Hearing Impairment Simulation Model using the standard MPEG-1 Audio Layer III

Alejandro José Uriz, Pablo Daniel Agüero, Francisco Denk, Jorge Castiñeira Moreira, and Juan Carlos Tulli

Facultad de Ingeniería, Universidad Nacional de Mar del Plata, ,
Juan B. Justo 4302, 7600 Mar del Plata, Argentina
{ajuriz,pdaguero,casti}@fi.mdp.edu.ar
http://200.0.183.36/pegasus

Abstract. *The MPEG standard for digital compression for high-fidelity audio provides two implementations of the psychoacoustic model: the Psychoacoustic Model I and the Psychoacoustic Model II. Both of these models are explained in detail in the ISO/IEC 13818-3:1194(E). In this work a model of the auditory system will be obtained with the purpose of simulating several types of hearing impairments. The system is able to simulate several types of auditory problems from curves that represent the Sound Pressure Level as a function of the frequency.*

Key words: Hearing impairment, Psychoacoustic Model I, MPEG-1 Audio Layer 3, auditory model, voice compression.

1 Introduction

In our Society exist a big number of people with a hearing impairment. Hence, (in order to devise or probe any sort of helping strategy) it is necessary to obtain an artificial model to simulate several types of pathologies. The main idea of this work is to develop a tool to test technological devices to help hearing impaired people without the necessity of human volunteers in early stages of research. An example of this approach is made in Uriz et al. [1]. Another aspect in the development of this tool is to know *how a sound is perceived by a hearing impaired person?*. In this way, researchers will take into account these special considerations to develop devices according to the disabilities of the users.

The purposed system is based on an implementation of the MPEG-I Audio Layer 3 compression standard [2, 3] using Matlab, based on the Psychoacoustic Model I [4]. In this model, the curve **Sound Pressure Level** (*SPL*) as a function of the frequency is entered. This curve gives the minimum level of power that a signal of a specific frequency needs to be perceived by a person. This curve can be adapted to obtain a synthesized voice affected by a specific simulated hearing impairment.

This work is developed as follow: In Section 2 a theoretical introduction of the auditory model used in this work is developed, with the purpose that the reader acquires information about this model. In Section 3 the hearing impairments

that will be simulated are presented. In Section 4, the proposed algorithms to face the problem are developed, analyzing the advantages and disadvantages of each one. In Section 5, the results of subjectives experiments are presented and discussed. Finally, the main conclusions are summarized in Section 6 where the guidelines to follow in future work are also presented.

2 Model of Auditory System

The proposed system is based on an implementation of MPEG-I Audio Layer 3 compression standard [2, 3], which obtains high compression levels by considering the perceptual limitations of the human hearing system. Then, it removes the unnecessary information that the human hear can not perceive. Figure 1 shows a schematic of a generic coding system of this model.

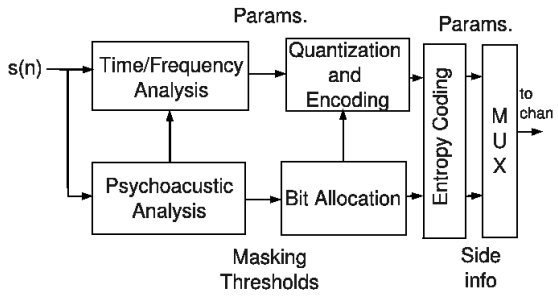

Fig. 1. Model of a perceptual audio encoder.

The system proposed in this paper only uses the Psychoacoustic model I and the filter banks of the coder [2, 3]. This decision is made because the main goal of the system is not to code the signal but to know the part of the information that is not necessary and must be eliminated. The most important aspects of this model are presented in the next Sections.

