Audio data hiding based on amplitude modulation and its application

西村, 明
Faculty of Informatics, Tokyo University of Information Sciences

https://doi.org/10.15017/18879

第6章　情報秘匿に基づく音響信号と同期した情報呈示

6.1　まえがき

第4章では，スピーカから再生された情報秘匿済み音響信号を，ユーザの手元の機器で受信して検出復号化し，秘匿情報を利用する形態に対する，従来の音響情報秘匿技術やその具体的な利用提案例を概観した．この場合，音響情報秘匿技術に対しては，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，電子透かし用途より埋め込む情報量を高めることが要求される．そして，振幅変調に基づく情報秘匿手法がそれらの要求を満たしていることを示した．

本章では，こうした情報秘匿済み音響信号の空間伝搬と受信を前提とした利用形態として，情報秘匿済み音響信号に同期してユーザへ情報呈示するシステムを提案する．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，伴奏音楽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．そして，本システムの使用条件を想定した残響音，反射音，背景雑音，マイクロホンへの入力過大によるクリッピング歪，および歌唱声によるシステム性能の低下を，定量的に評価する．

従来の再生音と同期した情報呈示システム，例えば，伴奏音に歌詞を同期して表示する，業務用カラオケボックス，マイク一体型家庭用カラオケシステムなどにおいては，歌詞と伴奏音が同一あるいはケーブルで接続された機器によって表示および再生される形態をとっている．本システムは，情報秘匿済み音響信号の空間伝搬によって字幕呈示情報を伝えるため，再生機器と字幕表示機器とを空間的に独立させることが可能で，表示機器の複数化および個人化，同時多数利用，屋外利用も可能となる．これは，従来のカラオケや字幕システムを置き換えるものではなく，従来にない利用方法をもつシステムの提案とも言える．

表 6.1: 埋め込むレコードのデータ定義.

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	情報のインデックス

6.2 音響信号と同期した情報呈示

本システムにおいて，音響信号に埋め込まれるデータは，主に呈示開始時刻と呈示終了時刻，そして呈示される情報のインデックスとなる情報であり，呈示される情報自体は事前に呈示システム内あるいはネットワーク上でアクセス可能な状態であることを前提としている．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 OS 上で実装している．なお，本章で用いる音響信号は全て 44.1 kHz サンプリング，16bit 直線量子化されたものである．

6.2.1 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 127 ビットあるいは 63 ビットの情報を，3 秒間のデータフレーム毎に埋め込む．BCH(127,29,21) を用いた場合，127 ビットあたり 21 ビットまでの符号誤りを訂正でき，伝送される情報量は 29 ビットとなる．この伝送情報量で 1 つの埋め込みレコードを構成し，表 6.1 のようにデータ定義する．レコードに含まれる各データ項目のビット長は，本システムの利用用途に応じて決定される．なお，1 データフレームあたりの埋め込み情報量は，BCH 符号化情報量より 1 ビット多いので，そのビットはパリティビット等に利用して信頼性を向上させることができるが，第 6.4 節以降のシミュレーションではこれを用いていない．オギジャリデータは機能拡張，例えば複数ビットを割り当てて字幕表示速度や表示位置を制御することを想定している．このような用途が無い場合には，使用しなくともよい．

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整数で表現する．つまりこのデータに 6 bit を割り当てれば，フレーム時間長 3 秒のとき，表示開始は最大で $3 \times 2^6 = 192$ 秒先まで可能になる．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示し，フレーム時間長を，2 のデータビット数の冪乗で割った分解能で指定する．つまり，フレーム時間長 3 秒の時に，開始時刻に 4 bit を割り当てれば最小の表示時間分解能は $3 \div 2^4 = 0.19$ 秒となる．相対終了フレームは，相対開始フレームに対して何フレーム先に表示を終了するかを指定する．終了時刻は

開始時刻と同じ定義である．情報のインデックスは，呈示するひとつの情報につき整数値を 1 つ割り当てる．

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 変換器から一定時間 (ここでは 4096 サンプル) の波形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．

なお，表示する情報の数に対して埋め込みフレーム数が多い場合は，表示開始フレーム以前のフレームに表示に必要なレコードデータを繰り返し埋め込むことによって冗長度を高め，検出率を高めることも可能である．しかし，この方策は以降のシミュレーションでは用いていない．

