

A TRAINABLE SPEECH ENHANCEMENT TECHNIQUE BASED ON MIXTURE MODELS FOR SPEECH AND NOISE

Ilyas Potamitis, Nikos Fakotakis, George Kokkinakis

Wire Communications Laboratory, Electrical and Computer Engineering Dept.,
University of Patras, 261 10 Rion, Patras, Greece, Tel:+30 2610 991722, Fax:+30 2610 991855
e-mail: potamitis@wcl.ee.upatras.gr

Abstract

Our work introduces a trainable speech enhancement technique that can directly incorporate information about the long-term, time-frequency characteristics of speech signals prior to the enhancement process. We approximate noise spectral magnitude from available recordings from the operational environment as well as clean speech from a clean database with mixtures of Gaussian pdfs using the Expectation-Maximization algorithm (EM). Subsequently, we apply the Bayesian inference framework to the degraded spectral coefficients and by employing Minimum Mean Square Error Estimation (MMSE) we derive a closed form solution for the spectral magnitude estimation task. We evaluate our technique with a focus on real, highly non-stationary noise types (e.g. passing-by aircraft noise) and demonstrate its efficiency at low SNRs.

1. Introduction

The primary objective of all enhancement methods applied in the context of speech processing is to reduce the effect of any signal that is alien to and disruptive of the message conveyed among participants in a communicative event (whether humans or ASR machines). Depending on the application, the objective of a speech enhancement algorithm is threefold:

- a) To improve speech quality and intelligibility by reduction of effort, fatigue and original message ambiguity. Despite the fact that almost every one-channel speech enhancement technique actually inflicts a small degradation to the intelligibility of a short waveform, listeners involved in extended listening tasks tend to find processed speech more intelligible after enhancement [1].
- b) To improve spectral parameter estimation of a speech coder when noisy speech is subjected to coding since, in general, speech coding algorithms assume clean speech.
- c) To achieve robust Automatic Speech/Speaker Recognition.

With regard to speech quality and intelligibility, it is essential that we respect the specific idiosyncrasies of human perception of speech and, therefore, reconstruct the time-domain signal. In spite of key contributions on the subject of *Short-Time Spectral Attenuation* algorithms (STSA) as applied to speech enhancement [2-5], there is still need for further work primarily on the problem of balancing the trade-off between noise reduction and speech distortion. The STSA family of algorithms attempts to uncover the underlying spectral magnitude of speech by applying a gain function to the observed, noisy short-time spectra, where the gain function is related to the noise power spectrum.

In this work we propose a novel STSA algorithm that incorporates into a Bayesian formulation the long term pdf of each spectral band of an ensemble of clean recordings

resulting in a better treatment of low energy spectral regions. The spectral magnitude of noise is modelled by a mixture of Gaussians. In order to build a Gaussian mixture model for the background noise the technique requires the availability of sample noisy recordings from the operational environment.

A different mixture of Gaussians is also employed to account for the representation of the magnitude of each spectral band of an ensemble of high quality speech. Three variations of the proposed technique are studied that correspond to the provenance of phonetically balanced speech, whether from gender independent/dependent or speaker specific speech corpora. The descriptive parameters of each mixture are derived from the observed spectral bands of the clean data by employing the EM algorithm. Under the Bayesian framework and MMSE it is shown that enhancement rule has an analytic form that allows for compensating the effect of time-varying noise types by associating each mixture to a different realization of a frequency variation.

The paper is an extension of [6] in a sense that it models noise with a mixture focusing on true, non-stationary noise types whose fast time-frequency variations are difficult to be tracked. Objective as well as subjective evaluation of signals degraded with additive passing by aircraft noise [8] at low SNRs confirm the benefit of our approach. *The proposed work is supported by the GEMINI (IST-2001-32343) EC project.*

2. Description of the algorithm

Let $s(m)$ denote the clean time-domain signal corrupted by noise $n(m)$ where (m) is the discrete time index. The observed signal $x(m)$ is given by $x(m)=s(m)+n(m)$ and is subjected to Short Time Fourier Transform (STFT). Based on the generalized spectral subtraction framework [2], we can derive a linear-spectral representation of a clean speech signal corrupted by additive noise using a 2N point FFT as:

$$x_{\kappa,1}^{\alpha}=s_{\kappa,1}^{\alpha}+n_{\kappa,1}^{\alpha}, \quad \kappa=0,\dots,N \quad (1)$$

