Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 263–287

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification☆

Stefan Scherer[a,c,*], John Kane[b], Christer Gobl[b], Friedhelm Schwenker[c]

[a] University of Southern California, Institute for Creative Technologies, 90094 Playa Vista, CA, United States
[b] Trinity College Dublin, Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences, Dublin 2, Ireland
[c] Ulm University, Institute of Neural Information Processing, 89069 Ulm, Germany

Received 6 October 2011; received in revised form 19 April 2012; accepted 1 June 2012
Available online 13 June 2012

Abstract

The dynamic use of voice qualities in spoken language can reveal useful information on a speakers attitude, mood and affective states. This information may be very desirable for a range of, both input and output, speech technology applications. However, voice quality annotation of speech signals may frequently produce far from consistent labeling. Groups of annotators may disagree on the perceived voice quality, 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. Further, the study looks to include these features as inputs to a fuzzy-input fuzzy-output support vector machine (F^2SVM) algorithm, which is in turn capable of softly categorizing voice quality recordings. The F^2SVM is compared in a thorough analysis to standard crisp approaches and shows promising results, while outperforming for example standard support vector machines with the sole difference being that the F^2SVM approach receives fuzzy label information during training. Overall, it is possible to achieve accuracies of around 90% for both speaker dependent (cross validation) and speaker independent (leave one speaker out validation) experiments. Additionally, the approach using F^2SVM performs at an accuracy of 82% for a cross corpus experiment (i.e. training and testing on entirely different recording conditions) in a frame-wise analysis and of around 97% after temporally integrating over full sentences. Furthermore, the output of fuzzy measures gave performances close to that of human annotators.

Keywords: Voice quality; Fuzzy-input fuzzy-output support vector machines; Fuzzy classification; LF-model; Cross corpus analysis

1. Introduction

The term voice quality refers to the timbre or coloring of a speaker's voice. It includes but is not limited to what is perceived by the listener as pitch and loudness. For an individual speaker their voice quality is composed of longer term settings of the vocal system combined with dynamic shifts in the system for communicative purposes (Laver, 1979; Mackenzie Beck, 2005).

☆ This paper has been recommended for acceptance by Simon King, Ph.D.

* Corresponding author at: University of Southern California, Institute for Creative Technologies, 90094 Playa Vista, CA, United States. Tel.: +1 310 448 0372.
E-mail address: scherer@ict.usc.edu (S. Scherer).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2012.06.001

In spoken communication voice quality is used in certain languages for contrastive linguistic function (e.g., in Gujarati the words meaning 'twelve' [bə̤ɾ] and 'outside' [baɾ] , and 'last year' [pɔ̤ɾ] and 'early morning' [pɔɾ] are contrasted solely on the presence or absence of breathiness in the vowels (Ladefoged and Maddieson, 1996)). A speaker's voice quality is also an important feature of paralinguistic signaling in speech and can provide the listener with information pertaining to the speaker's affective state (Gobl, 2003; Campbell, 2007). The use of breathiness has been studied in connection with politeness particularly among male speakers of Japanese (Ito, 2004). Breathy voice has also been generally observed in association with intimacy and familiarity (Laver, 1980). Tense voice on the other hand has been reported in more active affective states, e.g., anger and happiness (Gobl and Ní Chasaide, 2003; Yanushevskaya et al., 2005).

The use of voice qualities is also a tool used by speakers for managing spoken discourse. A study on Finnish interactive speech provided evidence of creaky voice qualities being consistently used by Finnish speakers for turn-yielding functions, in contrast with glottal stops which were frequently used for turn-holding (Ogden, 2001).

From the above examples it can be seen that voice quality can provide useful insights into the intentions and mood of the speaker, and indeed voice quality features have also been utilized in order to improve emotion classification (Lugger and Yang, 2008). It follows that robust characterization of voice qualities may be desirable for both input (e.g., recognition) and output (e.g., synthesis) ends of speech technology applications. Voice quality descriptions have been included in various speech synthesis systems in order to provide platforms for more expressive synthesis (see e.g., Campbell, 2004; Raito et al., 2008; Cabral et al., 2008). In terms of speech recognition systems robust parameters describing the speaker's voice quality would help determine the intention of the spoken utterance which may be ambiguous if only the linguistic elements are detected.

It follows that the purpose of this study is to put forward a framework for identifying voice qualities from speech utilizing robust acoustic features on a tense to breathy continuum. Furthermore we propose the use of a classification approach which is capable of leveraging the disagreement on the part of annotators as a source of information in the classification.

From a speech production point of view it is the mode of phonation, or manner in which the vocal folds vibrate, that is largely responsible for producing what is perceived as a person's voice quality (Laver, 1980). This is what some have called the narrower view of voice quality (Laver, 1980; Mackenzie Beck, 2005) and indeed for voice qualities on the breathy to tense dimension (i.e. those investigated in the current work), phonation plays a primary role (Laver, 1980). However, it is in fact the settings of the entire vocal apparatus that affect a person's voice quality and for some voice qualities (e.g., whisper) there may be no vocal fold vibration and, hence, phonation does not contribute to the perceived vocal timbre.

