The Effect of Fuzzy Training Targets on Voice Quality Classification

Stefan Scherer[1,3], John Kane[2], Christer Gobl[2], and Friedhelm Schwenker[3]

[1] Institute of Creative Technologies, University of Southern California, United States
[2] Phonetics and Speech Laboratory, Trinity College Dublin, Ireland
[3] Institute of Neural Information Processing, Ulm University, Germany

Abstract. The dynamic use of voice qualities in spoken language can reveal useful information on a speaker's attitude, mood and affective states. This information may be desirable for a range of speech technology applications. However, annotation of voice quality may frequently be inconsistent across raters. But whom should one trust or is the truth somewhere in between? The current study looks first to describe a voice quality feature set that is suitable for differentiating voice qualities on a tense to breathy dimension. These features are used as inputs to a fuzzy-input fuzzy-output support vector machine (F^2SVM) algorithm, to automatically classify the voice qualities. The F^2SVM is compared to standard approaches and shows promising results. Performances for cross validation, leave one speaker out, and cross corpus experiments of around 90% are achieved.

1 Introduction

The term voice quality (henceforth VQ) refers to the timbre or coloring of a speaker's voice. For an individual speaker their VQ is composed of longer term settings of the vocal system combined with dynamic shifts in the system for communicative purposes [1]. A speaker's VQ is an important feature of paralinguistic signaling in speech and can provide the listener with information pertaining to the speaker's affective state [2]. For instance, breathy voice has been generally observed in association with intimacy and familiarity [1]. Tense voice on the other hand has been reported in more active affective states, e.g., anger and happiness [3].

It has been widely observed that VQ can provide useful insights into the intentions and mood of the speaker, and indeed VQ features have also been utilized in order to improve emotion classification [4]. It follows that robust characterization of voice qualities may be desirable for both input (i.e. recognition) and output (i.e. synthesis) ends of speech applications.

The purpose of this study is to put forward a framework for identifying voice qualities on a tense to breathy continuum. Few studies have focused on automatic classification of voice qualities using combinations of features. The main work in this area has been done in the domain of pathological voice types [5]. Hidden Markov models (HMMs) and a regression approach were employed

to categorize speech signals, that were generally of a longer duration than the signals in this study. The task was to match the annotated degree (form 0 to 4) on three VQ scales, namely breathiness, roughness and deviance. Accuracies of about 50% within each of the three scales could be achieved in the study. However, the speech material used was mainly pathological voices which weakens its comparability with the present study. In this study we investigate fuzzy-input fuzzy-output support vector machine (F^2SVM) introduced in [6] for the task at hand and compare their performance to standard approaches, that do not make use of the fuzzy membership assignments provided by human experts.

The remainder of the paper is organized as follows: In Sec. 2 the utilized VQ features for the classification experiments are introduced. Along with the introduction of the speech dataset used, Sec. 3 introduces the annotations by experts, which are later used as training targets for the fuzzy classification experiments. Section 4 then briefly introduces the utilized F^2SVM, which compete against two standard non-fuzzy approaches. In Sec. 5 the results for the experiments are reported and discussed in Sec. 6. Finally, Sect. 7 concludes the paper and provides an outlook.

2 Voice quality features

The VQ features used in the current study were selected on the basis of being stated to be able to characterize voice qualities across the breathy to tense dimension. The features described in Sects. 2.1 - 2.5 describe aspects of the glottal source signal, which is derived using automatic inverse filtering. This is done using the pitch synchronous automatic inverse filtering (PSIAIF) method described in [7], with f_0 extracted using ESPS/*waves+* software package. The features described in Sects. 2.1 to 2.5 can then be measured on the output signal from this method. However, as the output of this method can sometimes contain uncancelled formant oscillations, which can negatively impact the features, we use one further feature which is measured without the use of inverse filtering (see Sec. 2.6).

2.1 Time based LF model parameters (*Ra*,*Rk*,*Rg*,*EE*)

The most commonly used glottal source model is the Liljencrants-Fant (LF) model [8]. It is a five parameter (including f_0) model of differentiated glottal flow (i.e. the residual signal after inverse filtering if lip radiation has not been compensated for). The model has two components. The first component, the open phase, is a sinusoid function that increases exponentially and the second component is an exponential function which models the return phase.

$$U'_g(t) = \begin{cases} E_0 e^{\alpha t} sin\omega_g t & for \;\; t_o \leq t \leq t_e \\ \frac{-EE}{\epsilon T_a}(e^{-\epsilon(t-t_e)} - e^{-\epsilon T_b}) & for \;\; t_e < t < t_c \\ 0 & for \;\; t_c \leq t \leq T_0 \end{cases} \quad (1)$$

