IMPROVED ESTIMATION OF FEMININITY USING GMM SUPERVECTORS AND SVR FOR VOICE THERAPY OF GENDER IDENTITY DISORDER CLIENTS

Chengshuo Wang†, Masayuki Suzuki†, Nobuaki Minematsu†, Kyoko Sakuraba‡, Keikichi Hirose†,

† The University of Tokyo, Tokyo, Japan
‡ Hospital of Dokkyo Medical University, Saitama, Japan

ABSTRACT

This paper proposes a new method of estimating perceptual femininity (PF) of an input utterance using Gaussian Mixture Model (GMM) supervectors and support vector regression (SVR). The method is used to develop a femininity estimation tool, which is introduced to voice therapy of Gender Identity Disorder (GID) clients, especially MtF (Male to Female) transsexuals. In our previous study [1], we developed a PF estimator, where a male GMM and a female GMM of spectral features and those of pitch features were built and their likelihood scores of an input utterance were combined by linear regression to estimate PF. In this work, inspired by recent speaker recognition models [2], we replace the four likelihood scores from the four GMMs with supervectors composed by a spectral GMM and a pitch GMM estimated from an input utterance. Further, instead of simple linear regression, we introduce SVR, which is discriminative linear regression. Experiments using an MtF speech corpus show that the proposed method improves correlation between human and machine scores of PF and also reduces squared prediction error.

Index Terms— Femininity, GID, MtF, supervector, SVR

1. INTRODUCTION

We can find a large number of studies attempting to apply advanced acoustic technologies to medical treatment, such as cochlea implants and artificial larynxes. Recently, advanced speech technologies have also been applied to medical treatment [3], such as on-line screening test of laryngeal cancer [4]. The present paper examines the use of speech technologies for voice therapy for GID clients.

A GID individual is one who strongly believes that his or her true psychological gender identity is not his or her biological or physical gender, i.e., sex. In most of the cases, GID individuals live for years trying to conform to the social role required by their biological gender, but eventually seek medical and surgical help as well as other forms of therapy in order to achieve the physical characteristics and the social role of the gender which they feel to be their true one. In both cases of FtMs (Female-to-Male) and MtFs, many of them take hormone treatment to make physical change of their bodies and the treatment is certainly effective for both cases. However, it is known that the hormone treatment brings about sufficient change of the voice quality only for FtMs. Considering that the voice quality is controlled by the physical conditions of the articulators, the vocal folds and the vocal tract are presumed to retain their pretreatment size and shape in the case of MtFs. To overcome this hardship and mainly to shift up the baseline F_0 range, some MtFs take surgical treatment. Although the F_0 range is certainly raised in the new voice, as far as the fourth author knows, it is a pity that the naturalness is decreased in the new voice instead. Further, many clinical papers and engineering papers on speech synthesis claim that raising the F_0 range alone does not produce good femininity [5, 6, 7]. Since the shape of the vocal tract has a strong effect on the voice quality, good control of the articulators has to be learned to achieve good femininity. Considering small effects of hormone treatments and surgical treatments on MtF clients, we can say that the most effective and least risky method to obtain good femininity is taking voice training from speech therapists or pathologists with good knowledge of GID.

2. WHY FEMININITY ESTIMATOR?

In the typical therapy, the following three methods are used. 1) raising the baseline F_0 range, 2) changing the baseline voice quality, and 3) enhancing F_0 dynamics to produce an exaggerated intonation pattern. One of the difficulties in the voice therapy lies not on a client's side but on a therapist's side, i.e., accurate and objective evaluation of the client's voice. It is often said that as synthetic speech samples are presented repeatedly, even expert speech engineers tend to perceive better naturalness in the samples, known as habituation effect. This is the case with good therapists. To avoid this effect and evaluate the femininity unbiasedly, listening tests with novice listeners are desirable. GID clients are also eager to know how they sound to novice listeners. But listening tests take unignorable cost and time. If we can develop a femininity estimator as listening test simulator, it is expected to help both therapists and clients in a therapy.

Among the above three methods, the first two ones focus on static acoustic properties and the last one deals with dynamic F_0 control. The dynamic control of F_0 for various

speaking styles is still a very challenging task in speech synthesis studies and, therefore, we only focus on the femininity created by the F_0 range and the voice quality.

