Audio Engineering Society
Convention Paper
Presented at the 119th Convention
2005 October 7–10 New York, New York USA

Quality Enhancement of Low Bit Rate MPEG1-Layer 3 Audio Based on Audio Resynthesis

Demetrios Cantzos[1] and Chris Kyriakakis[1]

[1] Integrated Media Systems Center (IMSC), University of Southern California, Los Angeles, CA, 90089-2564, USA

ABSTRACT

One of the most popular audio compression formats is indisputably the MPEG1-Layer 3 format which is based on the idea of low-bit transparent encoding. As these types of audio signals are starting to migrate from portable players with inexpensive headphones to higher quality home audio systems, it is becoming evident that higher bit rates may be required to maintain transparency. We propose a novel method that enhances low bit rate MP3 encoded audio segments by applying multichannel audio resynthesis methods in a post-processing stage or during decoding. Our algorithm employs the highly efficient Generalized Gaussian mixture model which, combined with cepstral smoothing, leads to very low cepstral reconstruction errors. In addition, residual conversion is applied which proves to significantly improve the enhancement performance. The method presented can be easily generalized to include other audio formats for which sound quality is an issue.

1. INTRODUCTION

The majority of MPEG1-Layer 3 (Mp3) audio encoded at low bit rate does not deliver high quality sound. On the other hand, high bit rate Mp3 segments, even though they deliver sufficient sound quality, are too large to transmit or store. With the emergence of high quality consumer audio systems and the prevalence of Mp3 as the standard audio coding scheme, the need for enhancing low bit rate Mp3 audio data without imposing excessive storage or transmission requirements, seems natural. In this work, we attempt to improve the quality of Mp3 encoded audio data based on a recently introduced concept termed audio resynthesis ([1]).

In audio resynthesis, a reference (source) channel is transmitted and then used to recreate the remaining (target) channels at the receiving end by deriving a small set of constant parameters. In order to apply this concept to Mp3 audio enhancement, we replace the source channel with a low bit rate Mp3 audio music segment and the target channel with the original uncompressed audio segment of the same music piece. Our main goal is to recreate the high quality target segment at the receiver end by transmitting a small set of constant parameters and by using the low quality source segment that is already stored at the receiver. This scheme is implemented in a post-processing stage or during decoding and thus both source and target are preconverted to the same lossless data format (e.g. WAV).

Recent work on audio resynthesis ([1]) has been based on previous spectral transformation algorithms ([2,3,4]). The basic assumption made in these algorithms is that the spectral parameters are of Gaussian nature and hence are modeled by a Gaussian mixture. This greatly facilitates the Maximum Likelihood (ML) parameters estimation since the popular Expectation-Maximization (EM) algorithm can be applied. As we show later, the actual nature of the cepstral coefficients of an audio signal is not strictly Gaussian and thus the Gaussian mixture model, although convenient, is not the best solution. We present a new approach on modeling the cepstral coefficients by employing the Generalized Gaussian mixture model. This model is very flexible and incorporates a large number of distributions including the Gaussian.

A new technique is also introduced which takes effect during the cepstral conversion step. Due to the linearity of the conversion function and the abrupt changes of the cepstral vectors during short time periods, the reconstruction errors are considerably high. We propose a method in which the cepstral vectors are smoothed and the number of mixture components increases to facilitate the task of the conversion function.

Finally, a novel technique related to residual processing is implemented. In many cases of low bit rate Mp3 sources, reconstruction in the cepstral domain is not adequate for distortion-free enhanced audio. For this reason we also apply residual conversion and even though it is not as accurate as cepstral conversion, it proves to significantly enrich the spectral details of the enhanced Mp3 music piece.

2. STATISTICAL CONVERSION

The approach followed is based on previous statistical conversion algorithms related to speech synthesis ([2,3,4]). In our application, the short term spectral parameters are selected to be the LPC cepstral vectors ([5]). The LPC analysis is carried out in overlapping frames through a sliding window and hence each frame is modeled as an AR filter excited by a residual. We extract the LPC cepstral vectors of the target (which is unknown at the receiving end) and source signals. Our goal is to modify the cepstral vectors of the source signal so that they would be close in the least squares sense to the target cepstral vectors of the same music piece. This is accomplished by deriving a mapping function that will convert each of the source cepstral vectors to the target ceptral vector of the same time frame (the two signals are time-aligned). The function is assumed linear and will be fully determined by a small set of constant parameters. As shown later, a similar conversion technique can be applied to the residual vectors in which the source residual is modified so that it better matches the target residual.

In order to implement the conversion function, we assume that the source cepstral (and residual) vectors are generated by a probability density function (pdf). The task of determining this pdf is effectively the system training. The audio segment used during training is chosen so that it is capable of modeling a large and diverse number of music pieces and is called the training set. The testing source and testing target signals are the particular signal segments on which we apply the conversion scheme and derive the specific conversion function. In the following subsection we present the probabilistic model associated with the training task.