2.1 Absolute Threshold of Hearing (ATH)

The compression process starts with the analysis of the response of the auditory system at each of the input frequencies. It is made using the **Absolute Threshold of Hearing** (ATH), that is a function which represents the minimum amount of energy needed by a pure tone of a given frequency to be detected by a listener in a noiseless environment. The ATH is typically expressed in terms of dB SPL. The dependence of this threshold with the frequency was quantified by Fletcher [5], who reported test results for a range of listeners. In the particular case of a quiet environment, this curve is modeled by the Eq. 1.

$$T_q(f) = 3.64(\frac{f}{1000})^{-0.8} - 6.5e^{(f/1000-3.3)^2} + 10^{-3}(\frac{f}{1000})^4 \qquad (dB \quad SPL) \quad (1)$$

This threshold is representative of a young listener with acute hearing. A plot of this function is showed in Fig. 2.

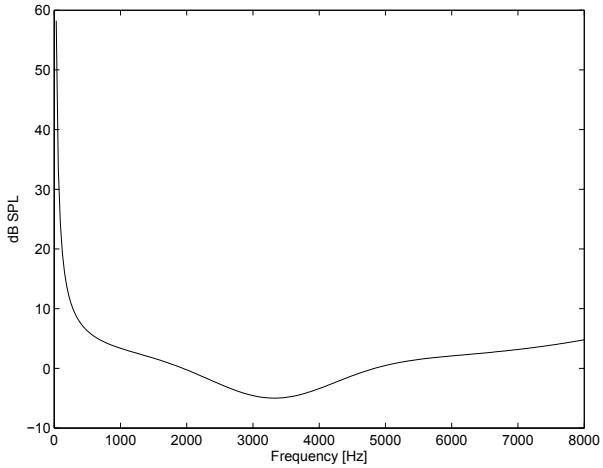

Fig. 2. The absolute threshold of hearing in quiet.

Function of Eq.1 is a general case, but in the case of a particular listener, it can be obtained through an audiometry, a medical study used to evaluate the frequency response of the ears of a listener. An audiometry is a test that can be used as a tool to obtain a model of an ear with a disability.

2.2 Hybrid filter bank

The hybrid filter bank is a polyphase filter bank composed by 32 subbands [6], which are narrower in the range between 2 and 4 kHz for the Layer 3. Due to this, the masking widths of the ear in these ranges are the slightest. The reason for this implementation is to reduce the *aliasing* [7] due to the laps between bands, which originates audible distortions.

2.3 Simultaneous Masking

Once the two most important parts of a auditory system are defined, it is necessary to present a process known as masking. This process refers to a process where one sound is rendered inaudible because of the presence of another sound. Simultaneous masking may also occur whenever two or more stimuli are simultaneously presented to the auditory system. The relative shapes in the magnitude

spectra of the masker and maskee determine to what extent the presence of certain spectral energy will mask another spectral energy. There are several types of masking between spectral components, but it is convenient to distinguish between only three types of simultaneous masking:

- **Noise Masking Tone (NMT):** In this case, a narrow band noise masks a tone within the same critical band, provided that the intensity of the masked tone is below a predictable threshold directly related to the intensity.
- **Tone Masking Noise (TMN):** In the case of TMN, a pure tone placed in the center of a critical band masks noise of any subcritical bandwidth or shape, provided that the noise spectrum is below a predictable threshold directly related to the strength of the masking tone.
- **Noise Masking Noise (Noise):** In this scenario, where a narrow-band noise masks another narrow-band noise, it is more difficult to characterize than in the case of NMT or TMN, because of the confounding influence of phase relationships between the masker and maskee.

The spectral masking is assimetrical, because the masking thresholds are not the same in each side of a component. This aspect is shown in Fig.3 where the masking limits are presented.

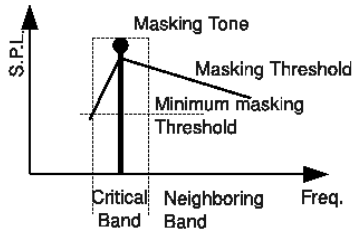

Fig. 3. Masker and its correspondent masking thresholds.

