音声の個人性と音響的特徴および話速と音響的特徴との関係の研究

Research on acoustic features in relation with voice individuality and speaking rates

帝京科学大学

Teikyo University of Science & Technology

桑原尚夫

Hisao Kuwabara

A variety of linguistic as well as non-linguistic information is carried through acoustic parameters in speech wave. The first experiment was designed to investigate acoustic features that reflect the voice individuality by altering acoustic parameters through analysis-synthesis method. Formant frequencies, their bandwidths and fundamental frequency were taken into consideration. Each parameter was independently changed and re-synthesized speech was presented to listeners for the individuality judgment. The result reveals that formant frequency has the greatest influence on the individuality and fundamental frequency the least. Waveform distortions, such as zero-crossing and center-clipping, were found to have little influence to the individuality. The second experiment was conducted to investigate how the speaking rates affect to acoustic parameters. It was found the durations of individual syllables in continuous speech were roughly proportional to the speaking rate but, for slow speech, the ratio between consonant and vowel parts within a syllable was quite different from that of the fast and normal speech. The first and second formant frequencies for vowels were found to be centralized almost proportional to the speed of speaking rate while third and fourth frequencies were not. Finally, a perceptual experiment was performed for individual syllables taken out from the running speech.

Key words: voice individuality, formant frequency, zero-crossing, center-clipping, speaking rate

1. まえがき

音声情報処理の進歩により、音声認識や合成技術に関する限り、語彙の種類、話者や発声スタイル等にある程度の制限を課せばほぼ実用段階に来ていると考えられる。残る問題は、今までの研究では避けてきた様々な点、例えば話者のバリエーション、雑音環境、発声速度、発声スタイルなど、極めて困難ではあるが克服しなければならない問題が残されている。本研究はこれらの問題点のいくつかに対し、解決には遠いものの、その糸口をつかむべく基礎的な研究を行い、ある種の基本的なデータを積み上げたものである。

まず、音声の個人性 [1-3] に着目し、種々の音響的特徴量のうち、個人性に最も強く関連しているものを聞き取り実験によって調べた。ピッチ同期による通常の分析・合成方式を用い、音源特性と共鳴特性をまず分離し、それらを独立に制御することによって音声を再合成して個人性がどの程度保存されるか、あるいは失われるかの実験を行ったものである。また、零交差波、センタークリッピングの2種類の波形歪を与えたとき、個人性情報と音韻性情報はどの程度保存されるかについても同時に実験を行った。実験方法とその結果を２節で述べる[4]。

次に、発声速度が個々の音響パラメータにどのような影響を与えるかについての研究を行った。“速い”、“普通”、“遅い”、の3種類の速さで発声した音声を分析し、まず、個々の音韻の継続時間が速さと共にどのように変化するか、次に母音のフォルマント周波数およびピッチ周波数がどのように変化するか、について調べた。その結果について3節で述べる。

さらに、連続音声中の個々の音節が、それらを単独に取り出したとき聴覚的にどの程度の音韻性を保存しているかを聞き取り実験によって調べた。また、速い発声の音声（f-speech）に関しては、１）音節を一つずつ取り出す１音節単位の切り出し、２）隣り合った二つの音節を切り出す２音節単位、３）連続した三つの音節を取り出す３音節単位、の３種類の切り出し方法で行った。

2. パラメータ制御による音声個人性の知覚実験

音声信号に含まれる音源と共鳴特性を分離し、それらを独立に制御して変化させることによって音声を再合成し、個人性に与える影響を調べた。なお、この研究結果についてはかなりの部分既発表のものが含まれる。

2.1 ピッチ同期分析・合成方式 [5]

図1に実験に用いた、ピッチ同期方式による分析・合成システムを示す。

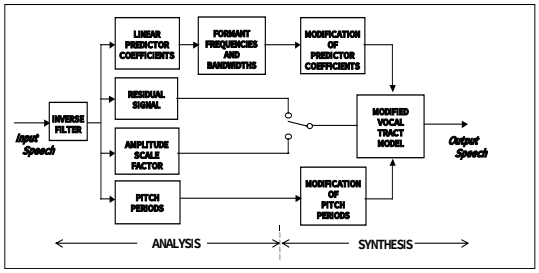

Fig. 1 Block diagram of pitch synchronous analysis-synthesis system.