2.1. The Generalized Gaussian Mixture Model

In the previous statistical conversion algorithms a common assumption is that the spectral vectors are of Gaussian nature and hence the Gaussian mixture model is employed. The Gaussian mixture model has been treated in numerous other applications and an algorithm to estimate its parameters (EM) is readily available. However, as we show later, the cepstral vectors of audio data are not strictly Gaussian and thus this model is not the best selection.

A more flexible model is adopted here, which includes the Gaussian mixture as a subcase, and is called the Generalized Gaussian mixture. Its component pdf, the Generalized Gaussian pdf, is more flexible and adapts to virtually any unimodal distribution. Its analytical form for a random variable z is:

$$g(z;\ ,\sigma,a) = \frac{a\beta}{2\sigma\Gamma(1/a)}\exp[-|\beta\frac{(z-\ )}{\sigma}|^{\alpha}] \qquad (1)$$

where μ is the mean, σ is the variance, α is the shape parameter, $\Gamma(\cdot)$ is the Gamma function and β is a dependent parameter:

$$\beta = \left[\frac{\Gamma(3/a)}{\Gamma(1/a)}\right]^{1/2} \qquad (2)$$

If α =2.0 we have the Gaussian pdf and if α =1.0 we have the Laplace pdf. When α >>1 the distribution tends to the uniform pdf and when α < 1 the distribution becomes impulsive.

We consider the training cepstral vectors (and the testing source vectors) to be generated by a mixture with component pdf as described in equation (1). The mixture formulation of the Generalized Gaussian case is shown below:

$$G(\boldsymbol{x}) = \sum_{k=1}^{K} p(C_k) \prod_{j=1}^{q} g(x^{(j)};\ {}_k^{(j)}, \sigma_k^{(j)}, a_k^{(j)}) \qquad (3)$$

where C_k denotes the cluster (component) k, K is the number of clusters and $p(C_k)$ denotes the prior probability that the cepstral vector $\boldsymbol{x}$ belongs to cluster k. The cepstral vector is q dimensional where q is the cepstral order and the jth coefficient is denoted by $x^{(j)}$. The vector coefficients are considered to be independent and thus the joint pdf is the product of the q coefficient pdf's. This diagonal formulation is favorable since it decreases the computational complexity during implementation.

2.2. Mixture Parameters Estimation and Clustering

The inclusion of a third independent parameter (the shape parameter α) incurs additional complexity when it comes to ML (Maximum Likelihood) estimation of the pdf parameters. This becomes more apparent in a mixture pdf where it is obvious that the model is considerably more difficult to manipulate than the Gaussian mixture and the EM algorithm cannot be applied easily because the Expectation step is very hard to compute. Also, even though the EM algorithm is guaranteed to approach a local maximum, it is uncertain how fast this can be reached.

We decide to follow a different path than the one used in the conventional mixture estimation methods by clustering the vectors and focusing on each cluster separately. This will divide the parameters estimation task into K simpler tasks. In order to perform this decomposition we employ fuzzy clustering techniques through the c-means algorithm ([6]) and cluster the training vectors into K groups. The c-means is known to avoid local minima better than the k-means and it also provides a 'fuzziness' option that regulates the occurrence of outliers.

The next step is to perform ML estimation on each cluster. The estimation is now straightforward because the mean for each component is known (it is the cluster center). We also compute $p(C_k)$ as the number of vectors that belong to cluster k divided by the total number of vectors. The ML estimator for the shape parameter $a_k^{(j)}$ of cluster k and coordinate j is given by ([7]):

$$\frac{\psi(1/a_k^{(j)}+1)+\log(a_k^{(j)})}{a_k^{(j)^2}} + \frac{1}{a_k^{(j)^2}} \log\left(\frac{1}{n_k} \sum_{t:\boldsymbol{x_t}\in C_k} |x_t^{(j)} - {}_k^{(j)}|^{a_k^{(j)}}\right) - \frac{\sum_{t:\boldsymbol{x_t}\in C_k} |x_t^{(j)} - {}_k^{(j)}|^{a_k^{(j)}} \log(|x_t^{(j)} - {}_k^{(j)}|)}{a_k^{(j)} \sum_{t:\boldsymbol{x_t}\in C_k} |x_t^{(j)} - {}_k^{(j)}|^{a_k^{(j)}}} = 0 \qquad (4)$$

where n_k is the number of vectors that belong to class k and $\psi(\cdot)$ is a function given by:

$$\psi(\tau) = -0.5777 + \int_0^1 (1-t^{\tau-1})(1-t)^{-1} dt \qquad (5)$$

The expression in (4) is solved by iterative methods. The variance parameter $\sigma_k^{(j)}$ of the kth cluster and jth coordinate is then estimated as follows ([7]):

$$\sigma_k^{(j)} = \left[\frac{a_k^{(j)} \beta^{a_k^{(j)}} \sum_{t:\boldsymbol{x_t}\in C_k} |x_t^{(j)} - {}_k^{(j)}|^{a_k^{(j)}}}{n_k} \right]^{1/a_k^{(j)}} \qquad (6)$$

Note that the zero[th] cepstral coefficients (energy coefficients) are discarded because they introduce strong bias during parameters estimation. Besides, the frame energy information (relative to the other frames) is already contained in the residual.

