音声認識を用いた歌声追尾歌詞表示システム
Automatic sync-lyric display system using speech recognition

森川 彰
morikawa akira
法政大学情報科学部ディジタルメディア学科
E-mail: akira.morikawa.ud@cis.hosei.ac.jp

Abstract

In the study, use speech recognition engine Julian[1] and speech recognition does the lyrics of the singing voice of the whole music. And of the lyrics take aligning it. Make a system displaying lyrics in accommodate the timing of the singing voice. By this system, may not need information of the timing of the lyrics into lyrics data. Even one's own songs use, even minor music and even indie music without lyrics data can display lyrics by only information of lyrics and how to read lyrics. Perform speech enhancement in music to get alignments of music. Do an experiment to control the performance of the music. As a result, the song was not able to be emphasized for music encompasses the musical performance and the chorus, etc. in a spectrum subtraction. Because a spectral subtraction can be remove only a regular noise. The sounds other than the vocal component were removed in the modulation spectrum. It succeeded in the suppression of the performance etc. Therefore, the alignment of music which the alignment was not able to be taken before the voice was emphasized was able to be taken.

1 まえがき

音楽の歌詞を覚えたいとき、ＣＤに付属している歌詞カードを見ながら音楽を聴き、歌詞を覚えるのは、屋外では歌詞カードを持ち歩くのが面倒である．また、室内でも歌詞カードでは、今どこの歌詞が歌われているのか分からなくなってしまうといったことがある．そこで、音楽の歌詞を楽曲の歌声に合わせて表示させるようなシステム構築を目標とする．

最近、株式会社シンクパワーが運営する歌詞表示機能「歌詞ピタ」サービス [2] により、歌詞データを購入して音楽プレーヤーに入れておけば、楽曲を再生すると歌詞が自動スクロール表示される機能が出た．実際に使用してみたところ、表示される歌詞は楽曲の歌詞が始まる瞬間にスクロールが行われる仕組みであった．そこで試しに同じ楽曲だが、イントロ部分が長くアレンジされた楽曲にイントロ部分が短い通常バージョンの楽曲の歌詞データを歌詞ピタでのせたところ、歌詞が歌声とずれたタイミングでスクロール表示された．その結果から歌詞データには歌詞情報とスクロールのタイミングの情報が入っていることが推測でき、歌詞ピタではアライメントの検出はされていないことが分かった．

本研究のシステムは、歌詞ピタとは違い、楽曲から音声認識エンジン Julian を用いてアライメントをとり、楽曲中の歌声のタイミングに合わせて歌詞を表示するシステム構成である．また、歌詞データも「歌詞ピタ」サービスでは有料だが、本研究のシステムでは歌詞とその読み方があればいいので、ユーザ自身が打ち込んでしまえば、お金をかけることなく歌詞表示を行うことができる．

Supervisor: Prof. Katunobu Itou

歌詞表示システムの先行研究として、文献 [3] があるが、本研究では歌詞アライメントにより歌詞表示を行うので、音響透かしを用いる際に起る音質の劣化をさせることなく歌詞表示を行うことが可能である．

2 音声認識を使った歌詞アライメント

2.1 Viterbi アライメント

本研究で行う歌詞表示システムで楽曲から絶対的に取得しなければいけない情報は、楽曲のどのタイミングでどの歌詞が歌われているのかという情報である．タイミングというのは歌詞の一単語が歌われる始めるタイミングと歌い終わるタイミングの二つである．そのタイミングを知るために音声認識を用いて、楽曲の歌声から歌詞のアライメントをとる．アライメントをとる手法としては音声認識処理で得られる Viterbi アライメントを用いる．尤度最大の HMM 状態のみを各時刻で選択し、モデルの最尤遷移経路を算出して近似的に確率を求める手法を Viterbi アルゴリズムと呼ぶ．Viterbi アライメントとは、その最尤遷移経路の状態割り当てのことをいう．