15kHz、12ビットで量子化された入力音声から、ピッチ区間毎に、線形予測係数、残差信号、振幅ファクター、ピッチ周期、の4種類のパラメータをまず抽出する。実験では、これらのパラメータの中からピッチ周期と予測係数の2種類を使い、それぞれを独立に制御することによって音声を作り変えた。ピッチ制御は極めて単純で、合成部分に入力する残差信号の長さを変えることによって実現し、フォルマント周波数およびバンド幅の変化は予測係数を変えることによって実現している。

2.2. ピッチ変化による個人性

2.2.1. ピッチ周波数の制御

ピッチ周波数の制御は極めて簡単で、入力する残差信号の長さを変えることで行った。図2にそのブロック図を示す。

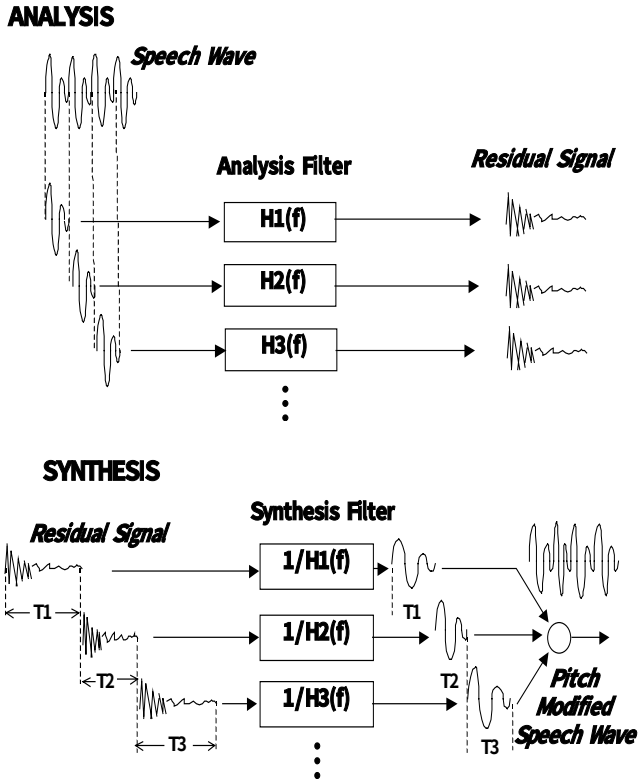

Fig. 2 Schematic illustration of pitch frequency modification.

分析部分では、入力音声はまずピッチ毎に残差信号を抽出して保存する。合成部分では、ピッチ周波数を上げるには入力残差の最後の部分をカットし、下げるには逆にゼロ信号を最後に追加して入力する。この操作により、共鳴特性（声道特性）は原音声の特性を保ったままピッチ周波数のみ変化させることができる。被験者はピッチ変化音声を聞き、原音声と同一人物の声か否かを判断する。

2.2.2. ピッチ変化による個人性判断実験結果

図３に実験結果を示す。８人の話者が発声した５母音を用い、ピッチパターンの形は原音声のそれを保ったまま、平均ピッチを４０％まで高く、および低くした場合の個人性判断を行った。図の横軸はピッチ周波数の変化であり（０％は原音声、負側はピッチを下げる方向)、縦軸は個人性判断の割合である。本図は８人の話者の結果を平均したものである。ピッチ変化に対

しては個人性は比較的頑健であることが分かる。特に、ピッチを低くする場合は個人性は殆ど変化がないことが分かる。また、話者による違いは多少見られるが、一般的な傾向は皆同じである。

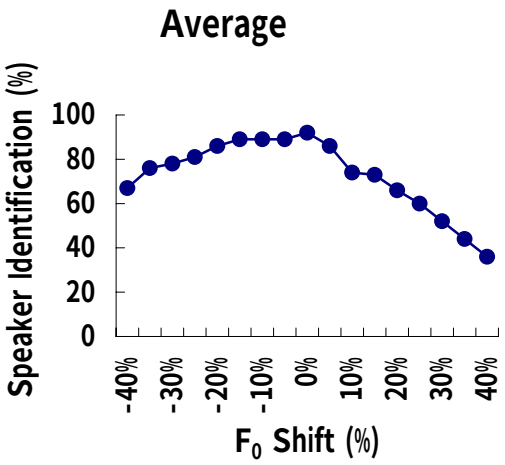

Fig. 3 Result of speaker identification for pitch frequency shift averaged over 8 speakers.

