社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－SLP－74（40）
2008 / 12 / 10

音声の高速再生のための話速推定法と高速発話時の特性解析

話速バリエーション型データベースの活用例

吉原　亨† 蔦木 圭悟† 髙橋 弘太†

† 電気通信大学情報通信工学科 〒182-8585 東京都調布市調布ヶ丘 1-5-1
E-mail: †{yosihara,tsutaki,kota}@ice.uec.ac.jp

あらまし できるだけ聞き落としを生じることなく，時間的な効率を上げて音声再生を行うためには，個々の音声に対して最適な再生速度を決定するための指標が必要となる．本稿では，この問題に関して得られた２つの成果について発表する．第一の成果は，２つの異なる狭帯域エネルギの時間変化に着目した話速推定法を提案し，正規化した誤差で16%の推定精度で話速推定が行えることを示したことである．第二の成果は，通常発話の音声と高速発話の音声を，それぞれ極めて速い話速に話速変換した結果を観察し，高速発話を話速変換した音声が，より調波構造の乱れが少ないことを示したことである．本稿は，我々で製作している話速バリエーション型音声データベース (SRV-DB) を用いて行った．我々は本研究をSRV-DBの有効な利用法の一例として発表する．
キーワード 話速推定，話速変換，音声データベース

Speaking rate estimation and utterance analysis of fast speech for high-speed reproduction

A practical example of speech database with speaking rate variations

Toru YOSHIHARA†, Keigo TSUTAKI†, and Kota TAKAHASHI†

† Department of Infomation and Communication Engineering, The University of Electro-Communications
Chofugaoka 1-5-1, Chofu-shi, Tokyo, 182-8585 Japan
E-mail: †{yosihara,tsutaki,kota}@ice.uec.ac.jp

Abstract A quantitative index is needed to estimate the optimum reproduction speed for high speed reproduction of various voices without missing voices. In this paper, we report two results concerning this problem. As the first result, we propose a method for estimating speaking rate and show that we can estimate the speaking rate with 16 % of root mean squared error using proposed method. As the second result, we show that the higher conversion rate becomes, the larger the inconsistency of the harmonic structure of a voice become. In this study, we have used a newly constructed speech data base called SRV-DB. We intend to announce this paper as a practical example of the SRV-DB.
Key words estimate speaking rate, speaking rate conversion, speech data base

1. はじめに

近年，ＨＤＤレコーダの普及やインターネットを通じた動画配信の増加により，人々が見ることのできる視聴覚コンテンツの量は爆発的に増えた．そのため，コンテンツを短時間で効率よく視聴する技術に注目が集まっている．

最も簡単に時間的な効率を上げるには再生速度を上げて視聴すればよい．理想的なコンテンツの高速再生とは，視聴内容の欠落を最小限に留めつつ，視聴時間の削減を最大とする再生である．こうした再生を実現するためには，様々な音声データに対して，どの再生速度が最適かを導き出す指標が必要となる．しかし，これまで高速再生時における再生速度の最適性について深く議論した研究は存在しなかった．そこで我々は，人間が聞き落としを生じる限界の再生速度を定式化することで，再生速度の最適性について評価しようとしてきた [1], [2].

聞き取り可否の境界条件を求めるためには，話速推定の実現

が必須である．正確な話速の推定を行うことで，時間的な効率を高めた高速再生が可能となる．だが，正確な話速が判るだけでは真の意味での最適な再生速度を求めることはできない．より精密な結果を得るためには，早口の音声や高速再生用に話速変換された音声の性質についても知見を深めておく必要がある．

話速推定法の定量的な評価や，早口音声の特性について調べる為には，話速が既知である様々な話速の音声が大量に必要となる．しかし，既存の音声データベースの中に我々の要求を満たすものは存在しなかった．そこで本研究室では，プロのアナウンサーに依頼し，話速が厳密に管理された音声の録音を行っている．録音した音声はデータベース化し，話速バリエーション型データベース (SRV-DB) として外部に公開している．本発表をこのデータベースを利用した研究の例としても参考にして頂きたい．

