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

2008－SLP－74 (11)
2008 / 12 / 9

音声認識システムの満足度評価におけるユーザモデル

原　直[†]　北岡 教英[†]　武田 一哉[†]

† 名古屋大学大学院 情報科学研究科　〒464-8603 名古屋市千種区不老町 1

あらまし　楽曲検索音声対話システムを用いたフィールドテストによるデータを用いて音声対話システムのユーザ満足度を推測する数学的モデルについて検討を行った．本研究ではユーザ満足度の背景となる心理尺度として，ユーザの主観に基づく体感認識精度を導入する．体感認識精度は満足度の指標や様々なユーザプロファイルとともにフィールドテストの事後アンケートを通じて収集した．まず，体感認識精度が対話データの書き起こしに基づいた客観認識精度指標よりも満足度指標と関係が高いことを示す．続いて，体感・客観認識精度の同時分布に対するトップダウンクラスタリングによってユーザのグループを認識精度に対する鋭敏さという観点により分類を行う．体感認識精度を与えうる客観認識精度の下限についてもこの同時分布を用いて計算を行う．最後に，ユーザプロファイルや環境条件を用いてユーザ満足度指標を推測するためのグラフモデルを構築し，分散の約 13%に相当するユーザ満足度分布の不確かさを削減することを示す．

キーワード　システム評価，性能指標，ユーザプロファイル，フィールドテスト

User modeling for a satisfaction evaluation of a speech recognition system

Sunao HARA[†], Norihide KITAOKA[†], and Kazuya TAKEDA[†]

† Graduate School of Information Science, Nagoya University, Furo-cho 1, Chikusa-ku, Nagoya, 464-8603, Japan

Abstract　A mathematical model for predicting the user satisfaction of a speech dialogue systems is studied based on a field trial of a voice-navigated music retrieval system. The *Subjective* Word Accuracy (*subjective*-WA), of the user is introduced as a background psychometrics for the satisfaction. In the field test, *subjective*-WA is collected through questionnaires together with satisfactory indexes and various user profiles. First we show that the *subjective*-WA is more significant to the user satisfactory than (*Objective*) Word Accuracy (*objective*-WA), which is calculated using the manually given transcriptions for the recorded dialogue. Then through top-down clustering of the joint distribution of *subjective*- and *objective*-WAs, we show that the user population can be grouped into several sub-groups in terms of sensitivity to recognition accuracy. The lower bound of the *objective*-WA for the given *subjective*-WA is also calculated from the joint distribution. Finally, a graphical model is build that predicts the user satisfactory index from user profiles and reduces the distribution uncertainty of user satisfaction by 13% of its variance.

Key words　System Evaluation, Performance Measure, User Profile, Field Test

1. はじめに

音声対話システムの設計において性能予測を行うことは重要な課題であり，音声認識精度が性能指標としてよく用いられている．しかし，発話誤りに対するユーザの感じ方は対話の流れの影響を受けるため，システムに対するユーザ満足度は音声認識精度の単純な関数とはならない．対話システムの性能を特徴づける一般的な指標を構築することは，現在でも重要かつ困難な課題である．

従来研究として，音声認識システムの評価は1970年代頃から盛んに行われており [1]，ユーザビリティの観点から音声対話システムのインタフェースとしての評価も行われている [2], [3]．DARPA Communicator project [4], [5] では，複数の旅行計画システムを構築，運用し多数の対話データを収集し，異なったシステム間の比較に用いるための性能指標構築に関する研究が行われている．Walker ら [6] は音声対話システムのユーザ満足度を評価するためのフレームワークとして PARADISE を提案し，その評価を行っている．PARADISE ではユーザ満足度を幾つかの指標からなる関数としてとらえて，予測モデルを導いているが，未知のユーザに対するシステム性能の推測については議論されていない．また，SNR や話速などのユーザ環境条件と音声認識性能の関係についてはこれまでにも様々な研究が行われており，(例えば, [7]〜[9] など)，話速に応じたモデル適応やデコーディング手法の切り替えによる性能向上 [10] や SNR に応じたモデル切り替えによる性能向上 [11] などが報告されている．しかし，ユーザ主観の音声認識性能の推測に関するモデルは十分に議論されていないため，これらの音声認識性能向上が音声対話システムにおいてユーザ満足度をどれだけ向上させるのかを議論することが難しい．

