Journal of Information Hiding and Multimedia Signal Processing ©2017 ISSN 2073-4212
Ubiquitous International Volume **8**, Number **6**, November **2017**

Enrichment of Audio Signal using Side Information

Akinori Ito

Graduate School of Engineering
Tohoku University
6-6-05 Aramaki aza Aoba, Sendai, 980-8579 Japan
aito@spcom.ecei.tohoku.ac.jp

Received January 2017; revised March 2017

Abstract. *This paper describes methods that add values to audio signals using side information. Many acoustic signal processing methods have been proposed for estimating the lost information from the original signal. Using the appropriate side information, we can enhance the estimation easily. In this paper, the principle of audio signal processing using side information is described first, and then three applications are described: packet loss concealment of audio signal, manipulation of mixed music signal and frequency band extension of telephone speech.*
Keywords: Audio signal processing, Side information, Packet loss concealment, Signal separation, Band extension

1. **Introduction.** Many research works have been conducted for developing advanced services and applications applying speech and audio signal processing, such as data compression applied to speech and audio coding [1, 2], blind source separation [3], speech enhancement [4], and packet loss concealment [5].

Many of these works have a common topic, in which we want to compensate for the lost information in the original audio signal by signal propagation, contamination by environmental noise or packet losses. For example, the problem of the single-channel blind source separation can be formulated as follows. Let $x_1(t)$ and $x_2(t)$ be the mutually-independent signals. When the signal

$$y(t) = x_1(t) + x_2(t) \tag{1}$$

is observed, we want to know the original signals $x_1(t)$ and $x_2(t)$ using only $y(t)$. As individual information of $x_1(t)$ and $x_2(t)$ is lost, we need some kind of assumption (such as individuality) to estimate $x_1(t)$ and $x_2(t)$. Another example is the packet loss concealment. Suppose we have contiguous packets of an audio signal $\boldsymbol{x_1, x_2, ..., x_N}$. When $\boldsymbol{x_i}$ is lost, we estimate the lost packet based on some kind of assumption (such as continuity of the signal).

In both cases (and also in most of the similar cases), we make assumptions on the signal to be estimated, such as the distribution of the signal, correlation between the signals, spectral shape, temporal fluctuation, etc. There are two points to be considered for designing an estimation method: how to make a good assumption on the signal to be estimated, and how to estimate the signal according to the assumption. The evaluation measure of these kinds of estimation methods is the quality of estimated signal (compared with the true signal), which is often measured by the signal-to-noise (or distortion) ratio or other signal quality indices such as PESQ [6] or PEAQ [7].

On the other hand, the same process can be viewed as communication between the sound source (sender) and the observer (receiver). The sender has all the information on the sent signal. Therefore, if the sender knows that parts of information of the signal will be lost in the communication channel, the sender can send additional information individually that can be used to recover the lost information during transmission. Using the additional information, the original signal can be estimated easily using simple algorithms. The additional information can be either appended to the original signal using a specific data format, or embedded into the original signal using data hiding methods [8, 9, 10, 11, 12]. In this case, the method (design of the additional information and the method of recovering the original signal) is evaluated by both the rate of the additional information and the quality of the recovered signal.

In this paper, I describe the audio signal transmission methods using additional information.

2. **Side information for signal estimation.** In general, design of the side information for signal estimation strongly depends on the problem we want to solve. In this section, let us consider a simple case where we want to estimate a scalar value x when it is lost.

The simplest side information that can be used for estimating x is x itself. If the value of x is lost randomly by probability p, the loss probability becomes p^n by repeating the value n times. Information to be transmitted grows n times.

A general method to compensate for the lost information using fewer side information is the forward error correction (FEC) using error-correcting code. Using the simple parity-based error correction method for packet loss recovery [13], we send one parity packet for every $n-1$ normal packets. If one packet among the n packets is lost, we can recover the lost packet by gathering all of $n-1$ packets. The bitrate becomes $n/(n-1)$ times larger than the original bitrate, while the virtual packet loss rate becomes $p(1-(1-p)^{n-1})$ when the physical packet loss rate is p. We can exploit more efficient FEC using more sophisticated error-correcting codes, such as Reed-Solomon code [14], Turbo code [15] and LDPC code [16].