2.4 Non-Simultaneous Masking

Another type of masking is the non-simultaneous masking. It is a phenomena that extends along the time. For a masker of finite duration, this type of masking occurs because the ear needs a finite time since the masker dissapears and the next component appears. The ear requires a recover time to return to its normal state. This recover time is about 5ms for a healthy ear.

3 Hearing Impairments under study

In this Section some types of hearing impairments, and their representation using the Absolute Threshold of Hearing curve in each case are presented. In the first place the case of a **healthy ear of a young person**, which is shown in Fig. 2.

Simulation of a **severe deafness** is wanted to be simulated [8–10], where the auditory system does not respond to frequencies higher that a cutoff frequency of around 300Hz to 500Hz, can be modeled by increasing at least 50dB the ATH curve [8], in the range of hearing disability, as shown in Fig. 4. This impairment causes the loss of the majority of the components of a phoneme.

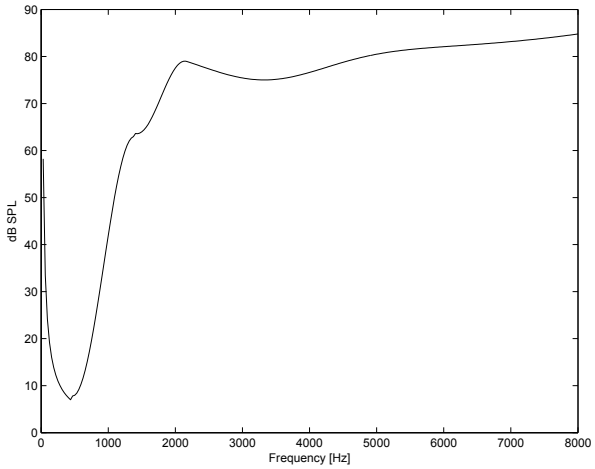

Fig. 4. An example of a curve of the absolute threshold of hearing with a severe deafness.

Another type of hearing impairment is shown in Fig. 5 where the ATH curve rises linearly as a function of the frequency due to a **sensorineural hearing loss** [11]. It causes a loss in the components of higher frequency.

Figure 6 shows the ATH curve of a person with a **bilateral acoustic trauma** [9, 10, 12], a hearing impairment that causes a rejection in a band of frequencies. It is a pathology characterized by a *dead zone*, and causes the loss of some types of critical sounds in phonemes like consonants as the /r/ and /s/, where the loss of critical components may cause that those two phonemes sound equal.

4 Proposed System

The system proposed in this paper is divided in two parts that work in parallel. The first stage takes an audio signal input, and processes it using the algorithm presented in Sec. 2. The block eliminates the unnecessary information[13, 14], by selecting the one that will be used by the second stage of the system to generate the simulated hearing impairment. On the other hand, the second part of the system takes the same audio input signal, and processes it by windowing the signal using a frame of **N=512** points (32ms, $f_{sampling}$ = 16KHz), then a **Fast Fourier Transform** (FFT) of 512 points is made. Once in the frequency

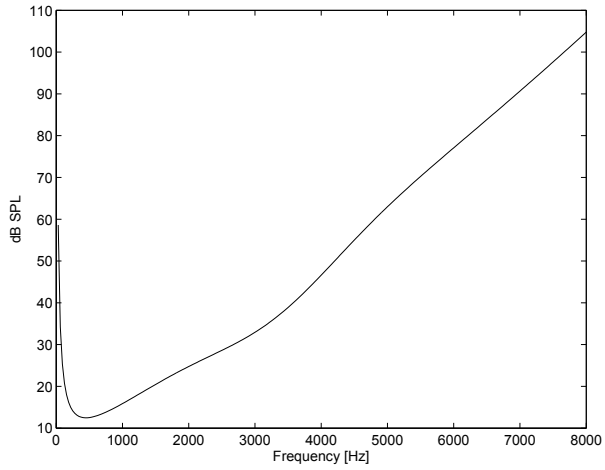

Fig. 5. The absolute threshold of hearing with a sensorineural hearing loss.

