スピーカ再生音に同期した音響電子透かしを用いる情報提示 ―カラオケ歌詞表示システム―

西村　明[†]　**坂本　真一**[††]

† **東京情報大学 総合情報学部 情報文化学科**
〒 265-8501 **千葉市若葉区谷当町** 1200–2
TEL 043-236-4658, akira@rsch.tuis.ac.jp
†† **株式会社オトデザイナーズ**
〒 351-0104 **埼玉県和光市南** 1–27–65
TEL 048-201-2229, sakamoto@otodesigners.com

あらまし　振幅変調に基づく音響透かし技術を用いて、スピーカから再生されるカラオケの伴奏に同期して歌詞を表示するシステムを、パソコン上で動作するソフトウェアとして試作した。システムの特徴は、伴奏音再生機器と歌詞表示機器を分離できる点である。振幅変調に基づく音響透かし技術の特徴は、数秒の埋め込み時間フレームを用いることによって残響や反射音に対して頑強で、付加雑音を検出時にキャンセルする仕組みにより雑音にも頑強な点である。音楽信号をスピーカより放射し、残響と雑音が付加された上でマイクロホンで受音する条件をコンピュータ・シミュレーションにより検証し、埋め込んだ情報を有効に検出して十分な時間精度で表示できることが分かった。

Providing information which synchronized with the audio signal reproduced by loudspeakers using audio watermark — A system for displaying Karaoke lyrics —

Akira NISHIMURA[†] and Shinichi SAKAMOTO[††]

† Department of Media and Cultural Studies, Faculty of Informatics,
Tokyo University of Information Sciences
1200–1, Yatoh-cho, Wakaba-ku, Chiba-city, Chiba 265-8501, Japan
†† Otodesigners, 1–27–65, Minami, Wakou-city, Saitama 351-0104, Japan

Abstract　A watermarking technique using subband amplitude modulation was applied to a prototype system that displays Karaoke lyrics synchronously with the watermarked audio signal reproduced by loudspeakers. A distinctive feature of the system is the separation of the reproduction system and the display system for lyrics. The watermarking technique is robust against reflections and reverberations, because the technique applies relatively slow amplitude modulation in a long embedding frame of several seconds. The technique can cancel additive noise in the detection stage. The robustness of the system was evaluated by a computer simulation in terms of the correct rate of data transmission under reverberant and noisy conditions reproduced by a loudspeaker. The results showed that the performance of detection and the temporal precision of synchronization of display were sufficient.

1. はじめに

音響信号へのデータハイディング技術は、音響信号自体に気づかれないようにデータを埋め込み、必要なときに検出して利用する技術である。従来は、著作権管理を目的として、様々な変形を経ても埋め込みデータを検出できる音響電子透かし技術の開発が進められてきた。また、埋め込むデータの方に価値のある、ステガ

ノグラフィとよばれる技術も開発されてきた。後者は、前者より埋め込みデータ量を多くする必要があるため、前者ほど変形に対する耐性が高くできないことが特徴である。

近年、スピーカから再生されたデータ埋め込み済み音響信号を、ユーザの手元の機器で受信して復号化し、埋め込まれた情報を利用する、といった利用形態を想定した技術が開発されている [1]~ [3]。この技術においては、空間伝搬に伴う反射音や残響、背景雑音、スピーカやマイクによる周波数特性の歪などに対する耐性を保ちながら、埋め込む情報量を高める必要がある。

本稿では、こうしたデータ埋め込み済み音響信号の空間伝搬と受信を前提とした利用形態のひとつとして、データ埋め込み済み音響信号に同期してユーザへ情報表示する応用技術を示す。具体的な応用として、カラオケ伴奏音楽に歌詞の呈示情報を埋め込み、伴奏信号におけるデータの埋め込みフレーム時刻の同期検出を元に、歌詞の呈示タイミングに合わせて表示を行う。この技術は、カラオケだけでなく、映画における字幕情報の呈示など、音響信号と同期した情報の呈示/活用が必要な場面において有効である。

2. 空間伝搬耐性のある音響透かし

データ埋め込み済み音響信号の空間伝搬と受信を前提とした場合、音響信号に加わる変形としては、

- 背景雑音の重畳
- 反射音や残響音の重畳
- 発信側の DA 変換器と受信側の AD 変換器のサンプリング周波数の相違
- スピーカやマイクロホンによる周波数帯域の制限や変形
- スピーカにおける歪や、AD 変換時の過大入力によるクリッピング歪