The FEC-based side information can be used on any kind of media, including audio, image, video and text. If x is a kind of media data such as audio and image that permits small amount of error, various kinds of side information can be exploited. In this case, we investigate a side information design that gives a good balance between the amount of side information and the quality of the recovered signal. For example, when sending a sequence of quantized samples, we can use side information based on a coarse quantizer [17]. In this method, we prepare two quantizers: the fine quantizer Q_F and the coarse quantizer Q_C. When sending a sample x, we calculate two quantized samples $Q_F(x)$ and $Q_C(x)$, and send them independently. When $Q_F(x)$ is lost, x is recovered using $Q_C(x)$. The finer the quantization of Q_C is, the better the quality of the recovered signal is, but the bitrate increases. If Q_F and Q_C are identical, it is the same framework as the method that sends the same data repeatedly.

The above-mentioned method recovers the lost sample by only using the side information that corresponds to the lost sample. If the data to be recovered is a sequence with some correlation between samples, we can use the information of the samples just before or after the lost sample. The framework of the packet loss concealment uses the content of the previously received packet as the estimation of the lost packet [5]. Our group has proposed a more general framework, where a pair of values with correlation (x_1, x_2) is transmitted [18]. This method calculates two side information $f(x_1)$ and $f(x_2)$ from x_1 and x_2, and sends two pairs of tuples $\langle x_1, f(x_2)\rangle$ and $\langle x_2, f(x_1)\rangle$ independently.

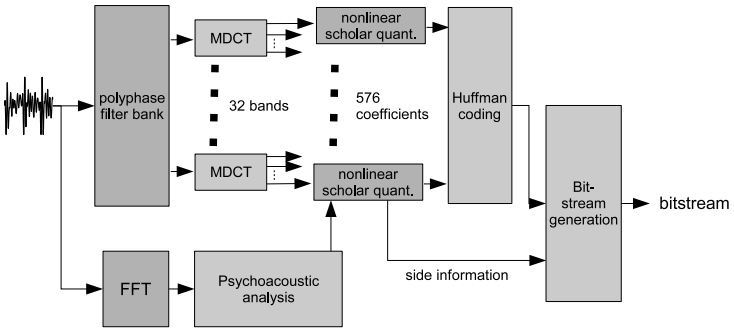

FIGURE 1. A schematic diagram of the MP3 encoder

If $\langle x_2, f(x_1)\rangle$ is lost, x_2 is recovered from x_1 and $f(x_2)$ by

$$\hat{x}_2 = g(x_1, f(x_2)). \tag{2}$$

Under this framework, we can control the total bitrate by changing the side information encoder f. Here, f is not necessarily invertible; in other words, we can only exploit the part of x for calculating f. For example, if x is a speech signal in a frame, $f(x)$ may contain only the spectral envelope, fundamental frequency or linear prediction residue. In this case, the information that is not contained in $f(x)$ is estimated from the surrounding signal of the lost signal. This method can be viewed as a continuous version of Unequal Error Protection [19] that protects only a part of signal using the FEC framework.

3. **Application.** Next, several examples of application are introduced that add the audio signal some kind of value using side information.

3.1. Packet loss concealment of MP3 audio.

3.1.1. *Side information using sign information.* When transmitting continuous signals such as speech or image, those signals are quantized and encoded first, and then divided into packets to transmit over the network. In most cases, TCP [20] is used as a transport protocol for transmission over the Internet. However, when the real-time factor is important or the system uses IP multicast, RTP [21] is often chosen as the protocol. While the RTP gives a mechanism to detect packet loss, it does not provide any method to recover the lost packet, and thus we need to prepare the packet loss concealment mechanism [5] in the application. Our research group developed an advanced packet loss concealment methods for MP3 audio packets [22, 23].

