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

2008－HCI－128 (8)
2008－MUS－ 75 (8)
2008／5／28

STRAIGHTを用いた簡易モーフィングによる印象変化の評価について

西田 沙織 †, 大西 壮登 †, 吉田 有里 †, 森勢 将雅 ††, 西村 竜一 ‡, 入野 俊夫 ‡, 河原 英紀 ‡

† 和歌山大学大学院システム工学研究科 †† 関西学院大学理工学部 ‡ 和歌山大学システム工学部

時間軸だけを整合させるという簡易な方法によりモーフィングした音声を対象として、自然性と話者性の主観評価実験を行った。その結果を、単語・モーフィング率・話者の組み合わせという3つの観点から分析した。単語別に見た場合、自然性・話者性の評価には、有意差は認められなかった。モーフィング率別に見た場合、モーフィング率が50 %に近づくほど自然性が低下し、モーフィング率が25 %から75 %では、話者性の正答率が60 %程度となった。話者の組み合わせ別に見た場合、組み合わせが同性か異性かで評価の傾向に差が見られた。組み合わせが同性のときは自然性は高くなるが話者性を判別しにくくなり、異性のときは自然性は低くなるが話者性を判別しやすいという傾向が認められた。これらの結果より、同性の話者の場合には、簡易なモーフィングを実用的な手法として利用できる可能性があることが分かった。

Effects on perceived impression of manipulated speech using a simplified morphing procedure based on STRAIGHT

Saori NISHIDA † Masato ONISHI † Yuri YOSHIDA † Masanori MORISE †† Ryuichi NISIMURA ‡
Toshio IRINO ‡ Hideki KAWAHARA ‡

† Graduate School of Systems Engineering, Wakayama University
†† School of Science and Technology, Kwansei Gakuin University
‡ Faculty of Systems Engineering, Wakayama University

A morphing procedure only relies on temporal axis alignment was tested subjectively in terms of naturalness and speakers' identity. Effects of contributing factors were investigated regarding on test words, morphing rates and used speakers. Naturalness of the morphed speech was deteriorated when the morphing rate nears 50%. Identification of mixing rate of two speakers was about 60% when the morphing rate is 25%, 50% or 75%. Naturalness of the morphed speech sounds were found higher when speakers' sex was identical while mixing rate identification were lower. These results suggest that the proposed simplified procedure is practically usable for morphing speakers having the same sexual distinction.

1 はじめに

高品質なVOCODERであるSTRAIGHT[1]と、それに基づいた音声モーフィング[2]は、感情音声の知覚[3, 4]、歌唱表現の知覚と制御[5, 6]、歌唱デザインの転写に向けたインタフェースの研究[7]などに広く応用されている[8]。音声モーフィングでは、対象とする属性（複数も含む）の値が異なる2つの音声試料から、属性と物理パラメタの対応関係についての明示的な知識に依存せずに、中間的な値を有する音声を合成することができる。音声モーフィングのこのような特長は、心理実験のための刺激連続体の準備や、歌唱や音声に新たな個性や表現を加えるなどのポストプロダクションへの応用の可能性を拓く。本資料では、このモーフィングを簡単に利用するための方法を紹介し、それを用いた実験結果について報告する。

2 音声モーフィングの簡単化

まず、従来の音声モーフィングの概要を説明する[2, 9]。STRAIGHTによる分析は、音声を、基本周波数、非周期性スペクトログラムからなる音源情報と、周期性の影響が取り除かれたスペクトログラム（STRAIGHTスペクトログラム）とに分解する。これらのパラメタは、時間と周波数の関数（基本周波数は時間のみの関数）であり、実数

値をとる。

音声モーフィングを行うためには、まず、STRAIGHT スペクトログラム上に重要な時間周波数座標（特徴点）を手作業により設定する。モーフィングプログラムは、与えられた二つの音声試料の対応する特徴点が重なるように時間周波数座標を変換し、その座標におけるパラメタの値を与えられた割合（モーフィング率）に基づいて補間／補外する。その後、パラメタの時間周波数座標を、モーフィング率に基づいて変換し、前述の処理で求められたパラメタを合成プログラムに渡すことにより、必要とする音声が合成される。