などが挙げられ、これらの変形に対する耐性を備えた技術を用いることが重要である。また、伝送情報量はなるだけ多くとりながら、音質劣化は最小限に抑える必要もある。しかし、音楽観賞用でないスピーカからも再生されることを前提としている時点で、ある程度の音質劣化は見込まれているため、著作権管理用途の音響透かしのように、高性能な再生機器を用いてさえ検知が困難なほどの高音質は必要ないと考えられる。

これらの要求を、従来の音響透かし技術が満たすかどうかについて検討してみる。パッチワーク法 [4] は強度変化を与える時間幅が 100ms 程度と短いので、反射音や残響の影響を受けやすい。エコー拡散法 [5] は、反射音や残響の影響を軽減するためには、埋め込み時間フレームを長く設定する必要があり、埋め込みデータ量に制限がある。2 チャンネル伝送を前提として片チャンネル毎に強度変調を与える手法 [1] は、反対側チャンネルの影響が少なくなるよう、透かし検出時にはマイクロホンを一方のスピーカに近づける必要がある。また、スペクトル拡散法 [2] は、付加雑音には強いが、埋め込みデータ量に制限がある。

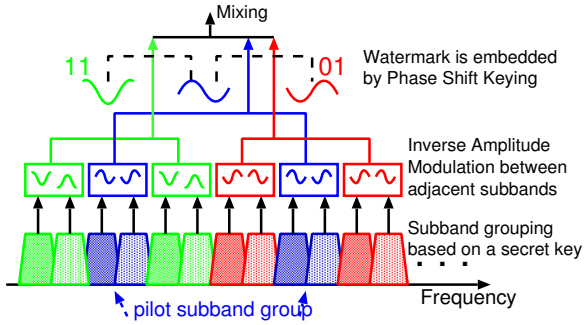

Fig. 1 埋め込み処理の概要

3. 振幅変調に基づく音響透かし

著者が考案した振幅変調に基づく音響透かし [3], [6], [7] は、空間伝搬を前提とする透かし技術に対する前述の要求を、ほぼ満たしている。ここでは、その埋め込みと埋め込み区間検出手法について簡単に触れる。詳細は文献を参照していただきたい。

3.1 埋め込み方法

本方式では、2 つの隣接する周波数帯域に分割された信号同士にそれぞれ逆位相の正弦振幅変調を与える。透かし埋め込み帯域を全て帯域分割し、このペアとなる隣接帯域を複数含む 2 つ以上のグループに各帯域を分類し、そのグループ間の変調位相差にすかし情報を埋め込む (図 1 参照)。埋め込みデータフレーム毎に、基本となるパイロット帯域グループの変調位相を反転させることによって、検出時に埋め込み区間の同期検出を可能とする。さらに、すべての帯域ペア間には、埋め込み時の鍵によってランダムに決定された初期変調位相差があらかじめ与えられる。透かしデータの符号化は、位相差 $\pi/2$ 毎に値を割り当てる 4 値の PSK 方式をとる。

3.2 埋め込み区間同期検出方法

透かし入り信号に対して、FFT をオーバーラップさせながら実行し、絶対値をとることによって振幅変動波形を得る。その後、パイロット帯域グループの変動波形に対して、埋め込み周期おきに加算と減算と繰り返して 1 周期分の累積変動波形を得る。これは埋め込み周期ごとに位相が反転しているため、反転分を補正して同

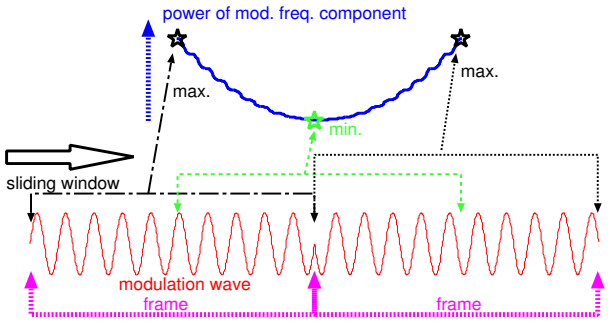

Fig. 2　埋め込み区間境界検出の概要.

