有声発音を伴わない発音による会話音声の創出入力システム

Voice Regeneration System to convert Whispering Voice to Pseudo-Real Voice

竹　内　康　人
Yasuhito TAKEUCHI

鹿児島大学工学部情報工学科
（〒890-0065 鹿児島市群元 1-21-40）
e-mail: ytake@ics.kagoshima-u.ac.jp

Abstract: A software based prototype of voice regeneration system to convert whispering voice talk to pseudo-real voice talk is developed. Its intention is to allow a person to talk to the other party via phone channel just by whispering, i.e. without vibrating his or her vocal fold to generate real voice. The kernel of the signal processing system is either FFT based or correlation based vocoder to regenerate a pitched-voiced signal from whispered-unvoiced signal spectrum or auto-correlation, at its tonal part of the sequence. The system, if implemented in mobile or other style of phone, may solve major reason to prohibit its use in closed public space (such as railway wagon) due to its "acoustical nuissance" of talking voice.

（1）開発の主旨

（1—1）目的意識

本件は発声を必要としない音声通話装置、特に話者の発するべき送信音声信号を話者に具体的な有声発声行為をせしめずに推定、合成し、送話の用に供する装置システムのプロトタイプ開発に関する。

携帯電話器の使用が旅客列車の客車の中などの“逃げ出す事の出来ない”閉鎖空間において嫌われ、または管理者により禁止される理由の一つに、周囲の者に通話者の発する声が“うるさい”もしくは“奇異な印象を与える”という事が挙げられる。また携帯電話に限らず有線電話であっても、会議中に出席者が静寂を守らなければならないにもかかわらず別な所と通話せねばならなくなった時などに、通話者の有声発声音声が邪魔になるという事態は多々あり得る。

そこで本件開発は、話者が故意に声帯を振動させないで語り掛ける行為を観測し、その観測情報に基づいて話者が通用通り声帯を振動させていたら得られた筈の音声信号を近似的に合成し、これを電気的手段による通話の用に供するために通信回線を介して送出する装置を実現せんとするものである。

この装置は、話者の口に近接して置かれるマイクロフォンを有し、このマイクロフォンにより話者の囁き声の音響信号を採取し、かくして採取された音響信号の分析結果から該話者が通常通り声帯を振動させていたら得られた筈であろう所の音声信号を近似的に合成し、このようにして合成された音声信号を電気的手段による通話の用に供するために通信回線を介して送出する。

またこの装置は、近接マイクロフォンの捕える囁き声の信号を分析して第一および第二のフォルマントなどを推定し、この推定された各フォルマントの情報を用いて送出用合成音声信号を合成する如く構成された装置である事が期待される。

しかしまたこの装置は、近接マイクロフォンの捕える囁き声の信号を分析して声帯の緊張度を推定し、この推定された声帯緊張度に関する情報を用いて送出用合成音声信号のピッチを決定する如く構成された装置である事が期待される。

しかしながらこの装置に関し最も期待される特徴は、近接マイクロフォンの捕える囁き声の信号を分析しはするものの、その結果をあえて認識や分類に付すことなく即その場で加工再利用する形でもって、受け入れた囁き声の音響信号を比較的簡素で単純な処理アルゴリズムにより通常の声に近似できる音響信号に変換する装置であることが期待される。

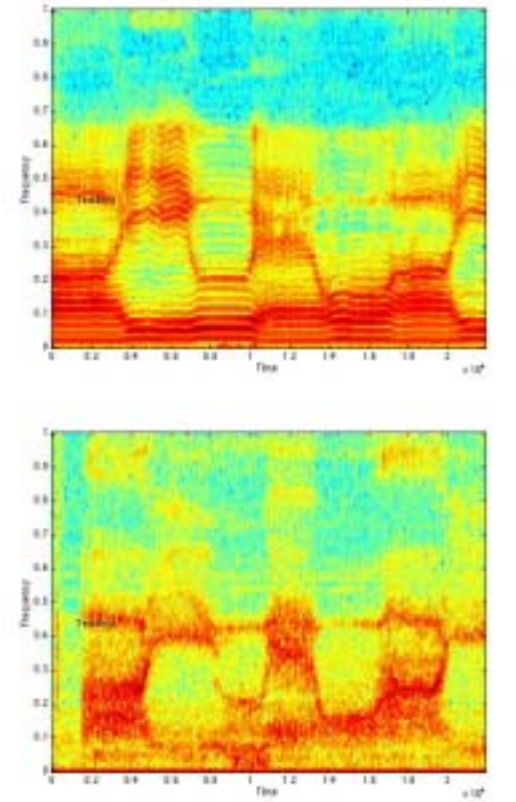
この装置を規定するこのような技術概念は、既に開発者により特許申請され、またそれは既に公開されている（1）。