Figure 1 shows the block diagram of an MP3 encoder. The input signal is analyzed into 32 frequency bands by the polyphase filterbank, and the signal of each frequency band is further analyzed by the modified discrete cosine transform (MDCT). The window length of MDCT differs according to the stationarity of the signal; typically the window size is 18, and finally the input signal is transformed into 576-dimensional frequency domain. The MDCT coefficients are quantized by the non-linear scalar quantizer dimension by dimension, and the levels of quantization are determined based on the permitted noise level calculated using the psychoacoustic model. Finally, the quantized MDCT coefficients and the side information that contains the gains of the critical bands are packed into a granule, and two granules are sent by a packet.

Since the most important information in a MP3 packet is the MDCT coefficients, it is important to estimate the MDCT coefficients precisely when a packet is lost. The

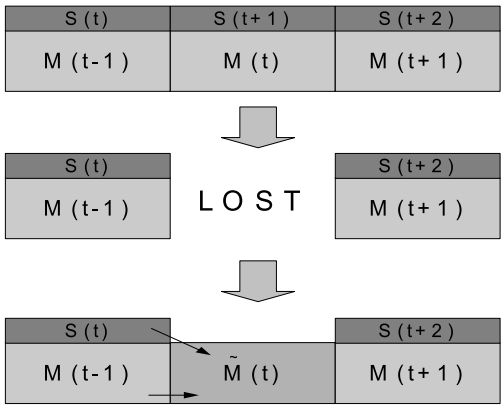

FIGURE 2. Packet loss concealment using sign information

packet loss concealment method for VoIP often uses the content of the previous packet as the content of the lost packet. However, using the content of the previous (or next) packet causes severe deterioration of the signal when using the codec based on MDCT such as MP3. When transforming the MDCT coefficients using the inverse MDCT, the transformed time-domain signal contains the time-domain aliasing noise. This noise is canceled by overlap-adding with the previous and next frame (time-domain aliasing cancellation) [24]. When a packet is lost and the lost packet is substituted by the previous (or next) packet, the time-domain aliasing noise is not canceled but emphasized by the overlapping with another aliasing noise. A method was proposed that estimates the time-domain aliasing noise using an iterative calculation [25], but it is not suitable for real-time processing.

Therefore, our group developed a method to improve the estimation of MDCT coefficients by using the MDCT coefficients of the adjacent packet and exploiting one-bit side information for one MDCT coefficient [22]. This method uses the sign (plus or minus) of an MDCT coefficient as side information. Let the MDCT coefficients at the t-th frame be

$$\boldsymbol{M}(t) = (M_1(t), \ldots, M_{576}(t)). \tag{3}$$

Then the side information of the t-th frame is

$$\boldsymbol{S}(t) = (\mathrm{sign}(M_1(t)), \ldots, \mathrm{sign}(M_K(t))) \ (K \leq 576) \tag{4}$$

where

$$\mathrm{sign}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{otherwise} \end{cases}. \tag{5}$$

Here, $0 < K \leq 576$ is a constant that determines the upper limit of the side information. The side information $\boldsymbol{S}(t)$ is appended to either the $(t-1)$-th or $(t+1)$-th packet. If the t-th packet is lost, the MDCT coefficients of the t-th packet is estimated as follows.

$$\tilde{M}_i^{(s)}(t) = \begin{cases} \mathrm{sign}(M_i(t))|M_i(t-1)| & \text{if } i \leq K \\ M_i(t-1) & \text{otherwise} \end{cases}. \tag{6}$$

Figure 2 shows the estimation of the lost packet using the proposed method. According to the experimental result, it was confirmed that $K = 50$ gives a sufficient quality of the recovered signal. When $K = 50$, bitrate of the side information becomes around 8 kbit/s.

3.1.2. *Concealment method switching.* To improve the quality of the packet loss concealment, our group proposed a method that adaptively selects the concealment methods [23]. This method uses two concealment methods; the first one is the method explained in the

TABLE 1. Average number of selection made as preferred quality by PLC method

Method	Classic	Jazz	Pop
Linear weighting	3.7	3.0	1.7
Method switching	6.3	7.0	8.3

previous section, and the other method is based on the one-bit quantization [18]. The one-bit quantization method estimates the MDCT coefficients as follows.

$$\tilde{M}_i^{(q)}(t) = \mathrm{sign}(M_i(t))\sqrt{\frac{2\sigma_i^2}{\pi}} \tag{7}$$

Here, σ^2 is the variance of M_i,

$$\sigma_i^2 = E\left[M_i(t)^2\right]_t \tag{8}$$

and it is estimated from the previously received frames. In [23], the variance was calculated using the previous 50 frames.