2.1 従来の方法の問題

特徴点を付す手作業は、モーフィングの条件の精密な制御を研究者に委ねることを狙って導入された [2] ものである。しかし、適切な位置に特徴点を付与するためには、信号処理と音声学および聴覚心理学の知識が同時に必要となる。そのため、手作業が与える自由度は、一般の利用者にとって逆に利用の妨げになる。また、コンテンツ加工など、日常的に大量のデータを処理する場合や、対話的に様々な音声試料間のモーフィングを行いたい場合などには、手作業による作業量の多さ（例えば 1.4 秒の音声試料「どげんかせんといかん」の場合に 52 個の特徴点の付与が必要であった）が障害となる。さらに、低い周波数領域では、聴覚の周波数分解能が相対的に高いため [10]、STRAIGHT スペクトル上での特徴点の僅かな誤りが、大きな品質劣化につながる [11]。

2.2 時間軸整合のみを用いたモーフィング

以下で説明するように、適用領域を限れば、（周波数方向の特徴点を付与せずに）時間軸の整合のみによるモーフィングが利用できる可能性がある。時間軸の整合のみであれば、様々な自動化手法を用いることにより、容易にリアルタイムモーフィング [12] が利用できるようになる可能性がある。

STRAIGHT スペクトルは、主要なピークの近傍では、二次曲線で近似される。二つの音声資料のスペクトル形状の違いの主な原因であるフォルマント周波数の差は、この二次曲線の近似が成立する近傍のサイズと比較すると、低い周波数領域では相対的に小さくなる。従って、低い周波数では、特徴点を省いて平均を求めるだけで、ピーク位置は適切に補間される。高い周波数領域では、聴覚の周波数分解能が低下している [10] ため、ピークの位置そのものよりも、聴覚フィルタの帯域内でのエネルギーの重心の位置が問題となる。この場合も、特徴点を省いて平均を求めるだけで、重心の位置が適切に補間される。これらは、二つの音声試料のスペクトルが類似している場合には、周波数方向に特徴点を設定する必要が無いことを意味する [11]。すなわち、時間軸を整合させるだけで、音声モーフィングを実現できる可能性がある。本資料では、この方法を検討する。

表 1　音声録音・加工環境

項目	機器／条件
マイク	NEUMANN U87A （単一指向性：コンデンサ）
マイクアンプ	M-Audio DMP3 Dual Microphone Pre Amplifer
ソフト (PC, OS)	Audacity ver 1.3.2 (iMac, Mac OS X 10.4.11)
標本化	44100Hz, 16bit
場所	実験室（150 m^2）

3 実験

これまで音声モーフィングを使用したことのない学部 3 年生 4〜6 名からなる 6 つのグループが、実験に参加した。被験者は、まず、この簡易な方法によるモーフィングについての 4 ページの説明書を渡され、実験全体の簡単な説明を受けた。その後、毎回 90 分のコマ一つをを 4 週にわたって使用し、音声の収録からモーフィング音声の合成、主観評価、報告までの一連の作業を行った。その結果、用意したプログラムに不備があった最初の組を除き、いずれのグループもモーフィング音声の合成から報告までの作業を完了することができた。以下では、その一つのグループが行った実験結果を紹介する。

音声の録音・加工環境は表 1 の通りである。

3.1 実験用単語の決定と録音

モーフィングの元となる音声として、話者 A、B、C、D から 4 種類の単語を録音した。単語は「こんにちは（濁音や促音が含まれていない）」「デザイン（濁音が含まれている）」「マック（促音が含まれている）」「プログラム（濁音と半濁音が含まれている）」の 4 種類である。話者は A、B が女性、C、D が男性である。

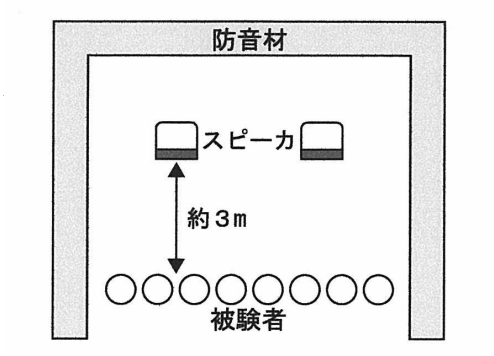

図 1　聴取実験時の配置

3.2　音声波形へのラベルの付与

非破壊サウンド編集ソフト Audacity [13] を用いて、視察によりラベル境界を設定した。ラベルの表記法は、文献 [14] を参考にして、実験用に簡略化したものを用いた。ラベルの付与では、Audacity の表示機能（スペクトログラムと波形）と試聴を繰り返し利用して作業を進めた。