3. OUR OLD ESTIMATOR

3.1. Method of femininity estimation

In our previous study [1], we built a male model and a female model separately for spectral features and pitch features. The four models were trained as GMMs using a large speech corpus containing biologically male speakers' utterances and biologically female speakers' ones. Here, the male model of spectrum is denoted as M_M^s and its female version is M_F^s. Similarly, the pitch models for the two genders are M_M^p and M_F^p. By using these models and likelihood ratio [8], we can introduce spectrum-based femininity F^s of utterance o as

$$F^s(o) = \log \mathcal{L}(o|M_F^s) - \log \mathcal{L}(o|M_M^s). \quad (1)$$

Pitch-based femininity can also be defined as

$$F^p(o) = \log \mathcal{L}(o|M_F^p) - \log \mathcal{L}(o|M_M^p). \quad (2)$$

Flexible combination of the above scores is linear regression.

$$F(o) = \alpha \log \mathcal{L}(o|M_F^s) + \beta \log \mathcal{L}(o|M_M^s) + \gamma \log \mathcal{L}(o|M_F^p) + \varepsilon \log \mathcal{L}(o|M_M^p) + C, \quad (3)$$

where parameters of α, β, γ, ε and C can be estimated if we have a training corpus of GID utterances labeled with perceptual femininity (PF) scores. To estimate the values of these parameters, we built an MtF speech corpus and assigned a PF score to each utterance through listening tests.

3.2. MtF speech corpus labeled with PF scores

A speech corpus of 111 Japanese MtF speakers was built, some of whom sounded very feminine and others sounded less feminine. Their age varied from 19 to 78. Each speaker read the beginning two sentences of "Jack and the beanstalk" with natural speaking rate. The two sentences had 39 words. Some clients joined the recording twice; before and after voice therapies. Then, the total number of recordings was 142. For reference, 17 biologically female Japanese read the same sentences. All the speech samples were digitized through 16 bit and 16 kHz AD conversion.

All the sentence utterances were randomly presented to 9 male and 9 female adult Japanese subjects through headphones. All the subjects were in their 20s with normal hearing and they were very unfamiliar with GID. The subjects were asked to judge subjectively how feminine each utterance sounded and write down a score using a 7-degree scale, where $+3$ corresponded to the most feminine and -3 did to the most masculine. Some speech samples of the 17 biological female speakers were used as dummy samples. Every subject joined the test twice and 36 femininity judgments by 18 subjects were obtained for each utterance. Figure 1 shows the histogram of the averaged PF scores of the individual MtF utterances. Although some utterances still sounded rather masculine, a good variance of PF was found in the corpus. The averaged PF of the 7 biological female speakers was 2.74.

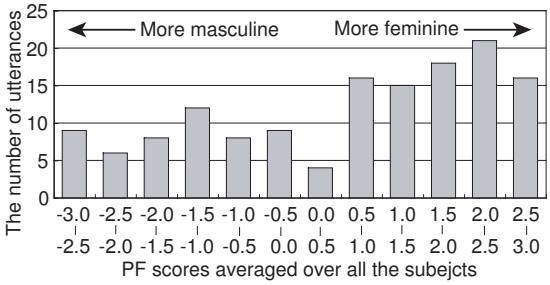

Fig. 1. Histogram of the averaged PF scores

Using this corpus, the five parameters in Equation 3 were estimated so that the squared prediction error was minimized.

4. OUR NEW ESTIMATOR

4.1. Features for femininity estimation

Our old estimator was developed using classical GMM-based speaker recognition technology. In the classical technology, utterances of a speaker was modeled as GMM and N GMMs were trained for N reference speakers. For a new utterance, the speaker identity of the GMM that produced the highest likelihood score was used as a result of speaker recognition. Recently, the methodology of speaker recognition has been shifted from MFCC + GMM to supervectors + SVM or i-Vector approach [2]. Inspired by this change, in this study, we test supervector-based features under the task of femininity estimation. First, using the utterances of a large speech corpus, which is not the MtF corpus, we build the universal background model as GMM (UBM-GMM). Then, the UBM-GMM is adapted through MAP adaptation into each speaker of the MtF corpus, where only mean vectors are modified in the adaptation process. Generally speaking, the mixture indices of two GMMs of the same number of mixtures have no direct correspondence. By using the UBM-GMM as background model, however, the GMM of each speaker comes to have the same structure among all the speakers and index i of a GMM corresponds directly to index i of another GMM. If we build a supervector from the GMM of an MtF speaker, which is obtained by joining all the mean vectors, the supervector is expected to characterize that MtF speaker well.