（1—2）先行技術と一般的知見

囁き声の音響学的研究は、その主観評価も含め、歴史的にさほど新しいものではない。例えば古くも 1957 年に萌芽的研究が見られる（2）。また最近では囁き声の音響構造の詳細な評価（3）、また囁き声そのものを音声認識にかける努力（4）等がいくつかの研究陣によりなされている。本件の課題は一種のいわゆる声質変換（5）の特殊な例と捉える事もできるであろう。しかしながらこのような装置への社会的要求はくだんの携帯電話の問題のみならず、例えば医原性声帯喪失者、呼吸困難患者などへの支援策としても有益であり得るものである。

ここで一般的知見として、ある語句をちゃんとしゃべった場合の音声のスペクトログラムと、同じ語句を

囁き声でしゃべった場合のそれを、比較提示しておく。



上：あいうえおあい（有声）、
下：アイウエオアイ（囁き声）

ともに有声音においては基本周波数の数倍までの低域のエネルギーが高く、またピッチ周期の縞模様が顕著に見られる。囁き声はこれらを除く高域が主体であるが、本件開発は要するに右の信号を受け入れて左の信号に遅滞なく変換して出力する装置の事である。

（2）成果

（2―1）FFT-iFFT 型ボコーダ- の思想と実現

囁き声に関しては、これは通常の発声に特徴的な声帯（声門）のインパルス駆動入力がなくなり、その代りに声道周辺の構造物に呼気があたって生ずる白色雑音様の信号が系の駆動原として効果している物である、と単純化して考える事ができる。もちろん厳密には声帯を周期的に閉じている時と、あけっ放しにしてただ呼気が通過するに任せている時とでは、この系の共振器としての特性は僅かながら有意に異なる。が、実用的見地から大雑把に言って、声道の可変寸法音響管共振器としての特性は母音のどれを発声しているかのみに応じて決り、駆動入力がパルスか白色雑音かには由らないと考えて良い。しからば囁き声を周波数分析して得られたスペクトラム像をそのフィルタパラメーターの代用として用いて、これに入力としてそれらしき声帯の駆動パルスを与えて再合成してやればこれを有発声音声に変換できると期待される。さらに別な研究家からは、声道の可変寸法音響管共振器としての特性の観測には呼気さえも不要で、ただ口元から極小寸法のスピーカーとマイクロホンを用いて音響学的に能動計測すれば足りる、旨の主張がなされている。が、本件開発においては、そこ迄は深入りせず、呼気を以て駆動される囁き声の発声過程を受動的なマイクロホンにより観測する事から出発する。故にこのモデルは以下のごとく要約される。

(1)　囁き声音声信号を通常のマイクロホンで受けて処理すべき音声信号を得る
(2)　A/D 変換すると同時に前後で適当なフィルタ、自動レベル調整などを行う。
(3)　処理の単位のフレームに区切る。例えば 20~30mSec 程度。オーバーラップも可。
(4)　フレーム毎に FFT する。以下周波数ドメインで処理を行う。
(5)　信号レベルなどから発声区間となるべき区間をフレーム単位で同定する。
(6)　上記発声区間　　フレームについてはピッチ周波数相当のビン群に修飾加工を行う。ここで修飾加工とはビンの数値（複素数）の間引き、追加、消去、反転など。
(7)　上記発声区間　　フレーム以外のフレームについては、修飾は行わない。
(8)　上記(6),(7)で得られた加工済みの FFT データを iFFT に付しつつまたもとの時間軸になる様に繋ぎ、また並べ直しをする。
(9)　事後処理フィルタで聴感を改良しつつ D/A 変換して変換音声信号を得る。