期加算することに相当する。そのフレーム開始時刻を、埋め込み周期分だけずらしながら累積変動波形を求め、埋め込み変調周波数において最大変動パワーが得られたフレーム開始時刻を埋め込み区間の先頭として特定する (図 4 参照)。

4.　カラオケ歌詞表示試作システム

ここでは、データ埋め込み済み音響信号に同期してユーザへ情報表示するシステムとして、カラオケ伴奏音楽に同期して歌詞を呈示する試作システムについて説明する。

本システムにおいて、音楽信号に埋め込まれるデータは、主に表示開始時刻と表示終了時刻、そして表示される歌詞情報へのポインタとなる情報であり、歌詞情報自体は事前に表示システム側で保持する。表示システムとしては、最終的には携帯電話や PDA が望ましいが、現時点では Windows パソコン上の MATLAB で実装している。

伴奏音楽に埋め込まない歌詞情報を表示システムへ取り込む方法としては、インターネットを経由した楽曲の販売時に、付帯的なダウンロードとして実現するか、カラオケトラック入り音楽 CD 販売時に、ダウンロードのキーを同梱することなどが考えられるが、この仕組みは試作システムには含まれない。

以下に、試作システムの仕様を簡単に説明する。

4.1　埋め込むデータの構造

BCH エラー訂正符号を用いて符号化された 128 ビットの情報を、3 秒間のデータフレーム毎に埋め込む。ここでは、BCH(127,29,21) を用いて、127 ビットあたり 21 ビットまでの符号誤りを訂正できる。一方、伝送される情報量は 29 ビットであり、これを表 1 のようにビットを割り当て、データに分割する。

相対開始フレームは、レコードを埋め込んだフレームに対する、表示を開始するフレームの相対的な位置を 1~ 64 (6 bit) の整数で表現する。 つまり、フレーム時間長 3 秒のときには、表示開始は最大で 192 秒先まで可

Table 1　カラオケ歌詞表示のために埋め込むレコード

オギジャリデータ	相対開始フレーム	開始時刻	相対終了フレーム	終了時刻	表示歌詞のインデックス
2 bit	6 bit	4 bit	6 bit	4 bit	7 bit

能になる。開始時刻は、相対指定されたフレーム中の表示開始時刻を、フレーム時間長に対して 0/16 ~ 15/16 で設定する。よって、フレーム時間長 3 秒の時は、最小の表示時間分解能は 0.19 秒 (テンポ 160 での 8 分音符) となる。表示終了時刻については、開始と同じ定義である。表示歌詞は、一回に表示する文にインデックスを 1 つ割り当てることにより、128 文の歌詞を指定することができる。なお、このようなレコード定義は、カラオケの歌詞表示に限ったものであり、他の用途では、異なったビット割り当てやレコード長を用いるべきであろう。

復号化されたレコードデータはバッファメモリに格納され、AD 変換器からデータを取り込む度に、各レコードの表示開始時刻データと終了時刻データを走査して、表示開始と表示終了処理を行なう。

4.2　データ埋め込みと同期

1 つの歌詞表示情報を埋め込む最小時間間隔はデータフレーム時間長であるが、演奏には歌の無い部分もあるため、フレーム時間長を 3 秒としても、実際にはデータフレームの数は表示する歌詞の数より数倍程度多いことが一般的である。途中から伴奏を再生したときにも歌詞が表示できるようにするには、歌詞表示の直前の複数フレームにデータを埋め込み、冗長性を確保するのが良いだろう。実用的には、伴奏のどこに歌詞表示データを埋め込むかを最適に設計する埋め込みシステムの構築が望ましい。

データフレームの検出には、歌詞のない伴奏が、最低 2 フレーム分の時間必要である。3 フレーム目以降では、検出されたフレーム境界を遡って、最初のフレームからデータの復号化と表示を行う。それ以前に歌詞の表示が必要な場合には、対応できない。フレーム境界を求める演算に利用する音楽の長さは、境界の検出精度にある程度影響を及ぼす。ここでは、6 フレーム分の入力信号を基にフレーム境界検出のための積算を行っており、次節の性能評価では、楽曲冒頭からフレーム数の関数としてデータ検出性能とフレーム境界検出精度を示す。

5.　試作システムの性能評価

試作システムの性能評価として、データ埋め込みに伴う音楽の品質劣化と、埋め込み済み音楽の空間伝搬に伴う、室内の反射や残響、背景雑音の影響を、コンピュータシミュレーションによって調べた。性能評価の

