RE-005

局所変化率変換に基づく有声音声の正弦波モデル
Sinusoidal Modeling for Voiced Speech Based on a Local Vector Transform

伊藤 仁† 伊藤 彰則†
Masashi Ito Akinori Ito

1. はじめに

音声の音響特性を詳細に分析する手法のひとつとして，入力信号を振幅と周波数が時間変化する正弦波成分の和として近似する正弦波モデルが挙げられる。McAuley と Quatieri は，短時間パワースペクトルの局所的なピークに基づいて，この正弦波成分のパラメータを推定する音声分析手法を提案した[1]。推定した正弦波パラメータから再合成した信号は，元の音声と知覚的に弁別が困難であると報告されており，この手法を応用した高品質の音声合成・変調アルゴリズムも検討されている[2,3]。

だが，いくつかの研究[4,5]で指摘されている通り，このスペクトルピークに基づく正弦波モデルは，音声の非定常部で分析精度が著しく劣化するという問題がある。通常の発話では，声帯の振動周波数や声道形状は時々刻々変化するため，出力音声を正弦波モデルで近似した場合に，その振幅や周波数が定常と見なせる区間はごく限られている。よって，非定常部で高精度のパラメータ推定が必要となる。

これを実現するために，推定した正弦波パラメータから再合成した信号と，入力信号との誤差を目的関数として，繰り返し演算によりパラメータを決定する手法が提案されている[5,6]。この手法は，初期値を適切に設定した場合には有効だが，一般的な最適化手法と同様，演算がローカルミニマムに収束してしまう危険性がある点と，スペクトルピークを用いる手法と比較して演算量が膨大になる[6]という点で，分析手法としては不十分である。

一方，比較的少ない演算量で非定常信号の音響特性を精度良く分析する手法として，入力信号の複素スペクトルを利用するアプローチがある。特に瞬時周波数(Instantaneous Frequency: IF)は非定常信号の音響特性を高い精度で分析できる手法として活発な研究が続けられており[7]，IF に基づく正弦波モデルも提案されている[8]。著者らは，周波数だけでなく振幅も非定常な正弦波信号を正確に分析するための手法として，局所変化率変換(Local Vector Transform: LVT)を提案した[9]。LVT は，局所的な分析区間における瞬時振幅と位相を単純な時変関数でモデル化することで，入力信号の複素スペクトルからこれらの関数を一意に決定できる点に特徴があり，有声音の低次の調波成分を高い精度で分析することができる。しかし高次の調波成分に関しては，特に基本周波数(F0)の非定常性が高い場合に，成分間の干渉により分析精度が劣化するという問題があった。IF 法でも同様の問題が指摘されており，これに対応するためには分析時間軸を F0 に応じて伸縮する所謂 Time-warping が有効であることが知られている[8]。本稿では，この時間軸変換を LVT に導入し，高次の成分を含む有声音声全体を分析するための新たな手法を提案し，その有効性を多数の音声信号を用いた性能評価実験により定量的に調べる。

† 東北大学大学院工学研究科 Department of Electrical Communication Engineering, Tohoku University

2. 提案手法

2.1 概要

Fig. 1 にLVTに基づく音声分析手法の計算処理を示す。この手法は，入力音声信号$x(t)$を下記の正弦波モデルで近似し，各正弦波成分の瞬時対数振幅$A_k(t)$と，瞬時位相$P_k(t)$を推定する。なお第k成分の瞬時角周波数は$dP_k(t)/dt$で表される。

$$x(t)=\sum_{k=1}^{K}\exp(A_k(t))\cos(P_k(t)) \tag{1}$$

これらのパラメータは 2 段階の計算処理で推定される。まずSTEP1 では，第 1 成分の瞬時位相（基本位相）$\varphi(t)$をLVTにより推定する。先行研究[9]で示した通り，LVTは有声音声の低次の調波成分など，パワースペクトル上での重なりが少ない正弦波信号を高い精度で推定できる。従って推定された基本位相は信頼性が高く，これを用いて入力信号の時間軸tを位相軸$\theta=\varphi^{-1}(t)$に変換することで，基本周波数の時間変化特性をキャンセルすることが可能である。

次にSTEP2 では，この変換信号 $\hat{x}(\theta)$ に対して再びLVTを適用し，今度は全ての正弦波成分の瞬時振幅 $\hat{A}_k(\theta)$ と位相 $\hat{P}_k(\theta)$ を推定する。$\hat{x}(\theta)$ を構成する全ての正弦波成分は，瞬時周波数が位相軸θに対してほぼ定常と見なせるため，パワースペクトル上での重なりが非常に小さくなる。よって，各成分の正弦波パラメータは，LVTにより精度良く推定できる。最後に，これらのパラメータの位相軸θを時間軸tに逆変換することで，式(1)の正弦波パラメータ$A_k(t)$と$P_k(t)$を決定する。以下では各計算処理の詳細について述べる。