6.2.2 検出処理演算量

検出時のフィルタバンク処理および埋め込み区間同期検出処理の演算には FFT を利用しており，それらの演算負荷が検出処理全体の中で最も高い．第 6.4 節以降での実施例では，フィルタバンク処理は 4096 点 FFT を 3/4 づつオーバラップさせて実施し，同期検出処理はオーディオデータが 4096 サンプル入力されるごとに変調周波数毎の強度計算のため 128 点 FFT を 129 回実施している．

携帯電話や PDA で検出復号化を行うには，演算量をさらに低減する必要がある．検出性能を大きく低下させずに演算量を低減するには，実際のマイクやスピーカの特性が高域まで伸びていないことからサンプリング周波数を半分の 22.05 kHz とし，フレーム境界検出精度がやや低下するものの，フィルタバンクにおける FFT のオーバラップを 1/2 とする方法がある．これにより単位時間あたり，フィルタバンク処理，同期検出処理ともに 半分のサイズの FFT を用いて半分の実行回数で実現できる．しかし，この演算量でも Windows Mobile OS 搭載の携帯電話あるいは PDA の処理能力の 2 倍程度となるため，フィルタバンク処理の演算量低減および，埋め込み時の変調周期をオーバラップ FFT 周期の整数倍とすることで強度計算を同期加算のみで実施し FFT 処理を無くすなどの改良が今後必要であろう．

6.3 データ埋め込みに伴う音質劣化の評価

情報秘匿済み信号をスピーカ再生とマイクロホン受音環境において利用する場合，より頑強なデータ検出のため，埋め込み強度を強めることになり，一般的な電子透かし用途の

表 6.2: データ埋め込み条件

Parameters	Normal	Robust
sampl. freq.	44100 Hz	←
embedding freq. region	≤ 11025 Hz	≤ 22050 Hz
bandwidth	86 Hz	←
subband pairs	64	127
subband groups	17	9
pilot subbands	12 pairs	36 pairs
pairs per group	3 — 4	11 — 12
frame period	3 s	←
mod. freq. [Hz]	1.67, 2.0, 2.33, 3.0	←
intensity	+12 dB	+15 dB
payload [bits/frame]	128	64
error correction	BCH(127,29,21)	BCH(63,18,10)
bit rate [bps]	9.7	6

情報秘匿に伴う音質劣化に比べて知覚可能な程度となる．

ここでは，Kabal[14] による PEAQ の基本バージョンの実装を用いて，情報秘匿済み音楽の音質劣化度合を測定した．対象音楽は，RWC ポピュラー音楽データベース (RWC-MDB-P-2001)[72] に含まれる 100 曲の左チャンネル冒頭 60 秒とした．また音質劣化の比較対象として，情報埋め込みを行わずに MP3 符号化後に復号化した場合も，同様に測定を行なった．

表 6.2 には 2 つの埋め込み条件が示されている．Normal は，一般的な雑音や残響のある環境で利用する場合であり，Robust は，第 6.5 節で示す，カラオケでの歌詞表示利用を前提として残響や歌声が重畳する環境で利用する場合の埋め込みパラメータ値である．表 6.2 中の mod. freq. は変調周波数であり，全ての帯域に対して 4 つの変調周波数で同時かつ独立に変調を行い，埋め込み処理を多重化していることを意味する．

図 6.1 には，2 つの埋め込み条件の結果と，MP3 で圧縮した音楽についての，音質劣化度合を表す ODG の平均値と ± 1 標準偏差の値を示した．この結果から，情報秘匿に伴う音質劣化は，平均的には「やや気になる (slightly annoying)」以下程度であることが分