2.3. Conversion Function

The conversion function $F(\cdot)$ acts on the vector sequence $[\boldsymbol{x_1},...,\boldsymbol{x_n}]$ and produces a vector sequence close in the least squares sense to the sequence $[\boldsymbol{y_1},...,\boldsymbol{y_n}]$. Since we have selected a diagonal implementation, this function will act on the individual vector components and minimize the error:

$$E = \sum_{t=1}^{n} \sum_{j=1}^{q} | y_t^{(j)} - F(x_t^{(j)}) |^2 \qquad (7)$$

as in [2]. This problem becomes possible to solve under the constraint that F is piecewise linear, i.e.

$$F(x_t^{(j)}) = \sum_{k=1}^{K} P(C_k \mid \boldsymbol{x}_t)[v_k^{(j)} + \frac{u_k^{(j)}}{\sigma_k^{(j)}}(x_t^{(j)} - \ _k^{(j)})] \quad (8)$$

for $t = 1,..,n$ and $j = 1,..,q$. The conditional probability that a given vector belongs to cluster k, $P(C_k|\boldsymbol{x}_t)$, is given by:

$$P(C_k \mid \boldsymbol{x}_t) = \frac{p(C_k)\prod_{j=1}^{q} g(x_t^{(j)};\ _k^{(j)}, \sigma_k^{(j)}, a_k^{(j)})}{G(\boldsymbol{x}_t)} \quad (9)$$

The unknown parameters set $[\boldsymbol{v},\boldsymbol{u}]$ can be found by minimizing (7) which reduces to solving a typical set of q independent least-squares equations ([2]) and hence the linear conversion function F is fully determined.

2.4. Conversion Optimization through Cepstral Smoothing and Data Overfitting

The cepstral conversion function will generally not provide the accuracy in results that is needed for audio reproduction. The cepstral vectors vary rapidly from frame to frame and many spikes occur. The conversion function, due to its linear form, cannot follow these abrupt changes and fails to produce the desired vectors.

A new technique is introduced here that improves the cepstral conversion performance. In essence, we smooth out the cepstral vectors to reduce the spikes by increasing the LPC analysis frame slide and length and at the same time increase the mixture groups number so that the conversion function has more components available. The frame slide and length increase is applied only on the testing source and target signals and not on the training signal. If we apply the frame slide and length increase on the training vectors too then their number will decrease considerably and the ML estimation will fail for a mixture of many components. The number of groups is around three times larger than the number determined by the MDL information-theoretic criterion ([8]) and thus the training data is overfitted. This overfitting does not affect the conversion stage since any unnecessary clusters are filtered out by the conversion function. This technique is proved to be extremely favorable since accurate reconstruction of the cepstral vectors is achieved.

2.5. Residual Modeling and Conversion

In many cases, an accurate cepstral reconstruction is not sufficient for acoustically undistorted enhanced Mp3 segments. Especially in the case of a very low bit rate source (e.g. 64Kbps), many audible artifacts are present because the source and target testing signals are simply too different. Instruments that are inaudible in the source signal will usually appear in the enhanced signal as distortions since the LPC coefficients alone fail to reproduce them. In such cases, the signal differences lie mainly in the residuals and therefore some residual processing is essential for better enhancement results.

We adopt the assumption that the residual vectors are correlated with their corresponding cepstral vectors ([9]) and thus share similar statistical properties. Therefore, we can apply the statistical conversion described in the previous sections to the residual vectors also. The probabilistic model used here is the same used for cepstral conversion (i.e. it is derived from the training cepstral vectors). However, the dimensionality of the residual vectors is much higher than that of the training cepstral vectors and therefore we have to divide them in subvectors of dimensionality equal to that of the training cepstral vectors. For instance, in the case of 30 training cepstral coordinates and 840 residual coordinates, we would divide the residual vectors in subvectors of 30 coordinates each and apply statistical conversion in each of the 28 subvectors sets separately.

Clearly, we do not expect a residual reconstruction with accuracy similar to that of the cepstral reconstruction because the residuals are too ‘spiky’. Furthermore, we have not derived a training set or a probabilistic model specifically for the residual vectors since the extremely high residual vector dimensionality would make this impractical. Besides, we would have to design a global mixture pdf that could efficiently model any set of testing residual vectors even though these are highly diverse and contain the fine details of the signal.