本稿は以下の流れに従って進めていく．まず，第2節で考案する話速推定法について詳しい説明を行う．次に，第3節で話速推定法の精度を求めるために行った実験について述べる．その後，第4節で実験結果の提示と考察を行う．次の第5節では，音声データベース (SRV-DB) を利用し，高速発話時の特性や，より速い話速に話速変換された音声の特性について調べた結果を述べる．最後に第6節で本稿のまとめを行う．

2. 話速推定法

2.1 話速の定義

始めに本研究における話速の定義を行おう．本研究では話速を1秒間に発話されるモーラ数で定義する．また8 [モーラ／秒] の音声を標準話速の音声と呼ぶことにする．

2.2 原　　理

話速推定にはこれまで複数の方法が提案されてきた．Xiaojuan ら [3] は，スペクトル変化，音声パワー，基本周波数の動的特徴量に着目し，これら3つの重み付け線形和から推定話速を求める方法を提案した．Morgan ら [4] は4つの狭帯域エネルギおよび全帯域のエネルギ変化から音節の検出を行い推定話速を求める手法を提案した．また，Wang ら [5] は Morgan らの手法を発展させ，19 の狭帯域エネルギを用いて頑健な話速推定を行う方法を提案した．

こうした中，我々は2つの異なる狭帯域エネルギの時間変化から話速推定を行う手法を提案してきた [2]．この手法ではまず，2つの狭帯域エネルギ変化の俊敏さを，話速と相関のある値として抽出する．その値を2次関数で [モーラ／秒] の話速値へと変換し，最終的な話速値を得る．以後，提案する話速推定法について詳しく説明する．

2.3 実　　装

ここでは，提案する話速推定法の具体的な手法について述べる．提案手法の処理の流れを図2に示す．

はじめに，サンプリング周波数 $F_S = 44.1$[kHz] の入力信号 $x(t)$ に対してフレーム長 W_N，フレームシフト長 W_{Shift} のフレーム分析を行う．各フレームで FFT を実行し，i 番目の帯域を f_{l_i}〜f_{h_i} としたとき，時間軸方向 n 番目のフレームの狭帯域エネルギ $X(n, f_{li}, f_{hi})$ を得る．ここで，f_{l_i}, f_{h_i} はフィルタ番号 i により一意に定まるので，以後は $X(n, f_{l_i}, f_{h_i})$ を $X_i(n)$ と略記することにする．以後，得られた狭帯域エネルギ $X_i(n)$ から時間変化の頻度を求める操作を行っていく．

まず，$X_i(n)$ から2種類の異なるデータ $Y_i(n)$, $Y_i'(n)$ を作る．$Y_i(n)$ は式 (1) に示す指数平滑法より求める．

$$Y_i(n) = (1-a)Y_i(n-1) + aX_i(n) \tag{1}$$

ここで，a は指数平滑法の係数であり，時定数 τ との関係は，F_S, W_{Shift} を使って次のように表すことができる．

$$a = 1 - \exp(-\tau \frac{W_{\mathrm{Shift}}}{F_S}) \tag{2}$$

また，$Y_i'(n)$ は $Y_i(n)$ から Q_N サンプル抽出し，昇順に並べ替えた後に，$q : q-1$ に分割する点の値として抽出する．こうすることで，$Y_i(n)$ が時定数の小さな平滑化処理，$Y_i'(n)$ が時定数の大きな平滑化処理となる．2つの曲線のゼロクロスから，$X_i(n)$ の時間変化の頻度を求めていく．

次に，式 (3) を用いて $Y_i(n) - Y_i'(n)$ のゼロクロス点 $p_i(n)$ を検出する．ここで R は $Y_i(n) - Y_i'(n)$ の微細な振動によるゼロクロスの誤検出を防ぐパラメータであり，ゼロクロスの検出

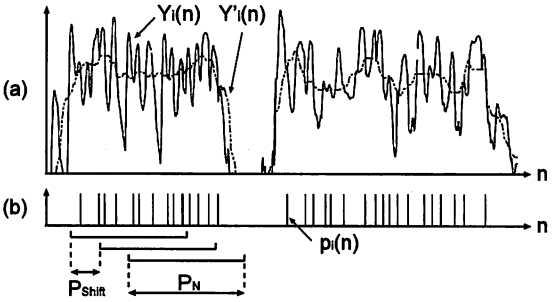


図1　$Y_i(n)$, $Y_i'(n)$, $p_i(n)$, P_N, P_{Shift} の波形の一例．

Fig. 1　Typical examples of the waveforms of $Y_i(n)$, $Y_i'(n)$, $p_i(n)$, and P_{Shift}.