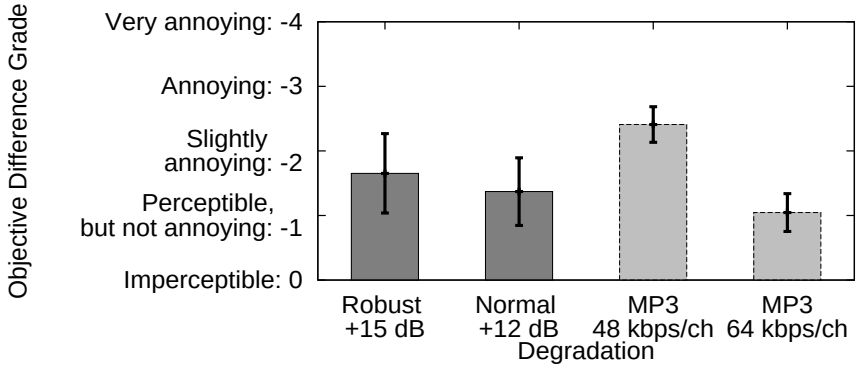

図 6.1: PEAQ **によって得られた客観品質劣化度合** (ODG) **の平均と** ± 1 **標準偏差．対象は** RWC-MDB-P-2001 **に収録された** 100 **曲．**

かった．また，MP3 **と比較すれば，**48 kbps/ch **と** 64 kbps/ch **の中間程度の音質であることが分かった．**

6.4 空間伝搬耐性

本システムが，残響あるいは入力過大，かつ雑音に対する耐性を持つことをシミュレーションによって確認する．埋め込み条件は，表 6.2 **における** Normal **条件を用いた．**

6.4.1 シミュレーション条件

対象楽曲は RWC-MDB-P-2001 **に含まれる** 100 **曲の左チャンネル冒頭** 60 **秒とした．これに，第** 4.2 **節で用いたものと同じ，**RWCP **実環境音声・音響データベース** [63] **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．このインパルス応答波形は既に図** 4.2 **に示した．この波形の絶対値ピークを中心とした** 128 **サンプルにハニング窓掛けを行って直接音とみなした場合，直接音に対するそれ以降の間接音のパワーの比率は** –0.4 dB **となる．この直接音成分のパワースペクトルは既に図** 4.3 **に示した．この畳み込みによって，再生機器の位相特性や反射音による位相特性の乱れも模擬できる．図** 6.2 **には，インパルス応答波形について** 110 **サンプル毎に求めた実効レベルの時間減衰を示した．**

その後，第 4.2 **節で用いたものと同じ，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**

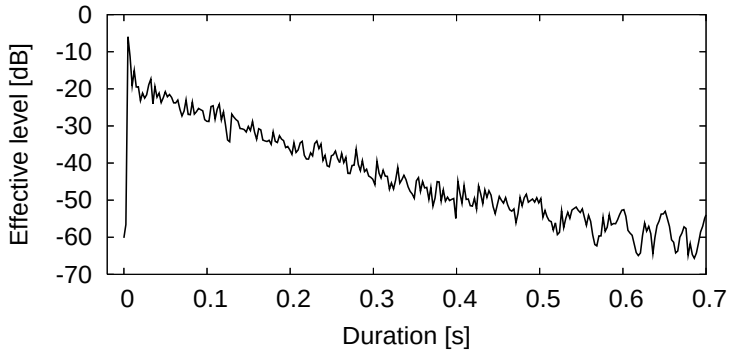

図 6.2: **インパルス応答におけるエネルギ減衰**.

した後，埋め込み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) **を** 15 dB **とした．雑音強度の時間変動は，**100 ms **の矩形窓毎に計算した実効値の標準偏差が，それぞれ** 4.6 dB, 1.9 dB, 3.3 dB, 4.0 dB, 0.49 dB **であった．これらの背景雑音は，本システムを屋内で利用する場合の雑音とはやや種類が異なるうえレベルも高いが，より困難な環境を模擬するために用いた．**

また，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，データ埋め込み後の音楽信号について，その最大振幅の 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．この場合残響は付加せず，環境騒音の** SNR **は** 30 dB **として，振幅制限の直前に加えた．**

情報秘匿済みの 100 **種の音楽信号と** 5 **種の背景雑音を組み合わせた** 500 **条件において，上記のシミュレーションを行った．**

6.4.2 フレームデータ検出率

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ランダムに** 1 **あるいは** 0 **の値を持つ** 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 6.3 **に残響と背景雑音付加の結果を，図** 6.4 **に背景雑音と振幅制限を与えた結果を示した．データ検出率** (Correct Detection Rate) **は，音楽の冒頭から数えたフレーム番号** (Frame Number) **に依存するため，横軸をフレーム番号として示している．残響が加わった場合は，楽曲の冒頭はいわゆるイントロであり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**