Using the mixture pdf derived from the training cepstral vectors shows that the converted residuals are much closer to the target residuals (than the source residuals are) and a large amount of information is conveyed to the enhanced Mp3 segment through this process. It was also observed that a high training cepstral order led to smaller residual reconstruction errors. Therefore we select a cepstral order for the training vectors that is higher than the cepstral order of the testing vectors.

3. IMPLEMENTATION

The algorithm described previously was applied and tested on a randomly selected music piece. The general scenario involves enhancing a 32sec long, 64Kbps Mp3 segment. This is the testing source signal. The testing target signal is the uncompressed WAV file of the same music piece. These two segments are time-aligned and since the algorithm is applied in a post-processing stage, the Mp3 source is also converted to a WAV format. Careful consideration has been taken to reduce the residual conversion parameters size as much as possible. As shown later in this section, the actual size of the conversion function is less than the size of the Mp3 source and much less than the size of the uncompressed, target file. Some objective enhancement results are also provided which prove the validity of this scheme.

3.1. Wavelet-Based Subband Coding

Due to the higher sampling frequency and richer content of an audio signal (compared to a speech signal) we follow a subband analysis. The subband separation is performed with wavelets ([10]) and in this case the 'Daubechies' filter of order 40 was a good choice since no audible aliasing effects were observed. Several differrrent wavelet tree structures were tested (e.g. equidistant subbands) but the most efficient structure proved to be one that emulates the critical bands of the human hearing system as in [11]. This choice is further justified by the fact that the Mp3 encoded source segment has passed through a critical filterbank also ([12]). The high number of subbands selected allows us, as we show later, to take advantage of the inter-band redundancy and also to process heavier the subbands that are the most significant (i.e. the ones that are more degraded or carry the audible parts of the signal). The actual wavelet filterbank is shown in Fig. 1 and is applied to both testing source and testing target signals leading to 17 testing subbands.

3.2. Training Model Derivation

A crucial part of the algorithm is to derive a Generalized Gaussian mixture pdf that does not have to adjust to the particular testing music piece. This probabilistic model should be global in the sense that it will include the statistical properties of all possible music segments and both transmitting and receiving ends will have access to it (e.g. pre-stored in both sides).

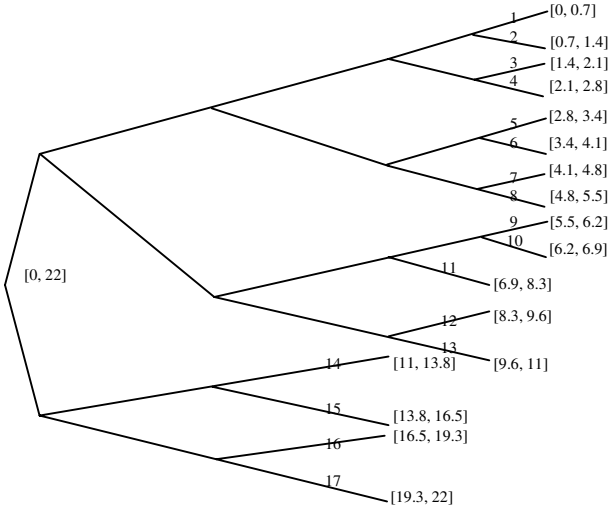

Figure 1: Wavelet tree structure used for subband analysis of the testing source and testing target signals (Numbers in brackets indicate the frequency region in kHz in each subband. Numbers on leafs indicate the subband index from 1 to 17)

Several candidate training sets were processed to produce a mixture pdf among which were the multichannel training set of [1] ,a white noise training set, a Brownian noise training set and a pink noise training set. Pink noise proved to be the most suitable training set and produced smaller cepstral reconstruction errors (up to 5% less in all subbands compared to the other sets).

In order to reduce the training model size and allow for the data diversity needed in the case of many mixture components ML estimation, we divide the training data set into 4 large equidistant subbands (instead of the 17 subbands shown in Fig. 1) covering the frequency range 0-22kHz (0-5.5kHz,5.5-11kHz,11-16.5kHz,16.5-22kHz) and each subband consists of 12,000 cepstral vectors of cepstral order 30.

Each of the 17 analysis subbands of the testing source and testing target signals acquires the training model parameters from one of the 4 larger subbands that it is part of. During cepstral conversion, the cepstral order of the training model is truncated appropriately for each testing subband to adjust to the lower cepstral order of the particular testing source and testing target cepstral vectors. During residual conversion, the large training cepstral dimensionality allows for more efficient division of the testing residual vectors into subvectors, as explained in section 2.5.

