食道音声の韻律及び声質の改善

Improvement of prosody and voice quality of esophageal speech

宇都宮大学工学部
Faculty of Engineering, Utsunomiya University
粕谷 英樹　　森 大毅
Hideki Kasuya　　Hiroki Mori

＜研究協力者＞
宇都宮大学大学院工学研究科
Graduate School of Engineering, Utsunomiya University
秋元 博樹　　藤井 圭
Hiroki Akimoto　　Kei Fujii

In order to improve voice quality of esophageal speech, methods based on analysis-conversion-resynthesis schemes are investigated. Listening experiments of the resynthesized speech by the ARX analysis-synthesis method and the mel-cepstral analysis-synthesis method showed that smoothing of F0 is effective in improving some aspects of the voice quality. In order to cope with esophageal speech of largely unstable fundamental frequencies, on the other hand, a method is also investigated to generate Fo contours automatically.

Keyword esophageal speech, ARX analysis-synthesis, mel-cepstral analysis-synthesis, smoothing

1. 研究の目的

喉頭癌などにより喉頭を摘出すると，無喉頭となり通常の発声 (phonation) 機能を喪失する。それによって喉頭摘出者は二つの問題を抱える。一つは音声の生成過程の第一段階である音源の生成に必要な声帯を失ってしまうこと，もう一つは肺･気管と構音器官が連結しなくなるので，肺からの呼気圧を発声のエネルギー源として使用することができなくなってしまうことである。現在，日本では高齢化社会を迎えて，喉頭癌患者が増加し，それに伴って喉頭摘出者の数も年々増加しており，喉頭に代わる適切な音源の生成（代用発声）の必要性が従来にも増して高まっている。

喉頭摘出者の代用発声法には食道発声法，シャント発声法，電気式人工喉頭，パイプ式人工喉頭を用いるものなどいくつかある。その中でも，食道に取り込んだ空気を逆流させて食道上部の粘膜を振動させて音源にする食道発声法は代用発声法の中でも，我が国ではもっとも広く普及している。ボランティア団体[1]などを通して，食道発声に習熟した喉頭摘出者の指導･訓練体制が確立している。食道発声法の訓練には困難を伴うものの，一端習得すれば補助装置なしで発声が可能であり，ある程度の抑揚や感情表現ができるなどの利点がある。しかし，健常者の音声に比べて，基本周波数が不安定である，抑揚が小さい，嗄声の声質を伴う，明瞭性があまり良くない，音量が不足する，呼吸時の気管孔雑音が混じる，などの問題点もある。

本研究では，基本周波数の不安定さと，それに伴う嗄声の特性を調べ，基本周波数の安定化法につ

いて検討した。方法論としては，食道音声の分析合成方式（ボコーダ）を採用して，食道音声を周波数スペクトル成分と音源成分に分離し，主として音源成分を修正･変換したり，自動的に生成したりすることによって，再合成した食道音声の聴覚印象がどのように改善されるかについて検討した。その際の分析合成法には，ARX 分析合成法[2, 3,4]とメルケプストラム分析合成法[5, 6,7]を用いた。

実用化を考えた場合に配慮しなければならない大きな問題の一つは“実時間性”である。これまでの研究で[8], 40 ms 程度の遅れは許容できるとされているので，40 ms 以内ですべての処理が終了することを前提にしなければならず，したがって，とるべき処理法も限られたものになることに留意する必要がある。

2. 食道発声のしくみと音響的性質

食道発声は食道へ空気を摂取し，その空気を逆流させて食道上部の粘膜（仮声門，新声門）を振動させることにより音源を作る方法である。食道発声の最初の行程である，食道へ空気を自然に摂取する行為が一番難しく，訓練を要する[9]。音源を生成することができれば，構音器官は正常者のそれと変わりないので発音が可能となる。しかし，摂取した空気を貯蔵する部分が食道のため，肺に比べて貯蔵量が少なくおよそ 150*ml*程度といわれている。したがって，呼気時間は数秒である。

一般に食道音声の音響的性質は，正常者との音声の生成過程の違いから推察できるように，次のようなものである。

- F0 が不安定
- 抑揚が小さい
- 嗄声を伴う
- 音量が小さい
- 気管孔からの雑音が多い
- 無声子音の発声が弱い
- 呼気段落が短い
- 明瞭性が低い