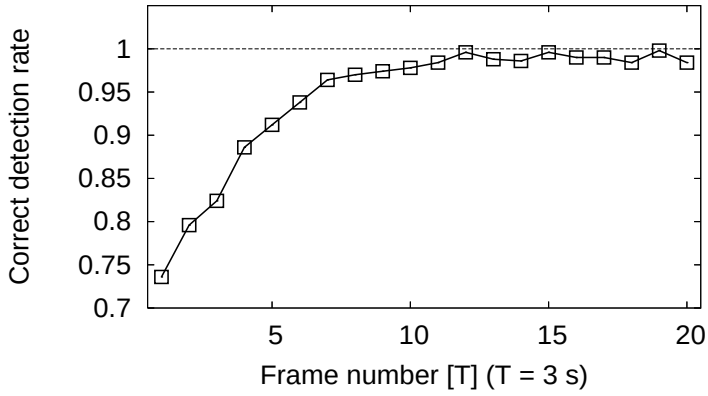

図 6.3: 残響と背景雑音が加わったときの，エラー訂正限界以内に収まったデータフレームの割合．それぞれの点は 500 条件のシミュレーションから得られた．

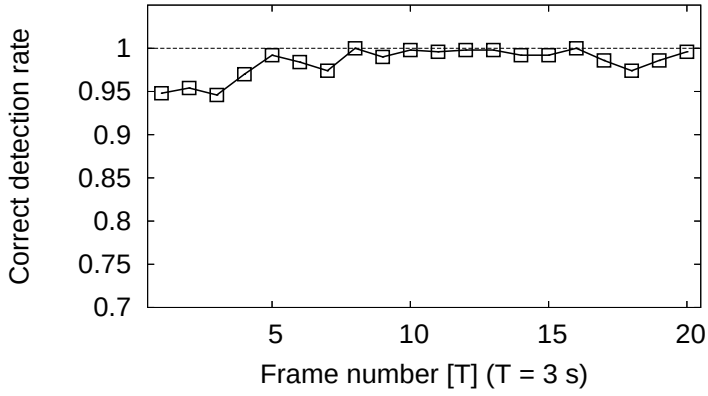

図 6.4: 背景雑音と振幅クリッピングが加わったときの，エラー訂正限界以内に収まったデータフレームの割合．それぞれの点は 500 条件のシミュレーションから得られた．

いる．振幅制限の場合には，より良好に検出ができることが分かった．

6.5 カラオケ歌詞表示システム

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呈示するシステムの応用として，カラオケ伴奏音楽に同期して歌詞を呈示するシステムについて説明する．伴奏音楽には歌詞情報そのものではなく，そのインデックスを埋め込む．歌詞情報本体を呈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，この仕組みは現時点でのシステムには含まれない．

表 6.3: カラオケ歌詞表示のために埋め込むレコードのビット割り当て.

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
3 bit	4 bit	2 bit	4 bit	5 bit

このシステムの性能評価シミュレーションは，Octave-3.0 上で動作するソフトウェアで実施した．実時間で動作するシステムは，当初は MATLAB R2008 にて Data Acquisition ToolBox, Signal Processing ToolBox, Communication ToolBox を用いて作成した．その後，Windows バイナリとして動作するシステムを委託製作した．Windows 版の動作中の画面の例は，図 6.5 にて示した．図中のスピーカおよびマイクの距離は，実際には，より遠くに配置しても動作することは確認している．