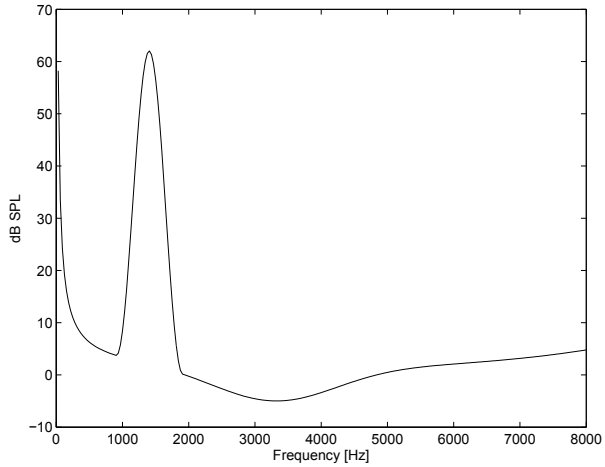

Fig. 6. The absolute threshold of hearing with a bilateral acoustic trauma.

domain, the unnecessary components are eliminated using the information of the first stage. Finally, an **Inverse Fast Fourier Transform** IFFT is made and the signal is resynthesized using the *Time Domain Overlapping and Add* method (TD-OLA). The resulting signal is the output of the system, affected by the corresponding hearing impairment. A scheme of this model is shown in Fig. 7.

When the system was implemented, a number of changes must be done. The signal under analysis is an audio signal, as described in [15, 16], that is analyzed using two types of windowing. For the zone in the middle of a phoneme, where it

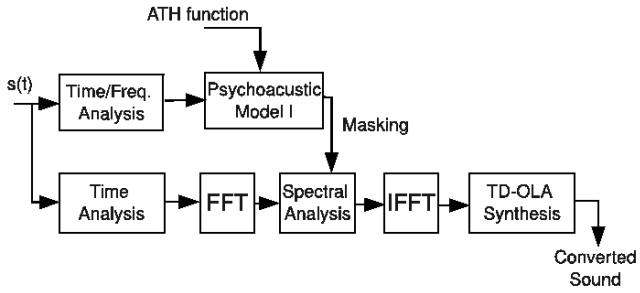

Fig. 7. Block diagram of the system proposed.

is assumed stationarity, the windowing is made using segments of around 32ms. On the other hand, in the border of the phoneme, narrower segments are used, because the signal is not as stable as in the center of the phoneme. The problems caused by this condition and the methods used to solve them are listed below.

4.1 Leakage and the Picket-fence Effect

The **leakage** appears due to a non-periodical signal is considerated as a periodical signal when the algorithm of the FFT is applied. This effect originates an spectral dispersion because in the analysis windows there are not an integer number of cycles of the signal. It is the case of the signals under study.

The picket-fence effect [15, 16] is a phenomena that appears in digital signal processing when the frequency of the component that is being processed is not an integer multiple of the sampling frequency divided by the quantity of bins of the FFT (N). Under these conditions, an error in the calculation of the FFT is appreciated because the relationship between frequencies is not a multiple of the resolution of the Fourier Transform of the signal. Such situation generates additional components due to that the original component is decomposed into components of different amplitudes. The frequency error e_f between the original component and the most important component originated by picket fence effect is shown in Eq. 2.

$$e_f = f_{IN} - \frac{f_{sampling}.k}{N} \tag{2}$$

f_{IN} is the frequency of the input component, $f_{sampling}$ the sampling frequency, N the bins of the FFT and k the index of the bin of the closest component generated by the picket-fence effect. On the other hand, the amplitude error e_A can be written as:

$$e_A = \frac{N}{2} \frac{\sin[\pi(k-m)]}{\pi(k-m)} \tag{3}$$

These effects are shown in Fig. 8.