本論文では楽曲検索システムの運用によって得られた対話データとアンケート結果を利用して，ユーザプロファイルとユーザ満足度の関係性を体感認識精度を介して推測モデルを構築する．体感認識精度とは，システムを利用したユーザが音声認識精度を主観的な指標として評価した値である．体感認識精度がユーザ満足度を示す心理尺度であると仮定し，楽曲検索システムを利用したフィールドテストにおけるアンケートによって，5段階のユーザ満足度指標とともに体感認識精度を収集した．

本研究では客観認識精度と体感認識精度とユーザ満足度をの関係性について図1に示すモデル構造を仮定する．このモデル構造では，客観認識精度はユーザの音響環境の影響を強く受けている．客観認識精度は体感認識精度の関係は認識誤りに対するユーザの敏感さの関数としてとらえられ，その関数はユーザプロファイルに関連している．そして，ユーザ満足度は体感認識精度として表現される．このモデルを用いることで陽に音声認識精度を計測することなくユーザに関する情報からのみユーザ満足度を推測する．

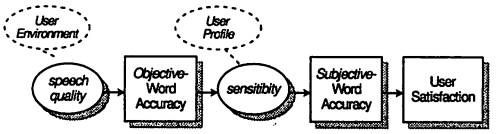

図1 ユーザ満足度のモデル
Fig. 1 User satisfaction model.

本論文は以下の節から構成される．第2節ではフィールドテストとデータ収集について概要と予備分析結果を述べる．第3節では，客観認識精度と体感認識精度の同時分布に関して，1) ユーザの分類と 2) 客観認識精度が与えられた場合の体感認識精度の下限について検討を行う．第4節では，ユーザ満足度に関するグラフィカルモデルを構築しユーザプロファイルからの満足度の推測精度について評価を行う．最後に第5節で本論文の結論を述べる．

2. 楽曲検索音声対話システムを用いたフィールドテスト

音声対話システム *MusicNavi* を用いたフィールドテストによってデータを収集した．*MusicNavi* は音声対話を通じて PC 内の楽曲を検索し再生することができるシステム [12] で，利用者の PC 内にダウンロード，インストールが行われ，インターネットを通じてサーバプログラムと通信が行われる．ユーザの PC 内に存在する楽曲のアーティスト名，アルバム名，楽曲名などに関係した単語リストのみを含んだ音声認識用の辞書を作成するために，*MusicNavi* は PC 内に含まれる楽曲ファイルのアーティスト名，アルバム名，楽曲名をサーバプログラムに送信し認識単語辞書を作成する．この機能によって不要な単語エントリを含まない認識単語辞書が作られる．このサーバ/クライアント構成を元にデータ収集機能も実装されている．

2.1 フィールドテストの概要

まず，被験者は WWW ページ上で事前アンケートの回答を行った．次に，1) 5曲以上を再生する，2) 20回以上システムとの対話を行うか 40 分以上システムを利用し続ける，という二つの条件を満たすまでシステムを利用した．最後に，WWW ページ上で事後アンケートに回答を行った．

以上の手順によって 500 名からなる音声対話データと

表 1 アンケート項目

Table 1 The items collected through the questionnaire

Age	
Gender	
Marital Status	
Address	47 prefectures
Job	14 classes
Experience	8 boolean variables
Noise Source	4 boolean variables
Microphone Type	text
Laudspeaker Type	text
Understanding	4 metrics of 5 classes
Quality of Dialogue	2 metrics of 5 classes
Subjective-WA	integer (0 to 100)
Satisfaction	5 classes
Good Impression	text
Bad Impression	text

アンケート回答を収集した．ただし，ユーザに対する教示として WWW 上に設置したシステム利用方法に関する説明文章を読むことを指示したが，収集されたデータには十分に理解していないと思われるユーザも多く含まれていた．また，劣悪な認識環境において，システムが誤認識と誤動作を繰り返したが最終的には上記 2 条件を満たしてしまったというユーザも含まれていた．そこで本研究では書き起こし内容を元に計算した認識精度が 60%以下のユーザは，利用方法を十分に理解していないか認識環境が劣悪であったと見なして，該当ユーザを除く 178 名の被験者のデータを利用して評価を行う．

アンケートの回答項目を表 1 に示す．アンケートの回答項目の「*MusicNavi* はあなたの声をどの程度間違えていると感じましたか？」という質問で 0%から 100%の間の回答を得ており，100 からこの数字を減ずることで体感認識精度の値を得た．一方でユーザ満足度は 5 段階の評価（5:満足，4:どちらかといえば満足，3:どちらともいえない，2:どちらかといえば不満，1:不満）として回答を得た．