The combined method adaptively adds the results of the two methods,

$$\tilde{M}_i^{(c)}(t) = w_i\tilde{M}_i^{(s)}(t) + (1 - w_i)\tilde{M}_i^{(q)}(t) \tag{9}$$

$$w_i = \frac{|\rho_{i1}|(|\rho_{i2}| - 1)}{2|\rho_{i1}\rho_{i2}| - |\rho_{i1}| - |\rho_{i2}|} \tag{10}$$

here ρ_{i1} is the correlation coefficent between $|M_i(t)|$ and $|M_i(t-1)|$, and ρ_{i2} is the correlation coefficient between $M_i(t)$ and $\mathrm{sign}(M_i(t))$. The combined method can reduce the total error; however, it causes large errors sporadically, which lowers the subjective quality of the signal. Therefore, the method switching fexamines both $\tilde{M}_i^{(s)}(t)$ and $\tilde{M}_i^{(c)}(t)$, and uses the method that minimizes the error. This method requires further additional side information that denotes which method was used for the t-th frame.

A subjective evaluation experiment was conducted to investigate the effect of concealment method switching. Three music clips were chosen from each of three genres (pop, jazz and classic) in RWC music database [26]. Three phrases were chosen from each of the clips, which was about 10 s long, as test stimuli. After encoding these signals using MP3 encoder, packet losses were simulated under the condition that packet loss rate was 10% and average length of burst packet loss was 3. In the subjective evaluation, two signals were presented to a subject in a random order, which were recovered using the above-mentioned two concealment methods. A subject chose one out of the two signals which was perceived as a high quality signal. 10 subjects participated the experiment.

Table 1 shows the result. The numbers in the table shows the average number of subjects who chose that method. This result shows that method switching gave better results compared with the previous method.

3.2. **Manipulation of music signal using side information.** There has been a number of research works that separates a mixed audio signal into individual sound [3]. Specifically, segregation methods of mixed music signals are extensively studied as a part of music information processing research [27, 28]. While many of the source separation methods exploit multiple microphones combined with a method such as independent component analysis [29], music signal separation methods do not assume multiple microphones because most of commercial music signals are produced by artificially mixing the recorded signals of the individual parts. Instead, many music signal separation methods assume a specific structure of the spectra, such as harmonics. For example, the music signal separation method proposed by Itoyama et al. [28] separates the input music signal into the harmonic signal and non-harmonic signal. Since it is not easy to completely separate

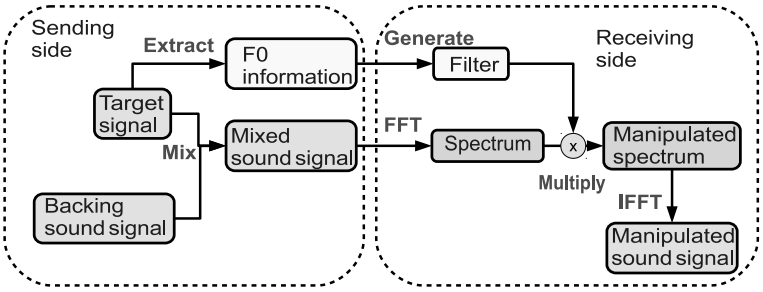

FIGURE 3. Signal manipulation using side information

the mixture of multiple instruments by only using the assumption of harmonics, several methods uses the score of the signal as a "clue" of separation [30, 31]. These kinds of methods are called "score-informed signal separation". These methods estimates the correspondence between the score and the input signal, and separates the input signal under the knowledge of the instrument and its sound.