The input tone is drawn in bold dashed line, the obtained component in bold continuous line, and the ommited components are drawn in continuous line.

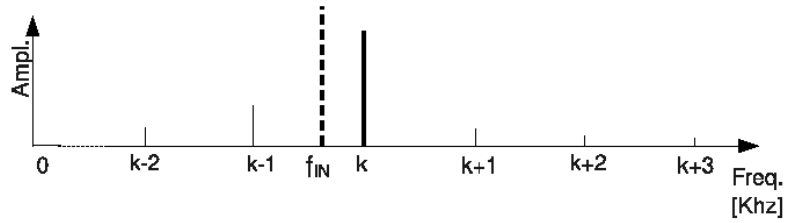

Fig. 8. Error due to picket-fence effect.

The problem of representing a pure tone affected by the picket-fence effect originates from the fact that the psychoacoustic model presented in the Sec. 2 eliminates some components by masking. These components that appeared due to the picket-fence effect, are necessary to reconstruct the original signal. Hence, the information required to resynthesize the signal is eliminated by masking and produces undesired artifacts (originated by phase and amplitude problems). For this reason, once the psychoacoustic model is applied, the adjacent components to a non-masked component are recovered from the original spectrum. Such heuristic is applied because it is impossible to know which component was originated by the the picket-fence effect. This procedure is shown in Fig.9.

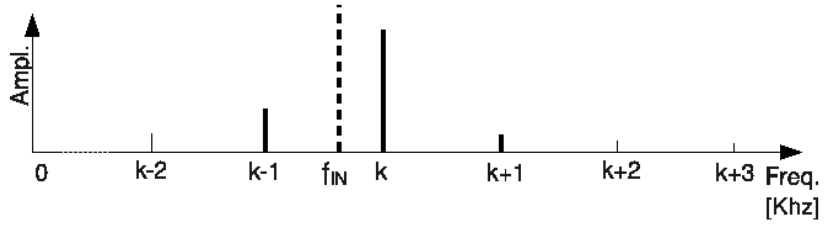

Fig. 9. Procedure to consider the picket-fence effect.

The graphic shows that the adjacent components (in continuous bold line) are taken into account to resynthesize the audio. Note that this correction only takes into account the picket fence effect. Due to in the proposed working conditions, it is possible to assume the periodicity of the most important components of the signal, for which the leakage error is much smaller than the effect of stakes. This is due to the the length of the segment **N** used (512 samples) it is small enough to ensure the periodicity of the components, because it provides a spectral resolution of 31.25Hz, with a sampling frequency of 16KHz.

4.2 Boundary Problems

Another problem that appears when the audio is resynthesized are discontinuities in the border of each frame. An analysis of the bibliography [15, 16] shows

that the MPEG-I Audio Layer 3 standard uses frames of 512 values, and 50% overlapped, separated by 16ms. This rate is not enough to cope with the problems in the border of each phoneme [7]. It produces undesired effects, like phase and amplitude discontinuities, due to temporal change between components of consecutive frames [17]. This problem is faced by considering three possible solutions:

1. **Reduce the distance between adjacent frames:** This is made by reducing the distance from 256 samples(16ms) to 128 samples (8 ms), 64 samples (4ms) and finally 32 samples (2ms). A similar approach is used in [10], where the overlap may be up to 99% . Then, the quality of the synthesized sound is the same than the original, the discontinuities are reduced and there are not any artifacts. The main disadvantage of this algorithm is that the computational load significantly increases, due to the greater quantity of frames, which rises inversely linearly depending on the distance between frames.
2. **Consider the components of adjacent frames:** This method is based on analyzing what components [12] are present in a frame but not in the preceeding and succeeding frames. A new vector is generated to smooth the changes between frames, increasing the quality of the resynthesized sound. The main disadvantage of this method is the artificial nature of the resulting audio. The computational load remains the same.
3. **Mixture between the two proposed methods:** This third model combines the other two proposals. Both the coarticulation effects and the distance between frames are considered. Such combination allows a better quality of resynthesized sound, similar to the quality of the input (distance between frames of 4ms).