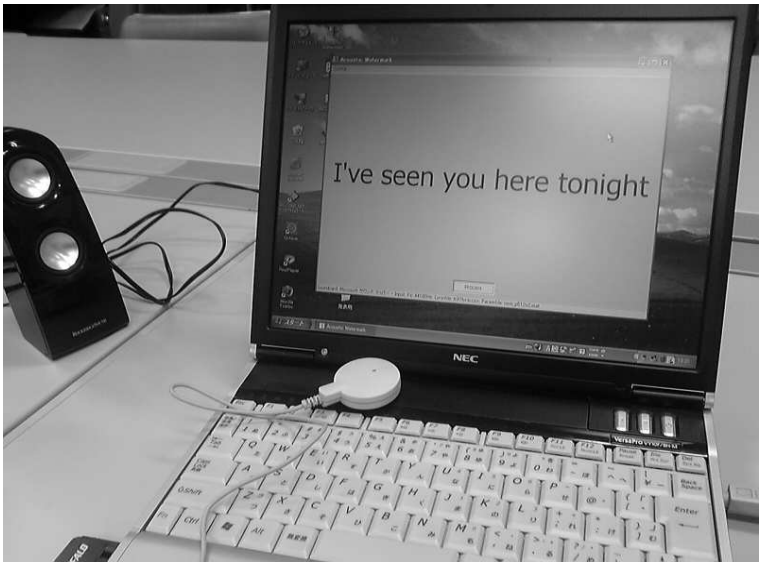

図 6.5: Windows 版同期表示ソフトウェアの実行中画面の例．

6.5.1 埋め込むデータ構造

表 6.2 中の Robust 条件を採用した．この条件では，63 ビットあたり 10 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 18 ビットであり，表 6.3 のようにビットを割り当てた．

表 6.3 のビット割り当てから，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は，最大で 24 秒先まで可能となる．同様に開始時刻の表示時間分解能は 0.19 秒

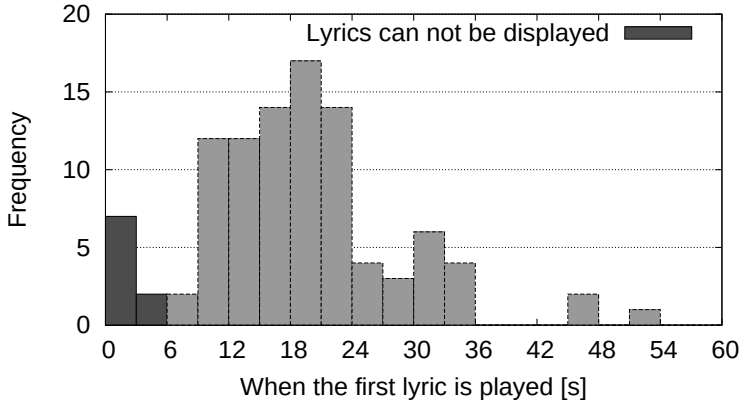

図 6.6: RWC-MDB-P-2001 **収録の** 100 **曲の歌詞の開始時刻と度数**.

(テンポ 160 **での** 8 **分音符) となる．表示時間は最大で** $3 \times 2^2 = 12$ **秒間である．表示歌詞数は** $2^5 = 32$ **であるが，リフレインなどの同じ歌詞には同じインデックスを割り当てることとする．**

6.5.2 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，楽曲中のフレーム数は表示する歌詞の数より数倍程度多いのが一般的である．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，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．**

データフレームの同期検出には， 2 **フレーム (**6 **秒) 分の伴奏時間が最低必要である．** 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．よって** 3 **フレーム目より以前に歌詞の表示は不可能であるが，このような楽曲がどの程度の割合で存在するかは明らかではない．図** 6.6 **に，**RWC-MDB-P-2001 **収録の，歌のあるポピュラー音楽** 100 **曲について，曲の開始から何秒後に歌詞が唄われるかの頻度 (**Frequency**) を示した．この結果から，** 1 **割程度は冒頭の歌詞が表示できない場合があると思われる．**

6.5.3 性能評価

歌詞表示システムの性能評価として，室内の反射や残響，および利用者の歌唱音声による影響を，コンピュータシミュレーションによって調べた．性能評価の対象楽曲は RWC-MDB-P-2001 に含まれる 100 曲の MIDI ファイルを，MIDI 音源 (timidity++) によって演奏した左チャンネル冒頭 60 秒とした．

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，RWCP 実環境音声・音響データベースに収録されているインパルス応答 (ファイル名: ir130.dat) を畳み込んだ．次に歌声合成ソフトウェアである VOCALOID2「鏡音リン・レン」を用いて作成した歌唱音 (男声はレン，女声はリンが歌唱) を，平均 SNR –5, 0, +5 dB の 3 条件で付加した．この平均 SNR 値は，A 特性フィルタリング後の残響つき伴奏音の実効値 (signal) と，A 特性フィルタリング後の歌唱音の実効値 (noise) を計算し，100 ms の矩形窓フレーム毎に歌唱音が含まれるフレームの SNR(dB) 値の算術平均値として定めた．