$\{x_{\kappa,1}^{\alpha}\}$ denotes the spectral magnitude of the degraded sub-band $\{\kappa\}$, $\{n_{\kappa,1}^{\alpha}\}$ the noise spectral magnitude, $\{1\}$ the frame index and $1 \leq \alpha \leq 2$ where α stands for the power index. For notational convenience we set $x=x^{\alpha}$, $s=s^{\alpha}$, $n=n^{\alpha}$ and we drop subscript $\{\kappa\}$, $\{1\}$ implying that the subsequent analysis holds for every time-trajectory of spectral sub-band $\{\kappa\}$ independently. We found that setting $\alpha=3/2$ optimises performance, though the subsequent analysis holds for every $\{\alpha\}$. Prior knowledge about the time frequency distribution of $\{s\}$ is provided by a mixture of Gaussians that model the undegraded spectral bands of the available clean speech corpora (Eq. 2). Practically, 2-3 minutes of clean speech, unrelated to the signals to be enhanced, were found sufficient to tune the free parameters of the algorithm.

Let $f(s)$ be the GMM representing the clean speech model derived from available training data:

$$f(s) = \sum_{r=1}^M p_m G(n; \mu_m, \sigma_m^2), \quad \sum_{m=1}^M p_m = 1 \quad (2)$$

The pdf of the spectral magnitude of noise is modeled by a mixture of Gaussian as:

$$f(n) = \sum_{r=1}^R p_r G(n; \mu_r, \sigma_r^2), \quad \sum_{r=1}^R p_r = 1 \quad (3)$$

where, M, R are the total number of mixture components for speech and noise respectively, p_m, μ_m and σ_m are the prior probability, mean and standard deviation of the m^{th} Gaussian speech mixture, while p_r, μ_r and σ_r are the prior probability, mean and standard deviation of the r^{th} Gaussian noise mixture. The descriptive statistics of the Gaussian mixture i.e $p_m, \mu_m, \sigma_m, p_r, \mu_r$ and σ_r are computed by the EM algorithm. Means are initialized uniformly over the interval of each spectral band magnitude, while weights are set to equal values and variance is lower-bounded to avoid picking narrow spectral peaks.

Subsequently we proceed to derive the MMSE estimation of the underlying spectral coefficients $\{s\}$ as $S_{\text{MMSE}} = E\{s|x\} = \int s f(s|x) ds$. The pdf of $\{s\}$ given the observation $\{x\}$ is derived by the Bayesian formula $f(s|x) = f(x|s)f(s)/f(x)$. Combining $f(s|x)$ and S_{MMSE} results in:

$$S_{\text{MMSE}} = \frac{\int s f(x|s) f(s) ds}{\int f(x|s) f(s) ds} \quad (4)$$

Substituting Eq. 2 and Eq. 3 into Eq. 4 and after completing the squares of the mixture components we derive the underlying spectral magnitude in terms of an integral I_v which is expressed in closed form through the parabolic cylinder functions D_v . (See [7])

$$I_v = \int_0^{+\infty} s^{v-1} \exp(-b_{m,r} s^2 - c_{m,r} s - d_{m,r}) ds = \exp(-d_{m,r}) (2b_{m,r})^{-\frac{v}{2}} \Gamma(v) \exp\left(\frac{c_{m,r}^2}{8b_{m,r}}\right) D_{-v}\left(\frac{c_{m,r}}{\sqrt{2b_{m,r}}}\right)$$

$$D_{v+1} - zD_v(z) + vD_{v-1}(z) = 0, \quad (\text{Eq. 3.462, [7]})$$

Setting $v=-2$ and solving with respect to D_{-3} leads to:

$$D_{-3}(z) = \frac{1}{2} [D_{-1}(z) - zD_{-2}(z)]$$

$$\text{where: } D_{-1}(z) = \exp\left(\frac{z^2}{4}\right) \sqrt{\frac{\pi}{2}} \left\{ 1 - \text{erf}\left(\frac{z}{\sqrt{2}}\right) \right\}$$