One application of these methods is remixing, where an artist splits an existing music clip into individual instrumental sounds and changes the instrument to create another version of the same music. Here, if the author of the music assumes that the produced music will be used for creating derivative works, the author can embed information useful for extracting individual instrumental sound into the produced music. Under this framework, several research group developed a method to manipulate the individual instrumental or vocal sound in a mixed music signal using side information [32, 33, 12].

Figure 3 shows the proposed framework. To employ this framework, we make the following assumptions.

- The sender side has signals of individual instruments or vocal signal and creates the final music signal by digitally mixing those signals.
- Before mixing the signals of individual parts, the sender extracts side information of a part to create the side information.
- The mixed music signal and the side information are sent to the receiver individually.
- Based on the side information, the receiver makes a filter to manipulate the mixed music signal.

The prototype system realized manipulating the volume of the vocal part in the mixed music signal. The most important side information is the fundamental frequency (F0) of the vocal signal.

The overview of the proposed system is as follows. The input music signal (in the frequency domain) $I(f)$ is created by adding the vocal signal $V(f)$ and the backing signal $B(f)$.

$$I(f) = V(f) + B(f) \tag{11}$$

Before mixing $V(f)$ and $B(f)$, the side information η is extracted from both of $V(f)$ and $B(f)$.

$$\eta = f_{side}(V(f), B(f)) \tag{12}$$

At the receiver side, a filter $G(f; \eta, A)$ is applied to the input signal $I(f)$ to create the output signal $O(f)$. Here, A is the target amplitude of the vocal signal in the manipulated signal.

$$O(f) = I(f)G(f; \eta, A) \tag{13}$$

To emphasize or suppress the vocal signal based on the fundamental frequency, the filter G was designed as a comb filter having the fundamental frequency of the vocal signal.

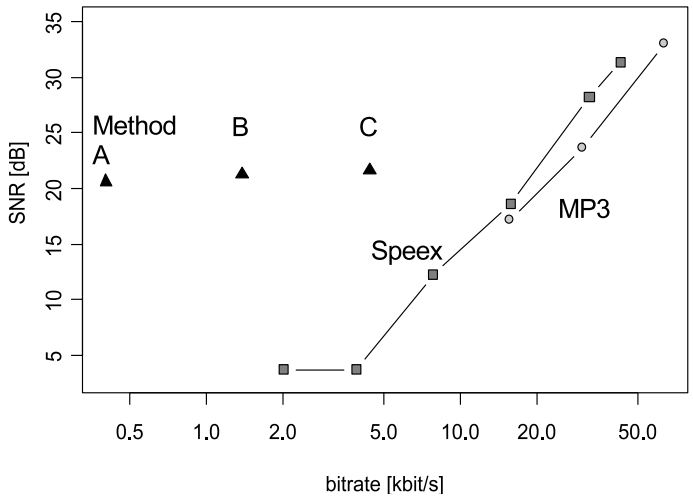

FIGURE 4. Quality of manipulated signal with respect to bitrate of the side information

The comb filter was created by superimposing the Gaussian functions such that its mean value is an integral multiple of the fundamental frequency.

$$G(f; f_0, \sigma, K, \alpha, \beta_x(k), A) \tag{14}$$
$$= 1 + \alpha(A-1)\beta(k) \exp\left(-\frac{(f - kf_0)^2}{2\sigma^2}\right)$$

Here, f_0 is the fundamental frequency, K is the number of harmonics, α is the parameter to control the amplitude, σ is the bandwidth of each Gaussian filter and $\beta(k)$ is the amount of manipulation of the k-th harmonic component. The side information is $\eta = (f_0, \beta(1), \ldots, \beta(K))$, and the other parameters are fixed to the optimum values.

Since the filter G is applied to the mixed music signal, not only the vocal signal but also the backing signal is affected by the filter. When the backing signal is stronger than the vocal signal within the bandwidth of a Gaussian component of the filter, it is known that the manipulated signal deteriorates. This situation happens when the vocal signal and the backing signal are correlated (e.g., the vocal and backing signal play the same melody), and the deterioration becomes larger.

To compensate the effect of the backing signal on the manipulated signal, the following three kinds of $\beta(k)$ were compared.