この工程の特徴とする所は、修飾加工されないフレームのデータはそのまま元の信号と同じ物に戻る点で、子音やアタックの部分の自然性は比較的良く保たれるという点である。が、欠点は、音声の機動性を維持するためにはフレーム長は数十 mSec 程度以上には出来ない事と、周波数ドメインでピッチパルスを埋め込む手続きがフレーム長の整数分の一を単位としてしか出来ない事から、発生させ得るピッチ（周波数）の候補が非常に限られる事で、ピッチの制御情報は別途創出ないし入手するとしても、結果として自由な抑揚を埋め込む事が出来ない事である。固定ピッチでの復元信号は丸でお経を読み上げている様に聴こえ、また声の質としても何やら金属的なキンキン声になり、決して好ましい印象とは言い難い。しかしこのようにして復元ないし代用された疑似音声信号は、処理の緒元パラメーターが適切に設定されさえすれば、不自然さはあっても耳にとってはかなり理解力がある信号となり、この限りにおいては“認識分類行程を経由しない”と言う本件開発の主旨には大変良く適っていると言える。

（2―2）自己相関型ボコーダーの思想と実現

音声の今一つの重要な性質として、聴覚は一種のスペアナであって、信号の位相には感じない、とされる

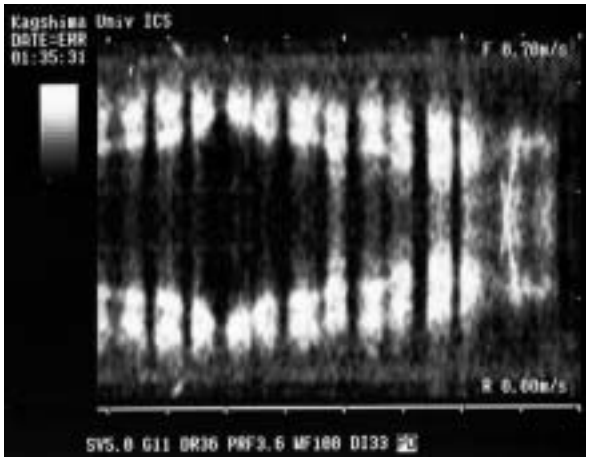
点である。この事に関連して、音声信号それ自身に代えてその部分ないし偏自己相関を繰り返し再生しても同じ様に聞こえる事が知られ、この点は Parcor 系の音声圧伸技術に大幅に取り入れられている。囁き声の有声復元に関してもこの思想が有益かと思われるため、先ず入力された囁き声音声信号を短時間形式の実時間自己相関に付し、その時系列変化を保ったまま繰り返し再生に付す事、またその間に必要な所はピッチパルス相当の信号を添加する事を試みた。ここで用いる Fano 型の短時間（実時間）自己相関はあらゆるサンプル点の時刻においてその前後一定の時間枠（窓枠）分の自己相関が得られる物で、その演算手法は後に示す。即ちこのプロセスは以下の如くになる。

(1) (2) データ取り込みおよび前処理に関しては先と同様。
(3) 信号を、今度はセグメント化せずに垂れ流し式に Fano 型の短時間（実時間）自己相関に付す。
(4) 有声区間の検出に関しては先と同様。
(5) ピッチパルスの時刻の情報を別途用意し、該当する時間（サンプルの番号）においてはその都度その時出来ていた相関関数の像を適宜フィルタないし重みづけして出力バッファの該当時刻の位置に足し込む。
(6) この足し込みの時、それが発声区間に該当したら足し込まれる自己相関像の原点近傍を持ち上げて強調する加工を施した上で足し込む。
(7) 発声区間でなければそのまま足し込む。
(8) 全てのピッチパルス時刻に関して上記の足し込みが完了したら、出力バッファの内容は出力すべき変換された音声になっているので、これを事後処理用フィルタで聴感を改良しつつ D/A 変換して目的の変換音声信号を得る。

この行程の特徴とする所は処理の流れが FFT-iFFT 式よりかえって簡素である点であるが、自己（相互）相関の演算は FFT、iFFT と異なり一括高速アルゴリズムが存在せず、演算対象区間内の全てのサンプルにその都度総当たりする、いわゆる八つ当たりプロセスでしか出来ないため、このような自己相関ボコーダ- の処理速度は FFT-iFFT 式ボコーダーより大略２桁遅い。しかし任意のピッチパルス列に従って再生出来る点は大きな特徴であり、これ故に本件開発の主旨に必須の要件を満たしている。