図 1 のように Viterbi アライメントを用いることにより、各音素の始端と終端を推定し、ユーザ歌唱の歌詞（音素）の時間的対応付けをおこなうことができる．

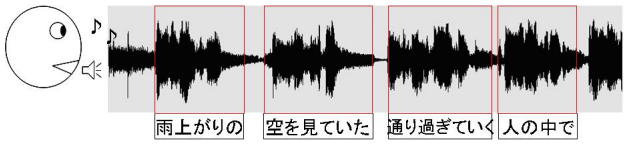

図 1. Viterbi アライメントの例

2.2 アライメントに必要な音声認識

2.2.1 認識用文法

Julian では,「認識用文法」（あるいはタスク文法）を構文制約を単語のカテゴリを終端規則として記述する grammar ファイルとカテゴリごとに単語の表記と読み（音素列）を登録する voca ファイルの 2 つのファイルに分けて記述する．

grammer ファイルは本研究では、アライメントをとることができればいいので、楽曲の歌詞を歌詞カードと同じように記述した．voca ファイルには、楽曲を聴いてブレスがはいるところで切った歌詞を記述した．文献 [4] を参考に、歌声を認識するので、長音は 「a:」 のように「母音+:」、格助詞の「へ」「は」はそれぞれ「e」「w a」というように 43 個の音素で記述した．表に登録した単語の一例を示す．

表 1. 登録単語の一例

単語	音素
コインのように	k o i N n o y o: n i
憎めど夏は今	n i k u m e d o n a t s u w a i m a
きっと	k i q t o

2.2.2 入力の無音区間・休止の扱い

音声入力においては、発声区間の前後に無音区間が含まれる．また、発話中の息継ぎなどにより、文中や単語間に短い休止が含まれることが多くある．歌詞のアライメントをできる限り正確にとるためには、楽曲中のブレスの部分も検出しなければならない．そこで、Julius の iwsp というコマンドを使用した．このコマンドを使うことによって、単語間ショートポーズへの対処の機能が ON になり、「任意の単語間にショートポーズが入りうる」という前提で特別な認識処理を行うことができる．

2.2.3 OPTION コマンド

本研究では、アライメントをとるのに設定ファイルに含まれているアライメントをとるコマンド walign を使用した．このコマンドを用いると認識結果に対して，単語単位の Viterbi アラインメントを行い、単語ごとにマッチした区間, およびフレームごとの平均音響尤度が出力される．以下に出力例を示す．

```
-- word alignment --
 id: from  to    n_score    unit
 ----------------------------------------
[   0  453]  -24.390158  2 [<s>]
[ 454  637]  -25.527058  0 [フラフープの輪が]
[ 638  706]  -26.231445  1 [棒に変わる、]
[ 707 1047]  -23.871397  3 [</s>]
 ----------------------------------------
```

この認識結果例では、一番上の単語列では、0 秒から 4.53 秒に相当する単語は”silB(文頭の無音)”でスコア (対数尤度) の平均は”-24.390158”ということを表している．このように始端と終端のフレームが出力され、歌詞表示システムではこの結果を用いてタイミングを合わせて歌詞表示を行っていく．また、状態単位の Viterbi アラインメントをとるコマンド salign を使用して、単語間ショートポーズ sp の入る部分を検出し、その部分を無音区間として扱うようにする．

2.3 歌詞表示システム

歌詞表示システムの構成のイメージとしては、図 2 のような流れになる．

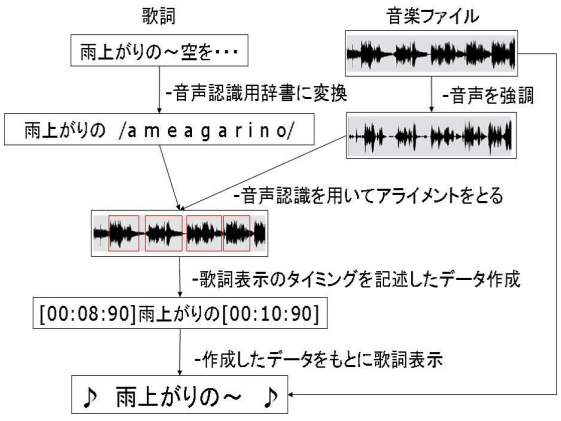

図 2. システム構成

3 背景音の影響

楽曲には、もちろん楽器演奏などの背景音が含まれている．これにより楽曲の音声認識を行うことが困難になる．

3.1 実験

背景音がどれだけ歌詞アライメントをとるのに影響してくるのか調べた．実験方法としては、ヴォーカルの歌声がないカラオケの楽曲に対して、自分で歌唱し、音声と背景音から SN 比を出し、アライメントをとった．これにより SN 比が違う楽曲からどの程度アライメントがとれるのか調べることができる．実験は Julian を使用して楽曲のアライメントをとり、成功した最大の秒数で比べていく．SN 比の出し方は、MATLAB を使用して音声データから雑音を割り、デシベルで表した数値を出すようなプログラムを書き、算出した．