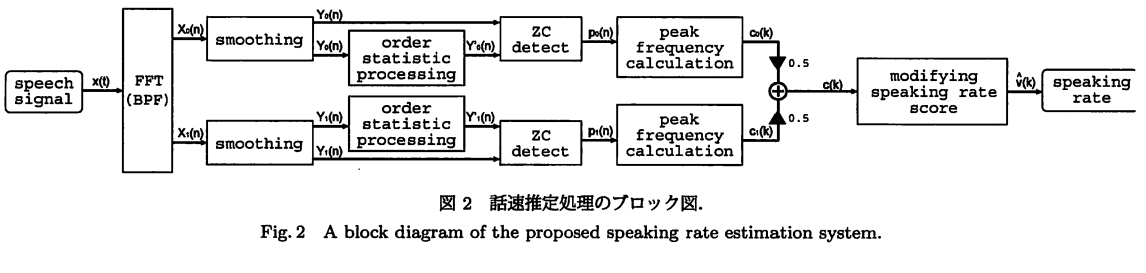


図2　話速推定処理のブロック図．

Fig. 2　A block diagram of the proposed speaking rate estimation system.

後 R サンプルは検出を行わないことを意味する．また，$r_i(n)$ は前回検出したゼロクロスのサンプル番号を意味する．

$$p_i(n) = \begin{cases} 1 & \left(\begin{array}{l} Y_i(n-1) - Y_i'(n-1) \geqq 0 \\ \quad \wedge\ Y_i(n) - Y_i'(n) < 0 \\ \quad \wedge\ n - r_i(n) > R \end{array} \right) \\ 0 & \left(\text{otherwise} \right) \end{cases} \tag{3}$$

次に，検出した $p_i(n)$ に対してフレーム長 P_N，フレームシフト長 P_{shift} のフレーム解析を行い，時間軸方向 k 番目のフレームのゼロクロス頻度 $c_i(k)$ を以下より求める．

$$c_i(k) = \frac{1}{P_N} \sum_{n=0}^{P_N-1} p_i(n) \tag{4}$$

なお，k のサンプリング周波数は $F_S/(W_{\text{Shift}}P_{\text{Shift}})$ と表すことができる．$Y_i(n)$, $Y_i'(n)$, $p_i(n)$, P_N, P_{Shift} の関係を図 1 に示す．(a) に $Y_i(n)$ と $Y_i'(n)$ の波形を，(b) に $p_i(n)$ の波形をプロットした．

その後，式 (5) から帯域ごとに得られる $c_i(k)$ の平均 $c(k)$ を求める．$c(k)$ は $p_i(n)$ を用いて式 (6) のように計算することもできる．

$$c(k) = \frac{1}{I} \sum_{i=0}^{I-1} c_i(k) \tag{5}$$

$$= \frac{1}{P_N I} \sum_{i=0}^{I-1} \sum_{j=0}^{P_N-1} p_i(j) \tag{6}$$

以上のようにして求められる $c(k)$ は，実際の話速値よりも小さな値として推定されることがわかっている．そのため式 (7) を用いて $c(k)$ に補正を行い推定話速 $\hat{v}(k)$ を求める．

$$\hat{v}(k) = f\big(c(k)\big) \tag{7}$$

ここで，$f(x)$ は $f(x) = \alpha x^2 + \beta x + \gamma$ で示される 2 次関数であり，α, β, γ は実験により決定するパラメータである．

