GMM スーパーベクトルと SVR に基づく MtF 話者のための女声度推定*

◎王程碩, 鈴木雅之, 峯松信明 (東大), 櫻庭京子 (独協医大病院), 広瀬啓吉 (東大)

1 はじめに

GID（Gender Identity Disorder，性同一性障害）者，特に，生物学的には男性であるが，社会的には女性としての生活を望む MtF(Male to Female) の方々にとって，声の女らしさ（女声らしさ）の獲得は困難であると言われている [1]。この場合，外科治療やホルモン投与に頼るよりも，ボイスセラピーがより効果的である [1]。MtF 者は，自分の声がどの程度女声として知覚されるのかを気にするが，これを客観的に調査するには，ある程度の規模の聴取者を対象にした聴取実験が必要である。これは多大の労力を要するので，セラピストの意見に頼ることとなる。しかし当該患者の声を聞き続けると，セラピストも馴化してしまい，純粋に客観的な判断が難しくなる。

このような背景から，従来より筆者らは「女声度聴取実験」シミュレータとしての女声度推定器の開発を行なってきた。先行研究では，男声モデル，女声モデルを GMM を用いて構築し，入力音声に対してこれらのモデルから計算される尤度スコアを用いて女声度を予測していた。本研究では話者認識の分野で近年標準技術となった GMM スーパーベクトル及び識別的な回帰モデルである SVR(Support Vector Regression) を使って女声度推定の高精度化を試みたので報告する。

2 MtF 音声に対する女声度ラベリング

「女声度聴取実験」シミュレータを作成する場合，様々な MtF 音声に対して女声度スコアが付与されたコーパスが必要である。本研究では先行研究において構築されたコーパスを用いるが [2]，これは MtF 話者（19〜78 歳）111 名から収録した 142 発声（セラピー前後の音声が別途収録されている話者もいる）に対する女声度ラベリング結果（−3〜+3 の 7 段階）である。評価者は 20 代の成人男女 9 名ずつで,142 発声の評定を 2 度行なった。なお，音声は全て「ジャックと豆の木」の冒頭の 2 文（単語数 39）である。

評定者内で二回の評定間の相関を求めると，平均 0.83 であった。また，各評定者と，その評定者以外の評定者の平均値で定義される平均女声度との相関を求めると，平均 0.87 であった。

3 先行研究

筆者らの一部は [2] において，JNAS データベースの全男性話者，全女性話者を用い，MFCC（s）及び F_0（f）を特徴量として，男声モデル（$P(s|M_M^s)$, $P(f|M_M^f)$），女声モデル（$P(s|M_F^s)$，$P(f|M_F^f)$）を GMM を用いて構築し，入力音声に対する声の女性らしさ（Voice Femininity, VF）を下記で定義した。

$$VF(s,f) = \alpha \log P(s|M_F^s) + \beta \log P(f|M_F^f) + \gamma \log P(s|M_M^s) + \epsilon \log P(f|M_M^f) + C \quad (1)$$

各尤度に対する重み係数や定数項は，女声度ラベリング結果を用いて最小二乗誤差基準で推定した。女声モデルに対する係数（α，β）は正の値を，男声モデルに対する係数（γ, ϵ）は負の値を持つことになる。

4 提案手法

先行研究における男声・女声モデリングは GMM を用いた話者認識技術の一応用として位置づけられる。近年話者認識において，多数話者の音声から学習した GMM，即ち UBM（Universal Background Model）を事前に構築し，これに対して各話者の音声サンプルを用いて UBM を話者適応し（平均ベクトルのみの適応），得られた UBM-based 話者 GMM から平均ベクトルを連結したスーパーベクトルを取り出し，これを話者の特徴量とする手法が広く使われている [3]。この手法は年齢推定などの領域にも近年使われている [4]。本研究でも同様の特徴量を女声度推定に応用する。

先行研究では GMM 尤度スコアに対して線形回帰の枠組みで女声度を予測していたが，本研究では識別的な回帰モデルである，SVR を用いて女声度の予測を試みる。

5 実験方法

5.1 音声資料と実験条件

GMM-UBM の構築は，先行研究同様 JNAS を用いた。このコーパスは，毎日新聞記事と ATR 音素バランス 503 文を，306 人の話者（男女各々153 名ずつ）が読み上げた音声コーパスである（一名あたり約 150 文, コーパス全体では約 45,000 文発話）。なお，以下の実験では音声中の無音区間は，自動検出により排除して実験を行なっている。表 1 に音響分析条件を示す。

5.2 実験手順

以下の手順に従って女声度推定実験を行なった。なお，女声度の予測は 142 個の MtF 発声に対して leave-one-out cross-validation を行い，推定された 142 個の予測値と聴取実験結果とを比較する。