3.3　モーフィング音声の作成

音声分析変換合成システム STRAIGHT を用いて、まず、ペアとなる二人の話者の音声を分析した。この分析された音声パラメタを、前の作業で作成されたラベルに基づいてモーフィングすることにより、変換されたパラメタが求められ音声が合成された。話者の組み合わせは AB、AC、AD、BC、BD、CD の 6 通り、モーフィング率は 0 %、25 %、50 %、75 %、100 %の 5 通りである。これらを各単語（単語数 4）について用意することにより、計 120 個のモーフィング音声を作成した。

3.4　評価実験

8 人の被験者に、作成したモーフィング音声とモーフィング元の音声 120 セットをランダムに提示し、評価してもらった。実験における被験者とスピーカとの位置関係を図 1 に、実験に用いた機材等を表 2 に示す。音声はモーフィング音声 X、モーフィング元の音声 A、モーフィング元の音声 B の順で提示した。被験者には自然な声に聞こえるかどうかの「自然性」を 1（非常に不自然）から 5（非常に自然）までの 5 段階で評価してもらった。また、併せて、A, B どちらの話者の声に近く聞こえるかを表す「話者性」を、用いたモーフィング率（5 段階）のいずれに該当するかで評価してもらった。120 個のモーフィング音声を 8 人の被験者に評価してもらい、話者性・自然性それぞれで 960 個の回答が得られた。

表 2　実験に用いた機材等

項目	機材／条件
スピーカ	BOSE 55 WER
D/A (PC, OS)	EDIROL UA101 MacBook pro, Mac OS X 10.5.2
ソフト	Matlab （sound 関数を使用）
提示音圧	70dBA
場所	実験室（100 m^2）

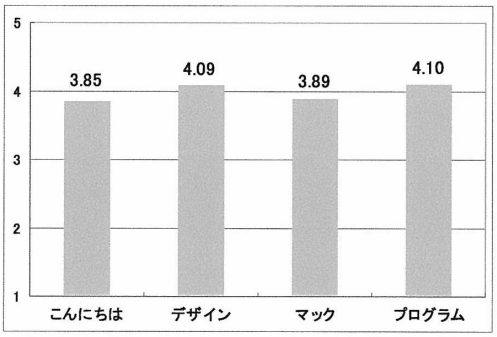

図 2　単語別自然性

4　実験結果の分析

被験者の回答を分析した結果を以下に示す。ここでは、被験者が回答した話者性に対応するモーフィング率と実際のモーフィング率とが一致していた場合、「話者性の評価が正答」であるとした。

4.1　単語別に見た自然性・話者性

自然性の評価の平均を図 2 に示す。有意水準 5 %で t 検定を行ったところ、4 つの単語の評価に有意差は認められなかった。

話者性の評価の正答率を図 3 に示す。正答率はどの単語でも 50%から 60%の間であった。有意水準 5 %で t 検定を行ったところ、4 つの単語の正答率の違いに有意差は認められなかった。

4.2　モーフィング率別に見た自然性・話者性

自然性の評価の平均を図 4 に示す。グラフは 50%のときが最も低い V 字型となった。有意水準 5%で、0%と 25%、25%と 50%など隣接したモーフィング率において t 検定を行ったところ、全てで有意差が認められた。この結果より、モーフィング率が