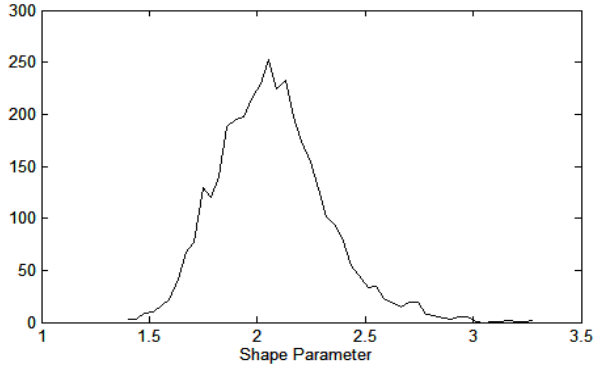

Figure 2: Histogram of shape parameters for the frequency band 0-5.5kHz of the pink noise training set

Fig. 2 shows the distribution of the mixture pdf shape parameters for all groups and vector coordinates of the first (0-5.5kHz) of the 4 training subbands. It is clear that the shape parameters, although strongly peaked at a =2.0, have the majority of their values in the intervals 1.4-2.0 (subgaussian) and 2.0-3.5 (supergaussian) which justifies the use of the Generalized Gaussian mixture as a more accurate model. Pink noise is random data rather than actual audio data but a similar histogram is obtained from the audio data set used in [1].

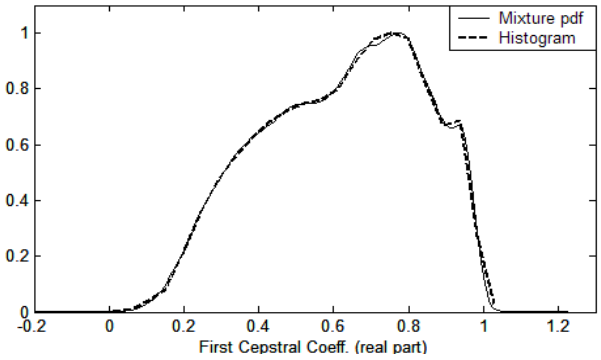

Figure 3: Fitting of mixture pdf (120 groups) to the normalized histogram of the first cepstral coefficients of the band 0-5.5kHz of the pink noise training set

In Fig. 3 the validity of the estimation algorithm, as described in section 2.2, is shown. Even though a mixture model of 40 groups would be sufficient (as determined by the MDL criterion), we increase this number to 120 and overfit the model for all 4 training subbands as explained in section 2.4. The fitting of the mixture pdf to the histogram is still very accurate which is attributed to the high modeling flexibility of the Generalized Gaussian pdf.

3.3. Cepstral Conversion Results

The cepstral conversion algorithm described in section 2 is implemented according to the experimental conditions of Table 1.

Analysis Subbands	Frame Slide/Length		Cepstral Order	
	Train(ms)	Test(ms)	Train	Test
1 - 13	10/15	50/76.2	30	8
14 - 17	10/15	50/76.2	30	15

Table 1: Experimental parameters.

The frequency regions of each of the analysis subbands in the left-most table column can be found in Fig.1. The frame slide and length are different for the training and testing segments as explained in section 2.4.

We now show the necessity of the conversion optimization scheme of section 2.4 by testing two scenarios where cepstral smoothing and data overfitting are not applied at the same time and which lead to increased cepstral reconstruction errors. In case A, resynthesis is applied with cepstral smoothing but no data overfitting (i.e. we derive a mixture pdf of 40 groups instead of 120), while in case B resynthesis is applied with overfitting (120 groups) but no smoothing (i.e. the training and testing recordings have both frame slide 10ms and frame length 15ms). The results are shown in Table 2.

Average Quadratic Cepstral Distance Between	Band 15	Band 16	Band 17
Target-Source (frame slide/length 50ms/75ms)	0.3537	0.4176	0.3597
Target-Resynthesis case A (no overfitting)	0.0840	0.0810	0.1035
Target/Resynthesis case B (no smoothing)	0.1695	0.1665	0.1920
Target-Resynthesis (smoothing+overfitting)	**0.0585**	**0.0556**	**0.0720**

Table 2: Two poor cepstral reconstruction scenarios A,B for subbands 15-17 and the case where cepstral smoothing and data overfitting are applied together.

The conversion results for the remaining subbands (1-14) are shown in Table 3. It is clear that the error reduction due to resynthesis varies across the subbands. However, the average cepstral distance between the

testing target and resynthesized segments is of the same order of magnitude for most of the subbands which means that the cepstral conversion technique has finite accuracy. By decreasing the duration of the testing segments and thus the number of cepstral vectors, the accuracy would increase but so would the conversion parameters overhead since more conversion parameters would have to be transmitted per unit length of testing segment.

Analysis Subband	Cepstral Distance Target-Source	Cepstral Distance Target-Resynthesis
1	9.4805E-4	5.5476E-4
2	0.0100	0.0057
3	0.0400	0.0179
4	0.0236	0.0139
5	0.0765	0.0269
6	0.1045	0.0260
7	0.0687	0.0230
8	0.0888	0.0259
9	0.2509	0.0709
10	0.2549	0.0723
11	0.9047	0.0463
12	0.0637	0.0157
13	1.9144	0.0305
14	1.1702	0.1145

Table 3: Average quadratic cepstral conversion results for subbands 1-14.