2.3. フォルマント変化による個人性

よく知られているように、線形予測係数には基本周波数の情報は含まれておらず、すべて共鳴特性のみである。他の音響特性を変えずにフォルマント周波数あるいはバンド幅を変えるには、従って、予測係数に変更を加えればよい。

2.3.1. フォルマント周波数・バンド幅の制御

各ピッチ周期毎に得られた線形予測係数から、まずフォルマント周波数とそれらのバンド幅を計算する。これはよく知られているように、予測係数を係数に持つ高次方程式の解で与えられる。得られた解から容易にフォルマント周波数とバンド幅が計算される。ここで、そのアルゴリズムについて簡単に述べる。

1） 分析時

$\{a_i\}\quad (i=1,2,\cdots,p)$ を予測係数とし、z を未知数として、ます高次方程式

$$a_p z^p + a_{p-1} z^{p-1} + \cdots + a_1 z = 0 \qquad (1)$$

の解を求める。それらの解を、$z_k = r_k e^{j\omega_k}$ とすれば、フォルマント周波数 F_k およびバンド幅 B_k は各々、

$$F_k = \omega_k / 2\pi T \text{ 、 } B_k = \log r_k / \pi T \qquad (2)$$

で与えられる。

2） 予測係数の変更

今、フォルマント周波数 F_k を F_k' に、バンド幅 B_k を B_k' に変更したとすれば、その変更に伴ってまず解 z_k を変更する必要がある。周波数あるいはバンド幅に一つでも変更を加えた場合、予測係数はすべて変更する必要がある。変更後の解を、変更しないものも含めて、$\tilde{z}_k$ と書き、変更後の予測係数を $\tilde{a}_i$ と書くと、$\tilde{a}_i$ は次式の恒等式を解くことによって得られる。

$$\tilde{a}_p (z-\tilde{z}_p)\cdots(z-\tilde{z}_1) \equiv \tilde{a}_p z_p + \cdots + \tilde{a}_1 z \qquad (3)$$

3） 合成時

フォルマントまたはバンド幅変更した音声を合成するには、変更した予測係数を持つ次のような声道フィルター $V(z)$ を構成し、分析時に得られた残差信号を入力することによって得られる。

$$V(z) = 1 \Big/ \sum_{i=1}^{p} \tilde{a}_i z^{-i} \qquad (4)$$

各ピッチ区間毎に得られたこれらの音声をつなぎ、同様に個人性判断の評価を行った。

2.3.2. フォルマント周波数の変化による個人性判断実験結果

フォルマント周波数全体（F_1～F_6）を一様にシフトさせたときの結果を図４に示す。図の横軸はシフトする周波数の割合であり、正は高くする方向、負は低い方向に変化させた場合である。縦軸は個人が同定できた割合である。また、０%の位置は原音声を意味している。

実験結果を見ると、個人性が保存される範囲は極めて狭く、高域、低域方向とも僅か３%程度である。特に、低域へのシフトでは５%でほぼ個人性は失われることが分かる。また、いずれの方向に対しても、１０%

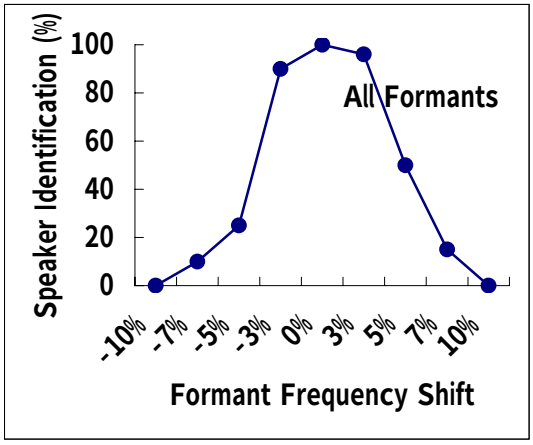

Fig. 4 Result of speaker identification for uniform shift of all formant frequencies.