$$\text{and } D_{-2}(z) = \exp\left(\frac{z^2}{4}\right) \sqrt{\frac{\pi}{2}} \left\{ \sqrt{\frac{\pi}{2}} \exp\left(\frac{z^2}{4}\right) - z \left[1 - \text{erf}\left(\frac{z}{\sqrt{2}}\right) \right] \right\}$$

m, r refers to the m^{th} and r^{th} Gaussian mixture for speech and noise respectively and

$$b_{m,r} = \frac{1}{2} \left(\frac{1}{\sigma_m^2} + \frac{1}{\sigma_r^2} \right), \quad c_{m,r} = - \left(\frac{\mu_m}{\sigma_m^2} + \frac{x - \mu_r}{\sigma_r^2} \right), \quad d_{m,r} = \frac{1}{2} \left(\frac{(x - \mu_r)^2}{\sigma_r^2} + \frac{\mu_m^2}{\sigma_m^2} \right)$$

Based on the Gaussian assumption for the spectral magnitude pdf of noise the MMSE estimation of the underlying clean spectral magnitude is:

$$S_{\text{MMSE}} = \frac{\sum_{r=1}^R \sum_{m=1}^M \frac{p_r}{\sigma_r} \frac{p_m}{\sigma_m} I_2(b_{m,r}, c_{m,r}, d_{m,r})}{\sum_{r=1}^R \sum_{m=1}^M \frac{p_r}{\sigma_r} \frac{p_m}{\sigma_m} I_1(b_{m,r}, c_{m,r}, d_{m,r})} \quad (5)$$

Based on the fact that the information of the speech signal is encoded in the frequency domain and that human hearing is relatively insensitive to phase information, we focus on the short-time amplitude of the speech signal leaving the noisy phase unprocessed. After the enhancement procedure has been applied, noisy phase is added back and the time-domain underlying frame $\{\hat{s}\}$ is subsequently reconstructed using inverse FFT and the weighted overlap-and-add method as:

$$\hat{s} = \text{IFFT} \left\{ S_{\text{MMSE}}^{1/a} \exp\left(j \cdot \arg(x^{1/a})\right) \right\}$$

2.1. Gender-Independent Speech Enhancement

The magnitude of each spectral band of the available clean speech is modeled by a Gaussian mixture model composed of six mixtures. Following the general concept of mixture models each mixture realizes a different cluster of magnitude values. By observing the mean value of each cluster in adjacent frequency bins (mixtures are sorted from bottom to top i.e. m1: the mixture of low magnitudes to m6: the mixture of high magnitude values) we can see that in the gender-independent case (see Fig. 1) the a-priori information that is incorporated in the enhancement rule is restricted to follow the general gross difference between the lower and upper frequency content of the speech spectrum. In the gender-dependent case (Fig. 2 and Fig. 3) each mixture evolves across frequencies in a more abrupt way showing some specialization on the localization of spectral energy around the formant area which is different for males and females. Finally, in the speaker-dependent case (Fig. 4) the means of each mixture in adjacent frequency bins specialize on the spectral idiosyncrasies of the speaker. A detailed analysis of each case follows in order to scrutinize the a-priori information one can obtain from the long-term distribution of the spectral magnitude of gender- and speaker-dependent speech corpora.

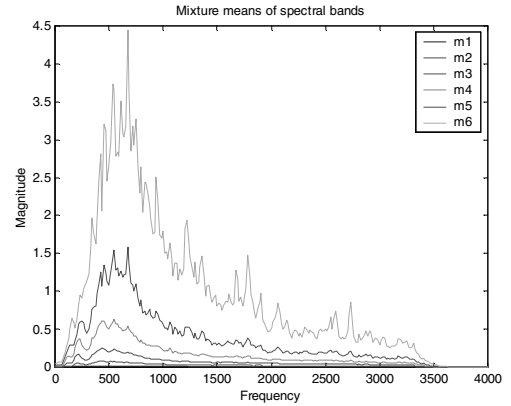


Fig. 1: Gender Independent case: Mixture Means over all spectral bands derived from an ensemble of 3 minutes recordings (Each gender Ninety seconds, 30 speakers each).