2.2 データの予備分析

インターネットを経由してサーバシステムに送信された発話から，音声信号の音響・音韻的特徴として，信号対雑音比 (SNR) [13] と話速 [14] を発話ラベリングなしで推測をおこなった．図 2 に SNR，話速それぞれの発話数頻度分布と各区間内の発話毎に算出した平均単語認識精度を示す．SNR の向上するにつれて認識精度向上の傾向が見られる．一方，話速については 0.5[syllable/sec] から 1.5[syllable/sec] の間でやや認識精度が高くなっている．この話速の区間に含まれている発話のうち約 60%が

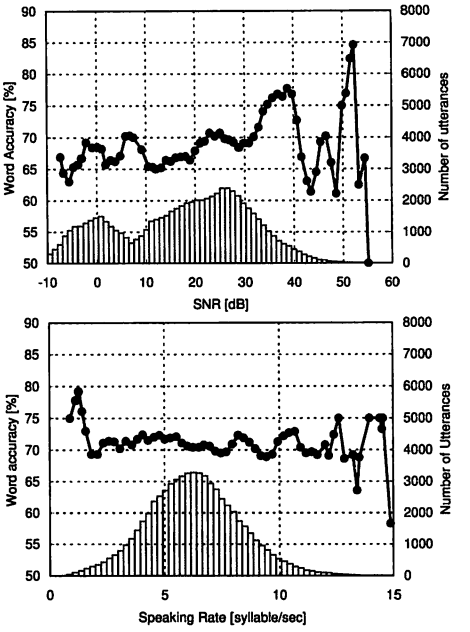

図 2 SNR(上) と話速 (下) の平均単語認識率．棒グラフは発話数を表す．

Fig. 2 Average word accuracies for bins of the SNR(top) and speaking rate(bottom). Box plots shows the number of utterances.

「はい」や「いいえ」であり，認識システムとして認識しやすい発話が多かったためと考えられる．

平均単語認識精度は 73.1% で，平均満足度は 3.51 であった．また，単語認識精度が 80%以上のユーザは全体の 30.3%，満足度が 4 以上と回答したユーザは全体の 44.3%であった．

ユーザ満足度を客観認識精度もしくは体感認識精度の関数として表した図を図 3 に示す．図より，ユーザ満足度との関係は客観認識精度よりも体感認識精度の方が高いことが読み取れる．

3. 体感認識精度と客観認識精度

本節では，体感認識精度と客観認識精度について，図 4 に示す同時確率分布に基づいてさらなる分析を行う．

3.1 ユーザクラスタリング

同時確率分布に対して混合ガウス分布 (GMM) をフィッティングした図を図 5 に示す．重み付けられた 4 つのガウス分布の組み合わせによって同時確率分布が表現されている．各混合要素は認識誤りに対して鋭敏であるか寛容であるかを特徴づけたユーザグループを表現していると考えられる．この図の場合では，第 1 のグループ (G_1) は客観認識精度によらず高い体感認識精度を答えたグ

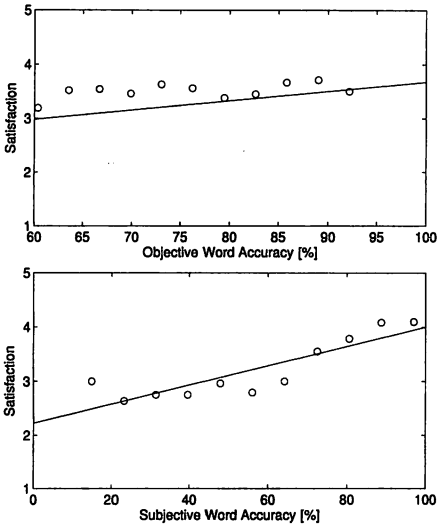

図 3 客観認識精度 (上) と体感認識精度 (下) に対するユーザ満足度指標関数. 図中の点は客観・体感認識精度それぞれの小区間におけるユーザ満足度指標の平均値を表す.

Fig. 3 Distribution of user satisfaction index as a function of objective word accuracy (top) and subjective word accuracy (bottom). Plotted user satisfaction indexes are average values of samples in certain range of subjective/objective word accuracy.