カラオケ音源の楽曲は、男性曲 1 曲、女性曲 2 曲を使用した．音声データはサウンドレコーダーを使用して、コンデンサーマイクで楽曲をイヤホンで聞きながら歌唱し、録音を行った．音声強調実験と同様に女性シンガーの楽曲は裏声で歌唱した．音声データはモノラルのサンプリングレート 16Khz である．その音声データとカラオケ音源を MATLAB を使って CD 音源で 2 チャンネルのステレオだったものを 1 チャンネルだけ取り出し、モノラルに変換した．また、サンプリングレートも MATLAB のコマンドを使って 44.1Khz から 16Khz にリサンプルした．後述の実験でも CD 音源には、同様の処理を行った．そこから正規化を行い、SN 比を-5dB、0dB、5dB という方で調整、合成していった．

カラオケ音源の各曲の特徴を以下に示す．

3.2 カラオケ音源の特徴

・楽曲 A

Perfume/Seventh Heaven

女性 3 人組の楽曲．演奏のほとんどが打ち込みの曲．歌声にもエフェクトがかかっている．

・楽曲 B

森山直太朗/生きてることが辛いなら

男性シンガーの楽曲．ストリングスなどの演奏が中心．

・楽曲 C

supercell/君の知らない物語

女性シンガーの楽曲．ギター、ベース、キーボードなどバンド演奏．

3.3 実験結果

表 2 は合成した楽曲からアライメントをとった結果である．SN 比が 0dB の状態がカラオケの音源と歌唱した音源の音量が同じで、5dB だと歌唱した音源が大きいことを表し、-5dB だとカラオケの音源が大きいことを表している．ここでの 0 秒というのは、一番短い一単語だけでもアライメントがとれなかったこと同じことを示す．

表 2. 背景音の結果

楽曲 \dB	-5	0	5
A(280 秒)	0	25	280
B(273 秒)	103	160	273
C(341 秒)	33	44	44

3.4 考察

実験から SN 比を 5dB にしたところ、楽曲 A、楽曲 B は楽曲全体のアライメントをとることに成功している．そこから、歌唱した音源がカラオケの音源より大きければ、アライメントはとりやすくなると考えられる．また、-5dB と 5dB の結果を比較したところ、平均で 2 分半アライメントが長くとれた．SN 比が 10dB 向上させれば、長くアライメントがとれるといえるであろう．

フレームの適合結果としては、楽曲 A で本来の歌詞とは 10 秒以上のずれが発声したが、楽曲 C ではアライメントがとれた範囲内ではずれは 1 秒未満であった．これは背景音が楽曲 C でアライメントをとれた範囲では小さかったのに対し、楽曲 A では大きく入っていたことが原因として考えられる．

また、歌声のみデータでアライメントをとったところ、楽曲 A、B は楽曲全体の歌詞のアライメントがとれたが、楽曲 C ではとれなかった．これは楽曲 A と B は男性ヴォーカルで楽曲 C は女性ヴォーカルだったため、C を裏声で歌ったことが影響していることが考えられる．このことから歌声によってもアライメントのとれる長さが変わってくる可能性がある．

4 音声強調の手法

楽曲から歌詞アライメントをとる際に問題になることは、楽曲にはもちろん歌声の音声だけではなく、バンドの楽器演奏

（ギター、ベース、ドラム、キーボードなど）やコーラスの音声など様々な音が入っていることである．そのため歌声の音声以外の成分を除去しなければ楽曲から歌詞のアライメントを正確にとることが難しくなってしまう．通常の音声認識において雑音とは、対象の音声以外の話し声や環境音などのことを示すが、本研究においての雑音とは歌声の音声以外の成分全てを示す．そこでスペクトルサブトラクション（以下 SS）と変調スペクトルの 2 つの音声強調の手法を用いて、楽曲に対してどの音声強調方法が有効かどうか調べた．
SS には、BollSS、BeroutiSS、Multi-bandSS のプログラムを使用して音声強調を行った．