（２―３）ピッチ情報の抽出と織り込み

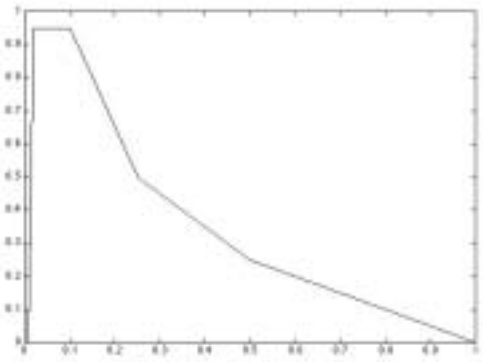
（ａ）　囁き声の母音相当区間にも声の高さすなわち音高（ピッチ、音楽用語）を示唆する情報がある事は、主観的にはほぼ自明、また同主旨の古い研究もある（６）。自験例を紹介すると、下記のごとく、同じ“ あ ”と聞こえる囁き母音にも、特にそのフォルマント F1, F2 あたりに“ つもり ”のピッチ情報がある。確かに高い声を出した積りで囁くと目を吊り上げ耳を引っ張り上げられた様な印象の囁き声になる。これが何に由来するかは完全に明らかではないが、振動はしないものの引っ張り上げられた声帯の所を呼気が通過する場合、引っ張り上げの力の程度に応じて、そのあたりの構造が変化するのか、声帯自体の緊張度が自由振動成分に貢献するのか、何らかの、しかし明らかな効果が認められる。他の母音に関しては確たる情報がないが、押して推察する事は許されよう。これを的確に、個人情報を含めても含めなくてもよいから兎に角も検出できれば“ つもり ”のピッチを復元できる可能性は十分存在する。しかし本研究開発においては与えられた期間と費用ではそこまで到達できなかった。



上記の図の説明：これは超音波診断装置のドプラスペクトラム分析部に囁き声の“ あ ”の音声信号を入力して分析させたもので、複素数入力を必要とする所に実数入力を与えているのでこの表示の中央を境に上下は折り返しの関係にある。横軸（時間軸）全長は約４秒、縦軸（周波数軸）の下端（零）から中央までは約 1.8KHz、低域と高域は分析前にフィルタでカットしてある。図の様に声の高さを変えた積りで囁くと、確かにその情報が発生する囁き声のスペクトラムの形に反映されるのがわかる。これは正当な手法で検出可能であると推察される。

（ｂ）前述の様に、開発された自己相関ボコーダ- には既にピッチパルス列さえ与えられればそれに従って１発１発の声帯パルスへの対応を模擬する同期再生を行なえる機構が実装されている。現在はこれに一定周期のパルスを内部で作って採用するか、またピッチ制御入力を外部から貰って採用するか、の方法でピッチないしイントネーションが創成できる。そこで、与えられた信号からあるポリシーでもって簡便に抽出でき

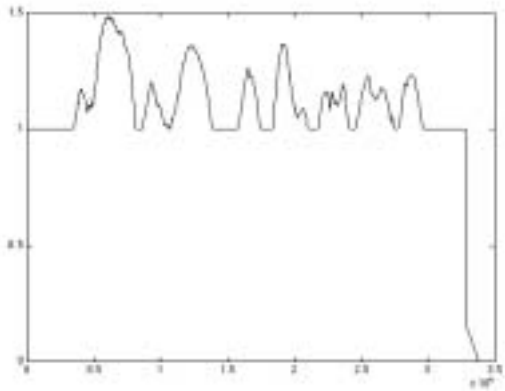
るピッチ類似の観測値（下図）をピッチコントロール信号に採用してみた所、ややおかしな感じの語り口ないし声色（こわいろ）もしくは聴き慣れない方言の様にになったが、一定ピッチの機械的お経より遥かに感触が良い、日本語らしい再生音声を得る事が出来た。ここで如何なるポリシーでこのピッチ制御信号を得たかは、知的所有権上の諸手続が完了した後に、別な機会に発表したい。



姑息代替え手法により得られたピッチコントロール信号の例

（２―４）自然性への配慮

この研究開発のどの段階でも問題化した特徴的な問題は、合成音声がどうしても金属的なキンキン声になってしまう点で、これは原因を追求するより姑息的ながらいわゆる妥協フィルタ(コンプロマイズフィルタ)を事後フィルタとして適用する事でもって改良した。一例を下記に示す。



上記の図の説明：　キンキン声を抑えるためのコンプロマイズフィルタの周波数特性の例。縦軸（振幅応答）は０から１まで、横軸は０からナイキスト周波数（この場合 5.5KHz）まで。