図 1 は男性話者の食道音声の波形と F0 軌跡を示した図である。発声内容は「パパもママもみんなで豆まきをした」である。図中の F0 軌跡には，話者の意図とは反した不安定な部分も見られる。不安定な F0 に影響されて，音声スペクトルや強度などの他のパラメータも同様に正常者のそれと比べ不安定である。このような部分を修正し再合成することにより，食道音声の嗄声の印象を改善するとともに，明瞭性を改善したい。

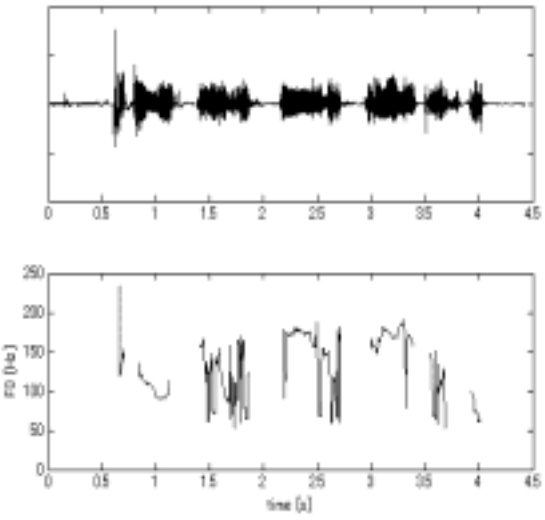

図 1 食道音声波形と F0 軌跡の例

3. 理想的な音源パラメータ変換の効果

ARX 分析により得られるパラメータには，フォルマント，有声音源パルスの振幅(a_v)，雑音源振幅(a_n)，有声音源パルスの声門開放率(q_o)，スペクトル傾斜補償パラメータ(g_v)，有声音源スペクトル傾斜調整パラメータ(g_s)，非周期的成分の境界周波数(f_a)がある[2,3,4]。

ここで検討するパラメータは F0 と a_v であり，それぞれについてスムージングを行った。スムージングの方法は，メディアン平滑化と線形平滑化を一組とし，それを 2 段に組み合わせものである[10]。なお F0 の

自動抽出が行えないほど不安定な部分については手動で修正を加えた。前に述べた実時間性を無視して，全体の分析結果を参照しながら手動によって変換を行う，という意味で，「理想的な変換」という言い方をしている。図2にF0の修正の様子を示す。図3はa_vのスムージングの様子である。図4はARX分析合成の流れを示したものである。この改善方法の有効性を確かめるために，ARXに基づく音声合成法を用いて再合成音声を作成し聴取実験を行った。

音声資料は習熟した食道発声者，男性 1 名の音声で，サンプリング周波数 10kHz，量子化精度 16bit で，発声内容は図 1 で示した「パパもママもみんなで豆まきをした」である。被験者は食道音声にあまりなじみのない健常者 7 名である。スピーカから原音声の後1秒おいて変換音声を聞かせ，“聞き易さ”を比較させた。原音声と変換音声の比較評価には表1のようにスコアを定めた。変換音声の詳細を以下に示す。

A) 原音声の無音区間の雑音を除去した音声
B) 分析合成音声（スムージングなし）
C) 変換音声（F0修正，a_vスムージング）
D) 変換音声（F0修正，a_v一定値）