2.2 短時間フーリエ変換

まず入力信号の，時刻t_Cにおける短時間フーリエスペクトル$X_0(\omega)$を次のように計算する。時刻t_Cは音声区間の開始から終了まで，一定の間隔Δで変化させる。

$$X_0(\omega)=\int_{-T}^{T}x(t-t_C)w(t)\exp(-j\omega t)dt \tag{2}$$

$$w(t)=\exp\left(-\gamma\cdot t^2\right)\Big/\sqrt{2\pi} \tag{3}$$

ここで$w(t)$はガウス型の窓関数で，γは定数である。また定数Tは積分範囲を表し，窓関数$w(\pm T)$が十分にゼロとみなせる範囲を選ぶ。さらに，各時刻t_Cにおいてスペクトル$X_0(\omega)$の角周波数ωに対する 1 次と 2 次の導関数$X_1(\omega)$と$X_2(\omega)$を次式のように計算しておく。

$$X_1(\omega)=\frac{dX_0(\omega)}{d\omega}=\int_{-T}^{T}x(t-t_C)(-jt)w(t)\exp(-j\omega t)dt \tag{4}$$

$$X_2(\omega)=\frac{d^2X_0(\omega)}{d\omega^2}=\int_{-T}^{T}x(t-t_C)t^2w(t)\exp(-j\omega t)dt \tag{5}$$

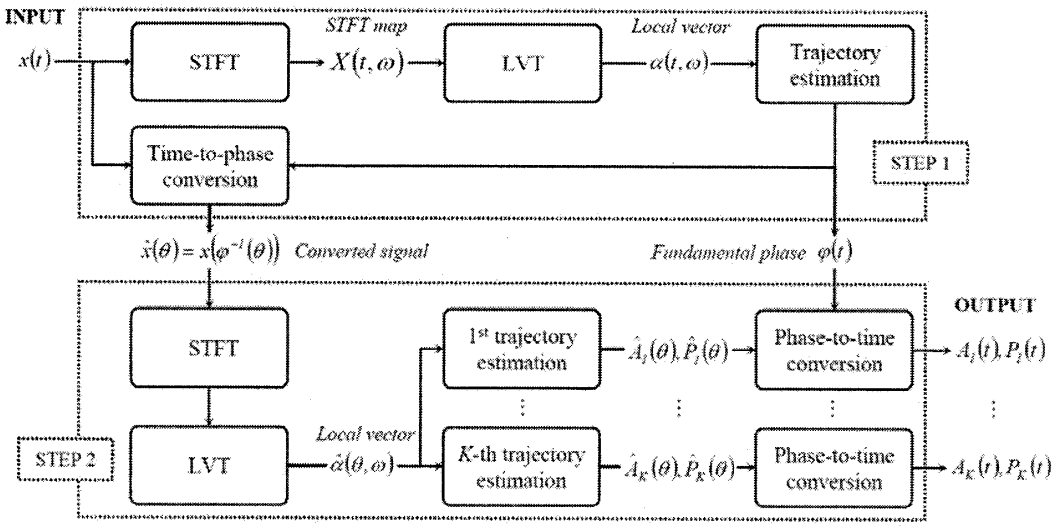

Fig.1 A block diagram of the proposed method.

2.3 Local Vector Transform

ここで角周波数ωに正弦波成分が存在し，その瞬時対数振幅a(t)と瞬時位相p(t)が時刻t_C近傍で，Taylor展開の 2 次までの項で近似できると仮定すると，これらのパラメータはフーリエスペクトルから次のように一意に決定できる[9]。

$$a(t)=\mathrm{Re}[\alpha(t-t_C)]\big|_{t\approx tC},\ p(t)=\mathrm{Im}[\alpha(t-t_C)]\big|_{t\approx tC} \quad (6,7)$$

$$\alpha(t)=\alpha_0+\alpha_1 t+\alpha_2 t^2/2 \quad (8)$$

$$\alpha_0=\log\left(2X_0^2/\sqrt{X_1^2-X_0X_2}\right)+X_1^2/2\left(X_1^2-X_0X_2\right) \quad (9)$$

$$\alpha_1=j\cdot\left(\omega+X_0X_1/\left(X_1^2-X_0X_2\right)\right) \quad (10)$$

$$\alpha_2=\gamma-X_0^2/\left(X_1^2-X_0X_2\right) \quad (11)$$

有声音声は調波構造を持つため，正弦波成分のエネルギーは基本周波数の整数倍近傍の周波数に離散的に分布している。この性質を利用して，上記の推定パラメータのうち，$|\mathrm{Im}[\alpha_i]-\omega|<\varepsilon$の条件を満たすものだけを選択することで，実際の正弦波成分に対応する候補を絞り込むことができる。