（２―５）実時間動作への考慮

最終的に実用的だろうと思われる本件開発の結果の自己相関ボコーダー型の囁き声変換システムは、実験的構築と試行錯誤の場では Mac 版の Matlab を用い、信号の入出力も Mac 固有のあてがい扶持の物を用いた。が、如何に Matlab が科学技術計算のための超高速インタープリタとして実績、定評のある証明済みの物であるとは言え、この体制で即実時間動作ができる物ではない。しかし処理量を見積ると、少なくともコンパイラレベルにて専用のプログラムを書き起こせば汎用のＰＣでもって実時間動作が十分可能という推察を得る。この場合、必須の処理時間は高々2 ないし数フレーム以内、つまり数十 mSec 以内と試算される。

（３）まとめおよび問題点と今後の展望および謝辞

a. まとめ：有声発音を伴わない発音による会話音声の創出入力システムの開発を行い、オフラインバッチ処理型の可変ピッチ再生自己相関ボコーダ- の形をした Matlab ソフトウエアを試作開発し、実用性の見通しが得られる成果を得た。この成果は汎用 PC を用いた実時間処理プログラムに移植作業中である。

b. 問題点：当初もくろんでいたピッチ情報の囁き声音声からの自動抽出には、目下の所、解析的に正しい手法としては成功していない。代りに姑息手段として信号から一義的に抽出した代替え情報でピッチコントロールを試行し、少なくとも日本語標準語においては実用性を大いに示唆する面白い結果を得ている。一方また老若男女各年齢層、また文化（方言）クラスターの相違、また外国語（外国人）など多様なテストサンプルによる試行と評価も未着手である。

c.今後の展望：上記の問題点に列記した積み残し事項の内、ピッチ情報の抽出の研究開発を最優先の課題としたい。また PC による実時間動作が可能になったら、老若男女外国人などに主旨目的通りの通話を試行をして見てもらって様子を見たい。

d.謝辞：先ず、本件研究開発に制度上のみならず多大な精神的なご支援ご鞭撻をいただきました竹内郁雄教授に謝辞を呈します。次に雑事一切をお引受け下さいました座間誠一、久保真紀子の各位に日常の謝辞を呈します。

（４）参考データ

(1) 試行錯誤時に採用した開発システム：

使用ソフトウエアツール：Matlab version5 Mac 版 (signal processing toolbox つき)

使用ハードウエア：Power Mac G3 および出荷時内蔵のおしきせサウンド入出力ボード、同用純正マイクロホン（多少改造)、汎用ポータブルＭＤ録再装置

信号仕様：入出力サンプリングレート：11.025KHz、
精度：16bit
取り込み語数：32Kword (約３秒)
録音、処理、再生とも手動のバッチ処理、

(2) 自己相関ボコーダーの要点：

音声信号の特徴として、ないしは聴覚の特徴として、（聴覚は位相を感じないという性質に基づき（勿論これには反論が多々あるが））原信号のかわりにその自己相関を聞かせても同じ様に聞こえる（フレーム間の編集の仕方如何ではあるが）という特徴がある。PARCOR系はこれを巧みに用いている。そこで囁き声の自己相関をピッチ周期に合せて繰り返し再生する事でもこの研究の主旨目的が達成される可能性が大である。またピッチパルスの強制挿入は再生に付す自己相関関数の原点周辺のピークを強調してやる事で簡単に実現できる。これが本手法の主旨であり、本研究ではこの方式がFFTボコーダ－より優れているという結論に至った。

処理のあらましは文章で述べた方が分かりやすいので以下に説明する。先ず原信号を自己忘却型の短時間自己相関の処理に付す。これはいわゆるFano型と言われる、のべつまくなし垂れ流し演算方式の実時間自己相関演算で、数式表現を借りると以下の様な形をしている。

$$F(t_0,\tau) = \frac{1}{\alpha}\int_{t=t_0}^{-\infty} \exp\left(-\frac{t}{\alpha}\right) \times f(t) \times f(t-\tau) \times dt$$