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ランダムに 1 あるいは 0 の値を持つ 64 ビットのうち，誤り訂正限界である 10 ビット以内にエラービット数が収まったフレーム数の率を調べた．なお，歌唱音が含まれないデータフレームに対しては，全てのフレームでエラー訂正限界以内に収まった．

歌唱音がわずかでも含まれるフレームは，全 2000 フレーム中 1369 であった．歌唱音が含まれるフレーム数を母数としたデータフレーム検出率 (Correct detection ratio for singing frames) を，各曲毎に 0.05 刻みで調べた．図 6.7 の横軸は検出率であり，縦軸はその検出率以下となる累積曲数 (Cummulative frequency of tunes) を示した．図 6.8 には，1 曲の全フレーム数 20 を母数とした，データフレーム検出率 (Correct detection ratio for all frames) を各曲毎に算出し，その検出率以下となる累積曲数を示した．

図 6.7 では SNR –5dB の最も厳しい条件で歌唱音が重畳した場合には，100 曲中 12 曲で歌唱音の含まれるフレームの検出率が 0.7 以下になることを示している．一方，歌詞表示情報の埋め込みは歌詞のあるなしに関わらず行われるため，曲全体の歌唱音のないフレームも含んだ場合の検出率を見ると，図 6.8 では，SNR –5 dB のとき，0.7 以下の検出率となるのは 6 曲である．実際の楽曲中のフレーム数は表示する歌詞の数より数倍程度多いのが一般的であり，同じ歌詞表示情報を複数フレームに繰り返し埋め込むことができるため，実際のデータ検出率はさらに高まるはずである．なお，歌唱音重畳時の検出率が低い楽曲は，伴奏がピアノのみ，あるいはギターのみの曲であった．

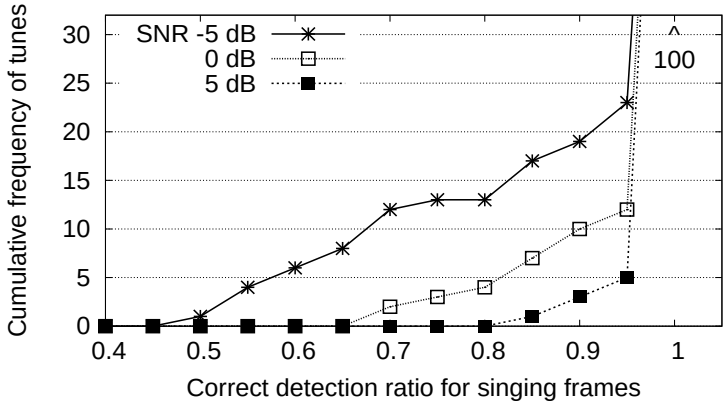

図 6.7: **歌唱音が含まれるフレームに対するデータ検出率と累積曲数**.

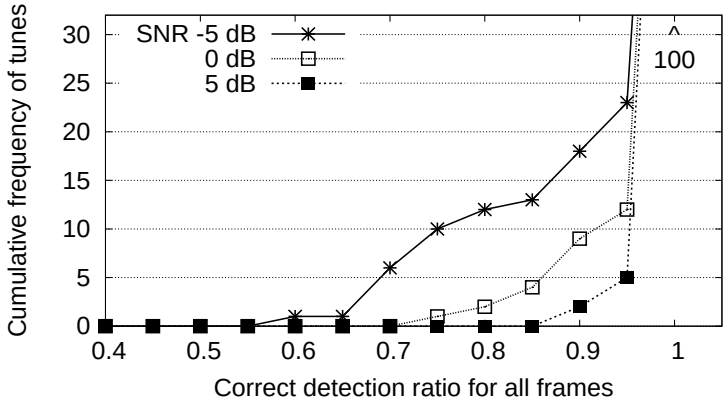

図 6.8: **曲の全フレームに対するデータ検出率と累積曲数**.