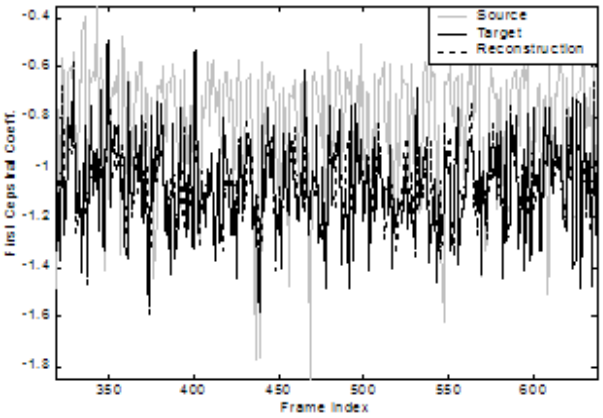

Figure 4: Cepstral reconstruction of the first coordinate for subband 17 (19.3-22kHz)

In Fig. 4, an example of cepstral conversion for subband 17 is shown. It is clear that the resynthesized first cepstral coefficients follow the corresponding target coefficients closely. Subbands 1-8 and 12 do not show observable errors since the initial distance between the source and target cepstral coefficients are small.

Finally, from Tables 2 and 3 we observe that the cepstral distance between the source and target signals greatly increases for subbands 9-17 (except subband 12). This is directly related to the fact that the 64Kbps Mp3 coding scheme severely degrades the signal content around the frequency region 5.5-22kHz while it retains the lower subbands. This will be taken into account during the residual conversion implementation presented in the next section.

3.4. Residual Conversion Results and Redundancy

The residual conversion scheme described in section 2.6 is implemented. We extract the residual vectors according to the 17 subbands analysis and apply the same 4 subbands training model used for cepstral conversion. The high cepstral order of the model (30) allows for the inclusion of low-valued vector coefficients which are necessary for modeling the residual valleys. Low cepstral orders were also tested and led to larger residual reconstruction errors. Therefore, the selection of a high training cepstral order is favorable. As mentioned, the testing source and target residual vectors acquire the model parameters according to one of the 4 training subbands the particular testing subband belongs to.

3.4.1. Residual Intra-Band Redundancy

The residual conversion scheme as described previously requires a large amount of conversion parameters to be created. For a full reconstruction of all the residual vectors of a particular subband, the size of the conversion parameters would be as large as 60% of the size of the target (uncompressed) signal and several times larger than the source Mp3 signal. For this reason, we decide to downsample the testing source and testing target residual vectors before conversion. We tested downsampling factors of 2, 4 and 8 and the best combination in terms of conversion parameters size and reconstruction accuracy proved to be a downsampling factor of 4. After conversion, the reconstructed residual is resampled to the original rate by using the previous two samples at each time instance. Under this scheme, the audio quality does not decrease noticeably compared to a full reconstruction and the size of the residual conversion function becomes four times smaller.

3.4.2. Residual Inter-Band Redundancy

In Fig. 4, the average quadratic residual distances between source and target residuals for all subbands are plotted.

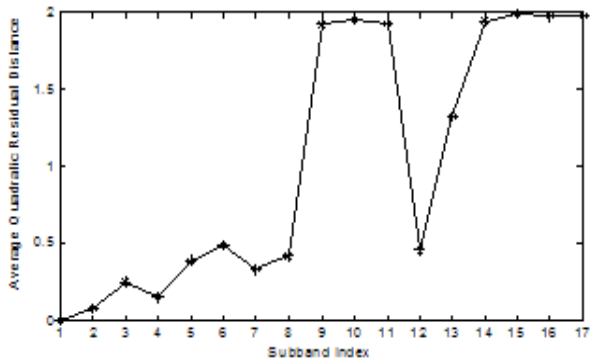

Figure 4: Average quadratic residual errors between source and target for all subbands.

It is clear that not all source subbands are heavily distorted. Subbands 1-8, 12 and 13 show small residual differences between the testing source and testing target segments. This means that we can apply residual processing to selected subbands only. Applying residual conversion to subbands 9-11 and 14-17 produced audible enhancement without deriving many conversion parameters or performing excessive computations. Processing the remaining subbands did not provide significantly better results that could justify the large amount of the resulting conversion parameters.

A further reduction in parameters is achieved by observing that the 4 highest testing subbands (14-17) show many residual similarities. By reconstructing only one of these residual signals and replacing all 4 residual signals with the particular reconstructed residual signal, a great reduction in the average quadratic residual distances for all 4 subbands is achieved. This is also attributed to the fact that the particular subbands content is not highly audible and the residual distances between source and target signals in these subbands are large. Thus, even a less accurately reconstructed residual is better than the original source residuals. This is shown below in Table 4 where the reconstructed residual is derived for subband 16 only and it is used for all 4 subbands. The residual conversion results for the remaining subbands are shown in Table 5. Each of these subbands has its own reconstructed residual since the lower subbands are very different to each other.