(A) Uniform $\beta(k)$

$$\beta(k) = 1 \tag{15}$$

(B) The component is manipulated when the vocal signal is larger than the backing signal

$$\beta(k) = \begin{cases} 1 & \text{if} \quad |V(kf_0)| > \theta|B(kf_0)| \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

(C) β is determined based on the ratio of the vocal and backing signal

$$\beta(k) = \frac{|V(kf_0)|}{|V(kf_0)| + |B(kf_0)|} \tag{17}$$

Figure 4 shows the result of the evaluation experiment. The evaluation data was the vocal signal by one male singer and the backing signal. The signal-to-noise ratio of the manipulated signal was measured when the amplitude of the vocal signal was emphasized twice. The proposed method was compared with the conventional method, where the vocal

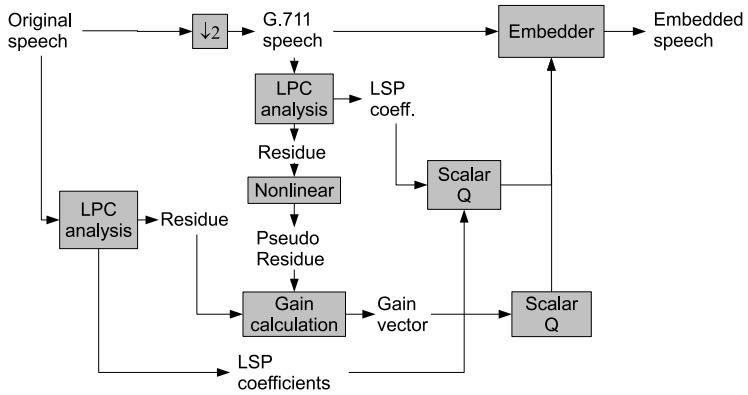

FIGURE 5. A block diagram of the encoder of telephone speech band extension

signal was compressed using existing codecs and used as the side information. Speex [34] and MP3 were used as the codecs. This result shows that the proposed method realizes manipulation of the vocal signal in a high quality with small amount of side information compared with using the existing speech codecs. The qualities of signal given by the proposed three methods (method A, B and C) are different, but the difference was not as large as expected.

3.3. **Frequency band extension of telephone speech.** Bandwidth of a normal telephone speech is less than 4 kHz to suppress bitrate. This narrowband speech is enough to convey linguistic information, but the wideband speech (about 8kHz bandwidth) is more suitable to transmit speaker individuality or to provide a good intelligibility under noisy environment. To achieve this, layered codecs have been proposed to switch the normal telephone speech and high quality speech [35, 36]. However, these codecs are not popular, especially as the codecs for home telephone line, because these layered codecs do not have backward compatibility to the codec of the fixed telephone line. To enable the home telephone line to communicate with a wideband speech, many works have been conducted to convert the speech through the normal telephone line into a wideband speech [37]. This task is to estimate the high frequency band from the low frequency band, and it is easily achieved using side information [38, 9, 39].

Figure 5 shows a block diagram of the frequency band extension method proposed by the author's research group [39]. This method calculates the side information from the high frequency gain and the LSP parameters of higher frequency band, and embeds it to the G.711 speech. The bitrate of the side information is 1.25 kbit/s. As a result of evaluation experiment, 3.27 MOS-LQO (Mean Opinion Score-Listening Quality Objective). The MOS-LQO [40] is an objective measure of sound quality that takes value of 1 (low) to 5 (high) similar to the subjectve evaluation, and is calculated from the value of PEAQ.

4. **Conclusion.** This paper introduced several research works, mostly conducted by the author's research group, which use the side information to achieve advanced signal processing and add value to the audio signal. Beside the works introduced here, there are several works such as embedding codes of lyrics in an audio signal [41] or embedding features of facial expression into the speech signal [42]. Omachi et al. proposed a method to embed side information into images to enhance the performance of character recognition [43]; similar method can be used for an audio signal.