Figure 10 shows the proposed model, where the block named *Articulatory System* represents each one of the possible solutions listed above.

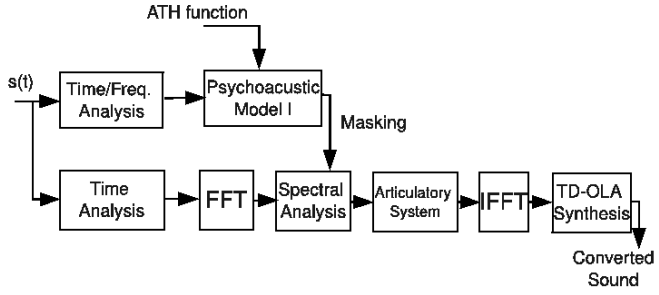

Fig. 10. Block diagram of the proposed system.

5 Experiments

The audio database used for this experiment contained 35 sentences in Spanish, uttered by two male and two female speakers. The sampling frequency was 16 KHz and the average duration of the sentences was 4 seconds.

Two subjective experiments were made with the latest proposed model of the system explained in Sec. 4.2. This algorithm was chosen because it combines the advantages of the other two proposals. In this stage 15 volunteers were asked to listen to the processed sentences. Listeners were asked to judge the quality of the processed sentences from 1 point (bad) to 5 points (excellent). On the other hand, the listeners were also asked to rate the intelligibility of the synthesized sound. The experiment is composed of audios processed using ATH curves in the following cases:

- Auditory model in a healthy condition. As is shown in Fig. 2.
- Auditory model with a severe hearing impairment. Which is shown in Fig. 4.
- Auditory model with a sensorineural hearing impairment. Which is shown in Fig. 5.
- Auditory model with a bilateral acoustic trauma. Which is shown in Fig. 6.

5.1 Quality

The first experiment is used to measure the **quality** of the processed audios. The main objective of this test is to validate the results of the other tests, due to if the audios have not a good quality, to evaluate the intelligibility is more difficult than in the case of an audio with a good quality. This problem is produced by the discontinuities that mask the acoustic problems due to the hearing impairment. Table 1 shows that the quality of the processed audio of a healthy ear is the

	Male	**Female**	**Joint**
Severe Hearing Impairment	3.0	3.2	3.10
Sensorineural Hearing Impairment	3.2	3.9	3.55
Bilateral Acoustic Trauma	3.0	2.7	2.85
Ear Healthy	3.8	3.7	3.75
Reference	5.0	4.8	4.90

Table 1. Mean values of the Test of Quality.

near to the original audio (used as reference). Then, it is possible to analyze the intelligibility of the processed audios.

5.2 Intelligibility

This experiment finds to show the problems in the intelligibility due to the hearing impairment. Results are shown in Tab.2, where the highest score are obtained by the audios without any hearing impairment. The following scores correspond to the audios processed using a bilateral acoustic trauma. This is due to that this impairment only eliminates few components, then the acoustic information is not reduced considerably. In the case of a sensorineural hearing loss the case is different, because the ATH curve rises linearly and exist a reject to the higher frequencies, which is appreciated in the consequently reduction of the intelligibility. Finally, the severe hearing impairment, obtains the worst score, due to under the conditions of simulation, only the first formant is perceived by the listener and there is not enough information to perceive the sound. Another

	Male	**Female**	**Joint**
Severe Hearing Impairment	1.5	1.1	1.30
Sensorineural Hearing Impairment	3.0	2.6	2.80
Bilateral Acoustic Trauma	4.0	3.5	3.75
Ear Healthy	5.0	5.0	5.00
Reference	5.0	5.0	5.00

Table 2. Mean values of the Test of Intelligibility.

aspect is that the intelligibility of the audios pronounced by a male speaker have a higher mean than the female speaker, because the formants of a male speaker are in lower positions, and consequently are not affected in the same way than the female speaker, for which the formants are placed in higher frequencies.