以上の処理より入力信号 $x(t)$ から推定話速 $\hat{v}(k)$ を得る．

3. 実　　験

ここでは考案した話速推定法の性能を評価するために行った実験について述べる．

3.1 実験データ

話速推定の性能評価を行う際，実験データをどのように用意するかは非常に重要な問題である．音声の真の話速と推定話速の誤差を求めるためには，音声データを人間が解析し，全ての時区間での話速を手作業でラベリングしなければならない．これは実験を進める上で大きな手間となる．

そこで本研究室では，話速研究用の話速バリエーション型音声データベース (SRV-DB) を新たに製作した [6]．この音声データベースでは，通常の発話で生じてしまう話速の揺らぎを，我々の開発した原稿提示システム (ReCoK5) を用いて厳密に制御している．そのため，一定の話速で文章を読み上げることや，指示通りに話速を変動させながら文章を読み上げることが可能

表 1　実験に使用したデータにおける話者と話速の組み合わせ
Table 1　The combination of seven speakers and seven speaking rates.

	PF00	PF01	PM00	AM00	AM01	AM02	AM03
5.00	√	√	√				
6.73				√	√	√	√
8.00	√	√	√	√	√	√	√
9.51				√	√	√	√
11.00	√	√	√				
11.31				√	√	√	√
13.45				√	√	√	√

表 2　M_j, V_j の値．
Table 2　Values of M_j and V_j.

j	0	1	2	3	4	5	6
M_j	3	5	8	5	3	5	5
V_j	5.00	6.73	8.00	9.51	11.00	11.31	13.45

となっている．

実験に使用した音声データは SRV-DB の ATR25文である．ATR25文では，各文章が一定話速で発話されているため，話速推定の精度を求めるのに適したデータとなっている．話者および話速の組み合わせを表 1 に示す．なお，話者につけられた 2 文字のアルファベットは 1 文字目がプロ (P) か一般人 (A) かを意味し，2 文字目が男性話者 (M) か女性話者 (F) かを意味している．

公開している ATR25文の音声データは，ほとんどが 4~10 秒の短いデータである．そのため，同一話者，同一話速の音声データを結合して 1 つの実験データとした．実験データは合計 29 種類用意した．なお，今回の実験では音声の結合の際，無音区間を挿入する操作は行っていない．もともとデータの冒頭・末尾には無音区間 0.3 秒が存在するので，その合計である 0.6 秒が文章間の無音区間となっている．

ここで，評価式を記述しやすくするために S_N, M_j, V_j の 3 つの記号を定義する．S_N は話速の種類の総数を意味する．M_j は j 番目の話速値の音声データ数を意味し，V_j は j 番目の話速値を意味する．j に対する M_j, V_j の値を表 2 に示す．なお，表 1 からもわかるように，今回の実験では $S_N = 7$ である．

3.2 実験方法

実験は 2 段階に分けて行った．はじめに話速推定に用いる最適なパラメータの推定を行った．次に得られたパラメータを用いて提案手法の推定精度を求めた．

3.2.1 パラメータの探索

本手法で利用するパラメータのうち，表 3 に挙げた 7 つのパラメータについて最適値の探索を行った．

探索を行う際の指標として，以下に示す評価関数 J を用いた．

$$J = \sum_{j=0}^{S_N-1} \sum_{m=0}^{M_j-1} \sum_{k=0}^{K-1} \big(\hat{v}_{j,m}(k) - V_j\big)^2 \tag{8}$$

ここで $\hat{v}_{j,m}$ とは話速 V_j の m 番目の音声データの推定話速を意味する．また，K は 1 つの音声データに対して誤差の計算

表 3 最適値探索を行ったパラメータ.
Table 3 Parameters to be adjusted.

記号	説明
f_{l_0}	第一帯域下側遮断周波数
f_{h_0}	第一帯域上側遮断周波数
f_{l_1}	第二帯域下側遮断周波数
f_{h_1}	第二帯域上側遮断周波数
Q_N	分位点検出区間長
q	分位数
τ	指数平滑法の時定数

表 4 各パラメータの探索範囲と最良値.
Table 4 Ranges of the evaluations and the optimal values for the parameters.