図5に評価結果を示す。評価値は被験者7名の平均値である。

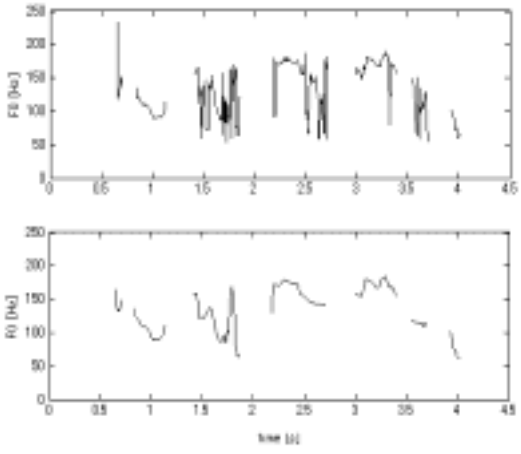

図2 食道音声のF_0軌跡のスムージングの様子

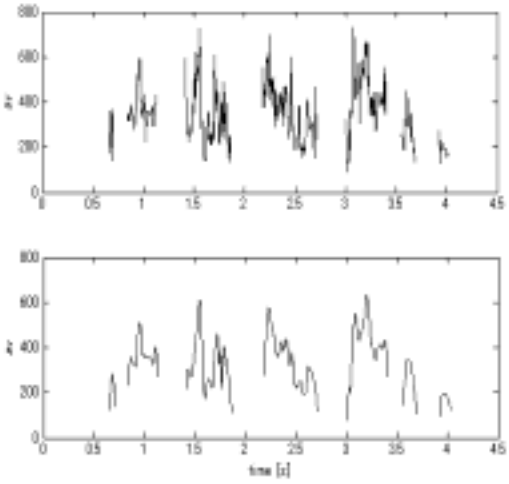

図3 食道音声のa_vのスムージングの様子

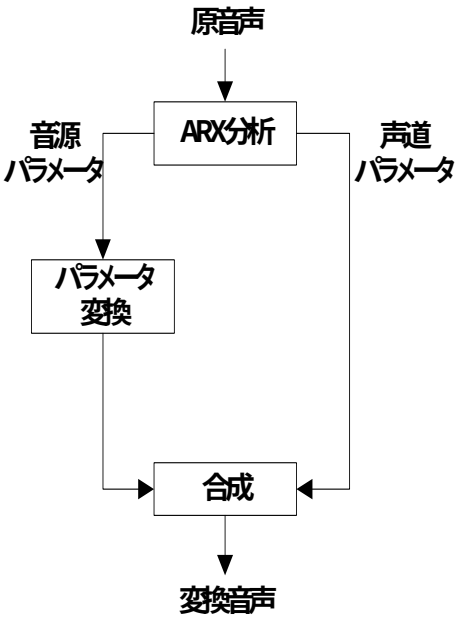

図4 ARX分析合成の流れ

表1 評価スコア

2	非常に良い
1	良い
0	どちらでもない
−1	悪い
−2	非常に悪い

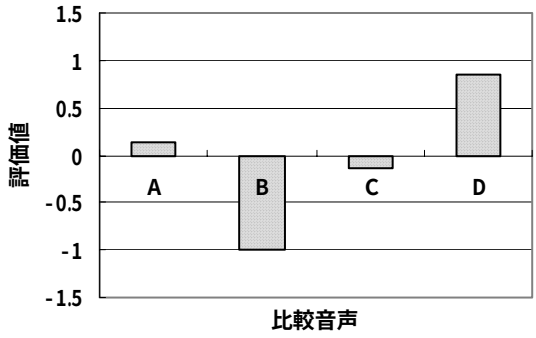

図5 原音声と変換音声の比較実験結果

聴取実験を行った結果, スムージングなしの分析合成音声Bでは原音声より評価が下がったが, F0を修正することによって, 変換音声 C では分析合成音声Bに比べて改善が得られた。さらに a_v を一定にすることによって, 変換音声 D では自然音声の強度の乱れを改善することができ, 評価があがった。変換音声Dは他の変換音声に比べ有意に差がみられF0, a_v は声質の改善に効果があることが確認された。

フォルマントパラメータに手を加えずに, ある程度の声質の改善が得られたことにより, 音源パラメータの修正・変換による改善方法が有効であることが示された。

4. 実時間条件下でのパラメータ変換

上記の実験により F0 のスムージングが声質改善に効果があることがわかった。将来的には実時間(約 30～40ms の時間遅れを目指す)で動作する声質改善補助装置の開発を目指している。そこで, 実時間処理で行える範囲のスムージングを施し, その再合成音声を作成した。5 点メディアン平滑化を施したF0を用いて, ARX 分析合成法およびメルケプストラム分析合成法による再合成音声を作成した。ARX 分析におけるパラメータである有声音源パルスの振幅(a_v), メルケプストラム分析におけるメルケプストラム係数も平滑化を施してある。図6と図7に男性話者と女性話者の音声波形と平滑化前後の F0 軌跡を示す。またメルケプストラム分析合成の流れを図8に示す。分析条件は表2の通りである。