Fig. 2 に，実際の音声から推定したパラメータの例を示す。入力音声は女性話者が発話した/menyu/であり，Fig. 2a がスペクトログラムを表す。図の点線はデータベースのラベル情報から得られた音韻境界を表す。この音声に LVT を適用し，各分析時刻において得られた瞬時対数振幅と瞬時周波数を Fig. 2b と Fig.2e にそれぞれ表す。式(6,7)で表現される各時刻の局所的な軌跡が，音声区間で連続的に並び，これらが入力音声の第 1～5 調波成分に対応することが確認できる。また Fig. 2c は，推定パラメータの微小領域を 3 次元で示したものであり，各曲線は振幅については t の 2 次関数，周波数については t の 1 次関数となる。実際の音声では，これらのパラメータはより複雑な関数で時間変化するが，少なくとも振幅と周波数を定常と見なした場合（振幅，周波数ともに t の 0 次関数）と比べると，この単純な非定常性を導入することで推定精度は飛躍的に向上する。Fig. 2 は推定パラメータから第 1 調波成分に対応する正弦波信号を再合成し，入力から除いた残差信号のスペクトルを表す。図から，この正弦波パラメータが高い精度で推定されていることが確認できる。

2.4 振幅・位相軌跡の推定

次に，各時刻で離散的に得られる多数の LVT パラメータに対して，単一調波成分に対応するものだけを選択し，連続関数である A(t)と P(t)を推定する。この選択には，以下のコスト関数 Ψ に基づく動的計画法を利用する。

$$\psi(n,m)=\min_l\left[\psi(n-1,m)+D(n,m,l)^2\right] \quad (12)$$

$$D(n,m,l)=\mathrm{Im}\left[\alpha_1(n-1,l)-\alpha_1(n,m)+\frac{\alpha_2(n-1,l)-\alpha_2(n,m)}{2\Delta}\right] \quad (13)$$

ここで，$\alpha_i(n,m)$は時刻$t_C = n\cdot\Delta$において得られたLVTパラメータのうち，前述した角周波数の条件を満たすm番目のものを示す。また$D(n,m,l)$は，パラメータ$\alpha_i(n,m)$と$\alpha_i(n\text{-}1,l)$の不連続性を表す関数であり，これを利用して音声区間内で連続的な候補だけを選択する。

この様にして候補が選択できれば，対応する連続関数 A(t)と P(t)は，それらの単純な重み付け和として次のように決定できる。

$$A(t)=\sum_{n=1}^{N}w(t-t_n)\mathrm{Re}[u_n(t)]\Big/\sum_{n=1}^{N}w(t-t_n) \quad (14)$$

$$P(t)=\sum_{n=1}^{N}w(t-t_n)\mathrm{Im}[u_n(t)]\Big/\sum_{n=1}^{N}w(t-t_n) \quad (15)$$

$$u_n(t)=\alpha_0^{SEL}(n)+\alpha_1^{SEL}(n)(t-t_n)+\alpha_2^{SEL}(n)(t-t_n)^2/2 \quad (16)$$

ここで，$\alpha_i^{SEL}(n)$は，nに対応する時刻で選択されたLVTパラメータを表す。また$w(t)$は式(3)の窓関数である。得られたA(t)とP(t)は，分析時間フレーム（$t = n\cdot\Delta$）近傍で選択したLVTパラメータの特性を反映し，フレーム間の特性は滑らかに時間変化する関数となる。