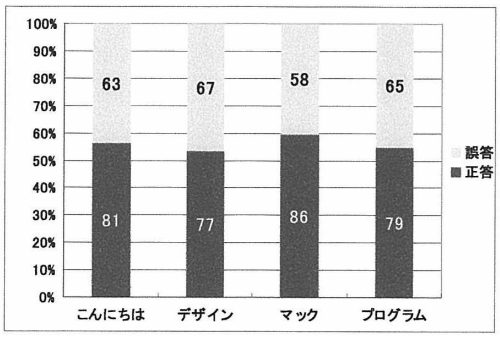

図 3　単語別正答率

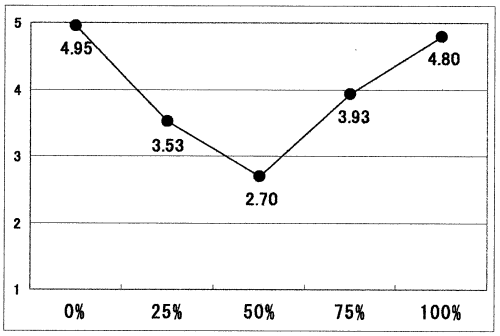

図 4　モーフィング率別自然性

50%に近づくほど自然性が下がることが分かる。

話者性の評価の正答率を図 5 に示す。モーフィング率が 0%のときに最も正答率が高く、100%のときはやや低くなった。25%から 75%までは同程

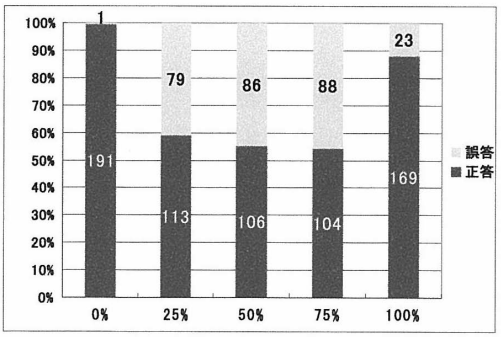

図 5　モーフィング率別正答率

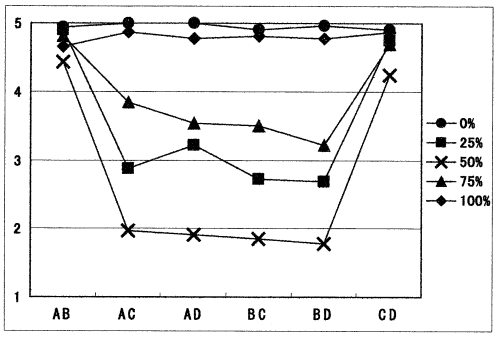

図 6　話者別自然性（モーフィング率ごと）

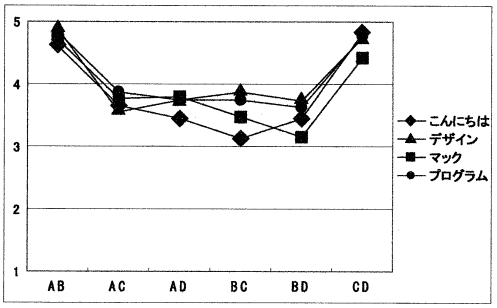

図 7　話者別自然性（単語ごと）

度の正答率であった。有意水準 5 %で t 検定を行ったところ、25%から 75%では正答率の違いに有意差は見られなかった。しかし、100%と 25～75%、100%と 0%とでは有意差が認められた。なお、モーフィング率 100%と 0%は共に他人の声が混ざっていない音声であるにもかかわらず正答率に差が現れた。これについては後程「課題」の章で述べる。

4.3　話者の組み合わせ別に見た自然性・話者性と回答の傾向

モーフィング率ごと、単語ごとの自然性の評価の平均を図 6 と 7 に示す。組み合わせが同性のときは、どのような条件でも自然性が高かった。組み合わせが異性のときは、いずれの単語もモーフィング率が 50%に近づくにつれて評価が著しく下がった。なお、モーフィング率が 25%（女性寄り）の音声よりも 75%（男性寄り）の音声の方が評価が高くなる傾向が認められた。