4.1 変調スペクトルに基づく音声強調

スペクトルやケプストラムなどの特徴パラメータの時間変化をフーリエ変換し、周波数次元でみたものを変調スペクトルと呼ぶ [5]．音声認識に必要な情報のほとんどが変調周波数の 1Hz〜16Hz の帯域に存在し、特に 2Hz〜4Hz の部分が最も重要であるとされている．文献 [6] において、パワースペクトルに対する変調スペクトルと対数パワースペクトルに対する変調スペクトルのどちらに対しても変調周波数の約 7Hz 以下の帯域のみ用いても音声認識の性能が低下しないということが示されている．

4.2 音声強調実験

音声強調の実験として、BollSS、BeroutiSS、Multi-bandSS と変調スペクトルを用いた音声強調を使用して、どの手法が楽曲の雑音を除去することができるか実験を行った．
変調スペクトルを用いた音声強調を行うプログラムの流れとしては、はじめに入力された音声信号からランニングスペクトルを求め、バスドラムやベースなどの 60Hz 以下になるような低域の成分を除去し、音声信号を変調スペクトルに変換し、変調周波数の 1〜7Hz 以外の成分を除去した．
楽曲データは、ヴォーカルの歌声がないカラオケの音源と自分で楽曲を歌唱した音声データを CD 音源と聞き比べて同じぐらいの音量バランスで合成したものを使用した．カラオケ音源には、日本のポピュラー音楽で、男性シンガーの楽曲として森山直太朗の「生きてることが辛いなら」と女性シンガーの楽曲として、supercell の「君の知らない物語」の 2 曲を使用．音声データはどちらもモノラルのサンプリングレート 16Khz で、サウンドレコーダーを使用して、コンデンサーマイクで録音した．なお、女性シンガーの楽曲は男性楽曲と違いを出すために、裏声で歌唱した．
評価方法は音声強調を行った楽曲データから歌唱した音声データを引いたものとカラオケの音源を用いて、セグメントごとに SN 比を出していき、最後にその SN 比の合計から平均をとって、どのぐらい雑音が改善しているかを数値で出した．図 3 は青色の波形がカラオケ音源で、赤色の波形が音声強調後の背景音である．

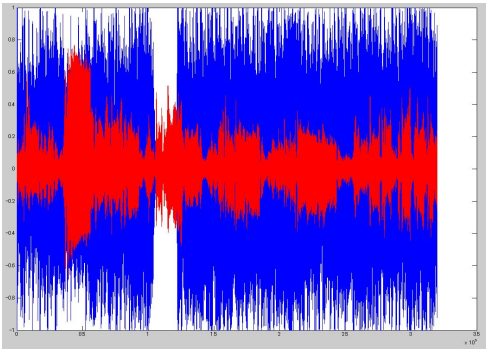

図 3. カラオケ音源と音声強調後の背景音

4.3 実験結果

表 3 は楽曲 2 曲に音声強調を行ってどれだけ雑音が改善されたかの改善度を表している．数値が 0 のときは音声強調前と後で違いがないことを表している．
表 3 を見る限り、変調スペクトルを用いた音声強調の改善度が一番高く、SS はどの種類もだいたい同じ改善度であった．
なぜ SS では楽曲の音声強調ができないかというと文献 [7] でも示されているように、SS は平均的なスペクトルを引くため定常的な雑音しか除去できないので、元信号が周期的である場合、音声強調の手法として効果的であるが、楽曲の演奏のような定常的とはいえない音に対しては十分な性能が得られないからだということを確かめることができた．

表 3. 評価結果

音声強調の手法	改善度（男性曲）	改善度（女性曲）
BollSS	-2.8108	4.0699
BeroutiSS	-2.7951	3.6918
Multi-bandSS	-2.7278	3.6893
変調スペクトル	4.8084	12.7124

5 アライメント実験

5.1 実験

実験では、日本のポピュラー音楽 5 曲を使用してアライメントをとった．楽曲は音声強調実験と同様にモノラルのサンプリングレート 16Khz に変換した．また、音声強調実験で一番効果が出た変調スペクトルを用いた音声強調をそれぞれの楽曲にかけ、雑音の除去を行った．今回の実験で選んだ楽曲はコーラスが入っていなかったり、歌声に派手なエフェクトがかかっていなかったりと比較的アライメントをとりやすいであろう楽曲を主観的に選んだ．楽曲のそれぞれの特徴は以下のとおりである．