のシフトで完全に個人性は失われることが分かる。

2.3.3.　バンド幅の変化による個人性判断結果

フォルマントバンド幅の音声生成面から見た物理的意味は必ずしも明らかではないが、直接的には声道壁での音響エネルギー損失が関与している。その意味で、バンド幅は個人性情報を担っている。図５に実験結果を示す。

バンド幅の変化は原音声のそれを１として、その倍率として与えた。広げる方向は、原音声の３倍、５倍、７倍、１０倍、狭める方向はそれらの逆数倍の合計８種類である。図から、一見個人性が急激に変化しているように見えるが、3倍あるいは1/3倍のスペクトル変化は大きく、事実波形も原音声のそれと大きく変わっている。その意味で、個人性はバンド幅には比較的鈍感であると考えられる。原音声の倍以上、あるいは1/5以下で個人性は完全に失われることが分かる。

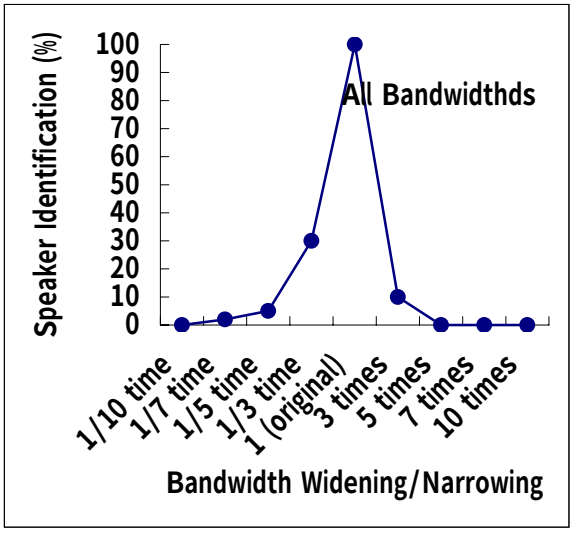

Fig. 5 Result of speaker identification for formant bandwidth manipulation.

2.4.　波形歪みによる個人性

次に、簡単な波形歪みを加えたとき個人性がどのように変化するかを調べた。波形歪みとは、零交差波とセンタークリッピングの２種類である。図６に零交差波に対する結果を示す。この実験は、個人性情報の変化と比較するため、同時に音韻情報等の言語情報の変化も調べた。従って音声資料は、１）単独発声の音節

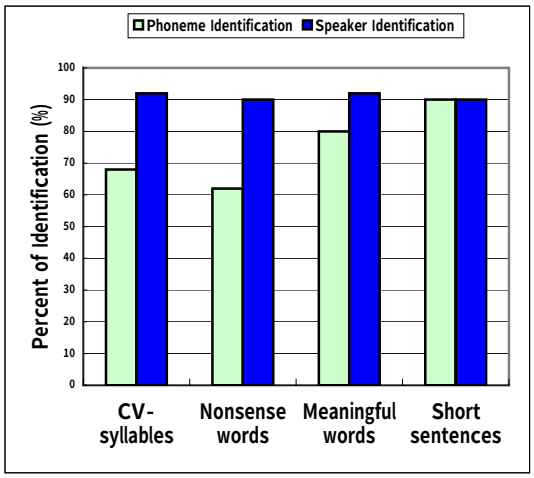

Fig. 6 Result of speaker identification for zero-crossing distortion.

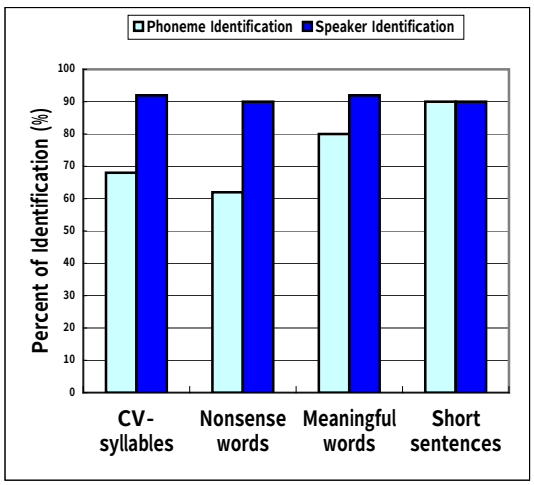

Fig. 7 Result of speaker identification for center-clipping distortion.