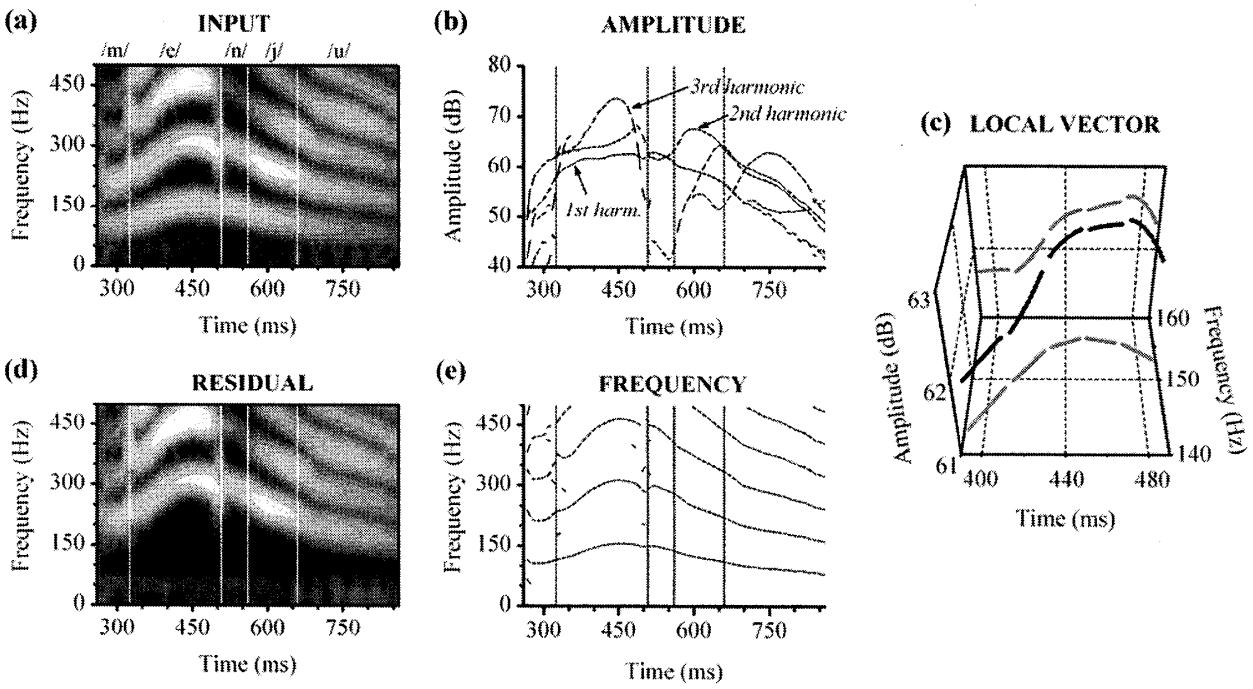

Fig.2 Local vector transform of voiced speech signal

2.5 時間-位相変換

Fig. 1 の処理STEP1 では, 第 1 調波成分の瞬時位相関数 $P_1(t)$を推定して基本位相$\varphi(t)$とする。Fig. 3 に, 実際の音声信号に対する基本位相の推定例を示す。入力信号は女性話者の単語音声で, 波形とスペクトログラムをFig. 3cと 3eに示す。Fig. 3eに示したように, 2 kHz以上の高周波数では, スペクトログラムにモアレ状の独特のパタンが観測できる。これは, 高次の正弦波成分の非定常性により, 各成分のエネルギーがスペクトル上で拡散し, 互いに干渉し合うことで発生する。この様な非定常性の影響を軽減するために, 入力信号の時間軸tを基本周期$\theta = \varphi(t)$を用いて次のように変換する。

$$\hat{x}(\theta) = x\left(\varphi^{-1}(\theta)\right) \tag{17}$$

この変換により, 入力信号 $x(t)$の時間軸 t は, 変換信号 $\hat{x}(\theta)$ の基本周期軸 θ に置き換えられる。これに伴って, 変換信号を短時間フーリエ分析した場合の角周波数軸は, 調波成分の番号(無次元量)となる。Fig. 3d に上記の音声信号から得られた変換信号波形を, また Fig. 3f に変換信号のスペクトログラムの例を示す。変換信号のスペクトログラムでは, 各調波成分に対応するエネルギーの存在位置が, 横軸(基本周期軸)の値によらずほぼ一定となっていることが分かる。これは, 推定された基本周期の妥当性を支持するものである。

単一の時刻(t = 400 ms)における入力信号と変換信号のパワースペクトルを比較すると, 入力信号のスペクトルでは第 10～13 次以下の低次の調波成分は, 互いに分離しているが, 20 次以上の高次の成分は干渉により明確なピークが観測できなくなっている事が分かる(Fig. 4g)。これに対して変換信号のパワースペクトルは, 30 次以上の高次の成分まで各調波に対応する明確なピークが観測できる(Fig. 4f)。このように成分間の干渉が小さい信号に対しては, 2.3 節で述べた LVT を適用することで, 正弦波成分のパラメータを容易に求めることができる。

Fig. 1 の STEP2 の処理では, まず変換信号に LVT を適用して式(8)の 2 次関数の係数群を推定し, 2.4 節で述べた動的計画法により, 各成分に対応する係数群を選択する。Fig. 4f に示したように, 変換信号では k 次の成分が存在する範囲は, 縦軸(調波番号軸)の k 近傍に限定されているため, 式(12,13)のコスト関数を計算する際に, 適切な範囲を設定することで選択誤りを抑制することが可能である。