記号	探索値	最良値
f_{l_0}	0, 500, 1000	500 [Hz]
f_{h_0}	1000, 2000, 3000	3000 [Hz]
f_{l_1}	1000, 2000, 3000,4000	3000 [Hz]
f_{h_1}	3000, 4000, 5000	5000 [Hz]
Q_N	0.1, 0.2, 0.4, 0.8, 1.6, 3.2	3.2 [s]
q	$\frac{2}{8}, \frac{3}{8}, \frac{4}{8}, \frac{5}{8}, \frac{6}{8}$	0.75
τ	0.001, 0.004, 0.016, 0.064, 0.256, 1.24	0.001 [s]

を行った回数を意味する．今回は $K = 90$ とした．$c(k)$ から $\hat{v}_{j,m}(k)$ へと補正を行う 2 次関数 $f(x)$ の係数 α, β, γ は，式 (9) の最小 2 乗解より求めた．

$$J' = \sum_{j=0}^{S_N-1} \sum_{m=0}^{M_j-1} \left(\overline{\hat{v}_{j,m}} - V_j\right)^2 \tag{9}$$

式 (8) の最小 2 乗解としなかったのは，$c_{j,m}(k)$ の分散が大きく，推定話速の平均値が真の話速と大きく異なってしまったからである．ここで，$c_{j,m}$ とは話速 V_j の m 番目のピーク頻度を意味する．(8) 式を用いた場合は，速い話速については低めに推定され，遅い話速については速めに推定される現象が起きた．そこで今回は，推定話速の平均値が真の話速に近づくように式 (9) を用いた．なお，$\overline{\hat{v}_{j,m}}$ は話速 V_j の m 番目のデータにおける平均推定話速を意味し，次の式 (10) で定義される．

$$\overline{\hat{v}_{j,m}} = \frac{1}{K} \sum_{k=0}^{K-1} \hat{v}_{j,m}(k) \tag{10}$$

各パラメータに対する探索の範囲および探索の結果得られたパラメータの組を表 4 に示す．なお，今回の探索では周波数の組み合わせのうち，$f_{l_0} = 1000$[Hz]，$f_{h_0} = 1000$[Hz] といった意味のない組み合わせについては排除してある．

3.2.2 推定精度の評価

表 4 に示したパラメータセットを用いて最終的な推定精度の評価を行った．使用した音声データは，パラメータ探索の時と同じ 29 種類のデータである．話速ごとに推定精度を評価するため，最終結果を E_j とし，推定話速との誤差の RMS を正規化した値として求めることにした．E_j は次の式 (11) で求めることができる．

表 5 各話速に対する推定誤差の正規化 RMS.
Table 5 Normalized RMS of error.

話速 [モーラ／秒]	推定時区間 8[s]	推定時区間 16[s]	推定時区間 32[s]
5.00	0.1751	0.1157	0.0774
6.73	0.1612	0.1090	0.0787
8.00	0.1831	0.1463	0.1305
9.51	0.1368	0.1033	0.0836
11.00	0.1787	0.1405	0.1236
11.31	0.1403	0.1089	0.0921
13.45	0.1398	0.1130	0.0988

$$E_j = \frac{1}{V_j} \sqrt{\frac{1}{KM_j} \sum_{m=0}^{M_j-1} \sum_{k=0}^{K-1} \left(\hat{v}_{j,m}(k) - V_j\right)^2} \tag{11}$$