Acknowledgment. The research works introduced in this paper have been done with many colleagues. I appreciate Prof. Shozo Makino (Professor Emeritus of Tohoku University, Professor of Tohoku Bunka Gakuen University), Prof. Yôiti Suzuki (Tohoku University), Prof. Motoyuki Suzuki (Osaka Institute of Technology), Dr. Seongjun Hahm (Capio) and other students and colleagues.

REFERENCES

[1] A. S. Spanias, Speech Coding: a tutorial review, *Proc. Of IEEE*, vol. 82, no.10, pp. 1541-1582, 1994.
[2] T. Painter and A. Spanias, A review of algorithms for perceptual coding of digital audio signals, *Proc. Int. Conf. On Digital Signal Processing*, pp. 179-208, 1997.
[3] M. S. Pedersen, J. Larsen, U. Kjems and L. C. Parra, A survey of convolutive blind source separation methods, *J. Benesty et al., Springer Handbook on Speech Communication, Springer-Verlag*, Berlin, 2008.
[4] Y. Ephraim and I. Cohen, Recent Advancements in Speech Enhancement, *Circuits, Signals, and Speech and Image Processing*, 2006.
[5] C. Perkins, O. Hodson and V. Hardma, A survey of packet loss recovery techniques for streaming audio, *IEEE Network*, vol. 12, no. 5, pp. 40-48, 1998.
[6] A. W Rix, M. P Hollier, J. G Beerends and A. P Hekstra, PESQ-the new ITU standard for end-to-end speech quality assessment, *emAudio Engineering Society Convention 109, Audio Engineering Society*, 2000.
[7] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends and C. Colomes, PEAQ-The ITU standard for objective measurement of perceived audio quality, *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3-29, 2000.
[8] N. Komaki, N. Aoki and T. Yamamoto, A packet loss concealment technique for VoIP using steganography, *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol. E86-A, no. 8, pp. 2069-2072, 2003.
[9] N. Aoki, A band extension technique for G.711 speech using steganography, *IEICE Trans. Communications*, vol. E89-B, no. 6, pp. 1896-1898, 2006.
[10] A. Ito and S. Makino, Data hiding is a better way for transmitting side information for MP3 bitstream, *Proc. Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 495-498, 2009.
[11] M. Parvaix, L. Girin and J.-M. Brossier, A Watermarking-Based Method for Informed Source Separation of Audio Signals with a Single Sensor, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1464-1475, 2010.
[12] M. Parvaix and L. Girin, Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 6, pp. 1721-1733, 2011.
[13] J. Rosenberg and H. Schulzrinne, An RTP Payload Format for Generic Forward Error Correction, RFC2733, 1999.
[14] I. S Reed and G. Solomon, Polynomial codes over certain finite fields, *Journal of the society for industrial and applied mathematics*, vol. 8, no. 2, pp. 300-304, 1960.
[15] C. Berrou, A. Glavieux and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo codes, *Proc. Int. Conf. on Communication*, pp. 1064-1070, 1993.
[16] R. Gallager, Low-density parity-check codes, *IRE Transactions on information theory*, vol. 8, no. 1, pp.21-28, 1962.
[17] W. Jiang and A. Ortega, Multiple description speech coding for robust communication over lossy packet networks, *Proc. IEEE Int. Conf. Multimedia & Expo*, vol. 1, pp. 444-447, 2000.
[18] A. Ito and S. Makino, Designing Side Information of Multiple Description Coding, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 1, pp. 10-19, 2010.
[19] A. R. Calderbank and N. Seshadri, Multilevel Codes for Unequal Error Protection, *IEEE Trans. on Information Theory*, vol. 39, no. 4, 1993.
[20] J. Postel, Transmission control protocol, RFC793, 1981.
[21] C. Perkins, RTP: Audio and Video for the Internet, *Addison-Wesley*, 2003.
[22] A. Ito, T. Sakai, K. Konno, S. Makino and M. Suzuki, Packet Loss Concealment for MDCT-based Audio Codec Using Correlation-based Side Information, *Int. J. of Innovative Computing, Information and Control*, vol. 6, pp. 1347-1362, 2010.