(CV-syllable)、２）無意味単語 (nonsense words)、３）有意味単語 (words)、および４）短文 (short sentence)、の４種類である。

実験結果を見ると、音韻性等の言語情報はかなりの程度失われるが、個人性はいずれの歪みに対しても殆ど影響を受けないことが分かる。この原因として考えられるのは、スペクトル領域で見るといずれもある特定の領域がダメージを受けるわけではなく、ダメージが領域全般に渡っているためであろう。上記以外にも、個人性に関与すると思われるパラメータは幾つか考えられる、その内の一つ、声帯波形（声門体積波形）[6]について調べたが、特に個人性の特徴は抽出できなかった。また、個人性に最も大きく関与すると思われる上記パラメータの時間変化（時間特性）[7] については今回は除いた。それらは今後の大きな課題である。

3. 話速による音韻の知覚と音響的特徴の変化

ここで問題にしている話速は、個人間での話す速さのバリエーションではなく、同一人物が速く話したり、ゆっくり話した時、文章中の個々の音韻の音響的特徴がどのように変化するかを、主として音韻の継続時間を中心に調べたものである。さらに、音声中の個々の音韻（音節）を単独に切り出したとき、それらが聴覚的にどれだけの情報を保存しているかを知覚実験によって調べたものである[8]。

3.1. 話速による音響的特徴の変化

3.1.1. 音韻継続時間の変化

１５種類の短い文章を、1）速い発声（以下、f-speechと呼ぶ）、２）普通の速さで発声（n-speech）、および３）遅い発声（s-speech）、の３種類で発声した音声資料を用いた。発声者は男性４名である。４人の話者毎に、最も話しやすい速度を“普通の速さ”とし、その２倍と感じる速度を“速い発声”、半分と感じる速度を“遅い発声”とした。特に計器を用いて話す速さを制御することなく、各自が自分の感覚に従って話すよう求めた。

日本語は基本的には拍（モーラ）が発声のリズムやテンポを構成しているが、これは幾つかの例外を除いて１拍は１音節に相当する。そこで、まず文章中の各音節区間を、波形、スペクトル、および聞き取り作業によって区分した。さらに、各音節を子音部分と母音部分に区分して、それぞれの長さを調べた。

（1） 平均音節長

図８に各発声速度での音節平均継続時間を示す。普通の発声では約156ミリ秒であり、速い発声で約94ミリ秒であった。また、遅い発声では約345ミリ秒である。

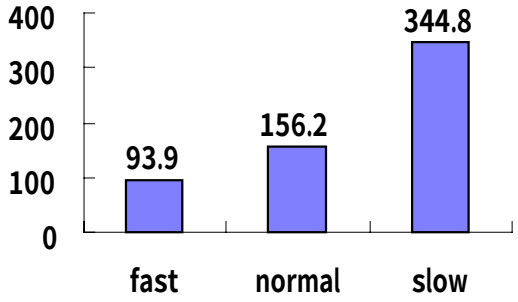

Fig. 8 Average syllable duration for three different speaking rates.

（2） 子音と母音の継続長比

１音節中に占める子音と母音の継続時間の割合を、各速度毎に平均すると図９のようになる。

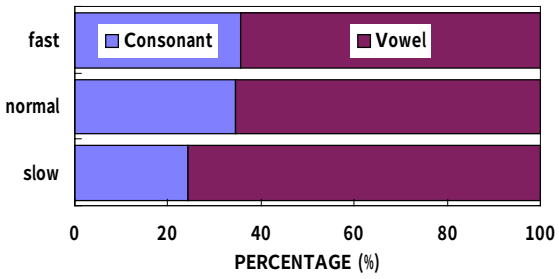

Fig. 9 Duration ratio between consonant and vowel in a CV-syllable.

９図を見ると、f-speech と n-speech とでは子音と母音の継続時間の比率は殆ど同じであるのに対し、s-speech では明らかに母音部分が長い。実際の数値で比較すると、f-speech と n-speech で母音の長さはそれぞれ64%と65%であるのに対し、s-speech では76%に急に上昇している。

１０図には、子音と母音部分のそれぞれについて、３種類の速さでの長さを比較したものを示す。子音部分は速さにほぼ比例して短くなっているが、母音部分

は、s-speech で急激に長さが伸びているのが分かる。

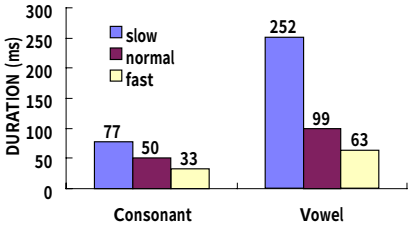

Fig. 10 Average duration for consonant and vowel parts of a syllable in three speaking rates

個別の子音毎に継続時間の変化を示したのが１１図である。この図は、n-speech の長さを 100(%)として、f-speech と s-speech とでどの程度の伸縮があったかを百分率で表したものである。一般的には摩擦性の子音の伸縮が大きいが、有声破裂音/b/の伸張が著しい。また、破裂子音の伸縮は一般に小さいことが分かる。