4. 実験結果と考察

4.1 パラメータの決定

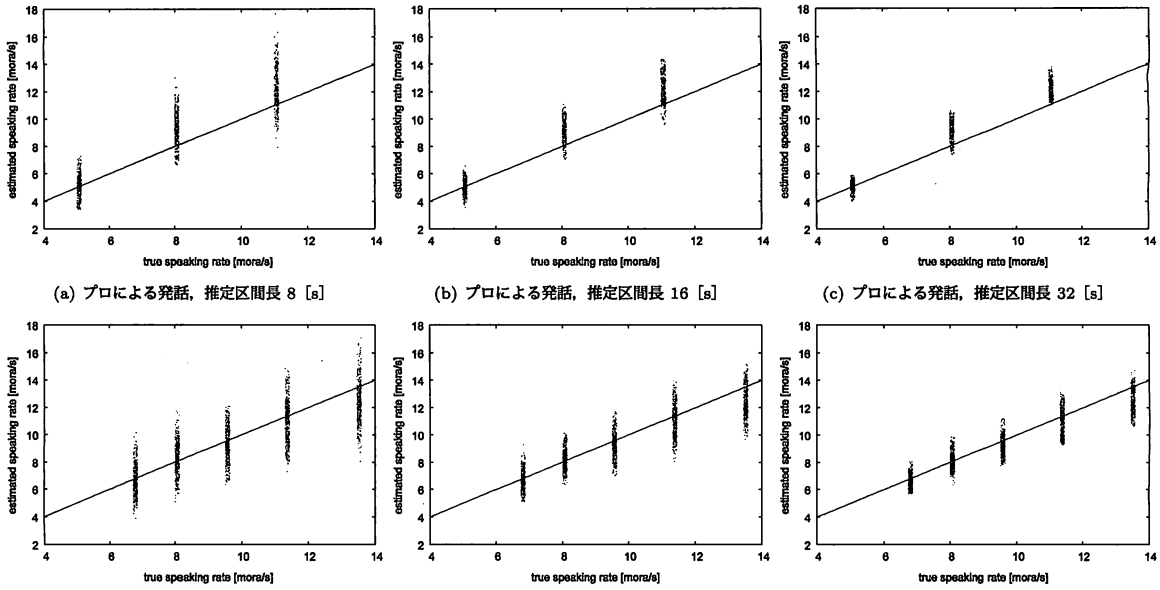
まず，探索の結果からどのように表 4 に挙げた最良値を決定したか述べておく．今回はまず，全探索結果の中で式 (8) の値を最小とするパラメータの組みを最良値として選んだ．その後，このパラメータが妥当であるか判断するため，合計 7 つあるパラメータのうち，6 つを固定して残りの 1 つのパラメータを変化させた様子を観察した．その結果，どの組み合わせでも，このパラメータセットが最適であったため，今回はこの組み合わせを使用した．

4.2 推定精度の評価

表 1 に示した 29 種類のデータに対して，式 (11) を用いて話速ごとに E_j を求めた結果を表 5 に示す．表には推定区間長 P_N を 16 [s]，32 [s] と変化させた場合の結果も掲載した．表 5 より 8 [s] の推定区間長で平均 16%の誤差で話速推定が可能であることが示せた．

次に，推定話速の分布をプロットした結果を図 3 に示す．横軸が真の話速，縦軸が推定話速となっている．図には誤差 0 を示す傾き 1 の直線が描いてある．プロット点から傾き 1 の直線までの，縦軸方向で測った距離が推定誤差となる．なお，この図ではプロットの重なりを避けるために，プロット点の位置をフルスケールの 1/80 の幅の一様分布で左右に振ってある．上段がプロによる発話音声の推定結果，下段が一般人による発話音声の推定結果である．また左端が 8 [s]，中央が 16 [s]，右端が 32 [s] の推定区間長となっている．

図 3 の上段に着目すると，分布が真の話速よりも大きい位置に生じている様子が観察できる．この結果から，プロの発話はエネルギ変化を検知しやすく，大きめの話速値が推定されると考えられる．次に図 3 の下段に着目する．一般人の推定結果は真の話速を中心としてほぼ上下均等に分布しているように見える．しかし，話者ごとに結果を観察した結果，推定話速が大きめに推定される話者と小さめに推定される話者とに分かれることが分かった．以上より，本手法では話者に依存する系統的誤差が少なからず存在することがわかった．今後は系統的誤差が生じる原因を解析し，話者への依存を極力低くする改善を行っ



(a) プロによる発話，推定区間長 8 [s]　(b) プロによる発話，推定区間長 16 [s]　(c) プロによる発話，推定区間長 32 [s]

(d) 一般人による発話，推定区間長 8 [s]　(e) 一般人による発話，推定区間長 16 [s]　(f) 一般人による発話，推定区間長 32 [s]

図 3　横軸に真の話速値，縦軸に推定話速値をプロットしたもの．
話速推定区間長を 8 [s]，16 [s]，32 [s] の三段階で変化させてある．

Fig. 3　Results of estimating speaking rates. The horizontal axis is true speaking rate and the vertical axis is the estimated speaking rate. The time interval of estimation is 8 [s], 16 [s], and 32 [s].