2.2. Gender-Dependent Speech Enhancement

In Fig. 2 we depict the mean-mixture results of applying the EM algorithm to the magnitude of spectral bands of an ensemble of 3 minutes of gender-dependent speech corpora. One can observe that the high-energy part of the spectrum, around 300 Hz to 1 kHz, is better separated from the low energy part of the spectrum than in the speaker-independent case and the formant area is analyzed in more detail. In Fig. 2 it is illustrated that for the male speech the energy is

concentrated in the lower bands (in contrast to the female case depicted in Fig. 3). This information is incorporated implicitly as a-priori information into the Bayesian inference framework of Eq. 4.

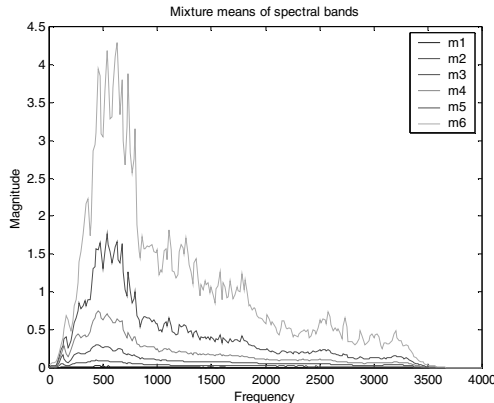


Fig. 2: Mixture Means over all spectral bands derived from an ensemble of 3 minutes recordings (30 male speakers).

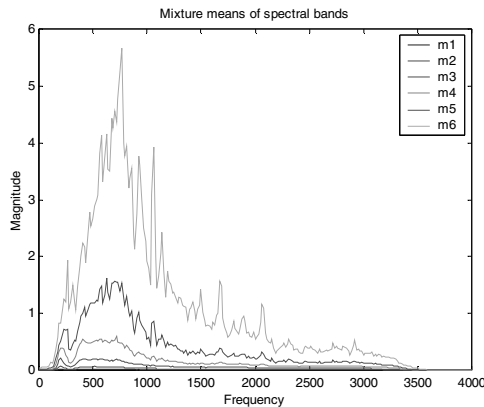


Fig. 3: Mixture Means over all spectral bands derived from an ensemble of 3 minutes recordings (30 female speakers).

2.3. Speaker Dependent Speech Enhancement

In Fig. 4 we depict the mean-mixture results of applying the EM algorithm to the magnitude of spectral bands of an ensemble of 2.1 minutes of female speech corpora from a sole speaker. One can observe that the formant area is clearly resolved and the mixtures are highly specialized to partition a certain area of the magnitude of the spectrum.

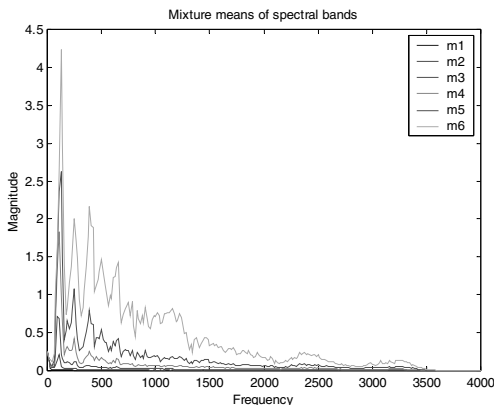


Fig. 4: Mixture Means over all spectral bands derived from an ensemble of 3 minutes recordings (male speaker).

3. Simulation and results

We built the Gaussian models for clean speech by using gender dependent/independent speech corpora taken from the SpeechDat database. During the evaluation procedure, each noise type [8] was added to a concatenation of 50 sets of clean speech files of 15 sec. mean duration so that the corrupted waveform ranged from -10 to 20 SNR dB. Each set of fifteen-second speech was produced by a concatenation of random waveforms with their silence part removed in order to ensure that each test file was corrupted with all forms of the degrading noise as it evolved with time. The FFT size is 512 points and the signals are hamming windowed with 50% overlap. The number of Gaussian mixtures is set to nine for noise and six for speech per spectral band as the objective measures indicated marginal gain by augmenting the number of mixtures.

The Itakura Saito (IS) and Weighted Spectral Slope (WSS) measures of the enhancement obtained by our technique are shown in Fig. 6 and Fig. 7 (IS autoregressive distance for DC-3 and Merlin aircrafts noise corruption respectively), Fig. 8 and Fig. 9 (WSS distance). Both measures were chosen because of their high correlation with subjective speech quality. The WSS measure decomposes the speech signal into a set of frequency bands (in our case a Bark-scale bank of 25 filters spanning the 4 kHz bandwidth), calculates the intensities within each critical band and a weighted distance between the measured slopes of the log-critical band spectra. The IS distortion measure is based on the spectral distance between AR coefficient sets of the clean and enhanced speech waveforms over synchronous frames of 15ms duration and is heavily influenced due to mismatch in formant locations. The proposed versions are compared against spectral subtraction.