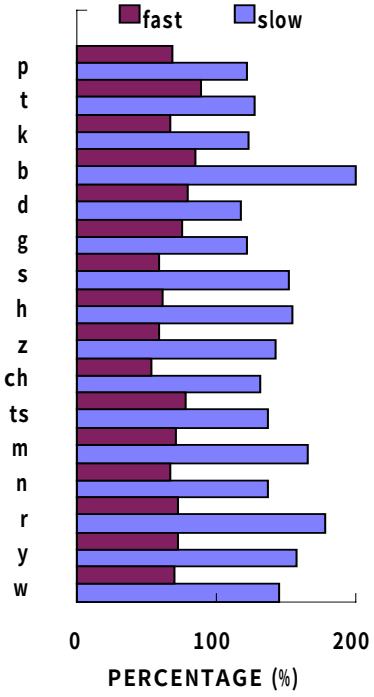

Fig. 11 Duration ratio for individual consonants.

3.1.2.　フォルマント周波数の変化

次に母音のフォルマント周波数の変化を調べた。調音結合の影響で、フォルマント周波数は一般に早口になるほど中性化が著しくなることが知られている。実際のフォルマント周波数はどのように変化するかを調べた。その結果を図１２に示す。

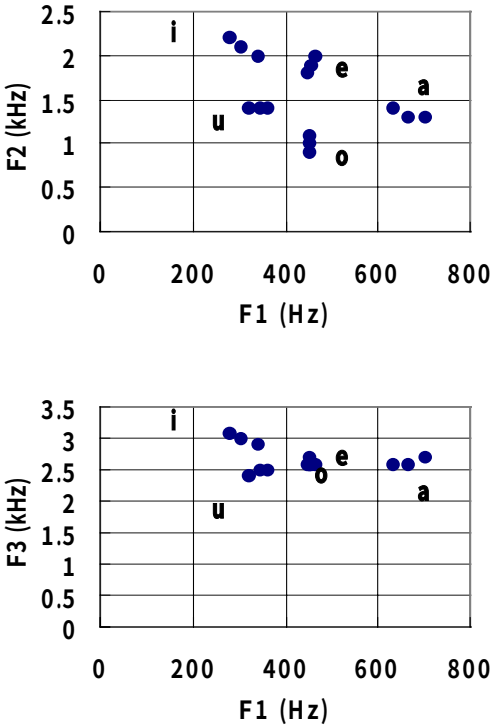

Fig. 12 F1-F2 and F1-F3 diagram of vowels for three different speaking rates.

図１２の上図は通常のF_1-F_2図であり、下図はF_1-F_3図である。各母音とも最も外側に位置する点がs-speechに対する点であり、次がn-speech、最も内側に位置する点はf-speechに対する点である。予想通り、ほぼ発声速度に比例して母音の中性化が起こることが分かる。ただし、第１、第２フォルマント比べて、第３フォルマントは極めて変化が小さいことが分かる。

3.2.　切り出し音節の音韻知覚

次に、３種類の速度で発声された連続音声から個々の音節を切り出したとき、それらがどの程度の音韻情報が保存されているかを知覚実験によって調べた。切り出し方法は、１）個々の音節を一つずつ取り出し１音節単位の切り出し、２）隣り合った二つの音節をまとめて切り出す２音節単位、３）連続する３音節を切り出す３音節単位、の３種類の切り出しを行った。それらをランダムに被験者に提示し、聞こえた通りに記

入してもらい、回答を音節ごとに集計を行った。

図１３は、１音節単位で切り出した時の音節同定結果であり、音節全体の正答率およびその内、母音部分および子音部分の正答率を分けて示したものである。音節全体で見ると、f-speech では平均40%以下と極めて低く、単独では音節としての情報は殆ど失われる。また、n-speech でも約60%で不十分であるが、s-speech では90%近い正答が得られ、ほぼ音節情報が保存されることが分かる。