て行きたい．

5. 高速発話時の音声特性

以下では，本稿のもう一つのテーマである高速発話時の特性解析ついて報告する．我々は，音声を聞き落すことなく短時間で聞き取らせることに興味があり，再生速度の最適性を判断する指標として話速を用いることを考えてきた．ところで，その指標は話速だけで十分だろうか．よりよい指標を求めるためには，高速発話時の特性はもちろんのこと，速い話速に話速変換された音声の特性についても詳しい知見が必要である．そうした知見を手にいれるための第一歩として，同一の文章を異なる話速で発話した音声同士の比較と，異なる話速で発話した音声を同一の話速に話速変換した音声同士の比較を行った．

比較に用いる音声には我々の作成した音声データベース (SRV-DB) の音声を使用した．SRV-DB はこうした利用に最も適した音声データベースである．本節を SRV-DB の典型的な利用例としてもとらえて頂きたい．

5.1 通常発話音声と高速発話音声との比較

人間が速い話速で発話するとき，観測される音声が通常話速の音声とは異なることがある．その一例を SRV-DB のデータから示そう．

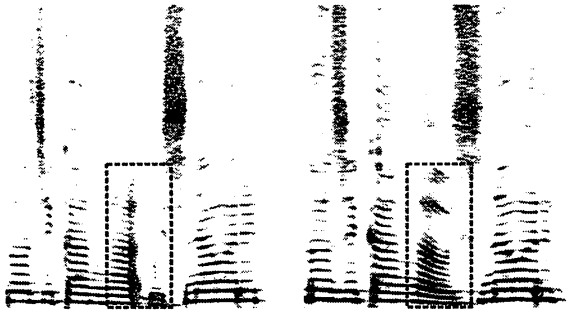
平均 8[モーラ／秒] および平均 11[モーラ／秒] で「…の時間が失われ…」と発話している部分のスペクトログラムを図 4 に示す．(a) では破線で囲った部分に 2 つのモーラが独立して存在していることを確認できる．一方，(b) では (a) で存在していた部分が直前のモーラと結合している様子が観察できる．(a) と (b) ではスペクトログラムが明らかに異なることがわかるだろう．

高速発話時には，より短い時間で必要な単語数の発声を行おうとするため，通常話速発話時とは異なる現象がいくつか生じてくる．したがって，通常発話音声のスペクトログラムを時間軸方向に収縮させても，高速発話時のスペクトログラムとは一致しない．我々が興味を持つのは，こうした変化が高速再生時に良い影響を及ぼすかどうかである．すなわち，モーラの連結が生じる 11[モーラ／秒] の音声は 18[モーラ／秒] まで話速変換しても聞き取れるが，通常発話の 8[モーラ／秒] の音声は 18[モーラ／秒] まで話速変換すると聞き取れなくなる，といった現象が起こるかどうかである．こうした現象の有無を検証することは今後の研究課題となっている．

5.2 話速変換音声と高速発話音声との比較

高速再生をするとき，人間の耳に届く音声は話速変換後の音声である．したがって，話速変換前と話速変換後で音声にどのような変化が生じるか観察することは，我々の研究を進める上で非常に重要な意味をもっている．ここでは，異なる話速を同一の話速に話速変換した音声を比較し，話速変換前の話速と話速変換後の話速との比が大きくなると，音声の劣化が激しくなる例を示そう．

8[モーラ／秒] の音声と 11[モーラ／秒] の音声を双方とも 16[モーラ／秒] に話速変換したときのスペクトログラムを図 5 に示す．発話内容は「がいじん」である．話速変換のアルゴリ



(a) 8 モーラ／秒での発話　(b) 11.31 モーラ／秒での発話

図 4　話速が速い時に起こるモーラの連結の例を示したもの.