5.2 各楽曲の特徴

1. コイントス/藍坊主
日本のロックバンド．ヴォーカル、ギター、ベース、ドラムからなる男性 4 人組．楽曲には打ち込みはほとんどないが、演奏自体が大きめで、その状態が楽曲の初めから最後まで続く．
2. ママのピアノ/風味堂
ピアノ、ベース、ドラムからなる男性 3 人組．楽曲はバラードのようなものなので、バンド演奏の音自体はそこまで大きくない．前奏はほぼなし．
3. たしかなこと/小田和正
シンガーソングライター．楽曲はギター、ベース、ドラム、その他の演奏．
4. まっ白/小田和正
上記と同様．楽曲はギター、ベース、ドラム、その他の演奏．
5. 歌舞伎町の女王/椎名林檎
女性シンガー．演奏はギター、ベース、ドラム、その他．楽曲の途中に約 15 秒ほど口笛のパートがある．

5.3 実験結果

表 4 は楽曲からアライメントをとった結果で、一番長くアライメントがとれた秒数を表している．
「コイントス」は、演奏の音量が大きいぶん音声強調前ではアライメントがとれる秒数が短かく、音声強調後も楽曲の演奏がある程度残っていたため他の楽曲と比べてアライメントは長くとれなかった
「ママのピアノ」は、演奏自体そこまで大きくなかったにも関わらず、思った以上にアライメントがとれなかった．そこで原因を調べるために音声強調前の楽曲を聞いてみたところ歌声にホールで歌っているかのようなエコーが若干かかっていることに気がついた．これは音声強調後の楽曲を聞いても他の楽曲と比べると音声にエコーがかかっているのが聞き取れた．そのことから音声にエフェクトがかかっているとアライメントをとることが難しくなる可能性がある．
「たしかなこと」は、演奏より歌声が大きかったので音声強調しない状態でもこれだけアライメントをとることができた．
「まっ白」は、音声強調前では一単語もアライメントをとることができなかった．

「歌舞伎町の女王」は、同じ種類の楽器演奏が含まれていた「コイントス」より音声強調後のアライメントが長くとれたことから、男性と女性の歌声に違いがあり、女性の歌声のほうがアライメントをとりやすい可能性が考えられる．

表 4. 様々な楽曲をアライメントをとった結果

楽曲	音声強調前	音声強調後
コイントス (235 秒)	10	91
ママのピアノ (286 秒)	34	83
たしかなこと (299 秒)	191	238
まっ白 (268 秒)	0	268
歌舞伎町の女王 (173 秒)	0	153

5.4 考察

実験では、どの楽曲も音声強調後はする音声強調前よりアライメントをとれる秒数が長くなった．プログラムでは、雑音を除去するために定常な成分を除去し、非定常な成分を残すようにしているため、楽器演奏やコーラスなど定常的とはいえない成分は大きく除去できないのではと考えていた．
しかし、実験結果から楽器の音などはランニングスペクトルにおける振幅・パワーの時間変化が大きく、今回音声強調した変調周波数よりも高いところに成分が分布していたため、除去されたという可能性が考えられる．
また、楽曲の歌声自体にエコーなどのエフェクトがかかっている場合、音声の成分自体が変わってしまう可能性があり、音声強調したときにうまく残すことができず、アライメントがとりにくいと考えられる．
カラオケ実験と同様にフレームがどの程度適合しているか調べたところ、「たしかなこと」のアライメント結果と楽曲の歌詞を比べたらだいたい 1 秒前後のずれがあった．歌詞表示システムを作成する際にはある程度のずれを考慮して歌詞を表示させなければいけない可能性がある．

6 歌詞表示システム試作

歌詞表示システムをグラフィカルなコンポーネントを使ったアプリケーションを開発する為に、Java で用意されている Swing というものを使用．現在作成途中ではあるが、完成している部分のプログラム構成は以下の通りで、図 4 は今回作成したプログラムの画面イメージである．
1) wav ファイルの読み込み
2) MATLAB 用設定ファイル (テンプレート) の読み込み
3) MATLAB 用設定ファイルのカスタマイズ (入出力ファイルの設定)
4) MATLAB に処理を回す
5) Timer コントロールを使って出力された wav ファイルがあるのを確認
6) Julius を使って出力された wav ファイルからアライメントを取得
7) ログファイルを出力
8) Timer コントロールを使ってログファイルがあるのを確認