As indicated in Fig. 6 to Fig. 9 there is always a variant of the proposed method consistently effecting smaller error rates in both objective measures over all SNRs compared to spectral subtraction. The difference is more subtle in the gender-dependent case than in the gender-independent case and becomes even more subtle in the speaker-dependent case. A consistent gain is observed when the proposed technique is tuned based on speaker-dependent speech data. Our technique belongs to the family of STSA techniques but differs in its ability to incorporate gender- or speaker-specific information in its formulation. This is accomplished by including a probability fitting stage that allows the tuning of its parameters

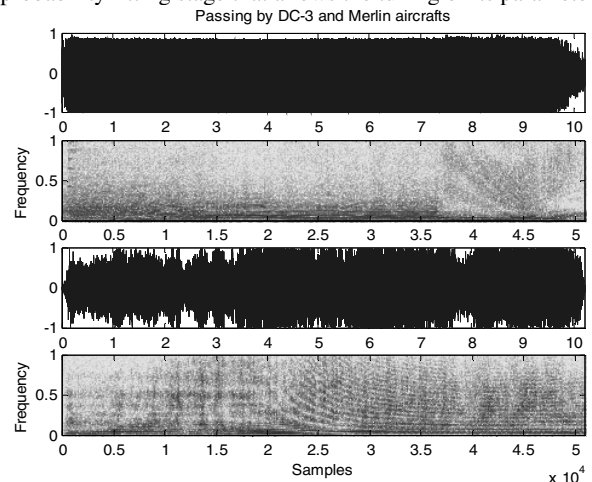


Fig. 5: First and Second Row from top: Passing-by DC-3 aircraft, Last two rows: Passing Merlin aircraft (see also [8]).

in a data-dependent fashion, rather than making general model assumptions. Compared to spectral subtraction the proposed technique is fully parametric without involving empirical tuning of thresholds.

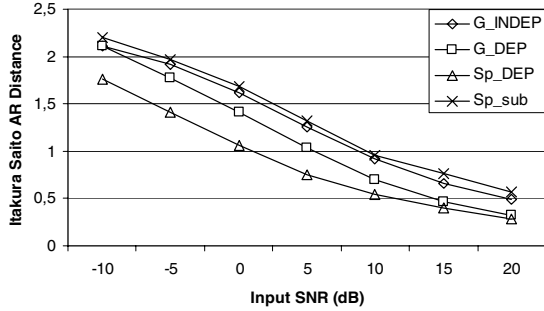


Fig. 6: Itakura-Saito AR-distance comparative measurements between all training scenarios of the proposed technique and spectral subtraction (Sp_sub). Noise corruption from a DC-3 aircraft. G_INDEP, G_DEP and Sp_DEP stands for the proposed algorithm trained with gender-independent, gender-dependent and speaker-dependent clean speech respectively.

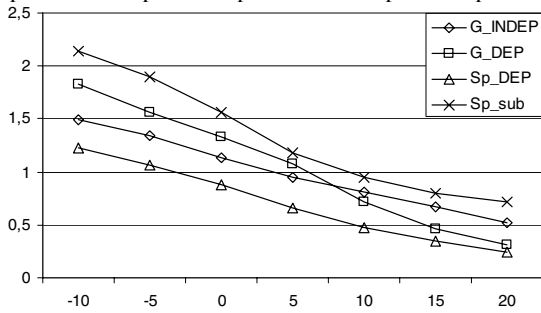


Fig. 7: Itakura-Saito AR-distance comparative measurements between all training scenarios. Noise corruption from a Merlin aircraft. G_INDEP, G_DEP and Sp_DEP as in Fig. 6.