F＝相関関数、f＝源信号、t＝実時間の進行、t_0=現在時刻、τ＝時間差軸、α＝減衰時定数要因。

連続信号をサンプリングしたサンプル点列にこの関数の離散表現を適用するには、新しいサンプルが得られる都度、以下の演算を行う。

$$P(n,k) = D(n) \times D(n-k)$$

$$C(n+1,k) = C(n,k) + \frac{1}{2^N}\left\langle P(n,k) - C(n,k)\right\rangle$$

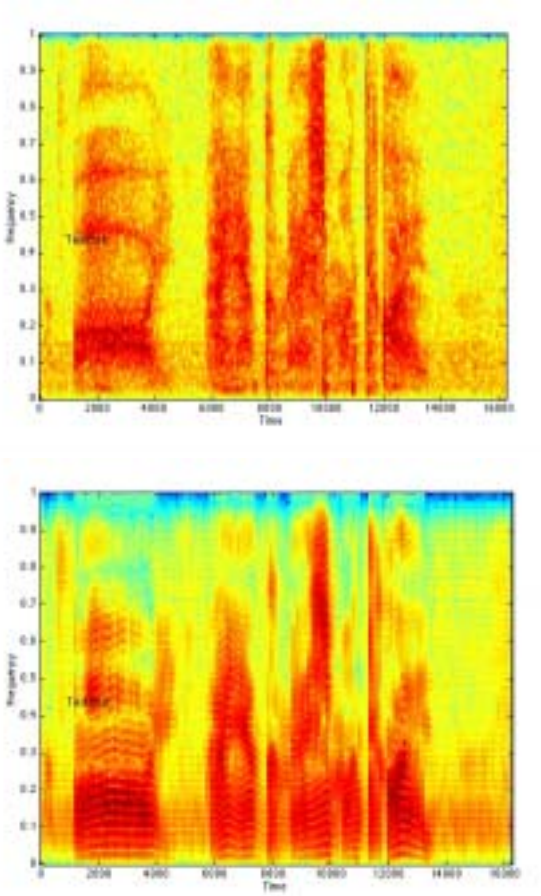
P=積の項、D＝源信号、C＝相関関数、n＝実時間の進行、k＝時間差軸。

ここで（１）式のα、また（２）式のNは現時点から過去に向けての等価的な観測ウインドウ幅を決める所の減衰時定数、ないし俗に自己忘却時定数と言われる設定値を支配するパラメーターである。数式表現の詳細は別の機会に譲るとして、今回は演算区間長 256サンプル、自己忘却時定数64サンプル相当を採用している。が、この辺はとみに主観的ないしカットアンドドライ的に決めるのが正解だろう。

ここで、この様に常時updateされて得られてるFano型短時間自己相関関数は、それ自身を再度セグメントを解いて連続信号に見える様に重みづけ再接続の編集をして再生聴音にかけると、既にあたかも声帯パルスが添加されて有声音になったかのごとく聞こえる。これは自己相関は必ず原点にピークが発生するからで、その原点のピークが編集作業上発生する擬似的な周期をピッチとして“聞こえる”からである。プロトタイプの場合、自己相関を２対１のハニング窓オーバーラッピングで重みづけ再接続の編集をして再生しているので、ベースとなるピッチ周波数は128サンプル相当、約140Hzとなっている。この最接続編集の周期を前述の姑息的代替え手法により作られたピッチコントロール信号により相関フレーム毎に修正して再生音声の抑揚を実現している。

しかし自然発生的な原点ピークを含めて再生するだけでは細声キンキン声の程度は従前の例より更に悪い。これより、有声音らしさを増すために、有声区間となるべき区間において自己相関の原点を太らせて持ち上げる人工的な処理を追加する。有声区間となるべき区間の決定には前回同様囁き声自体のレベルを用いた。即ちこの情報（信号の電力）は自己相関の原点のピークの高さそれ自身（もちろん上記の持ち上げ太らせ処理以前の）にあらわされているので、これが山勘で決めた閾値より大な場合において選択的に上記の持ち上げ太らせ処理を実施する。

(3) 可変ピッチ自己相関ボコーダ－による変換の例（自験例）



上：源囁き声信号のスペクトラム（オーイ、ハヤクメシモッテコイ）
下：本手法により変換されて得られた疑似発声音声のスペクトラム（おーい、はやくめしもってこい）

（５）参考文献

（１）竹内、特開 2000-276190 (特願平 11-124685)
（２）W. Meyer-Eppler, Realization of Prosodic Features in Whispered Speech, J. Acoust.
Soc. Am., 29, pp.104-106. (1957)
（３）松田、粕屋、ささやき声の音響特性と音声合成法、信学技報 SP99-6 (1999)
（４）伊藤、武田、板倉、“ 囁き声の音声認識のための音響分析に関する検討 ”
2001 年日音秋期大会　演題 1-Q-33.
（５）（例えば）
http://www.klab.ee.utsunomiya-u.ac.jp/~takahiro/research.html.
（６）J. B. Thomas, Perceived Pitch of Whispered Vowels, J. Acoust. Soc. Am., 46, pp.468-
470, (1969)