Table 2　データ埋め込み条件

Parameters	Values
bit rate	43 bps
sampl. freq.	44100 Hz
freq. region	≦ 11025 Hz
subband pairs	64
subband groups	17
frame period	3 s
mod. freq. [Hz]	1.67, 2.0, 2.33, 3.0
watermarking intensity	+12 dB

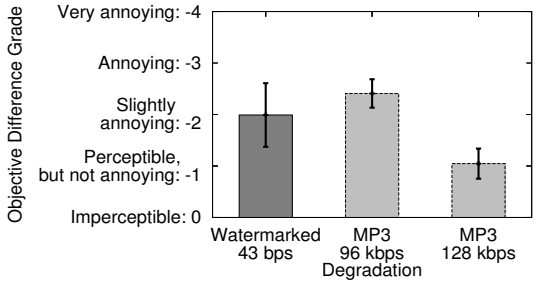

Fig. 3　PEAQ 測定によって得られた平均客観品質劣化度合と± 1 標準偏差。素材は RWC-MDB-P2001 に収録された 100 曲。

対象楽曲は RWC ポピュラー音楽データベース (RWC-MDB-P2001) [8] に含まれる 100 曲の左チャンネル冒頭 60 秒とした。データ埋め込み時のパラメータは表 2 に示した。

5.1　透かし埋め込みに伴う客観的音質劣化

音質劣化を客観的に測定する手法のひとつとして、ITU-R 勧告 BS.1387-1、いわゆる PEAQ(Perceptual Evaluation of Audio Quality) がある。これは原音 (劣化なし音) と加工音 (劣化あり音) をそれぞれ、聴覚フィルタを模したフィルタ群で帯域分割した上で、絶対閾値、周波数マスキングや時間マスキングを考慮した興奮パターン上での相違の度合を計算し、主観評価実験 (ITU-R BS.1116-1) で得られる劣化度合を予測する手法である。

ここでは、Kabel [9] による PEAQ の基本バージョンの実装を用いて、データ埋め込み済み音楽の音質劣化度合を測定した。また比較対象として、MP3 符号化後に復号化した場合も、同様に測定を行なった。図 3 には、本システムにおいて必要な埋め込みデータレートである 43 bps で埋め込みを行った場合と、MP3 の 96kbps(48kbps/ch)、128kbps(64kbps/ch) で圧縮した音楽についての、音質劣化度合の平均値と± 1 標準偏差の値をプロットした。この結果から、データ埋め込みに伴う音質劣化は、平均的には「やや気になる」程度であることが分かった。また、MP3 と比較すれば、96kbps と 128kbps の中間程度の音質であることが分かった。

ただし、PEAQ による測定値は、MP3 などの知覚符合化圧縮に伴う音質劣化を比較的よく予測できることは明らかになっているが、振幅変調に伴う音質劣化も同等に予測できるかどうかについては、いまだ明らかになっていない。この結果はあくまで参考であり、今後主観的劣化と客観的劣化指標との対応について検証を進めるべきであろう。

5.2　使用環境シミュレーション

残響のある室内でシステムを使用した場合に、マイクによって収音した音に同期して情報が表示できるかを定量的にシミュレーションする実験を行なった。

残響のある室内を想定して、データ埋め込み済み音楽信号に、RWC 実環境音声・音響データベースより選んだ、残響時間 1.3 秒の可変残響室で収録されたインパルス応答 (ファイル名: ir130.dat) を畳み込んだ。これにより、スピーカやマイクの特性による、フラットでない伝送特性も模擬できる。

その後、背景雑音として 4 種類の環境騒音 (収録場所: 駅のホーム、地下連絡通路、空港ロビー、混雑した交差点)、あるいはローパスノイズ (カットオフ 500 Hz、–9 dB/oct. : 他の環境騒音の平均的スペクトルに近い) のいずれかを付加した後、透かし情報を検出する処理を行った。5 種類の背景雑音は、オーバーオールの音楽信号パワーに対して、信号対雑音比 (SNR) は 15dB とした。現実の使用場面で重畳されるノイズとやや種類は異なり、ノイズレベルも幾分高いと思われるが、システムに困難な環境を模擬するために採用した。