- JNAS 中の全発話を用いて GMM-UBM を構築する。MFCC，F_0 に対して別個に構築する。

*Voice femininity estimation for MtF patients using GMM supervectors and SVR, by C. Wang, M. Suzuki, N. Minematsu, K. Sakugaba, and K. Hirose

Table 1　実験条件

サンプリング	16bit/16kHz
フレーム長，周期	25ms, 10ms
プリエンファシス	$1 - 0.97z^{-1}$
特徴パラメータ	12MFCC + 12ΔMFCC +Δ パワー，及び，対数 F_0
GMM	対角分散共分散行列，混合数 64

Table 2　従来手法と提案手法による女声度推定結果

	相関	二乗誤差
GMM-LR	0.85	0.70
SV-SVR	0.89	0.59

- 各 MtF 話者の音声（39 単語）を用いて GMM-UBM を話者適応（MAP 適応）し，各 MtF 話者の GMM を，MFCC, F_0 に対して得る。
- 64 個のガウス分布からスーパーベクトルを得る。MFCC が 25 次元，F0 が 1 次元であり，各々の GMM の混合数は 64 であるので，スーパーベクトルの次元数は 1664 となる。
- このスーペーベクトルを話者特徴量として，SVR の枠組みで女声度を予測する。

以上の手順により得られた結果を，同一条件で行なわれた従来手法の結果と比較する。

6　実験結果

従来手法（GMM-LR）と提案手法（SV-SVR）によって予測された女声度と，聴取実験により得られた女声度との相関係数及び平均二乗誤差を表 2 に示す。提案手法は，相関，二乗誤差の両者において先行研究よりも高い精度で女声度を推定できることが分かる。また聴取実験より，評定者内相関は平均 0.83，各評定者とその人以外の評定者の平均との相関は平均 0.87 であったが，上記の結果は，SV-SVR による女声度推定器は，「もう一人の評定者」として認定するに足る精度を有していると言える。

従来手法，提案手法による女声度推定結果と，聴取実験結果との散布図を図 1，図 2 に示す。図中の直線は散布図の直線近似ではなく，$y = x$（正解直線）をプロットしたものである。これらの図からも提案手法の方が，正解直線の周りにデータがより集中していることが分かる。

7　まとめと今後の課題

本稿では，性別の移行を希望する MtF 者に対するボイスセラピーの技術的支援を目的として，入力音声がどの程度女声として聞こえるのかを自動推定する「女声度推定器」の高精度化を検討した。先行研究では，GMM による男声・女声モデルを構築し，尤度スコアの線形回帰で女声度を推定していた。本研究ではこれを，特徴量として GMM スーパーベクトルを利用し，また回帰の枠組みを SVR に変更することで，聴取実験との相関値及び平均二乗誤差において，より高精度な結果を得ることができた。

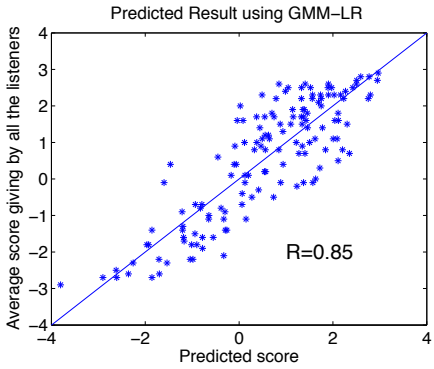

Fig. 1　GMM-LR による女声度と聴取実験結果

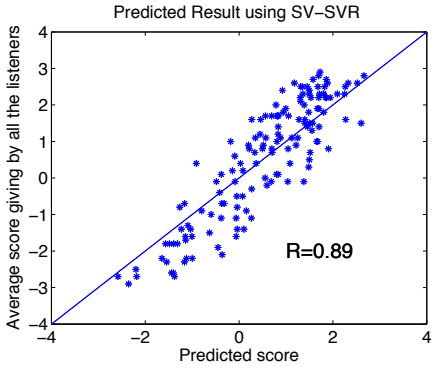

Fig. 2　SV-SVR による女声度と聴取実験結果

実際の臨床では，音高の制御，音色の制御，更には話し方の制御がセラピーの焦点となっている [1]。本研究では話し方については一切触れておらず，これについても検討する予定である。

参考文献

[1] 櫻庭京子他，“性同一性障害者（MtF）の音声に対する知覚的性別の自動推定”，信学技報 SP2005-189, 29–34, 2006.

[2] N. Minematsu *et al.*, “Development of a femininity estimator for voice therapy of gender identity disorder clients,” in Speaker Classification II, LNAI4441, 22–33, Springer, 2007.

[3] T. Kinnunena *et al.*, “An overview of text-independent speaker recognition: from features to supervectors,” Speech Communication, 52, 12–40, 2010.

[4] W.Toshiya *et al.*, “Investigations of features and estimators for speech-based age estimation,” in Proc. of the Second APSIPA Annual Summit and Conference, 470–473, 2010.