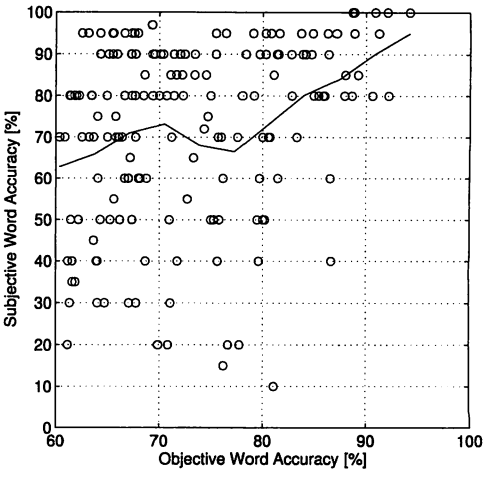

図 4 体感認識精度と客観認識精度の同時分布. 実線は与えられた客観認識精度 ± 2%の小区間で算出した体感認識精度の平均を表す.

Fig. 4 Joint distribution of *Subjective*- and *Objective*- WAs. Solid line shows average *objective*-WAs of samples at given *objective*-WA ± 2%.

ループであり, 認識誤りに寛容な評価を与えている. 第 2 のグループ (G_2) は体感認識精度と客観認識精度はほぼ比例しており客観認識精度に対する鋭敏さも高いことを示している. 第 3 のグループ (G_3) は平均客観認識精度は小さいが体感認識精度は広く分布している. これは, 低い客観認識精度のユーザにとって体感認識精度と認識性能は独立であることを示唆している. 一方で第 4 のグループ (G_4) は全体に客観認識精度よりも低いと感じている. 従ってこのグループは認識誤りに対して厳しく評価を行っている. ここで挙げた議論はデータに基づく仮説ではあるが, ユーザクラスタを認識誤りに対する鋭敏さについて特徴づけることはユーザ満足度推測という問題を考える手助けとなるだろう.

3.2 認識精度の下限

音声認識システムの設計を行う上で, 音声認識性能の品質保証を行うことは重要な課題である. 庄境ら [15] は, 認識性能によって話者を降順に並べたときの認識性能分布の屈曲点と全話者の最低性能点に着目し分析を行っている. この場合, 屈曲点は認識性能の急激な低下が見られる点での「全ユーザの x% が認識率 y% 以上を達成する」という品質保証の値を与える.

本研究では,「全ユーザの R% 以上が結果に満足する (α% 以上の体感認識精度と答える) ような客観認識精度の下限 β% はどれだけか?」という, ユーザの主観評価を考慮した品質保証問題としてとらえる. 具体的には前節で示した客観・体感認識精度の同時分布に関するパラメトリックモデル (GMM) を利用し以下の式を考えることで客観認識精度の下限 β を求める.

$$\Pr\{\alpha \leqq \mathbf{X_s} \mid \beta \leqq \mathbf{X_o}\} = \frac{\int_{\alpha}^{1}\int_{\beta}^{1} f_{X_o,X_s}(x_o,x_s)dx_o dx_s}{\int_{0}^{1}\int_{\beta}^{1} f_{X_o,X_s}(x_o,x_s)dx_o dx_s} > R \qquad (1)$$

ここで X_s と X_o はそれぞれ体感認識精度と客観認識精度を表す確率変数である.

図 6 は体感認識精度 α を 60%, 70%, 80%, 90%と設定したときに客観認識精度の下限 β に対するユーザ数の割合 R を描いた図である. 図より, 客観認識精度が 100% であっても体感認識精度が 90% 以上であると感じるユーザは 88% ほどであることが読み取れる. 体感認識精度 *alpha* が 80%, 70%, 60%以上であるユーザの割合 R が 90%以上であることを保証するためには客観認識精度の下限 β がそれぞれ 89%, 83%, 78%であることを示している.

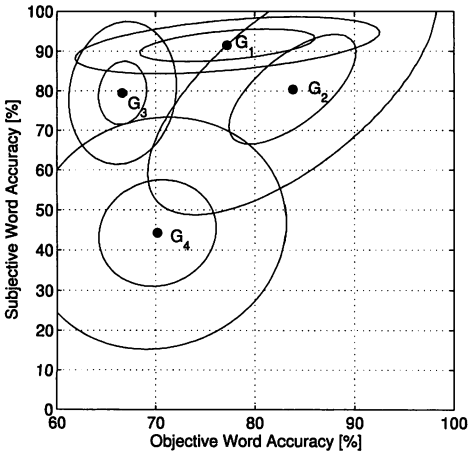

図 5 混合ガウスモデル (GMM) によってモデル化した客観認識精度と体感認識精度との同時分布の確率密度関数の等高線．各混合要素の平均値，1σ と 2σ の範囲が示されている．

Fig. 5 Contour plot of joint probability density function of *objective*-WA and *subjective*-WA modeled by a four-mixture Gaussian mixture model (GMM). Mean, one and two sigma areas are plotted for each mixture component.