Analysis Subband	Residual Average Quadratic Distance Target-Source	Residual Average Quadratic Distance Target-Reconstruct.
14	1.9504	1.0682
15	1.9919	1.1139
16	1.9789	0.8743
17	1.9762	1.1072

Table 4: Average quadratic residual conversion results for subbands 14-17 when using the reconstructed residual of subband 16 for subbands 14-17.

Analysis Subband	Residual Average Quadratic Distance Target-Source	Residual Average Quadratic Distance Target-Reconstruct.
9	1.9206	0.8694
10	1.9520	0.8809
11	1.9268	0.8574

Table 5: Average quadratic residual conversion results for subbands 9-11 when using the corresponding reconstructed residuals.

The results of Tables 4 and 5 prove the validity of the residual conversion scheme. Subbands 9-11 and 16 have reduced their original residual errors more than 50%. Subbands 14, 15 and 17 have reduced their original residual errors around 45% but this reduction could be even more if each subband had its own reconstructed residual instead of sharing the residual derived from subband 16. Achieving an error reduction of 50% or more for these subbands does not actually provide any acoustical improvement of the enhanced waveform since, as mentioned, they do not contain the highly audible parts of the signal.

3.5. Overall Performance

Several objective similarity measures were tested among which the Mutual Information in the time domain proved to be the most suitable. Fig. 5 illustrates the effectiveness of the selected wavelet structure against wavelet trees of 2, 4 and 8 equidistant subbands. These cases are further subdivided in cases of cepstral reconstruction only and cepstral reconstruction with residual reconstruction. In the case of 2 subbands, residual conversion is applied in both subbands. In the case of 4 subbands, residual conversion is applied in the upper 3 subbands and in the case of 8 subbands residual conversion is applied in the upper 6 subbands.

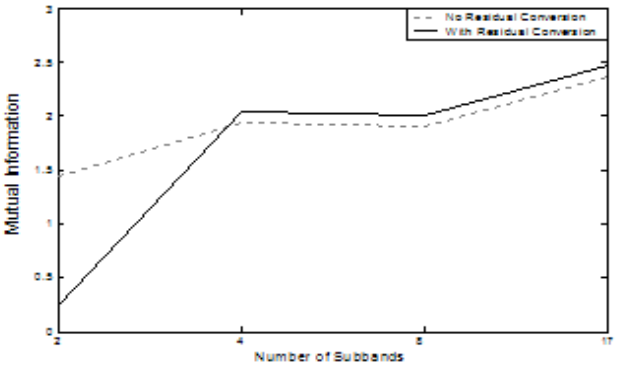

Figure 5: Mutual Information between testing target and resynthesized signals for various wavelet structures with and without residual conversion.

From Fig. 5 it is clear that audio enhancement is more efficient -in terms of conversion parameters size and quality improvement- when applying 17 bands wavelet separation with residual reconstruction. Even though residual processing does not increase dramatically the Mutual Information metric, the differences acoustically are very sharp and the resynthesized segment without residual conversion contains many periodic and random distortions. In contrast, audio enhancement with residual conversion does not cause any audible distortions as preliminary subjective tests show. The audio quality increase in the enhanced segment compared to the source segment is also easily perceptible.

To further illustrate this we provide some time domain waveform results of selected subbands when applying residual conversion and cepstral conversion under the 17 subband analysis. It is obvious from Fig. 6 that some subbands are severely degraded because the source waveform is almost non existent. The resynthesized signal follows much closer the target signal but as mentioned before there still exist residual differences between the target and resynthesized segments (see Tables 4 and 5) and therefore the two signals cannot be identical for subbands 9-17. Subbands 1-8 are not degraded enough (see Table 2 and Fig. 4) to show noticeable differences between the source and target waveforms and hence are not illustrated.

Table 6 shows the transmission requirements of our scheme when transmitting the cepstral conversion and residual conversion parameters under the 17 subbands separation. No arithmetic coding is applied to compress the conversion parameters set and therefore it is possible that the transmission size can be further reduced. Some