図 4. 画面イメージ

ある程度エラー処理は入れて、wav ファイルなどがない状態ではプログラムが動かないようになどしている．プログラムでは、入力 wav ファイル名の部分には、'INFILENAME' という文字列、出力 WAV ファイル名の部分には、'OUTFILENAME' という文字列にしておく．Java プログラム側でこの文字列の部分を、実際の wav ファイルの場所に差し替えるためである．出力 wav ファイルは実際の wav ファイル名が'c:\test.wav' だとすると'c:\test_out.wav' とファイル名の最後に自動的に'_out' が付くようになっている．

7 あとがき

本研究では、アライメントをとることで歌声に合わせて歌詞を単語単位で表示するシステムを提案した．それぞれの実験から、音声強調により楽曲の SN 比が 5dB 以上高くなることが分かった．また、歌詞アライメントは最低でも 83 秒、最高で 268 秒とることができた．実験結果から歌詞のアライメントをとって歌詞表示するシステムは有用であると考えられる．システムを作成する際は、実験した変調スペクトルを用いた音声強調を行って、アライメントがとりやすく、出力される情報がある程度楽曲の歌詞のタイミングと適合するように楽曲を加工する必要がある．アライメント実験から音声強調後どんなに結果が悪くても 83 秒以上ではアライメントがとれたので、楽曲をある区間で切り分け、それぞれからアライメントをとり、全ての情報をつないで 1 曲にすれば、長い楽曲であっても曲全体のアライメントをとることができると考えられる．

今後の課題としては、音声強調を行ってアライメントとり、取得したフレーム情報が楽曲の歌詞のタイミングとどのぐらいずれがあるのかを評価し、歌詞表示システムとして問題がない程度のずれになるようにする、もしくは全体でずれの平均を求めてその平均の分だけ歌詞の表示をずらして行う必要がある．アライメントがとるのが困難だと思われる楽曲（歌声にエフェクトがかかっているなど）の場合、アライメントがどの程度とれるか調べなければならない．また、楽曲を切ってアライメントをとった場合、切った位置が一単語の途中だと、切った部分をつなぎ合わせたときうまくつなぐことができない可能性がある．そこで、歌詞カードでは一般的に普通 A メロ、B メロ、間奏のように部分ごとに歌詞が書いておりその間が 1 行空いている、そのところで歌詞のデータは区切る．楽曲データもそれに合わせて A メロと B メロの間や間奏部分の長さからより長く音声が途切れている部分で区切って歌詞を照らし合わしていけば、楽曲全体を区切りつつもアライメントがとることが可能ではないかと考えられ、それを調べるような実験を行っていかなければいけないだろう．

参考文献

[1] "大語彙連続音声認識エンジン Julius" http://julius.sourceforge.jp/
[2] "「歌詞ピタ」サービス" http://ss.kashi-ism.jp/kashipita/
[3] 西村 明、坂本 真一 "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" 情報処理学会研究報告. pp.7-12 20070801 社団法人情報処理学会 (2007.8.01)
[4] 細谷徹、鈴木基之、伊藤彰則、牧野正三 "歌声から得た歌詞を用いた楽曲検索に関する検討" 日本音響学会講演論文集 (2004.9)
[5] 藤田 匡彦、早坂 昇、宮永 喜一 "耐雑音音声認識における変調ケプストラム操作の効果に関する一考察" 電子情報通信学会技術研究報告. pp.29-33 20070302, 社団法人電子情報通信学会 (2007.3)
[6] 早坂 昇、和田 直哉、宮永 喜一、畑岡 信夫 "ランニングスペクトルフィルタを用いた雑音にロバストな音声認識" 電子情報通信学会技術研究報告. pp.31-36 20030619 (2003.6)
[7] 岡崎雅嗣、国本利文 (ヤマハ)、小林隆夫 (東京工大 大学院総合理工学研究科) "多段スペクトルサブトラクション法を用いた楽音の強調" 電子情報通信学会論文誌 Vol.J88-D-2, No.12, Page2301-2310 (2005.12.01)