このようにして得られた各成分の LVT パラメータから, 式(14-16)を用いて, 振幅関数 $\hat{A}_k(\theta)$ と位相関数 $\hat{P}_k(\theta)$ が決定できる。これらの関数の基本位相軸 θ を, 下記のように逆変換することで, 各成分の正弦波パラメータを推定する。

$$A_k(t) = \hat{A}_k(\varphi(t)) \tag{18}$$

$$P_k(t) = \hat{P}_k(\varphi(t)) \tag{19}$$

なお, 式(17)の時間-位相変換と, 式(18,19)の逆変換の計算には, SINC 関数を利用する。逆変換は分析すべき調波成分の数 K だけ行う必要があるが, SINC 関数の係数は全ての調波成分で共通となるので, この変換による計算負荷はさほど大きくはならない。

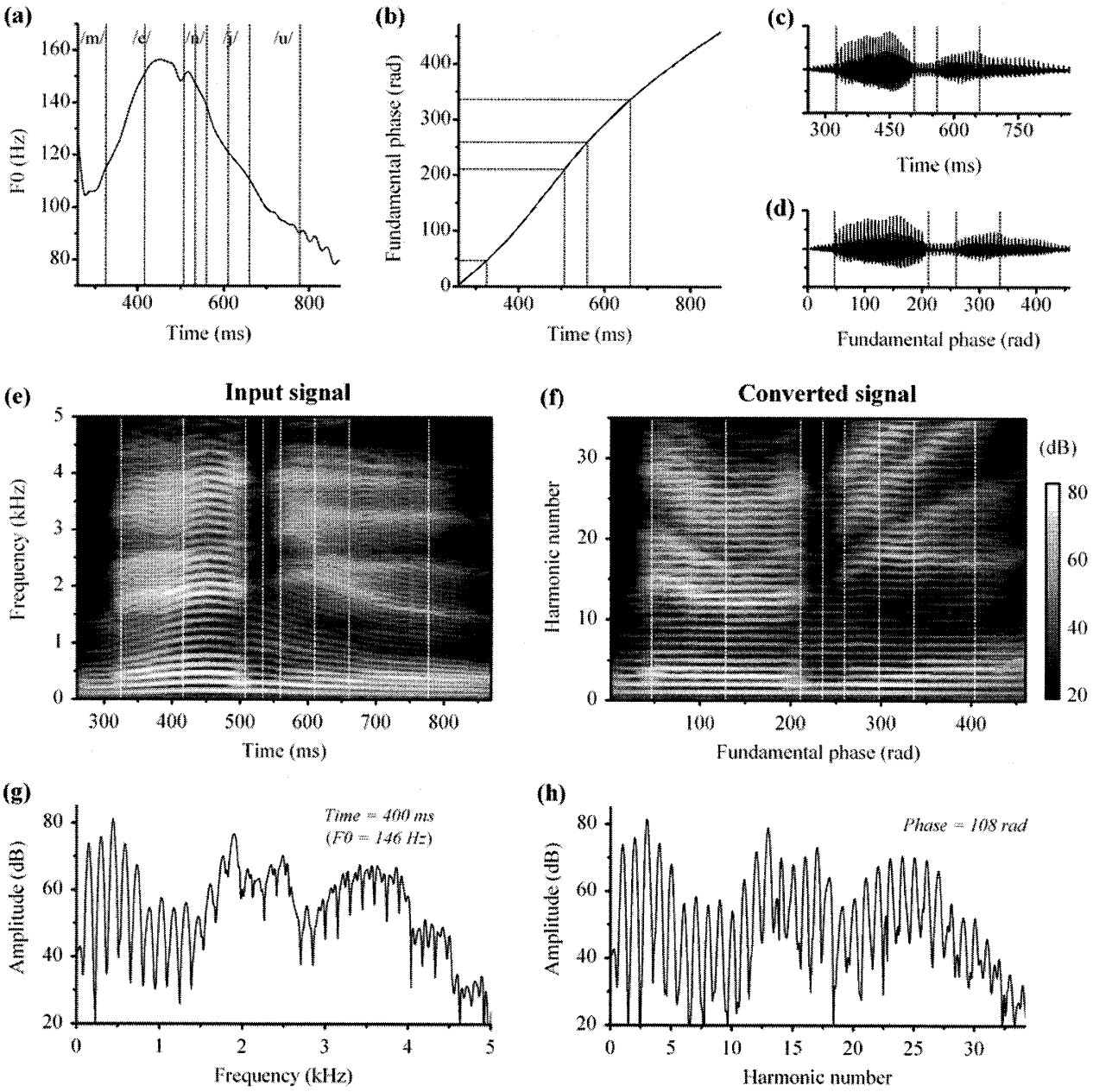

Fig.3 Time to phase conversion for voiced speech signal

3. 性能評価実験

以下では, 前節で述べた分析手法の性能を, 多数の話者が発話した単語音声を入力として評価する。入力信号が自然発話であるため, 基準となる真の正弦波パラメータは未知であり, この手法の推定精度を振幅や位相などのパラメータに基づいて直接的に評価することは難しい。そこで以下では, 推定されたパラメータから再合成した信号と入力信号の残差を計算することで, 推定精度を間接的に評価する。また比較のため, 同じ入力音声を既存のアルゴリズムで分析した場合の結果についても同様の評価を行う。

3.1 入力信号

テスト用の音声信号として, ATR ディジタル音声データベース[10]に含まれる音素バランス単語 216 語から, 有声音素のみで構成される 12 単語（/iyoiyo/, /aruminiumu/, /uyamau/, /eiyu/, /nyuiN/, /meue/, /yumoa/, /reNai/, /oNnamyori/, /menyu/, /meiryo/, /ou/）を抽出する。