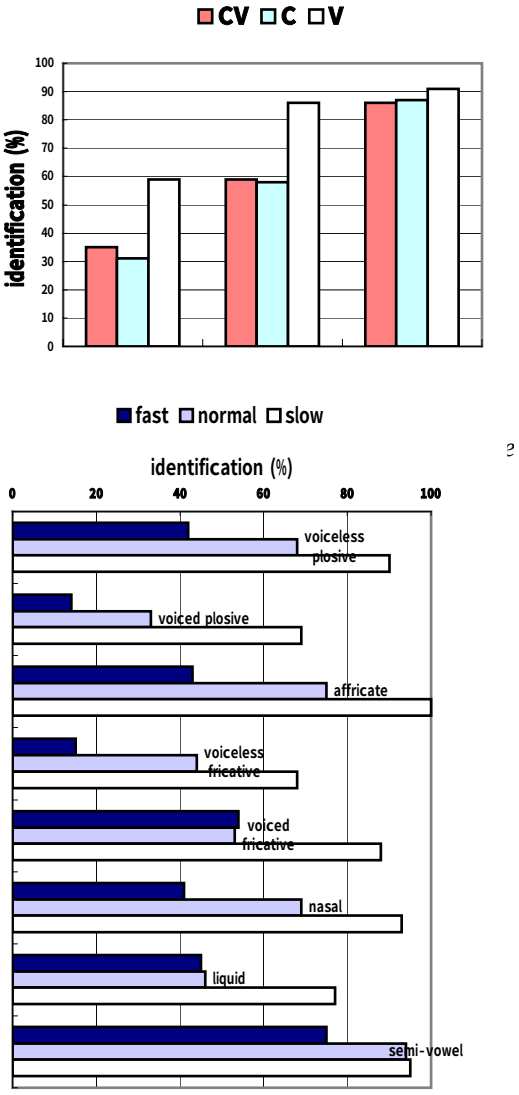

Fig. 14 Identification of CV-syllables for every consonant category.

図１４は、図１３の結果を子音クラス別に表示したものである。f-speech では特に有声破裂音と無声摩擦音の正答率が低いが、発声速度が遅くなるに従って急に正答率が上がることが分かる。また、発声方法から当然予想されたことであるが、半母音はいずれの速さでも正答率は高いことが分かる。

２音節単位の分節での実験結果を図１５に示す。上の図は２連音節の前の音節に対する結果であり、下の図は後の音節に対するそれである。全般に後の音節に対する正答率が高い。これは、前の音節では子音部分が部分的に削られるため、子音情報が一部失われるのに対し、後の音節ではそれがないことが最も大きな要因と考えられる。

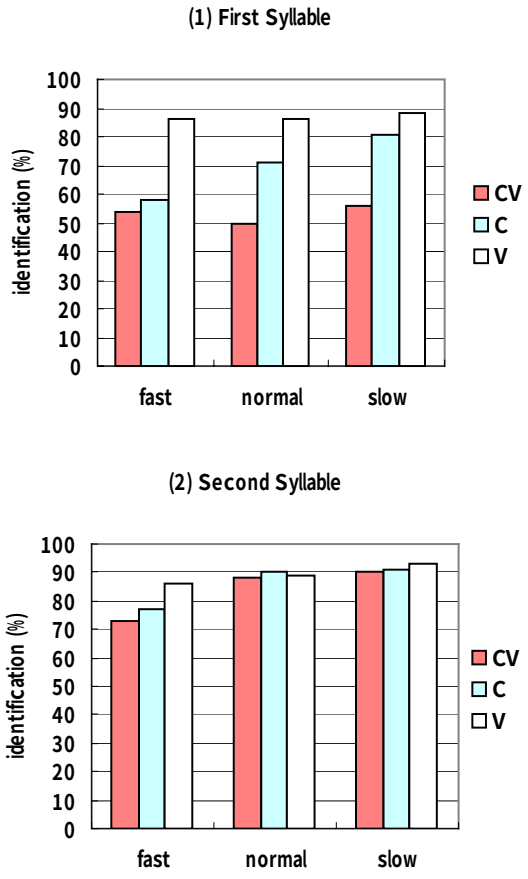

Fig. 15 Syllable identification for 2-syllable segmentation.

最後に、３音節単位の分節に対する結果を図１６に示す。最上部の図は３連音節の1番目（前）の音節に対する結果であり、真中は2番目（中央）の音節に対する結果、最下部は3番目（後）の音節に対する結果である。

最後に、３種類の切り出し方法での結果を比較するため、f-speech のときの実験結果を図１７に示す。この図から、同じ速い発声でも、３音節連鎖の中央の音節は91%の正答率であり、ほぼ完全に音節情報が保存