The model is generated using the time-points shown in Fig. 1, along with the parameters E_0, α and ϵ which are solved implicitly to ensure area balance

above and below the zero-line (see [8] for full details of the model). The model can be fit to an inverse filtered speech signal in the time domain using the method described in [9]. From the given model configuration, one can obtain four parameters: the amplitude parameter EE (shown in Fig. 1) and three shape parameters; Rg, Rk and Ra (see Eqs. 2). These parameters have been shown to be suitable for characterizing a range of voice qualities including breathiness and tenseness [10]

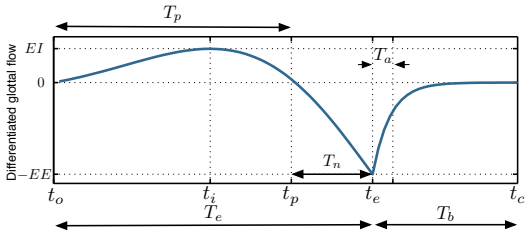

Fig. 1. Example LF model pulse for the glottal flow (above) and the differentiated glottal flow (below)

$$Rg = \frac{1}{2T_p \cdot f_0}; \quad Rk = \frac{T_e - T_p}{T_p}; \quad Ra = T_a \cdot f_0 \tag{2}$$

2.2 LF parameters frequency domain (Ra_f,Rk_f,Rg_f,EE_f)

An alternative approach for deriving LF model parameters in the frequency domain was initially described in [11] and has since been further developed. The method involves using the amplitudes of the first eight harmonics from the glottal source spectrum as inputs to a feed forward neural network, previously trained on a large volume of LF model configurations and their spectral information, in order to derive the four parameters stated above. Harmonic amplitudes are measured from the narrowband spectrum, obtained by taking a three pulse length segment of the glottal source signal, centered on a GCI, and windowed using a Hamming window. This approach was developed in order to improve the robustness of the extracted parameters to the presence of noise and phase distortion.

2.3 Normalized amplitude quotient (NAQ)

The normalized amplitude quotient (NAQ) parameter was introduced as a global glottal source parameter capable of differentiating breathy to tense voice qualities [12]. NAQ was shown to be more robust to noise disturbances than time based parameters and has, as a result, been used widely in applied work on VQ.

2.4 $\Delta H_{1,2}$

The difference in amplitude levels (in dB) between the first two harmonics of the narrowband glottal source spectrum ($\Delta H_{1,2}$) is thought to be a rough correlate of

the open quotient parameter and hence useful at discriminating breathy to tense voice qualities [13]. The narrowband spectrum is obtained by using three-pulse length sections, centered on a GCI and using a Hamming window.

2.5 Voice quality spectral gradients (OQG, GOG, SKG, RCG)

Lugger and Yang [14] described a set of spectral gradient parameters for characterizing voice qualities from glottal source signals. The parameters, comprising Open Quotient Gradient (OQG), Glottal Opening Gradient (GOG), Skewness Gradient (SKG), and Rate of Closure Gradient (RCG), were stated by the authors to be strongly correlated with typical glottal pulse shape parameters. They have been shown to be useful in the classification of voice qualities, gender and emotion, as well as relatively robust [14].

2.6 PeakSlope

A final feature is included which has recently been shown [15] to be able to separate breathy to tense voice qualities without the use of inverse filtering.

$$g(t) = -\cos(2\pi f_n t) \cdot \exp(-\frac{t^2}{2\tau^2}) \tag{3}$$

The speech segment $s(t)$ is convolved with $g(\frac{t}{s_i})$, where $s_i = 2^i$ and $i =$ 0,1,2,....,5. This essentially is the application of an octave-band filter bank with the centre frequencies being: 8 kHz, 4 kHz, 2 kHz, 1 kHz, 500 Hz and 250 Hz. Then the local maximum is measured at each of the signals and a regression line is fit to these peaks and the extracted parameter is simply the slope coefficient of this regression line [15].

3 Speech data

There is a distinctive lack of available speech data with VQ annotation. Further, as VQ annotation schemes differ and as the annotator's interpretation of VQ labels may not be consistent, this makes large scale data collection difficult. The speech data for this study comes from the recordings used in [13]. The original data were speech recordings of 6 female and 5 male speakers aged between 18 and 48 years (with a mean of 30). The speakers were asked to produce eight Finnish vowels /ɑ e i o u y æ ø/ using breathy, normal and tense phonation types. Participants were trained with producing the voice qualities before recording. While conducting the recording speakers were asked to repeat the utterance with stronger emphasis on the VQ when it was necessary. Each utterance was repeated three times resulting in 792 speech segments.