話者は成人男性 37 名, 成人女性 38 名の 75 名で, 合計 900 サンプルを用いて評価実験を行う。各サンプルの音声区間は, データベースのラベルにある単語の先頭の音韻の開始時刻と, 最後の音韻の終了時刻から決定する。

3.2 分析アルゴリズム

これらの入力音声信号の正弦波パラメータを，2 節で述べたアルゴリズムにより推定する。ここでは基本位相φを求める際の窓関数のパラメータγ = 7.8 Hz, フレーム刻みΔ = 5 ms, 基本周波数の存在範囲は 50～500 Hzとしている。また，時間軸変換後の推定では，γ = 0.031 rad^{-1}, $\Delta = \pi/4$, 第k成分の周波数存在範囲は$k \pm 1/2$ である。なお，正弦波パラメータは，最大 160 次までの調波成分について推定する。

この手法と比較するために，既存手法に基づく 2 種類のアルゴリズムについても同様の実験を行う。ひとつはスペクトルの局所ピークの振幅と周波数から正弦波パラメータを推定する手法[1]である(Peak-picking)。先行研究では，各時間フレームで得られた局所ピーク群を適切な調波成分に割り当てるために，birth-and-death と呼ぶヒューリスティックな処理が適用されているが，ここでは信頼性の高い基本周波数が LVT により得られているため，単純にこれを利用して割り当てを実現する。この Peak-picking 法は，時間軸変換を利用しないため，Fig. 3e に示すような成分間の干渉が強い音声で推定精度が劣化すると予測される。

また比較対象とする第 2 のアルゴリズムは，瞬時周波数に基づく IF-attractor を用いる手法である[8]。この手法は，提案手法と同様，基本周波数の時間変化に基づいて時間軸を変換し，IF-attractor を用いて正弦波パラメータを推定する。先行研究では，時間軸変換を分析フレームごとに実行しているが，これは推定精度を劣化させる要因である可能