In our previous study [1], two different kinds of GMMs were trained for each gender, which were a spectrum-based model and a pitch-based model. Similarly in this work, we build two UBM-GMMs, i.e., spectrum UBM-GMM and pitch UBM-GMM and they are adapted to each MtF speaker in the corpus. Then, we can get two supervectors, spectrum super-

Table 1. Acoustic analysis conditions

sampling	16bit / 16kHz
frame and shift	25 ms / 10 ms
acoustic features	12MFCC+12ΔMFCC+ΔP, $\log F_0$
GMM	64 mixtures with diagonal matrices
dim. of supervector	1,664 (=25×64 + 1×64)

vector and pitch supervector, for each MtF speaker and the final supervector is obtained by joining the two vectors. In stead of using the likelihood scores as in Equation 3, the elements in the final supervector are used as predictor variables for regression to perceptual femininity scores.

4.2. Method of femininity estimation

For regression to perceptual femininity, simple linear regression of four scores was used in [1] (See Equation 3). When we use supervectors for regression, the number of parameters used in regression becomes remarkably large, which easily leads to the well-known overfitting problem. To avoid this, in this work, we test SVR [9] to estimate PF.

Generally speaking, the goal of regression is to compute an estimate value $\hat{y_i}$ from n dimensional feature vector $\boldsymbol{x_i}$. In SVR, unlike simple linear regression, the regression model is obtained so that $\hat{y_i}$ deviates by ϵ at most from its original value y_i. This leads to the following:

$$\hat{y_i} = \boldsymbol{w}\boldsymbol{x_i} + b. \tag{4}$$

Here, $\boldsymbol{w}$ and b are obtained by solving the problems:

$$-\epsilon \leq y_i - (\boldsymbol{w}\boldsymbol{x_i} + b) \leq \epsilon \tag{5}$$

To allow deviations larger than ϵ, slack variables ξ_i and ξ_i^* (> 0) are introduced and Equation 5 is changed to

$$-(\epsilon + \xi_i^*) \leq y_i - (\boldsymbol{w}\boldsymbol{x_i} + b) \leq (\epsilon + \xi_i). \tag{6}$$

Finally, the SVR parameters are obtained by solving the following problems:

$$\text{minimize} \quad \frac{1}{2}|\boldsymbol{w}|^2 + C\sum_i(\xi_i + \xi_i^*) \tag{7}$$

$$\text{subject to} \quad -(\epsilon + \xi_i^*) \leq y_i - (\boldsymbol{w}\boldsymbol{x_i} + b) \leq (\epsilon + \xi_i) \tag{8}$$

5. EXPERIMENTS

5.1. Speech material

JNAS (Japanese News Article Sentences) speech database [10] is used to train four GMMs of M_M^s, M_F^s, M_M^p, and M_F^p for our old estimator and two gender-independent UBM-GMMs for our new estimator. In the corpus, 153 males and 153 females read out about 150 sentences. Before training the GMMs, silent segments are automatically detected and removed. Table 1 shows the acoustic analysis conditions. The number of mixtures of a GMM is fixed to be 64 in this work.

5.2. Agreement among subjects in rating PF

Correlation analysis is done by using the results of subjective rating of PF for the MtF speech corpus. As each subject rated each utterance twice, we calculate the correlation between the two trials within each subject. The averaged correlation over the subjects is 0.83. Next, we calculate the correlation between subject i's averaged scores over the two trials and the averaged scores over all the other subjects. The averaged correlation obtained by assigning each subject to subject i is 0.87. Considering this result, if the correlation between the outputs of our estimator and the averaged PF scores over all the subjects is close to 0.87, we can say that our estimator can resemble human perception of femininity very well.

5.3. Results and discussion

By using speaker-based leave-one-out cross-validation on the MtF corpus, we test our old estimator and our new estimator with three different acoustic feature sets: 1) pitch only, 2) spectrum only, and 3) both of them. In each test, the correlation and the average of squared prediction errors are calculated as performance metric for comparison

Figure 2 shows the correlations of four cases: a) new pitch-based estimator, b) new spectrum-based estimator, c) new estimator using both features, d) old estimator using both features. Table 2 summarizes the results of all the six cases in terms of correlation and prediction error.

From the figures and the table, it is clear that the estimation performance is increased by using both pitch and spectrum. As described in Section 1, this is very reasonable because of the claims from speech therapists and speech engineers in [5, 6, 7]. In the proposed method using both features, the correlation reaches 0.90 and we can say that our estimator is a very good simulator of human judgment of femininity.

It is somewhat surprising to us that femininity estimation by using only pitch works rather well ($R = 0.8$). This indicates that, in voice therapy for MtF clients, modification of manner of pitch control and that of manner of spectrum (timbre) control are done somewhat synchronously. So, even if we focus on either feature only, the estimation performance is reasonably good. In the MtF corpus, however, the voices of some clients are found to be "falsetto" voices, where the pitch is high enough but the timbre is rather masculine and the PF scores for these voices are very low. This is a typical case of mismatch between pitch control and timbre control and this is one of the reasons why we consider that both features are needed in femininity estimation.