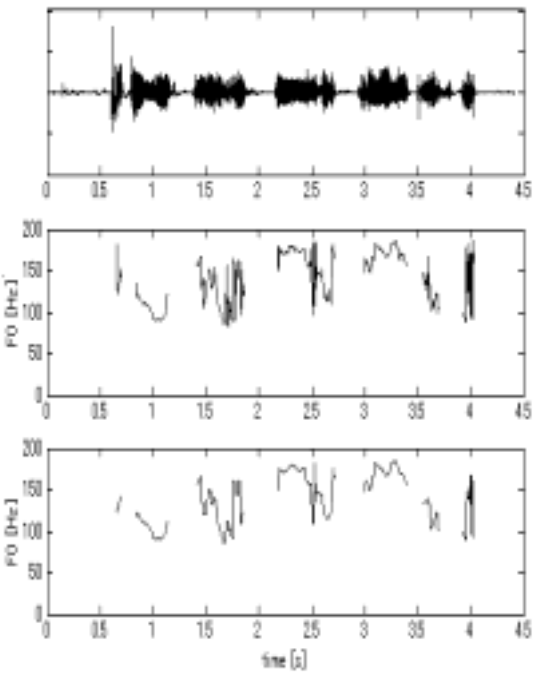

図6 男性話者の音声波形(上)と F0 軌跡
スムージング前(中), 後(下)

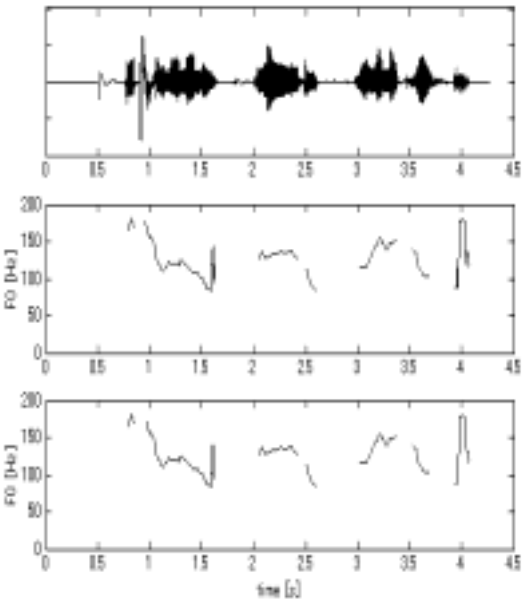

図7 女性話者の音声波形(上)と F0 軌跡
スムージング前(中), 後(下)

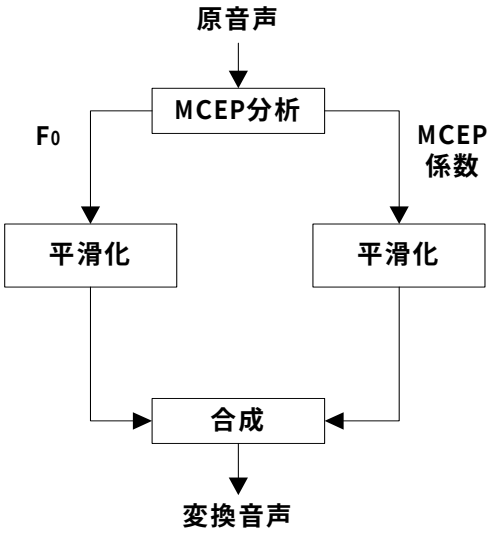

図 8 MCEP 分析合成の流れ

表 2 分析条件

	ARX 法	MCEP 法
分析次数	12	20
分析フレーム長	35 ms	35 ms
フレームシフト長	5 ms	5 ms

原音声と変換音声を比較する聴取実験を行った。音声資料は習熟した食道発声者，男性，女性各1名の音声で，サンプリング周波数 10kHz，量子化精度 16bit，発声内容は「パパもママもみんなで豆まきをした」である。聴取者はあまり食道音声になじみのない健常者 5 名である。スピーカから原音声の後 1 秒おいて変換音声を聞かせた。ARX 分析合成法とメルケプストラム分析合成法による変換音声についてそれぞれ 6 回繰り返し，以下の 6 項目について評価を行った。

- 明瞭性
- 自然性
- 安定性
- こもる感じ
- ざらざら感
- 総合的評価

それぞれの項目について，表1の5段階評価を行った。

男性話者の原音声と変換音声の比較結果を図 9 および表3に，女性話者の原音声と変換音声の比較結果を図 10 および表 4 に示す。評価結果は 5 人の聴取者の平均値である。チャートは 0 を基準にして外側にいくほど声質改善されたことを示している。