されることが分かる。

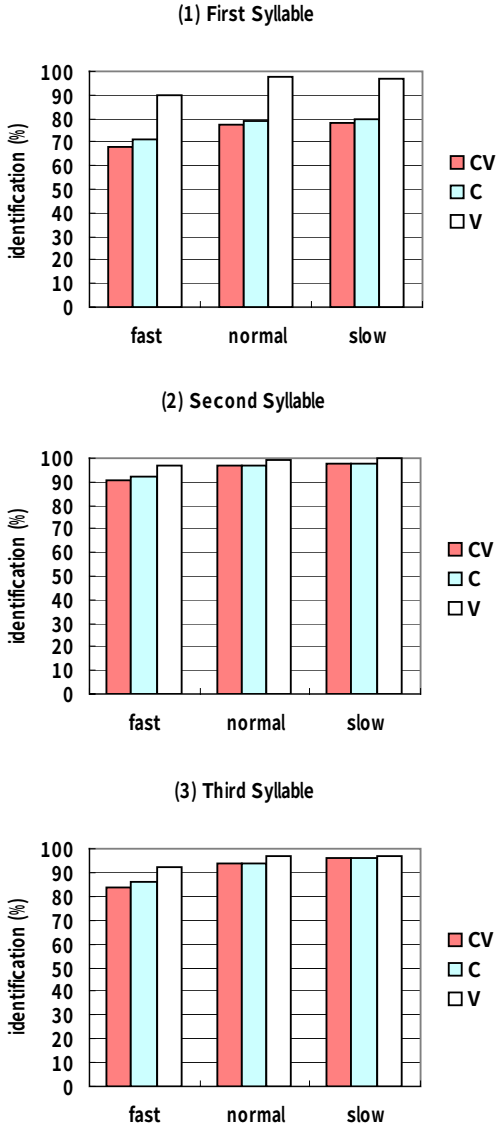

Fig. 16 Syllable identification for 3-syllable segment.

4. まとめ

音声の個人性についてまず音響パラメータとの関係を聞き取り実験によって調べた。ピッチ同期による分析・合成方式を用い、声帯、声道特性を独立に制御して個人性判断実験を行った。その結果、フォルマントが最も個人性に敏感であることが分かった。さらに、話速と音響的特徴および音節の知覚実験を行い、遅い発声では、母音部の継続時間が著しく長いこと、速い発声でも３連音節の中央の音節はほぼ完全に知覚できることが分かった。

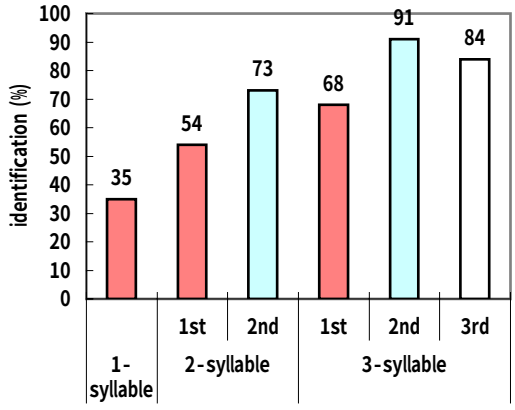

Fig. 17 Comparison of syllable identification for the 3 segmentation schema.

参考文献

[1] H.Matsumoto, et al , " Multidimensional representation of personal quality of vowels and its acoustical correlates," IEEE Trans., Vol.AU-21,.pp.428-436 (1973)

[2] 古井貞熙、" 音声個人性パラメータの時期的変動と話者認識 "、電子情報通信学会論文誌、Vol.67-A, No.12, pp.880-887 (1974)

[3] K.Itoh and S. Saito, "Effects of acoustical feature parameters of speech on perceptual identification of speaker," Trans. IECE Japan, Vol.J65-A,.pp.101-108 (1982)

[4] H.Kuwabara, "Contributions of vocal tract resonant frequencies and bandwidths to the personal perception of speech," ACUSTICA,Vol. 63, pp.120-128 (1987)

[5] H.Kuwabara, "A pitch-synchronous analysis/synthesis system to independently modify formant frequencies and bandwidths for voiced speech," Speech Communication, Vol.3, pp.211-220 (1984)

[6] D.Y.Wong, et al, "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE Trans., ASSP-27, .pp.350-355 (1979)

[7] 桑原尚夫、大串健吾、" アナウンサー音声の音響的特徴 "、電子情報通信学会論文誌、Vol.J66-A, No.6, pp.545-552 (1983)

[8] H.Kuwabara, "Acoustic properties of phonemes in continuous speech for different speaking rate, " Proc. of ICSLP, pp.2435-2438 (1996)