また、データ検出用のマイクロホンをスピーカに近づけた場合は、入力過大により振幅がクリッピングする事態も考えられる。このような状況での耐性を調べるため、透かし入り音楽信号の最大振幅の 0.125 倍以上の振幅を制限する変形 (+18 dB の入力過大) も模擬した。この振幅制限のシミュレーションでは、残響は付加せず、環境騒音の SNR は 30 dB として、振幅制限の直前に加えた。

埋め込みの強度は、いずれも埋め込みなしの信号に対して検出処理を行って得られる振幅変動の強度に対して、+12dB とした。100 種の透かし入り音楽と 5 種の背景雑音を組み合わせて 500 条件において、残響付加あるいは振幅制限がシミュレーションされた。本システムでは、6 フレーム分の入力信号を基にフレーム境界検出のための積算を行っているため、フレーム境界の検出精度とそれに依存するデータ検出率は楽曲の先頭からの時間に多少依存する。よって、結果は横軸を先頭からのフレーム番号として示した。

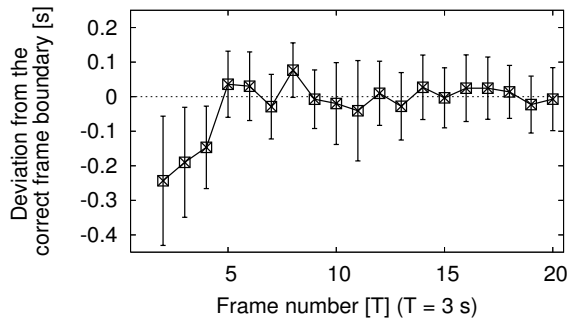

Fig. 4　残響と背景雑音が加わった場合での、フレーム境界時刻からの、検出時刻のずれ. 誤差棒は ± 1 標準偏差を示す。

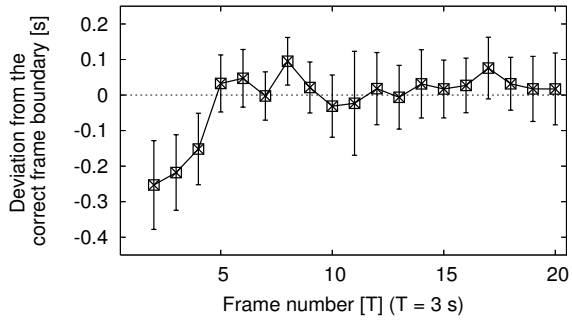

Fig. 5　背景雑音と振幅制限が加わった場合での、フレーム境界時刻からの、検出時刻のずれ. 誤差棒は ± 1 標準偏差を示す。

5.2.1　フレーム検出時刻の精度

フレーム境界時刻をどの程度正確に検出できているかを調べた。フレーム境界時刻の検出精度を、正しいフレーム境界からのずれ時間とした。図 4 に残響と背景雑音が加わった場合の結果を示した。図 5 には、背景雑音と振幅制限が加わった場合の結果を示した。

第 4 フレーム目 (冒頭から 12 秒) までは、検出されたフレーム境界は 0.1~ 0.3 秒程度を早めとなるが、それ以降は 0.1 秒前後の標準偏差でほぼ正確にフレーム境界が検出できることが分かった。また、表示における時間分解能は約 0.19 秒であるので、実用上ほぼ問題無く指定時刻に表示できるであろうことが分かった。

5.2.2　データ検出率

データ検出率の指標としては、3 秒間のデータフレームに埋め込まれた 128 ビットのうち、誤り訂正限界である 21 ビット以内にエラービット数が収まったフレーム数を、全体のフレーム数で割った正検出割合とした。

図 6 に残響と背景雑音付加の結果を、図 7 に背景雑音と振幅制限を与えた結果を示した。楽曲の冒頭では、音楽信号レベルが相対的に小さいため、検出率が低くなっているが、第 5 フレーム (15 秒) 以降では、90%以上の検出率となっている。