図 9 の男性話者の原音声と変換音声の比較結果では，ARX 分析合成音声は明瞭性，自然性，こもる感じで原音声より評価が下回ったが，

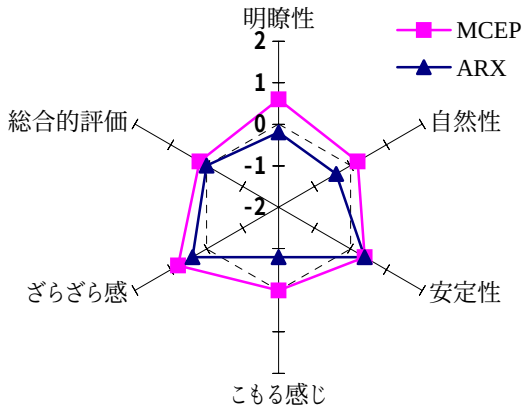

図9. ARXとMCEPの比較結果（男性）

表 3 ARX と MCEP の比較結果（男性）

	ARX	MCEP
明瞭性	**-0.2**	**0.6**
自然性	**-0.4**	**0.2**
安定性	**0.4**	**0.4**
こもる感じ	**-0.8**	**0**
ざらざら感	**0.4**	**0.8**
総合的評価	**0**	**0.2**

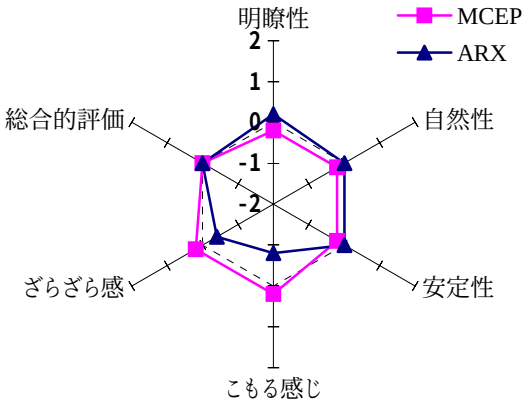

図10. ARXとMCEPの比較結果（女性）

表4 ARX と MCEP の比較結果（女性）

	ARX	MCEP
明瞭性	**0.2**	**-0.2**
自然性	**0**	**-0.2**
安定性	**0**	**-0.2**
こもる感じ	**-0.8**	**0.2**
ざらざら感	**-0.4**	**0.2**
総合的評価	**0**	**0**

メルケプストラム分析合成音声はほぼ全ての項目において基準値の 0 を上回っている。

図 10 の女性話者の原音声と変換音声の比較結果からは，ARX 分析合成音声の明瞭性，自然性，安定性は原音声と変化がないか，または若干改善されており，こもる感じ，ざらざら感では原音声より評価が下回っている。一方，メルケプストラム分析合成音声は ARX 分析合成音声とは対照的に明瞭性，自然性，安定性においては原音声より下回っているが，こもる感じ，ざらざら感については改善されていると評価された。

ARX 法の評価が低い理由として以下のようなことが考えられる。ARX 分析法は，分析窓内に推定された F0 に従って音源パルスを配し，音源と ARX モデルのパラメータの反復推定を行っている。すなわち F0 抽出がうまくいかなければ，他のパラメータの推定精度にも大きく影響を及ぼしてしまう。

図 6 と図 7 からわかるように，男性話者の F0 は女性話者のそれに比べ大きな不規則的変化がみられる。ARX 分析法はパラメータ推定が F0 に強く依存するため，その影響が評価にあらわれたと考えられる。

5. F0 の自動生成による声質改善の検討

F0 のスムージングによる改善は，元の食道音声からある程度安定に F0 が抽出できることを前提としていた。しかし，食道発声者にも習熟レベルによって音声の性質にかなりの違いがあり，安定した F0 が得られない場合が大半である。そのような場合の改善方法として，F0 軌跡を人工的に生成し，スペクトル情報はそのまま利用する方法を検討している。しかしこれには多くの課題がある。それらの一部を列挙すると以下の通りである。

- 有声無声判別の困難
- 個人性の喪失
- アクセントの実現の困難

有声無声の判別は，低周波成分のパワー情報や識別関数の導入によってある程度可能であり[10]，高周波成分を合成音声に結合させることにより，子音の明瞭性と自然性を増すことができる[11]。