From Table 2, we can say that the prediction error is reduced significantly by the proposed method. We consider that this is largely due to the training algorithm of SVR, where $\boldsymbol{w}$ is estimated with some constraints so that the squared error in each sample should be less than $\epsilon + \xi$.

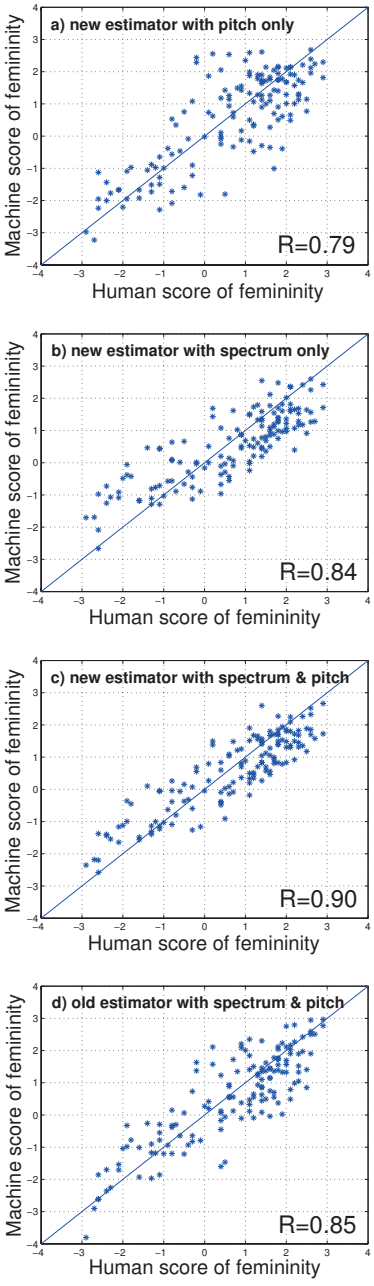

Fig. 2. Correlations between human and machine scores

6. CONCLUSIONS

In this work, we proposed a new method of femininity estimation by using GMM-based supervectors and discriminative regression of SVR. Compared to our old estimator, which is based on classical speaker recognition model of MFCC + GMM, improvement was found in two metrics of correlation and prediction error. Especially, prediction error reduction was found to be very large. As future work, i-Vector approach will be introduced and we consider that the use of HMM-based supervectors is interesting because it may be able to capture the temporal aspect of utterances better than GMM-based ones. Currently, our new estimator is actually used in the fourth author's voice therapy for MtF clients.

Table 2. Correlations and averaged prediction errors

estimator	feature	correlation	error
old	pitch	0.80	0.95
	spectrum	0.74	1.17
	both	0.85	0.70
new	pitch	0.79	1.01
	spectrum	0.84	0.83
	both	**0.90**	**0.59**
human agreement		0.87	—

7. REFERENCES

[1] N. Minematsu, *et al.*, "Development of a femininity estimator using speaker recognition techniques for voice therapy of gender identity disorder clients," Proc. ICASSP, 297–300, 2007

[2] T. Kinnunen, *et al.*, "An overview of text-independent speaker recognition: from features to supervectors," Speech Communication, 52, 1, 12–40, 2010

[3] E. Nöth *et al.*, "Medical speech processing – pathologies, treatment assistance, clinical trials," Tutorial of INTERSPEECH2010.

[4] H. Mori *et al.*, "Internet-based acoustic voice evaluation system for screening of laryngeal cancer," J. Acoustic Society of Japan, 62, 3, 193–198, 2006 (in Japanese)

[5] R. C. Bralley *et al.*, "Evaluation of vocal pitch in male transsexuals," J. Communication Disorder, 11, 443–449, 1978

[6] L. E. Spencer, "Speech characteristics of MtF transsexuals: a perceptual and acoustic study," Folia phoniat., 40, 31–42, 1988

[7] K. H. Mount *et al.*, "Changing the vocal characteristics of a postoperative transsexual patient: a longitudinal study," J. Communication Disorder, 21, 229–238, 1998

[8] H. Franco *et al.*, "Automatic detection of phone-level mispronunciation for language learning," Proc. EUROSPEECH, 1999

[9] A. Smola *et al.*, "A tutorial on support vector regression," Royal Holloway University of London, Tech. Rep., 1998, nC2-TR-1998-030

[10] http://research.nii.ac.jp/src/en/JNAS.html