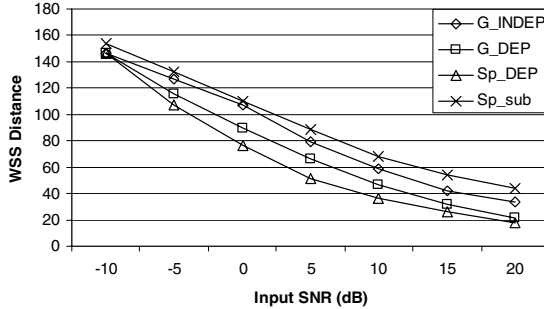


Fig. 8: WSS distance comparative measurements. Noise corruption from DC-3 aircraft. G_INDEP, G_DEP and Sp_DEP as in Fig. 6.

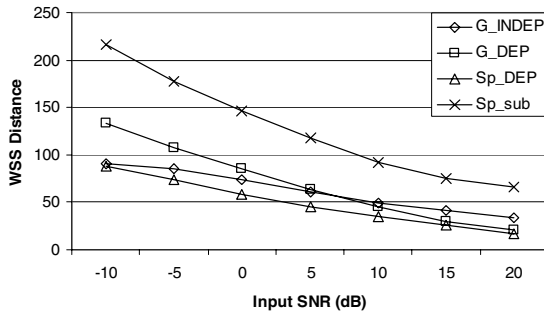


Fig. 9: WSS distance comparative measurements. Noise from Merlin aircraft. G_INDEP, G_DEP and Sp_DEP as in Fig. 6.

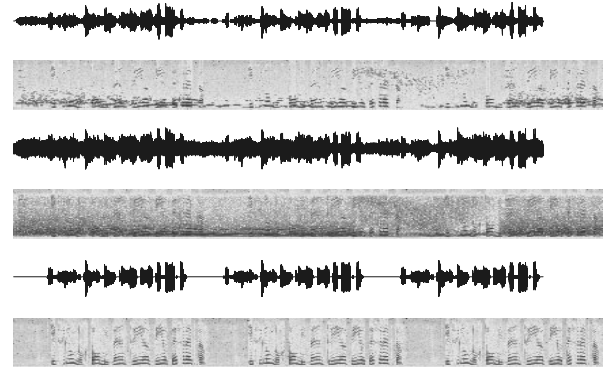


Fig. 10: First and Second Row from top: Time domain signal and spectrogram of the enhanced speech for noise corruption imposed by a DC-3 aircraft, Third and Fourth Row from top: noisy signal, Last two rows: clean signal.

4. Conclusions

The application of Weighted Spectral Slope and Itakura-Saito measures confirmed the benefit of modeling clean spectral bands and the spectral bands of noise with a mixture of Gaussians. We demonstrated that the proposed enhancement technique is able to compensate the effect of true, fast varying noise types typically experienced in airports. The key idea is to independently model the multimodal spectral distribution of the magnitude of spectral bands of speech with a mixture of Gaussians and to model noise variations the same way as speech variations. The a-priori information from available clean data and sample noise recordings combined into a MMSE formulation, supplied an analytic solution to the speech enhancement task. The solution proved to be a weighted subtraction between speech and noise mixtures. We also suggested that the incorporation of a-priori information from gender-dependent and speaker-dependent speech corpora leads to estimators that better treat the low energy time-frequency regions more effectively. Future work must deal with the channel effect that is not compensated in this paper.

5. References

- [1] Lim J., Oppenheim A., "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no 12, pp 1586-1604, 1989.
- [2] Berouti M., Schwartz R., Makhoul J., "Enhancement of speech corrupted by acoustic noise," *Proc ICASSP*, 1979.
- [3] McAulay R., Malpass M., "Speech enhancement using a soft decision noise suppression filter," *IEEE Trans. Speech & Audio Proc.*, vol. 28, no. 2, pp. 137-145, 1980.
- [4] Ephraim Y., Malah D., "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109-1121, 1984.
- [5] Ephraim Y., Malah D., "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 443-445, 1985.
- [6] Potamitis I., Fakotakis N., Kokkinakis G., "Gender Dependent and Speaker Dependent Speech Enhancement", *Proc ICASSP*, vol. I, pp. 249-252, 2002.
- [7] Gradshteyn I., Ryzhik M., Jeffrey A. editor, Fifth edition, "Table of Integrals, Series & Products," *Academic Press*, pp. 1094-1095, Eq. 9.247, Eq. 9.254, Eq. 3.462, 1994.
- [8] URL:<http://www.centercomp.com/cgi-bin/dc3/sounds?13>