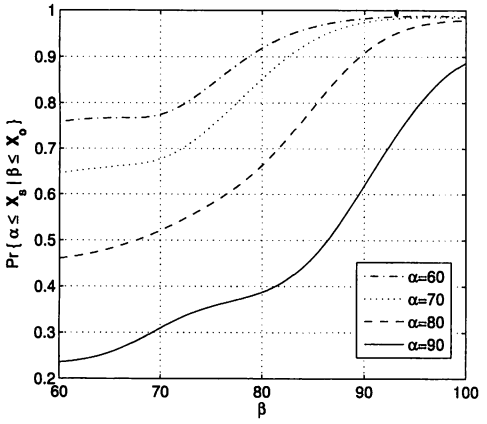

図 6 客観認識精度 β% 以上の時に体感認識精度 α% 以上と感じる母集団の割合 R%

Fig. 6 Relative population size R% that feel *subjective*-WA of α% or heigher under the β% of *objective*-WA performance.

4. ユーザ満足度の推論モデル

4.1 ユーザ満足度のグラフィカルモデル

序論で述べたように，本研究の目的はユーザの利用環境やプロファイルから図 1 に示すような依存性を持っていると仮定したユーザ満足度を推論することである．ユーザの満足度，環境，プロファイルの依存関係をグラフィカルモデルによって表現するために，図 7 に示すベイジアンネットワークを設計した．推論においては客観認識精度と体感認識精度は隠れ変数として扱われる．このモデルを使ったユーザ満足度の推論としては，ユーザ環境とプロファイルのみを与えたときに最も高い事後確率を与えるユーザ満足度のカテゴリの探索と定式化される．ただし，学習時には体感認識精度，客観認識精度，ユーザ満足度のカテゴリを利用している．

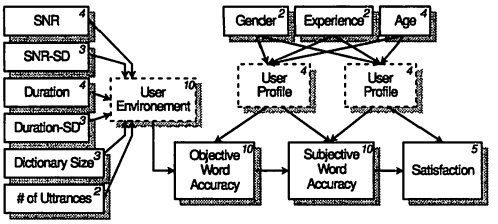

図 7 体感認識精度推論のためのネットワーク構造図

Fig. 7 Network diagram for prediction of *Subjective* Word Accuracy

このモデルは，環境条件とユーザ発話に基づく指標が認識精度の要因と仮定し，ユーザプロファイル毎に体感認識精度と客観認識精度との関係である認識精度に対する鋭敏さが存在していると仮定している．最後にユーザ満足度は推定された体感認識精度によって決定される．環境条件として扱われた変数は，SNR の平均と標準偏差，話速の平均と標準偏差，辞書の語彙数，一曲再生あたりの発話数であり，ユーザプロファイルとして扱われた変数は，性別，年齢，音声システムの利用経験である．体感・客観認識精度を含む全ての変数は離散カテゴリに符号化し，Bayes Net Toolbox for Matlab [16] を用いてベイジアンネットワークを実装した．

4.2 評　　価

前節の図 7 に示したグラフィカルモデルについて，環境条件とユーザプロファイルのみからユーザ満足度を推定することで評価を行った．被験者は 178 名で，7-fold クロスバリデーションを行った．被験者 20 名の評価セットに対して残り 158 名のデータを学習セットとして利用した．

評価実験の結果を表 2 に示す．比較対象として，ユーザ満足度に対する多変量線形回帰関数による推論結果をベースラインシステムとして表示している．表より，ベースラインシステムは 58 人 (32.6%) の満足度指標を正しく推定していたのに対して，提案手法では 80 人 (44.9%)

表 2 提案手法とベースラインの推論精度．ベースラインシステムは線形回帰関数による推論結果に基づく．

Table 2 Prediction accuracy of proposed graphical model. Baseline system is implemented as a linear regression function of *objective*-WA .

	baseline	proposed
# of correct	58	80
RMSE	1.27	1.11

の指標を正しく推定することができた．従って，体感認識精度を考慮してネットワークを構築することで，推定誤りの約 18%が改善された．満足度指標が数値的に意味があると仮定して残差の二乗平均平方根 (RMSE) を計算すると，ベースラインシステムと提案手法はそれぞれ 1.27 と 1.11 であった．この RMSE の差は全体のユーザ満足度の標準偏差のうちの 13%に相当する．言い換えれば，提案手法によってユーザ満足度の不確かさが 13%削減されたことから，ユーザ満足度の推定に関して提案手法の有効性が確認された．