The speech was recorded using a unidirectional Sennheiser electret microphone with a preamp (LD MPA10e Dual Channel Microphone Preamplifier) and a digital audio recorder (iRiver iHP-140). Audio was digitized at 44.1 kHz.

In order to describe three independent sets of voice qualities we carried out listening test with three expert judges. All participants were experienced in VQ

research and were also familiar with Laver's labeling framework [1]. The participants rated the speech samples on a five point Likert scale from breathy (1) to tense (5). Samples were presented to the participants in a randomized order, with an inter-rater agreement of $\kappa = 0.526$. For the present study we excluded all recordings for which the maximal membership assignment did not coincide with the intended class. 478 vowel recordings were left for analysis (with an inter-rater agreement $\kappa = 0.717$).

Also, included in the current study were 10 sonorant-only (all voiced) sentences, produced in three voice qualities (breathy, modal and tense) by one male speaker (i.e. 30 sentences in total). The utterances were produced in a semi-anechoic room and audio was captured using a B&K 4191 free-field microphone and a B&K 7749 pre-amplifier.

4 Fuzzy-input Fuzzy-output Support Vector Machines

Support vector machines (SVM) have become one of the most popular classifiers in many different machine learning or pattern recognition applications [16]. Extended architectures like one-against-one SVM, one-against-all SVMs or tree structured SVM [17] have been developed for the classification of crisp or hard labeled data in the more recent past.

While dealing with naturalistic data, like voice qualities or user states in natural recordings, however, labels or categories might not be clear or crisp at all, but rather subjective to the perception of the annotator. Since the ground truth or the correct class might be unknown or fuzzy, the so called fuzzy SVMs (FSVM) assigning memberships to several classes to single observations have been developed by [18]. Though, the output of those FSVMs is still crisp and no fuzzy output is generated. Therefore, so called fuzzy-input fuzzy-output SVMs (F^2SVM) capable of receiving soft labeled data and producing soft outputs with memberships assigned over multiple classes have been developed [6]. The fuzzy output of the F^2SVM is required, as in the case of a multi-class one-against-one SVM (three classes in the present study) a fuzzy output is required for the proper combination of the decisions of the single SVM. Consider, for instance, that all three one-against-one SVM (i.e. in this study: breathy vs. modal; tense vs. modal; breathy vs. tense) would have different crisp opinions. Then, it would not be possible to find a sound solution for the given input. If, however, the output were fuzzy such a stalemate is unlikely.

5 Experiments and Results

In the following we have listed the results of the recognition experiments that we conducted. The standard methods of choice for comparison were naive Bayes classifier (NB), giving a rough baseline, and standard crisp SVM utilizing the same radial basis function (RBF) kernel as the F^2SVM. The approaches were compared using a standard ten fold cross validation (X-VAL; 90% training /10% test data split) as well as leave one speaker out (LOSO; for each fold one of the

Table 1. Error (in %) comparison of **NB, standard SVM and crisp F^{2}SVM** outputs for **X-VAL** and **LOSO** experiments. The error (Err.) and standard deviation (Std.) are calculated. Significant results are marked with * or **.

	X-VAL		**LOSO**	
	Err. (%)	**Std.**	**Err. (%)**	**Std.**
NB	21.54**	6.58	23.94**	10.35
SVM	16.09*	4.59	18.33*	6.99
F^{2}SVM	**12.14**	**3.11**	**13.88**	**3.89**

eleven speakers was left out of the training set and was solely used for testing) paradigms. Additionally, the generalization ability of all three methods, i.e. NB, SVM, and F^{2}SVM, is compared in a cross corpus experiment using the sentence dataset (see Sec. 3).

For the F^{2}SVM experiments it was necessary to generate fuzzy targets resembling the degree of membership of each sample towards all of the three classes. For each of the recordings these membership values were calculated using the labels (i.e. five point Likert scale), as indicated by all the experts. These newly calculated values were then used as the target signal for the F^{2}SVM in the experiments. If no clear VQ was perceived by the annotator (i.e. mixed labels 2 and 4) the same amount of membership was assigned to both voice qualities. After normalization to the number of annotators the sum of all memberships of each sample adds up to 1.