次に、話者性の評価の正答率を図 8 に示す。モーフィング率が 0 %と 100 %の音声に対する回答は

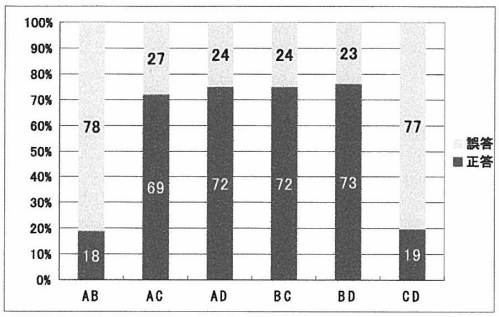

図 8　話者別正答率

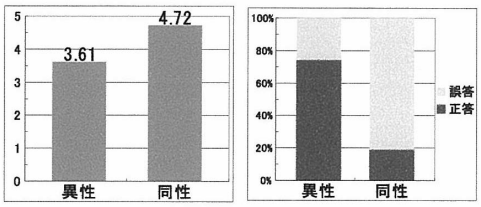

図 9　異性か同性かで見た自然性

図 10　異性か同性かで見た正答率

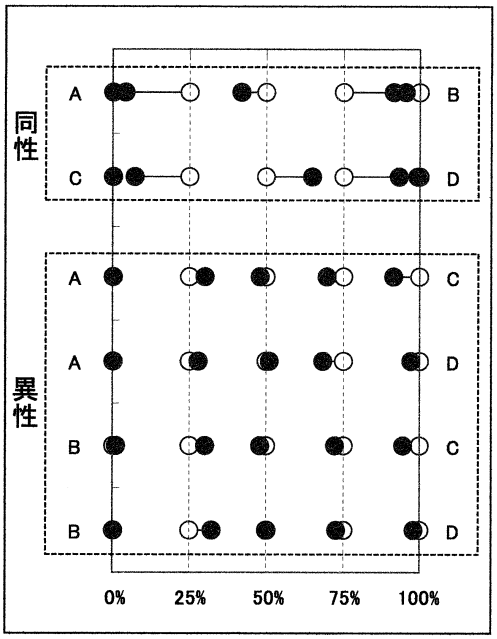

図 11　話者の組み合わせ別に見た回答の傾向

ほとんどが正答であるため、分析対象から除外した。正答率は同性の組み合わせである AB、CD のときが低く、異性の組み合わせである AC、AD、BC、BD のときが高くなった。有意水準 5 %で有意差検定を行ったところ、AB と CD の間、AC・AD・BC・BD の間には有意差はなく、AB・CD と AC・AD・BC・BD との間には有意差があるという結果が出た。よって、話者性の評価の正答率は話者の組み合わせが同性か異性かによってのみ違いが現れるといえる。

自然性の評価と話者性の評価の正答率を同性・異性ごとにまとめたのが図 9、10 である。自然性の評価は同性での組み合わせのときのほうが高くなるが、逆に話者性の評価の正答率は異性での組み合わせのときのほうが高い。これについても後程「課題」の章で述べる。

続いて、各組み合わせにおける回答の傾向を図 11 に示す。図の縦軸は、話者の組合わせを表し、横軸は、モーフィング率を表す。図中の○印は、合成された刺激を表す。図中の●印は、被験者の回答の平均値を表す。全ての被験者が全ての単語について正しくモーフィング率を回答した場合には、○と●とが重なる。図では、正答と回答の平均値とが離れている場合、対応するものを線分で結んだ。

同性での組み合わせの場合、モーフィング率が 25%と 75%のときは近いほうの話者の声に聞こえると回答する傾向が見られた。モーフィング率が 50%のときは喋り方や声が特徴的な話者のほうに回答が偏るという傾向が認められた。異性での組み合わせの場合、モーフィング率が 25%と 75%のときは平均値が 50%に近づく傾向が認められた。

5　課題

これらの実験により、ある程度の広さの会場でのスピーカ聴取であれば、時間軸を整合させるだけの簡易なモーフィングが、実用的なものとして利用できる可能性が示された。今回の実験では、可能性は同性の場合について示されただけである。また、実験手続きにも省略した部分があり、結果の信頼性は十分ではない。この方法の適用範囲を明らかにするためには以下に挙げる項目だけではなく、更に検討を加えることが必要である。

5.1　分析合成のみの音声の正答率

モーフィング率 0%と 100%は、単なる分析合成音である。しかし、100%のモーフィング率の場合

に、有意に誤答が生じている。今回の実験では、1コマ内に実験をおさめるために、提示順序のカウンターバランスを取っていない。提示は、(1) モーフィング音声、(2) モーフィング率 0%に相当する音声、(3) モーフィング率 100%に相当する音声、の順であった。(1) と (2) が隣接していることと、0%での誤答が皆無であることは、声の記憶の保持が正答率に関与していると仮定すると整合する。提示順序のカウンターバランスを取った実験による検証が必要である。