5. ま と め

音声対話楽曲検索システムのフィールドテストに基づいた音声対話システムのユーザ満足度推定のための数学的モデルに関する研究を行った．客観認識精度よりも体感認識精度の方がユーザ満足度との関係が強いという結果から，満足度という心理尺度の背景にはユーザの単語認識精度に関する主観的な印象が存在することが示唆された．体感・客観認識精度の同時分布のトップダウンクラスタリングによって認識精度に対する鋭敏さに関するユーザクラスタについて検討した．同時分布から所望の体感認識精度を達成するために必要な客観認識精度の下限を算出した．最後にユーザの満足度指標を環境条件やユーザプロファイルから推測するためのグラフィカルモデルを構築した．モデルの評価実験によってユーザ満足度分布の標準偏差のうち，13%に相当する曖昧さを削減することが示された．

以上によって提案モデルの効果を実験的に示した．しかし、本論文では誤認識発話を含む対話の流れがユーザ満足度に与える影響について考慮していないため、今後さらなる検討が必要である．

謝辞 本研究の一部は文部科学省リーディングプロジェクト「e-Society 基盤ソフトウェアの総合開発」及び総務省戦略的情報通信研究開発推進制度 (SCOPE) によるものである．

文 献

[1] R.K. Moore, "Evaluating speech recognizers," IEEE Trans. ASSP, vol.25, no.2, pp.178–183, 1977.

[2] 石川泰，"音声対話システムの評価法，" 日本音響学会誌，vol.54, no.11, pp.807–811, 1999.

[3] 石川泰，"Ui 設計とユーザビリティ: 音声インタフェースの課題，" 情報処理学会研究報告，SLP-68-7(3), pp.33–34, Oct. 2007.

[4] M. Walker, J. Aberdeen, J. Bol, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, S. Seneff, D. Stallard, and S. Whittaker, "Darpa communicator dialog travel planning systems: The june 2000 data collection," Proceedings of Eurospeech 2001, Sept. 2001.

[5] M.A. Walker, A. Rudnicky, R. Prasad, J. Aberdeen, A. Potamianos, R. Passonneau, S. Roukos, G. Sanders, S. Seneff, and D. Stallard, "Darpa communicator: Cross-system results for the 2001 evaluation," Proceedings of ICSLP2002, pp.269–272, Sept. 2002.

[6] M. Walker, D. Litman, C. Kamm, and A. Abella., "PARADISE: A framework for evaluating spoken dialogue agents," Proceedings of ACL 97, 1997.

[7] D. Pearce, and H.G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," Proceedings of in ISCA ITRW ASR2000, pp.181–188, Sept. 2000.

[8] M.A. Siegler, and R.M. Stern, "On the effects of speech rate in large vocabulary speech recognition systems," Proceedings of ICASSP1995, pp.612–615, May 1995.

[9] S. Tsuge, T. Fukuda, and K. Kita, "Frame-period adaptation for speaking rate robust speech recognition," Proceedings of ICSLP2000, pp.718–721, Oct. 2000.

[10] H. Nanjo, and T. Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition," Proceedings of ICASSP02, pp.I-725–I-728, May 2002.

[11] H. Fujimura, C. Miyajima, K. Itou, K. Takeda, and F. Itakura, "Analysis of a large in-car speech corpus and its application to the multimodel ASR," ICASSP2005, pp.I-445–I-448, March 2005.

[12] 原直，宮島千代美，伊藤克亘，武田一哉，"汎用 PC 上で利用された音声対話システムによる音声収集と評価，" 情報処理学会研究報告，SLP-64-29, pp.167–172, Dec. 2006.

[13] T.H. Dat, K. Takeda, and F. Itakura, "On-line gaussian mixture modeling in the log-power domain for signal-to-noise ratio estimation and speech enhancement," Speech Communication, vol.48, pp.1515–1527, Nov. 2006.

[14] D. Wang, and S. Narayanan, "Speech rate estimation via temporal correlation and selected sub-band correlation," Proceedings of ICASSP2005, pp.413–416, March 2005.

[15] 庄境誠，加藤智之，岡本淳，"自動車運転行動中発話の分析，" 情報処理学会研究報告，SLP-68-7(2), pp.33–34, Oct. 2007.

[16] K. Murphy, "The Bayes Net Toolbox for Matlab," Computing Science and Statistics, vol.33, pp.331–350, June 2001.