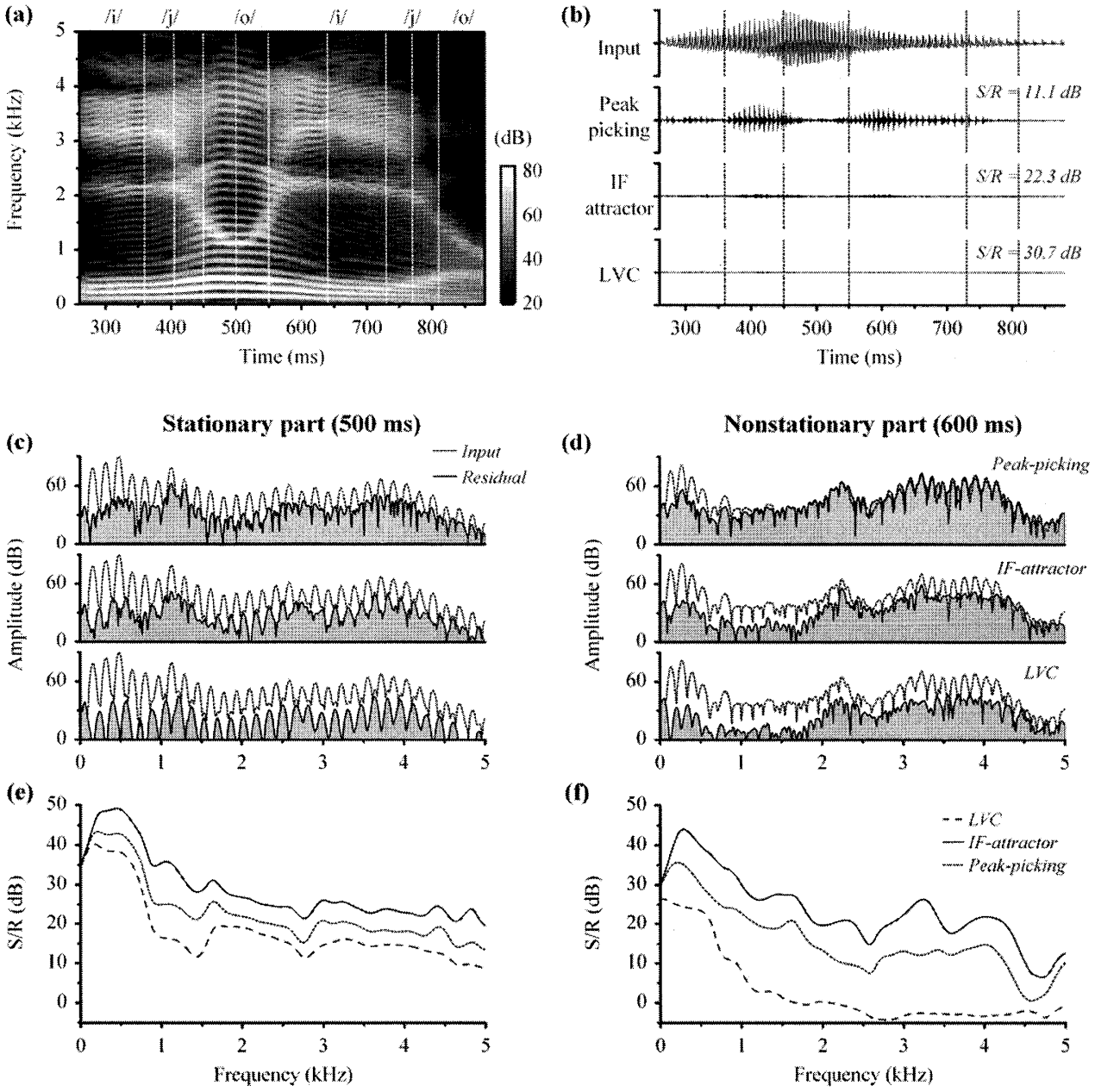

Fig.4 Input and reconstructed signals using the proposed and the conventional methods.

性があることが指摘されているため[8], ここでは提案手法と同じ音声区間全域に対する時間軸変換を行った後, IF-attractor で正弦波パラメータを推定し, 時間軸の逆変換を実行する。これら 2 種類の手法(Peak-picking, IF-attractor)と提案手法を用いて, 振幅と位相が一定の合成調波信号の正弦波パラメータを推定すると, アルゴリズムによる違いは殆ど見られず, すべてのアルゴリズムで高い分析精度が得られた。

3.3 結果

Fig. 4 に推定結果の例を示す。入力音声のスペクトログラムが Fig. 4a に, 信号波形が Fig. 4b の上段に対応する。図から, 基本周波数は時刻 400 ms と 600 ms で急激に時間変化し, 500 ms では変化が少ないことが確認できる。この音声信号の正弦波パラメータを, 上で述べた 3 種類の手法で推定した後, それぞれに対して式(1)により信号を再合成し, 入力信号との差をとることで残差信号が得られる。Fig. 4b の第 2～4 段は, Peak-picking 法, IF-attractor 法, 提案手法による残差信号波形を表す。残差信号のエネルギーは, Peak-picking 法で最も大きく, 提案手法が最も小さかった。パラメータの推定精度は, 入力信号と残差信号のエネルギー比(S/N)として定量化することが可能であり, この入力信号に対する S/N は, Peak-picking 法で 11.1 dB, IF-attractor 法で 22.3 dB, 提案手法で 30.7 dB であった。従って, この例では提案手法により推定された正弦波パラメータが, 最も正確であると言える。

ここで Peak-picking 法の残差信号に注目すると, 時刻 t = 400, 600 ms 近傍で大きなエネルギーを持つことが分かる。これは基本周波数の非定常性が強い区間とほぼ一致する。この点について詳細に調べるために, 時刻 500 ms(定常部)と 600 ms(非定常部)で, 入力信号と残差信号のパワースペクトルを計算した(Fig. 4c, 4d)。基本周波数の定常部では, 残差信号のスペクトルは, 入力信号と比較して 20～30 dB 程度低く, アルゴリズムによる違いは小さかった(Fig. 4c)。周波数帯域ごとの S/N も, 全てのアルゴリズムで良く似た傾向を示し, アルゴリズムの性能差の周波数依存性は殆ど見られなかった(Fig. 4e)。