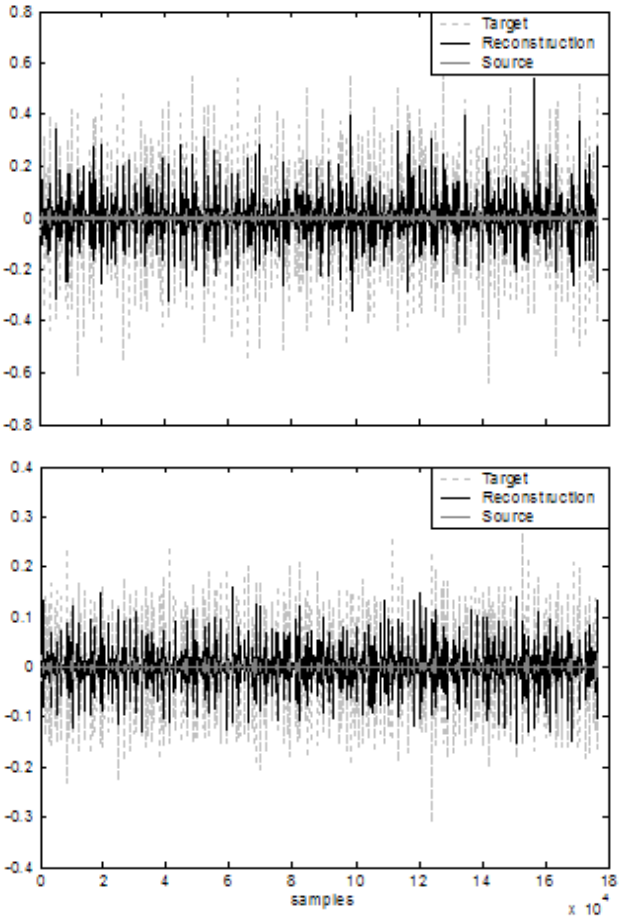

Figure 6: Time domain resynthesis results for subbands 11 (upper plot) and 17 (lower plot).

of the lower subbands can also not be processed at all (no cepstral conversion) since for these the source and target cepstral differences are very small.

Mp3 Source 64Kbps size (kbytes)	Conversion Parameters size (kbytes)	Target WAV size (kbytes)
252	195	2744

Table 6: Amount of transmitted conversion parameters compared to the source and target segment sizes.

As shown in Table 6, the conversion function size is smaller than the Mp3 source signal (77% of the source size) and much smaller than the target segment size. If we do not apply cepstral conversion for subbands 1-8 then the parameters size would be 155kBytes (61% of the source size).

4. CONCLUSIONS AND FUTURE RESEARCH

We presented a novel technique on audio quality enhancement of low bit rate Mp3 signals. Subjective tests are currently underway but the quality improvement is particularly audible since the source segment is Mp3 encoded in very low bit rate and therefore it is severely degraded. We have shown through objective means that the resynthesized signal is closer to the target (than the source is) in terms of cepstral and residual distances and also in the time domain by illustrating some subband waveforms.

The selection of subbands that need residual or cepstral conversion can be determined robustly by processing only the subbands that contain the highest residual or cepstral errors, respectively. Further investigation is needed on determining the optimal number of subbands since it is clear that a high number of subbands improves the enhancement performance and can also allow for detecting more redundancies (e.g. source subbands that are not degraded). The residual conversion scheme could be possibly further improved by selecting a higher cepstral order for the training model.

Finally, if we apply the resynthesis scheme to a 128kbps Mp3 source (which has double the size of the currently used source) the relative reduction in conversion parameters would be double the current one (38% of the source size) or more since it is possible that fewer subbands would need residual (or cepstral) conversion. Higher bit rate Mp3 source segments are currently being tested and naturally the algorithm performance is better since the overall differences between the source and target audio segments are smaller.

5. ACKNOWLEDGEMENTS

Research presented in this paper was funded in part by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152 and in part by the US Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or policy of the National Science Foundation or the US Government and no official endorsement should be inferred.

6. REFERENCES

[1] A. Mouchtaris, S. S. Narayanan and C. Kyriakakis, "Multi-resolution spectral conversion for multichannel audio resynthesis", *IEEE Proc. Int. Conf. Multimedia and Expo (ICME)*, vol.2, (Lausanne, Switzerland), pp.273-276, August 2002.

[2] Y. Stylianou, O. Cappe and E. Moulines, "Continuous probabilistic transform for voice conversion", *IEEE Trans. Speech and Audio Processing*, vol.6, no.2, pp.131-142, March 1998.

[3] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis", *IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, WA, May 1998, pp.285-289.

[4] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Trans. Speech and Audio Processing*, vol.3, no.1, pp.72-83, January 1995.

[5] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[6] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, NY, 1981.

[7] F. Muller, "Distribution shape of two-dimensional DCT coefficients of natural images", *Electronics Letters*, vol.29, no.22, pp.1935–1936, 1993

[8] J. Rissanen, "Modeling by shortest data description", *Automatica*, vol.14, pp.465–471, 1978.

[9] B. Gillett and S. King, "Transforming Voice Quality", *Eurospeech*, pp.1713-1716, 2003.

[10] G, Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge, 1996.

[11] D. Sinha and A.H. Tewfik, "Low Bit Rate Transparent Audio Compression using Adapted Wavelets", *IEEE Trans. Signal Processing*, vol.41, pp.3463-3479, December 1993.

[12] P. Noll, *MPEG Digital Audio Coding Standards*, CRC Press LLC, 2000.