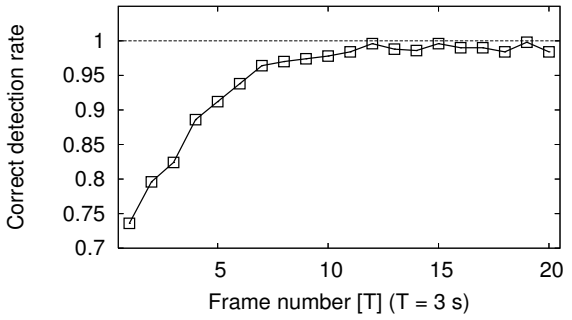

Fig. 6　残響と背景雑音が加わったときの、エラー訂正限界以内に収まったデータフレームの割合。それぞれの点は 500 条件のシミュレーションから得られた。

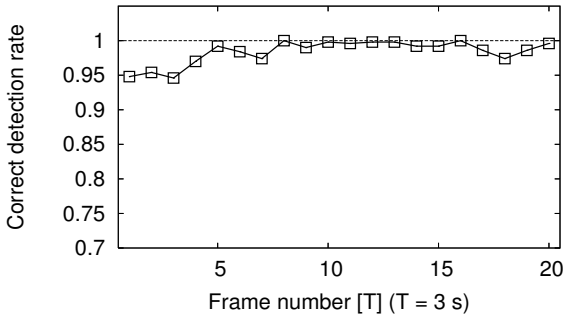

Fig. 7　背景雑音と振幅クリッピングが加わったときの、エラー訂正限界以内に収まったデータフレームの割合。それぞれの点は 500 条件のシミュレーションから得られた。

6. 考　　察

本システムの性能評価は、埋め込みデータ検出用マイクに、伴奏音楽と同時に歌声が収音される場合については、行なっていない。しかし、本システムが採用した音響電子透かし手法では、隣接狭帯域に逆位相の振幅変調を与え、検出時にそられの帯域の振幅エンベロープの除算を行なうため、両帯域に共通する振幅変動成分 (ここでは歌声) が加算されても、この除算にもいてキャンセルされる、という特徴がある。実際にシステムを動作させた場合にも、ランダムな雑音の付加より、コヒーレントな雑音の付加に対して頑強であるという、透かし検出処理の特徴をうかがうことができる。今後は、検出性能への歌声の影響を、定量的に調べることを考えている。

7. ま と め

振幅変調に基づく音響透かし技術を用いて、カラオケの伴奏に同期して歌詞を表示するシステムを、パソコン上で動作するソフトウェアとして試作した。システムの性能評価として、PEAQ を用いた音質の客観評価を行った結果、大きな音質劣化は無く、MP3 符号化に伴

う音質劣化と比較すると 96kbps と 128kbps で圧縮された場合の中間程度であることが分かった。また、データ埋め込み済み音楽信号への残響および背景雑音付加、あるいは入力過大による振幅制限のシミュレーションの結果、音楽信号に埋め込まれたデータは十分検出可能であることが分かった。システムはデータフレーム境界時刻の検出結果を基に、表示の時間制御を行うが、その検出精度も十分であることが分かった。よって、試作システムは、スピーカ再生される音楽に同期してリアルタイムに歌詞の表示の制御ができることが示された。

文　　献

[1] 茂出木敏雄, "非接触抽出可能な音楽への電子透かし埋め込み技術の開発," 電子情報通信学会技術研究報告, 19–24 (2005).

[2] 松岡保静, "サブバンド位相シフトを用いた音響電子透かし埋め込み法," 電子情報通信学会技術研究報告, 529–533 (2006).

[3] Akira Nishimura, "Data hiding for speech sounds using subband amplitude modulation robust against reverberations and background noise," in *Proceedings of IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 7–10 IEEE, (2006).

[4] Ryuki Tachibana, "Improving audio watermarking robustness using stretched patterns against geometric distortion," in *Advances in Multimedia Information Processing, PCM2002, (LNCS) 2532*, 647–654 Springer, Hsinchu, Taiwan, (2002).

[5] Byeong-Seob KO, Ryouichi Nishimura, and Yoiti Suzuki, "Robust Watermarking Based on Time-spread Echo Method with Subband Decomposition," IEICE Trans. Fundamentals, **E87-A**, 1647–1650 (2004).

[6] 西村明, "帯域分割と振幅変調に基づく音響電子透かし," 暗号と情報セキュリティシンポジウム 2006, No. 3F4-2 電子情報通信学会, (2006).

[7] Akira Nishimura, "Audio watermarking based on sinusoidal amplitude modulation," in *Proceedings of ICASSP 2006, IV*, 797—800 IEEE, (2006).

[8] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka, "RWC Music Database: Popular, Classical, and Jazz Music Databases," in *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, 287—288, (2002).

[9] P. Kabal, "An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality," TSP Lab Technical Report, Dept. Electrical & Computer Engineering (2002).