[23] A. Ito, K. Konno, M. Ito and S. Makino, Robust Transmission of Audio Signals over the Internet: an Advanced Packet Loss Concealment for MP3-based Audio Signals, *Interdisciplinary Information Sciences*, vol. 18, no. 2, pp. 99-105, 2012.
[24] Y. Wang and M. Vilermo, Modified discrete cosine transform — its implications for audio coding and error concealment, *AES Journal*, vol. 51, pp. 52-61, 2003.
[25] H. Ofir, D. Malah and I. Cohen, Audio packet loss concealment in a combined MDCT-MDST domain, *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 1032-1035, 2007.
[26] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, RWC Music Database: Popular, Classical and Jazz Music Databases, *ISMIR*, vol. 2, pp. 287-288, 2002.
[27] Y. Kitano, H. Kameoka, Y. Izumi, N. Ono and S. Sagayama, A Sparse Component Model of Source Signals and Its Application to Blind Source Separation, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4122-4125, 2010.
[28] K. Itoyama, M. Goto, K. Komatani, T. Ogata and H. Okuno, Instrument Equalizer for Query-by-Example Retrieval: Improving Sound Source Separation based on Integrated Harmonic and Inharmonic Models, *Proc. Int. Society of Music Information Retrieval Conf. (ISMIR)*, pp. 133-138, 2008.
[29] S. Choi, A. Cichocki, H.-M. Park and S.-Y. Lee, Blind Source Separation and Independent Component Analysis: A Review, *Neural Information Processing - Letters and Reviews*, vol. 6, pp. 1-57, 2005.
[30] J. Woodruff, B. Pardo and R. Dannenberg, Remixing Stereo Music with Score-Informed Source Separation, *Proc. Int. Society of Music Information Retrieval Conf. (ISMIR)*, pp. 314-319, 2006.
[31] R. Hennequin, B. David and R. Badeau, Score informed audio source separation using a parametric model of non-negative spectrogram, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 45-58, 2011.
[32] Y. Sasaki and A. Ito, Manipulating vocal signal in mixed music sounds using small amount of side information, *Proc. Int. Conf. Intelligent Information Hiding and Multimedia Signal Processing*, pp. 298-301, 2011.
[33] L. Girin and J. Pinel, Informed audio source separation from compressed linear stereo mixtures, *Proc. AES 42nd International Conference: Semantic Audio*, pp. 159-168, 2011.
[34] J. M. Valin and C Montgomery, Improved noise weighting in CELP coding of speech-applying the Vorbis psychoacoustic model to Speex, *Proc. 120th AES Convention*, 2006.
[35] International Telecommunication Union, G.722.1 : Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss, *ITU-T Recommendation*, 2005.
[36] International Telecommunication Union, G.722.2 : Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB), *ITU-T Recommendation*, 2007.
[37] P. Jax and P. Vary, Bandwidth Extension of Speech Signals: A Catalyst for the Introduction of Wideband Speech Coding?, *IEEE Communications Magazine*, vol. 44, no. 5, pp. 106-111, 2006.
[38] A. Kataoka, T. Mori and S. Hayashi, Bandwidth extension of G. 711 using side information, *IEICE Trans. Inf. & Syst.(Japanese Edition)*, vol. J91-D, no. 4, pp. 1069-1081, 2008,
[39] A. Ito, H. Handa and Y. Suzuki, A Band Extension of G.711 Speech with Low Computational Cost for Data Hiding Application, *Proc. Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 491-494, 2009.
[40] International Telecommunication Union, P.862.1: Mapping function for transforming P.862 raw result scores to MOS-LQO, ITU-T Recommendation, 2006.
[41] A. Nishimura, Presentation of Information Synchronized with the Audio Signal Reproduced by Loudspeakers using an AM-based Watermark, *Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007)*, vol. 2, pp. 275-278, Nov, 2007.
[42] Y. Abe and A. Ito, Multi-modal Voice Activity Detection by Embedding Image Features into Speech Signal, *2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 271-274, Oct, 2013.
[43] S. Omachi, M. Iwamura, S. Uchida and K. Kise, Affine Invariant Information Embedment for Accurate Camera-Based Character Recognition, *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 2, pp. 1098-1101, 2006.