5.2 同性と異性のモーフィングの差

異性の音声のモーフィングでは、基本周波数が大きく異なる。従って、モーフィング率を判定するための手掛かりが、同性の音声のモーフィングの場合よりも多くなる。正答率の違いは、この要因による可能性が高い。同様に、異性のモーフィングにおける 50%付近での大きな自然性の低下も、性別の判定があいまいになることの副作用である可能性もある。基本周波数の影響を排除した場合に、自然性とモーフィング率の正答率がどのようになるか、検討する必要がある。スペクトル包絡の AR モデルパラメタの推定 [15] から計算することのできる平均的な声道長の補償が有効である可能性も、検討すべき重要な課題である。

6 まとめ

一般の利用者が容易に音声モーフィングを行うことのできる簡易なモーフィング手法を提案し、合成されたモーフィング音声の知覚的印象について調べた。その結果、この方法によれば、簡単な教示と短時間の作業で、事前の知識を持たない利用者であっても、モーフィング音声を合成することができることが分かった。また、合成されたモーフィング音声は、同性の話者の音声を試料とした場合には、自然性の低下が少なく、実用的に用いることのできるものであった。これらは、実時間あるいはそれに近い時間で動作するモーフィングの実現の可能性を示しており、モーフィングを利用した歌唱デザインの実現のための有用な結果である。

謝辞

本研究は、科学技術振興機構による戦略的創造研究推進事業のデジタルメディア領域 crestMuse プロジェクトの支援を受けて行われた。実験に参加したメンバーに感謝するとともに、特に筆頭著者と同じグループで実験に参加した、川口静香、庫内亮輔の両氏に感謝する。

参考文献

1) Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction," Speech Communication, 27(3–4): 187–207, 1999.

2) Kawahara, H. and Matsui, H., "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," ICASSP'03, Hong Kong, I: 256–259, 2003.

3) Matsui, H. and Kawahara, H., "Investigation of Emotionally Morphed Speech Perception and its Structure Using a High Quality Speech Manipulation System," Eurospeech'03, 3157–3160, 2003.

4) 筧 一彦, 曽我部 優子, 河原 英紀, 表情と感情音声の知覚, 信学技報, **TL2005-13**, pp.61–66 (2006).

5) Yonezawa, T., Suzuki, N., Abe, S., Mase, K., and Kogure, K., "Cross-modal coordination of expressive strength between voice and gesture for personified media," ICMI'06, Banff Ca, pp.43–50 (2006).

6) Yonezawa, T., Suzuki, N., Abe, S., Mase, K., and Kogure, K., "Perceptual continuity and naturalness of expressive strength in singing voices based on speech morphing," EURASIP J. Audio Speech Music Process., 2007(3), pp.1–9 (2007).

7) 河原 英紀, 生駒 太一, 森勢 将雅, 高橋 徹, 豊田 健一, 片寄 晴弘, "モーフィングに基づく歌唱デザインインタフェースの提案と初期的検討," 情報処理学会論文誌, 48(12), pp.3637–3648 (2007).

8) 河原 英紀, "Vocoder のもう一つの可能性を探る," 日本音響学会誌, 63(8), pp.442–449 (2007).

9) 河原英紀, 西雅史, 森勢将雅, 野口美咲, 高橋徹, 入野俊夫, "STRAIGHT に基づくモーフィングのオブジェクト化による拡張と部分モーフィングの応用について," 音講論, pp.505-506, Mar. 14-16, (2006).

10) Moore, B. C. J., "An Introduction to the Psychology of Hearing," Academic Press (2003).

11) 鈴田 健太郎, 森勢 将雅, 高橋 徹, 河原 英紀, 入野 俊夫, "低周波数領域での区分線形補間の弊害についての一検討," 音響学会春季講演論文集, I–Q–6, pp.275–276 (2007).

12) 森勢将雅, 河原英紀, 片寄晴弘, "STRAIGHT によるリアルタイム歌唱モーフィングシステムの実装," 音楽情報研究会, 2008. (本研究会)

13) `http://audacity.sourceforge.net/`

14) 前川喜久雄, 菊池英明, 藤本雅子, 米山聖子, "『日本語話し言葉コーパス』の文節音ラベリング version 1.0," 国立国語研究所 (2004).

15) 板倉文忠, 斎藤収三, "統計的手法による音声スペクトル密度とホルマント周波数の推定," 電子通信学会誌, 53A(1), pp.35–42, 1970.