In Tab. 1 the error rates of all of the crisp classification experiments are listed. The F^{2}SVM outperforms the other baseline approaches in all experiments significantly. For the X-VAL experiments using all the available speakers 12.14% error (standard deviation $\sigma = 3.11$) was achieved, and only a slight decrease was observed while leaving one speaker out (13.88% error; $\sigma = 3.89$). In contrast to these results the standard SVM receiving the actual label as target in training resulted in 16.09% error ($\sigma = 4.59$) in the X-VAL and 18.33% ($\sigma = 6.99$) in LOSO. Both times the F^{2}SVM outperforms the standard SVM statistically significant in paired t-tests (X-VAL $p = 0.02$; LOSO $p = 0.04$). The baseline performance of the NB results in errors slightly over 20% for both the X-VAL and the LOSO experiment. Both times the NB is strongly outperformed by the F^{2}SVM with significant differences (X-VAL $p < 0.001$; LOSO $p = 0.008$). No statistically significant difference between the standard SVM and the NB was found.

The confusion matrices of these experiments can be seen in Tab. 2 (X-VAL experiment and LOSO experiment). All approaches result in very similar confusion matrices where almost no confusion between breathy and tense voice qualities are present. For the F^{2}SVM and the NB these errors are not reported in the X-VAL experiments, further, in the LOSO experiment they do not exceed 1%. In the standard SVM case breathy is confused with tense in 6% of the cases for the LOSO experiment (only 3% in the X-VAL experiment). The errors of the NB between neighboring voice qualities are, however, more frequent as in the other approaches.

Table 2. Comparison of confusion matrices using **NB, standard SVM and F^{2}SVM** approaches for **X-VAL** and for **LOSO** experiments with **all speakers** (eleven speakers). Numbers are hit rates and lines sum up to one for each confusion matrix modulo rounding errors.

		NB			**SVM**			**F^{2}SVM**		
		Breathy	**Modal**	**Tense**	**Breathy**	**Modal**	**Tense**	**Breathy**	**Modal**	**Tense**
X-VAL	**Breathy**	0.87	0.13	0.00	0.89	0.10	0.01	**0.90**	0.10	0.00
	Modal	0.19	0.65	0.16	0.13	0.78	0.09	0.08	**0.85**	0.06
	Tense	0.01	0.14	0.85	0.03	0.12	0.85	0.00	0.12	**0.88**
LOSO	**Breathy**	0.86	0.14	0.00	0.85	0.13	0.02	**0.88**	0.11	0.01
	Modal	0.20	0.62	0.18	0.13	0.78	0.09	0.09	**0.83**	0.08
	Tense	0.01	0.14	0.84	0.06	0.13	0.81	0.01	0.11	**0.88**

Table 3. Error (in %) comparison of **NB, standard SVM and F^{2}SVM** outputs for **cross corpus** experiments with frame-wise error rates as well as temporally integrated errors over full sentence length. The classifiers are trained on the Finnish vowel set and tested on the sentence data (compare Sec. 3). The error is calculated by comparing to the true label.

	Frame-wise	**Temporally integrated**
NB	29.53	30.00
SVM	33.33	30.00
F^{2}SVM	**17.66**	**3.33**

In order to further check the generalization ability of the approach a cross corpus experiment was conducted. All the mentioned methods, i.e. NB, standard SVM, and F^{2}SVM, were trained on the Finnish vowel set data and tested on the sentence dataset. The errors in % are listed in Tab. 3 comprising the errors on a frame-wise basis including vowels and consonants and the errors achieved after integrating the decisions of the approaches over the whole sentences, which were recorded in a constant VQ. It is seen, that the F^{2}SVM approach (frame-wise error 17.66%; sentence level 3.33%) again outperforms the other two reference approaches clearly. The two perform around 30% error for all cases. In the case of the sentence level integration of the decision the F^{2}SVM only mistakes one breathy sentence as a modal sentence.

6 Discussion of statistical evaluation

The most striking result from the experiments is the capability of the F^{2}SVM to classify the voice qualities more accurately than a standard SVM with the same features as input and kernel function (RBF kernel), in the classification experiments shown in Tab. 1. Therefore, it seems quite obvious that there is relevant information present in the fuzzy targets during training that improves the generalization capabilities of the classifier. As these experiments were conducted on the reduced dataset with an inter-rater agreement of $\kappa = 0.717$ the training of all approaches was conducted on a set for which the maximum of the annotators' membership assignments always coincides with the actual target label, in order to render a fair comparison. Furthermore, the underlying model employed

during expert annotation, described in Sec. 3, allowing the annotator to assign a label between breathy and modal (the value 2 in the Likert scale) and a value between modal and tense (the value 4 in the Likert scale) seems proven by the classification results shown in Sec. 5. This conclusion can be drawn since all the classifiers, comprising NB, standard SVM, and F^2SVM, confuse neighboring classes more often than the two extreme classes, breathy and tense.