一方, 非定常部では 3 種類のアルゴリズムで顕著な差が見られた(Fig. 4d, 4f)。Peak-picking 法は, 他の 2 つの手法と比較すると特に高周波数成分の推定精度が低くなった。これは 2.5 節で述べた成分間の干渉に起因すると考えられ, この結果は成分間干渉の影響を抑制するために時間軸変換処理が有効であることを示唆する。また IF-attractor 法は, 時間軸変換後のパラメータ推定において瞬時振幅の非定常性を考慮していないが, その推定精度は提案手法より約 10 dB 低かった。この結果は, 正弦波成分の瞬時振幅の非定常性を明示的にモデル化することで, パラメータ推定精度が向上することを意味する。

Table 1 に本実験で用いた全 900 個の信号に対する, 各アルゴリズムの S/N の平均と標準偏差を示す。アルゴリズムによらず, 男性より女性の音声の方が正弦波パラメータの推定精度がやや高い傾向が見られた。この結果は, 男性と比較して女性話者の基本周波数が低いため, 単位周波数あたりに含まれる調波成分の数が少ないことに起因すると考えられる。全データに対する平均 S/N は, Peak-picking 法で 14.4 dB, IF-aatractor 法で 23.4 dB, 提案手法で 28.4 dB であり, 提案手法の推定精度が最も高かった。なお, 各アルゴリズムの推定精度に関するこの順位は, 今回調べた全ての入力音声で不変であった。

	Male (444)	Female (456)	All (900)
Peak-picking [2]	13.8 ± 3.9	15.1 ± 3.7	14.4 ± 3.8
IF-attractor [8]	21.9 ± 3.2	24.9 ± 3.3	23.4 ± 3.6
Proposed method	27.3 ± 3.7	29.4 ± 3.8	28.4 ± 3.9

Table 1. S/N (dB) for the examined methods.

4. まとめ

非定常部を含む有声音声信号を高い精度で分析するための手法として, LVT により推定した第 1 調波成分の瞬時位相に基づいて入力信号の時間軸を変換する正弦波モデルを提案した。成人男女 75 名が発話した 900 個の単語音声に対して, この提案手法は既存の手法より高い分析精度を示し, 入力信号と残差信号のエネルギー比は平均 28.4 dB であった。この手法は, 音声知覚や生成など幅広い研究に信頼性の高いデータを供給できると期待される。

謝辞

本研究は文部科学省科学研究費補助金(21700282, 17075003)の助成を受けた。関係各位に深く感謝致します。

参考文献

[1] McAulay, R. J., and Quatieri, T. F. "Speech analysis/synthesis based on a sinusoidal representation," IEEE Trans. Acoust. Speech Signal Processing 34, 744–754, 1986.
[2] Quatieri, T. F., and McAulay, R. J. "Speech transformations based on a sinusoidal representation," IEEE Trans. Acoust. Speech Signal Processing 34, 1449–1464, 1986.
[3] Quatieri, T. F., and McAulay, R. J. "Shape invariant time-scale and pitch modification of speech," IEEE Trans. Signal Processing 40, 497–510, 1992.
[4] Marques, J. S., and Almeida, L. B. "Frequency-varying sinusoidal modeling of speech," IEEE Trans. Acoust. Speech Signal Processing 35, 763–765, 1989.
[5] George, E. B., and Smith, M. J. T. "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," IEEE Trans. Speech Audio Processing 5. 389–406, 1997.
[6] Robel, A. "Adaptive additive modeling with continuous parameter trajectories," IEEE Trans. Audio Speech Language Processing 14, 1440–1453, 2006.
[7] Fulop, S. A., and Fitz, K. "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," J. Acoust. Soc. Am. 119, 360–371, 2006.
[8] Abe, T., and Honda, M. "Sinusoidal model based on instantaneous frequency attractors," IEEE Trans. Audio Speech Language Processing 14, 1292–1300, 2006.
[9] Ito, M., and Yano, M. "Sinusoidal modeling for nonstationary speech based on a local vector transform," J. Acoust. Soc. Am. 121, 1717–1727, 2007.
[10] Kuwabara, H., Sagisaka, Y., Takeda, K., and Abe, M. "Construction of ATR Japanese speech database as a research tool," ATR Technical Report, TR-I-0086, 1989.