音声の個人性は F0 平均や F0 レンジによっても大きく左右されるものであり[12]，なるべく話者の F0 を反映するように生成する必要がある。

もっとも困難な課題はアクセントの実現である。実時間処理を考慮した声質改善補助装置において，つぎの瞬間に発声される言葉のアクセントを予測することは不可能であり，ある規則に基づく F0 軌跡の生成を行うしかない。発声内容とアクセントとの不一致は聞き手に違和感を与える可能性がある。しかし，F0 軌跡の生成によって，アクセントの不一致と引き替えに合成音声の安定性の改善が得られるはずで

ある。そこで，ここでは F0 軌跡の生成による合成音声の作成法について検討する。図 11 にその流れを示した。分析方法は，メルケプストラム分析法を用いている。F0 生成例を図 12 に示す。

このモデルは F0 軌跡の開始端と頂点の値，開始端から頂点までの時間，F0 下降の傾きで決定される。F0 のピーク位置はこの時間のみによって決定され，この例でははおおむね 1 型のアクセント核を実現するように調整されている。図 12 において中段は F0 下降の傾きが小さい場合である。下段は F0 下降の傾きを大きくした場合である。下段は呼気段落が長い場合 F0 が下がりすぎてしまうため，閾値を設けて下がりすぎないようにしている。

F0 軌跡の生成を用いた合成音声は，発声内容とのアクセントの不一致部分では機械的な印象を与える。しかし，合成音声の安定性は改善されている印象を受ける。今後はアクセントの不一致が聴取者に与える影響も考慮して評価していく必要がある[13]。

6. 今後の課題と展望

食道音声の声質改善において，F0 と音源振幅のスムージングが有効であることを示し，ARX 分析合成法ならびにメルケプストラム分析合成法による再合成音声の比較評価を行った。その結果，不安定な F0 においてメルケプストラム分析合成法はロバストな分析法であるといえ，その再合成音声は，明瞭性，自然性，安定性，ざらざら感の項目について改善がみられ有効性が示された。

今後の課題として 3 点を挙げる。第 1 点は，分析・変換・合成技術の一層の改善である[14,15]。そのなかでも，特に F0 がかなり不安定な食道発声者が多いことを考慮すると，元の音声から F0 を抽出し平滑化するというよりは，F0 軌跡の自動生成が現実的である。どのような元音声の情報に基づいてどのような F0 軌跡を生成すべきかについて，一層の検討が必要である。場合によっては，不規則的な F0 軌跡から信頼できるデータのみを取りだし，得られたデータから F0 軌跡を修正する方法や，不安定な F0 の定量的評価法などを検討する必要がある。

第 2 点は，実時間性を考慮したハードウエアの設計と開発である。幸い最近のＤＳＰチップの進歩は著しく，高速・高度な機能を有するものを比較的安価に入手できるようになった。このような技術の進歩を考慮しながら，利用者が大きな経済的負担を負うことなく入手できるような装置を開発する必要がある。

第 3 点は，食道発声者の多様なニーズに応えられるような柔軟な設計指針をまとめることである。これまで行った食道発声者からの意見聴取によれば，苦難を伴う訓練の末に獲得した食道発声によって得られる食道音声には強い愛着があり，別人のような声になって欲しくないという意見が強くある一方で，なんとかざらざら感のない音声にして欲しいという要求や，電話などで不便さを感じている女性からは，女性らしくピッチを上げてもらいたいという要求もある。このような多様なニーズに応えられるように，現状の技術レベルで解決が可能なこととそうでないことを明確にした上で，コストも念頭に入れながら，利用者に納得できる設計指針をまとめる必要がある。