データ検出時に得られるフレーム境界時刻は，伴奏音と歌詞表示の同期の基準時刻となる．この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平均ずれ時間 (Mean deviation from the correct frame boundary) **として，最も厳しい条件である** SNR –5 dB **での結果を図** 6.9 **に示した．**

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平均で** 0.1 **秒程度早めとなるが，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の検出が可能であった．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．**

6.6 考察

スペクトル拡散法などの一部を除いて，音響電子透かし技術の多くは，本技術もそうであるが，音の無い部分にデータを埋め込むことはできない．本システムの用途としては，

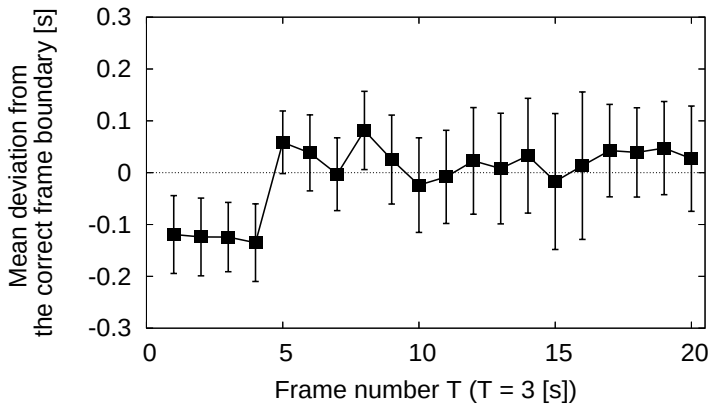

図 6.9: 残響と歌声が加わった場合での，フレーム境界時刻からの，検出時刻のずれ. 誤差棒は ± 1 標準偏差を示す．

カラオケ歌詞表示以外にも，映画のサウンドトラックに字幕表示情報を埋め込んで，セリフに同期して字幕を表示するシステムが考えられる．これは，前席の背もたれ背面にパネル画面 PC を掛け，字幕と画面への視線と同方向となる利用形態である．この場合，映画サウンドトラックには，無音部分が多く含まれるという問題がある．これには，データ埋め込みフレーム時間長を長くとる対応が考えられる．本技術では，データフレーム時間の一部にしか信号が含まれていなくても，その部分を有効に活用してフレーム同期とデータ検出を行うことができる．このことは，第 4 章において示したように，ポーズを含んだ音声信号でも，本技術は有効にデータ検出が可能であることからも分かる．

本技術をカラオケ歌詞表示システムとして利用する場合，楽曲の冒頭部分の歌詞は表示できず，検出表示端末の設置位置によっては第 6.5.3 節で想定した以上の高雑音および残響の多い環境となることも考えられる．よって，性能的に従来技術を置き換えることは困難である．一方，ここではシステムへの妨害となる環境の評価を定量的に行ったため，使用環境に応じた表示不能率を算出することが可能である．例えば，検出率 x の楽曲および環境において，n 個の異なる情報を m 回繰り返して埋め込む場合，最大 1 つまでの表示情報欠落を許容するときの表示可能率は $(1-(1-x)^m)^n + n(1-(1-x)^m)^{n-1}(1-x)^{(n+m)}$ で与えられる．再生系と情報表示系の空間的分離や，表示端末の個人化，同時多数端末による表示など，従来のシステムにない機能に価値をおき，使用条件の制限となるシステム動作環境と上述の表示可能率とのバランスを考えながら実使用に望むべきであろう．

6.7 あとがき

振幅変調に基づく音響情報秘匿技術を用いて，データを埋め込んだスピーカ再生音をマイク受音して，その信号に同期して情報を呈示するシステムを作成した．システムの性能評価として，情報秘匿済み音楽信号に対して PEAQ を用いた音質劣化の客観評価を行った結果，音質劣化は MP3 48 kbps/ch と 64 kbps/ch で符号化された音楽信号の中間程度であることが分かった．情報秘匿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．

さらに，カラオケ歌詞表示システムとして利用時の，埋め込みデータ検出への残響や歌唱音の影響も調べた．その結果，SNR 0 dB で歌唱音が付加されても，90%以上の楽曲の，90%以上の歌唱音を含むフレームでデータ検出が可能なことが分かった．また，歌詞表示の時間制御の基となるデータフレーム境界時刻の検出精度も十分であることが分かった．