Overall, the approach is sufficiently stable over untrained speakers and generalizes well. This, however, is not only the case for the fuzzy approach but also for the two baseline approaches, indicating that the features are representing the voice qualities quite well and are quite independent of the speakers (compare leave one speaker out results in Tab. 1).

The generalization capabilities of the approaches were further compared in a cross corpus experiment. The classifiers were trained using the features extracted from the Finnish vowel set data and tested on the features of the sentence data, including features corresponding to voiced-consonants and vowels alike. The F^2SVM clearly outperformed the reference approaches, with an accuracy of around 82% for the frame-wise decisions. Further, after integrating the decisions over the whole sentences the accuracy rose to more than 95%, meaning that one of the thirty sentences was confused.

7 Conclusion

In the present study we investigated the capability of F^2SVM to classify VQ samples from a vowel corpus, as well as in a cross corpus study using data taken from full sentences. The results in Sec. 5 show high accuracy rates including cross validation and leave one speaker out validation conditions. Additionally, we have shown strong generalization capabilities in cross corpus analysis and leave one speaker out experiments. The proposed method outperformed its competitors (standard SVM, and NB) in crisp classification experiments clearly, by only utilizing the information present in fuzzy labels during training. This is a very encouraging result supporting the value of fuzzy interpretations of VQ data and annotations. The results are very promising for future work including the extension of the approach to running speech and more naturalistic data.

One of the shortcomings of the present study is, that we only considered acted VQ samples. However, we believe the findings here help pave the way to improved VQ analysis in realistic speech data. The analysis of the sentence corpus is a first step into that direction and it seemingly worked very well.

8 Acknowledgements

The presented work was developed within the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG). The second and third authors are supported by the Science Foundation Ireland, Grant

07/CE/I1142 (Centre for Next Generation Localisation, www.cngl.ie) and Grant 09/IN.1/I2631 (FASTNET).

References

1. J. Laver, *The Phonetic Description of Voice Quality.* Cambridge University Press, 1980.
2. C. Gobl, "The voice source in speech communication," *Ph. D. Thesis, KTH Speech Music and Hearing, Stockholm*, 2003.
3. C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Comm.*, vol. 40, pp. 189–212, 2003.
4. M. Lugger and B. Yang, "Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters," in *Proc. of ICASSP*, 2008, pp. 4945–4948.
5. M. Wester, "Automatic classification of voice quality: Comparing regression models and hidden Markov models," in *Proc. of VOICEDATA'98*, 1998, pp. 92–97.
6. C. Thiel, S. Scherer, and F. Schwenker, "Fuzzy-input fuzzy-output one-against-all support vector machines," in *Proc. of KES*, vol. 3. Springer, 2007, pp. 156–165.
7. P. Alku, T. Bäckström, and E. Vilkman, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Comm.*, vol. 11, no. 2-3, pp. 109–118, 1992.
8. G. Fant, J. Liljencrants, and Q. Lin, "A four parameter model of glottal flow," *KTH, QPSR*, vol. 4, pp. 1–13, 1985.
9. H. Strik, B. Cranen, and L. Boves, "Fitting a LF-model to inverse filter signals," in *Proc. of Eurospeech.* ISCA, 1993, pp. 103–106.
10. C. Gobl, "A preliminary study of acoustic voice quality correlates," *KTH, QPSR*, vol. 4, pp. 9–21, 1989.
11. J. Kane, M. Kane, and C. Gobl, "A spectral LF model based approach to voice source parameterisation," in *Proc. of Interspeech.* ISCA, 2010, pp. 2606–2609.
12. P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parameterization of the glottal flow," *J Acoust Soc Am*, vol. 112, no. 2, pp. 701–710, 2002.
13. M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types," in *Proc. of Interspeech.* ISCA, 2007, pp. 1410–1413.
14. M. Lugger and B. Yang, "Classification of different speaking groups by means of voice quality parameters," in *Proc. of Sprach-Kommunikation.* VDE, 2006.
15. J. Kane and C. Gobl, "Identifying regions of non-modal phonation using features of the wavelet transform," in *Proc. of Interspeech.* ISCA, 2011, pp. 177–180.
16. K. P. Bennett and C. Campbell, "Support vector machines: hype or hallelujah?" *ACM SIGKDD Newsletter*, vol. 2, no. 2, pp. 1–13, 2000.
17. F. Schwenker, "Solving multi-class pattern recognition problems with tree-structured support vector machines," in *DAGM-Symp.* Springer, 2001, pp. 283–290.
18. C. F. Lin and S. D. Wang, "Fuzzy support vector machines," *IEEE Trans Neural Net*, no. 13, pp. 464–471, 2002.