The scores obtained in the test are in agreement with expectations.

6 Conclusions

In this paper it is presented a model that allows the simulation of several types of hearing impairments. Another advantage is that the hearing impairment modelled can be modified easily, by only changing the Absolute Threshold of Hearing curve by a new ATH function.

Also, by using techniques of digital signal processing developed for voice signaling, it was possible to optimize the performance of the system, due to the discontinuities were reduced significantly. Subjective results show that the later proposed technique achieves a performance according to the expected.

Future work will be devoted to improve the quality of the voice resynthesized.Another interesting use to be developed is to test hearing impairment devices without any human tester. In this way the model will be used to reduce the number of volunteers or in some cases to simulate hearing impairments for which there is no volunteer for experiments.

References

1. Uriz, A., Agüero, P., Tulli, J.C., Gonzalez, E. and Denk, F.: Desarrollo de un Sistema de Compresion de Voz portatil para Pacientes Discapacitados, Proceedings of SABI 2009, (2009).
2. ISO/IEC JTC1/SC29/WG111 MPEG IS13813(1994): Generic Coding of Moving Pictures and Associated Audio, (1994).
3. Painter, T. and Spanias, A.: Perceptual Coding of Digital Audio, Proceedings of the IEEE. vol.88(4), pp. 451–513 (2000).
4. Slaney, M.: Auditory toolbox, Technical Report 010, (1998).
5. Fletcher, H.: Auditory patterns, Rev. Mod. Phys..pp. 47–65 (1940).
6. Zwicker, E., Flottorp, G. and Stevens, S. S., Critical Band Width in Loudness Simmation, Journal of the Acoustical Society of America. vol.29(3), pp. 548–557 (1957).
7. Huang, X., Acero, A., and Hon, H.W.: Spoken Language Processing. A Guide of Theory, Algorithm, and System Development, (2001).
8. Vilchur, E.: Signal processing to improve speech intelligibility in perceptive deafness, Journal of the Acoustical Society of America. vol.53(6), pp. 1646–1657 (1973).
9. Baer, T., Moore, B.C.J. and Kluk, K.: Effects of low pass filtering on the intelligibility of speech in noise for people with and without dead regions at high frequencies, Journal of the Acoustical Society of America, 112(3), pp. 1133–1144 (2002).
10. Timms, O.: Speech Processing strategies based on the sinusoidal speech model for the profoundly Hearing Impaired, Ph.D. Thesis., (2003).
11. Kim, D.W.,Park, Y.C., Kim, W.K., Park, S.J., Doh, W., Shin, S.W. and Youn, D.H.: Simulation of hearing impairment with sensorineural hearing loss, Proceedings of the 19th International Conference IEEE/EMBS. pp. 1986–1989 (1997).
12. Calupper, J. and Fastl, H.: Simulation of hearing impairment based on the Fourier Time Transform, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 857–860 (2000).
13. Pan, D.: A Tutorial on MPEG/Audio Compression, IEEE Multimedia. pp. 60–74 (1995).
14. Arelhi, R. and Campbell, D.R.: A MATLAB Simulink Implementation of Speech Masking Based on the MPEG Psychoacustic Model I, Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis. pp. 543–548 (2003).
15. Proakis, J.G. and Manolakis, D.G.: Digital Signal Processing: Principles, algorithms and applications. (1996).
16. Oppenheim, A.V., Schafer, R.W. and Buck, J.R.: Discrete-time signal processing. (1999).
17. Fant, G.: Acoustic Theory of Speech Production: The Hague, NL, Mouton, (1970).