このような課題を一つ一つ解決することによって，目的を達成できるものと確信する。

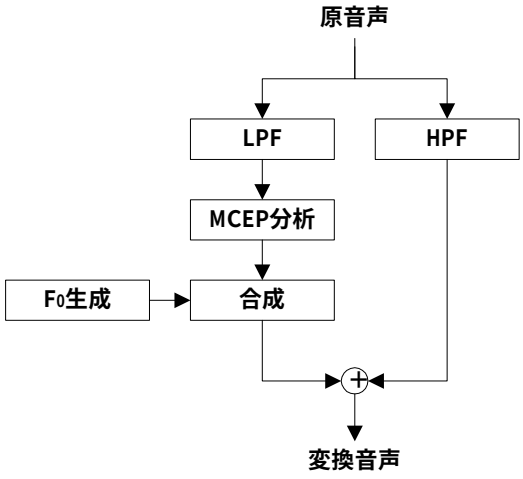

図11　F0 自動生成による合成音声作成の流れ

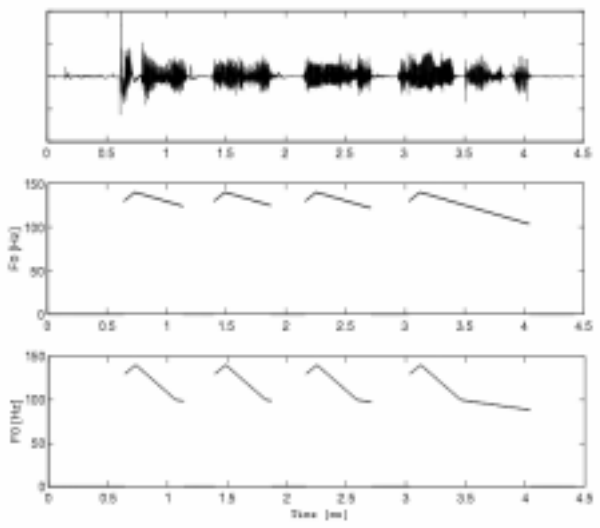

図 12 自動生成した F 軌跡の例

快く食道音声資料を提供された松井謙二氏(松下電器産業(株))に厚くお礼申し上げる。

文 献

[1] 例えば銀鈴会, http://ginreikai.or.jp/

[2] 大塚貴弘, 粕谷英樹, “音源パルス列を考慮した頑健な ARX 音声分析法,”日本音響学会誌, Vol. 58, No. 7, pp. 386-397, 2002.

[3] T. Ohtsuka and H. Kasuya, “Aperiodicity control in ARX- based speech analysis-synthesis method,” in *Proc. Eurospeech* 2001, Vol. 3, pp. 2267-2270, 2001.

[4] T. Ohtsuka and H. Kasuya, “An improved speech analysis-synthesis algorithm based on the autoregressive with exogeneous input speech production model,” in *Proc. ICSLP* 2000, Vol. Ⅱ, pp. 787-790, 2000.

[5] 小林隆夫, “音声のケプストラム分析, メルケプストラム分析,” 信学技報, SP98-56, pp. 33-40, 1998.

[6] 徳田恵一, 小林隆夫, 深田俊明, 斎藤博徳, 今井聖, “メルケプストラムをパラメータとする音声のスペクトル推定,” 信学論(A), Vol. J74-A, No. 8, pp. 1240-1248, 1991.

[7] 今井聖, 住田一男, 古市千枝子, “音声合成のためのメル対数スペクトル近似(MLSA)フィルタ,” 信学論(A), Vol. J66-A, No. 2, pp. 122-129, 1983.

[8] K. Matsui, N. Hara, N. Kobayashi and H. Hirose, “Enhancement of esophageal speech using formant synthesis,” *Acoustical Science and Technology*, Vol. 23, No. 2, pp. 69-76, 2002.

[9] 佐藤武男, 食道発声法, 金原出版株式会社, 2001.

[10] L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, “Applications of a Nonlinear Smoothing Algorithm to Speech Processing,” *IEEE Trans. Acoust., Speech, and Signal Proc.*, Vol. ASSP-23, No. 6, pp. 552-557, 1975.

[11] 原紀代, 松井謙二, ”識別関数を用いた食道発声音声の有声/無声/鼻音判定法,” 日本音響学会春季講演論文集, 2-P-21, pp. 335-336, 1998.

[12] 箕輪有希子, 後藤和貴, 粕谷英樹, “個人性知覚に影響を及ぼす音響パラメータの既知/未知話者における差異,” 日本音響学会秋季講演論文集, 3-3-7, pp. 261-262, 1999.

[13] 笠井友美, 森大毅, 粕谷英樹, “発話速度の違いがアクセント知覚に与える影響,” 日本音響学会秋季講演論文集, 2-1-15, pp. 203-204, 2000.

[14] Y. Qi, B. Weinberg, and N. Bi, “Enhancement of female esophageal and tracheoesophageal speech,” *J. Acoust. Soc. Am.*, Vol. 98, No. 5, pp. 2461-2465, 1995.

[15] Y. Qi, “Replacing tracheoesophageal voicing sources using LPC synthesis,” *J. Acoust. Soc. Am.*, Vol. 88, No. 3, pp. 1228-1235, 1990.