Fig. 4　An example of the connection of two moras in the case of high speaking rate.

ズムには WSOLA 法を使用した. (c) から (d) への変換に比べて, (a) から (b) への変換では調波成分の乱れが大きくなっていることが分かる.

実際に話速変換後の音声を聞くと, (b) よりも (d) の方が聞き取り易かった. 話速変換前と話速変換後の話速差が大きくなると音質が劣化するという事実は, 次のような状況が起こりうることを意味している. すなわち, 12[モーラ／秒] から 18[モーラ／秒] に話速変換した音声は聞き取り可能だが, 6 [モーラ／秒] から 18[モーラ／秒] に話速変換した音声は聞き取り不可能である, といった状況である. 本当にこうした状況が起こるのか, また起こるとしたらどのような条件の下で起こるのか, 今後調べていく予定である.

6. おわりに

高速再生時において, 再生速度が最適かどうかを判断する基準を作成するために, 音声の話速を推定する手法を考案した. 2つの狭帯域エネルギ変化を狭帯域エネルギから作成する2つの信号のゼロクロスの頻度としてもとめ, 正規化した誤差で平均 16%の誤差で話速推定可能であることを示した.

また, 話速に関する研究のために作成した音声データベース (SRV-DB) を用いて, 高速話速に関する特性, および話速変換後の音声の特徴を挙げて, データベースの利用例を示した.

今後は, 推定話速を利用し最適な再生速度を得るための研究に着手していきたい.

文　献

[1] 高橋弘太：“フレキシブルな時間軸による最適な速度曲線での音声再生”, 信学技報, vol.107, no.116, pp. 37–42 (2007).

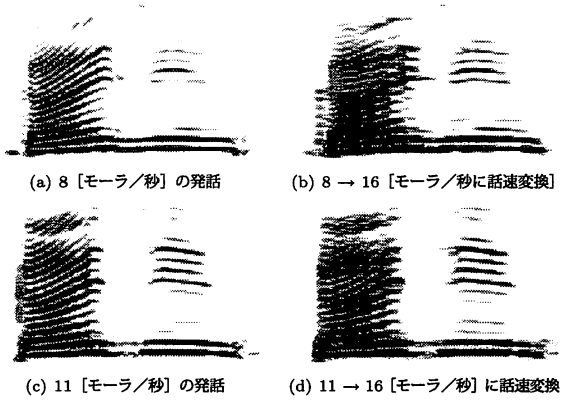
[2] 吉原亨, 高橋弘太：“話速適応性を有するフレキシブルな時間軸による音声再生”, 信学技報, vol.107, no.234, pp. 19–24 (2007).

[3] X. Gong, 広重真人, 荒木健治, 栃内香次：“発話速度推定ための多次元音響特徴量について”, 信学技報, vol.103, no.263, pp. 25–30 (2003).

[4] N. Morgan and E. Fosler-Lussier: “Combining multiple estimators of speaking rate”, Acoustics, Speech, and Signal Processing, 1998. ICASSP'98. Proceedings of the 1998 IEEE International Conference on, 2, (1998).

[5] D. Wang and S. Narayanan: “Robust Speech Rate Estimation for Spontaneous Speech”, Audio, Speech and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on], 15, 8, pp. 2190–2201 (2007).

[6] 高橋弘太, 蔦木圭悟, 吉原亨：“話速管理機能を持った原稿提示収録システム (recok5) と話速バリエーション型音声データベース (srv-db) の公開について”, 信学技報, 本件の前の発表 (2008)



(a) 8 [モーラ／秒] の発話　(b) 8 → 16 [モーラ／秒に話速変換]

(c) 11 [モーラ／秒] の発話　(d) 11 → 16 [モーラ／秒] に話速変換

図 5　異なる 2 種類の話速の音声を WSOLA 法を用いて同じ話速に変換したときのスペクトログラム.

Fig. 5　The spectrograms of time scale modified speeches. The original speaking rates are 8 [mora/s] and 11 [mora/s], modified speaking rate is 16 [mora/s].