAulay, R.J., Quatieri, T.F.: Speech Analysis/Synthesis Based on a Sinusoidal Representation. In: IEEE Trans. Acoust., Speech, Signal Processing. Volume 34(4). (1986)
2. Prandom, P., Goodwin, M.M., Vetterli, M.: Optimal time segmentation for signal modeling and compression. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. (1997)

3. Goodwin, M.M.: Matching pursuit with damped sinusoids. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. (1997)
4. Nieuwenhuijse, J., Heusdens, R., Deprettere, E.F.: Robust Exponential Modeling of Audio Signals. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. (1998)
5. Jensen, J., Jensen, S.H., Hansen, E.: Exponential Sinusoidal Modeling of Transitional Speech Segments. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. (1999)
6. Jensen, J., Jensen, S.H., Hansen, E.: Harmonic Exponential Modeling of Transitional Speech Segments. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. (2000)
7. Gribonval, R., Depalle, P., Rodet, X., Bacry, E., Mallat, S.: Sound signal decomposition using a high resolution matching pursuit. In: Proc. Int. Computer Music Conf. (1996)
8. George, E.B., Smith, M.J.T.: Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis-synthesis of musical tones. In: J. Audio Eng. Soc. Volume 40(6). (1992)
9. Edler, B., Purnhagen, H., Ferekidis, C.: ASAC – Analysis/Synthesis Audio Codec for Very Low Bit Rates. In: 100th Conv. Aud. Eng. Soc., preprint 4179. (1996)
10. Vafin, R., Heusdens, R., Kleijn, W.B.: Modifying transients for efficient coding of audio. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. (2001)
11. Maragos, P., Kaiser, J.F., Quatieri, T.F.: Energy Separation in Signal Modulations with Application to Speech Analysis. In: IEEE Trans. Signal Processing. Volume 41(10). (1993)
12. Bovik, A.C., Havlicek, J.P., Desai, M.D., Harding, D.S.: Limits on Discrete Modulated Signals. In: IEEE Trans. on Signal Processing. Volume 45(4). (1997)
13. Quatieri, T.F., Hanna, T.E., O'Leary, G.C.: AM-FM Sepration using Audiotory-motivated Filters. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. (1996)
14. Oppenheim, A.V., Schafer, R.W.: Discrete-Time Signal Processing. 1st edn. Prentice-Hall (1989)
15. Bedrosian, E.: A product theorem for Hilbert transforms. In: IEEE Signal Processing Lett. Volume 44(1). (1963)
16. Goodwin, M.M.: Nonuniform filterbank design for audio signal modeling. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. (1997)
17. Zou, M.Y., Zhenming, C., Unbehauen, R.: Separation of periodic signals by using an algebraic method. In: Proc. IEEE Int. Symp. Circuits and Systems. (1991)
18. Santhanam, B., Maragos, P.: Multicomponent AM-FM demodulation via periodicity-based algebraic separation and energy-based demodulation. In: IEEE Trans. Communcations. Volume 48(3). (2000)
19. Moore, B.C.J.: An Introduction to the Psychology of Hearing. 4th edn. Academic Press (1997)
20. Mallat, S., Zhang, Z.: Matching pursuit with time-frequency dictionaries. In: IEEE Trans. Signal Processing. Volume 40. (1993)
21. European Broadcasting Union: Sound Quality Assessment Material Recordings for Subjective Tests. EBU (1988) http://www.ebu.ch.