Nevertheless, as the phonation mode is critical for producing breathy to tense voice qualities it seems intuitive to exploit acoustic features derived from the voice source (i.e. the residual signal from inverse filtering the speech signal with an estimate of the vocal tract transfer function). It has been shown in previous studies that developing feature sets separately for both the voice source and vocal tract filter components can provide better modeling of speech (Krishnamurthy and Childers, 1986). Further, although some voice quality measurements can be made directly from the speech waveform (e.g., Hillenbrand et al., 1994; Ishi et al., 2008) voice source-based feature sets have been shown to be crucial in the fine-grained modeling of an array of voice quality types (Gobl and Ní Chasaide, 1992).

To separate the voice source component from the speech signal, we need to remove the impact of the vocal tract from the signal. For this, researchers usually apply knowledge from the acoustic theory of speech production (Fant, 1960). The theory, which provides a simplified model of the speech production process regards the speech signal $S(z)$ (in the z-domain), as the end result of a linear combination of the glottal flow, $G(z)$, with the vocal tract filter, $V(z)$, and lip radiation $L(z)$ (see Eq. (1)):

$$S(z) = G(z)V(z)L(z) \tag{1}$$

The vocal tract filter can be described using an all-pole model by considering it to be a combined set of lossless tubes[1] (Markel and Gray, 1982). If such an all-pole vocal tract model can be derived this can facilitate the design of an all-zero filter to be used for removing the effect of the vocal tract from the speech signal. The lip radiation component is

[1] Of course this is a simplification and does not properly model certain aspects of the vocal tract system, for instance the presence of zeros in nasal regions.

typically modeled as a first order differentiator, however, as it can be convenient to work with the differentiated glottal flow signal ($G'(z)$ in Eq. (2)) no further compensation needs be applied. Indeed in the present work we will refer to the differentiated glottal flow signal as the voice source signal:

$$G'(z) = \frac{S(z)}{V(z)} \tag{2}$$

However, as neither the glottal flow nor the vocal tract components are directly observable and as both components vary dynamically in running speech the decomposition of speech into vocal tract filter and voice source is known to be problematic. Algorithms have been developed to separate the vocal tract influence from the speech signal using a variety of approaches (e.g., closed-phase methods (Alku et al., 2009), phase based methods (Drugman et al., 2009), iterative methods (Alku et al., 1992)). Despite the attention this problem has received from the research community it is yet to be considered a solved problem and most methods produce incomplete formant cancellation when analyzing running speech and non-modal voice qualities.

Given an estimate of the voice source, the signal can then be parameterized in order to describe the perceptually important aspects of the source pulses. Indeed many of the acoustic features used in the present study (see Section 2) are used to characterize these important aspects.

Next we address the issue of hidden ground truths. The annotation of breathy to tense voice qualities can be difficult and annotators need to rely on their own subjective perception of the utterance. As a result it is likely that annotators may disagree somewhat on the voice quality label of a given speech segment. Mixed voice qualities could be present in the speech signal and a gradual answer may be necessary as opposed to a discrete label. Therefore, we elaborate methods for handling this sort of fuzzy labeling in the present work. The fuzzy-input fuzzy-output support vector machine (F^2SVM) introduced in (Thiel et al., 2007; Borasca et al., 2006; Thiel, 2009) is an ideal candidate for this type of task receiving a fuzzy membership label as input with the features for training and producing fuzzy memberships[2] as output. We compare the performance of this F^2SVM in manifold experiments with standard methods, that receive crisp[3] targets and produce crisp outputs, in our studies.

The usefulness of the F^2SVM for the classification of emotion from speech and facial expressions using fuzzy teacher signals has already been shown in (Thiel et al., 2007; Thiel, 2009) where they outperformed standard support vector machines (SVMs) that relied on crisp labels. The accuracy gain was solely achieved by introducing fuzzy labels as teacher signals, as in the present work. This study further extends the experiments of (Thiel et al., 2007; Thiel, 2009) by fitting the fuzzy output of the F^2SVM to the mixed labels of human annotators. We then evaluate the results using fuzzy measures such as the D_1 distance measure presented in this work. The D_1 measure is adapted from the well known S_1 similarity measure introduced by (Dubois and Prade, 1980). We chose this measure for this study, as it is one of the most common to compare fuzzy labels, e.g., in multiple classifier systems, where groups or ensembles of classifiers (e.g., SVM) are combined using fusion schemes to collaboratively predict certain targets (Kuncheva, 2004).

Although there has been considerable research done on describing the acoustic and physiological characteristics of voice qualities (see e.g., Gobl and Ní Chasaide, 1992; Hillenbrand and Houde, 1996; Childers and Lee, 1991; Blomgren et al., 1998) there has been far less work done 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. Some comparable work, however, was conducted in Wester (1998). 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 (from 0 to 4) on three voice quality 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.

A further study involving voice quality classification of non-pathological voices was conducted in Lugger et al. (2008). The study looked at the use of HMMs for representing voice quality contours and compared results with K-means classification. Results showed crisp classification accuracy of 39% for modal, 57% for breathy, 77% for creaky and 61% for rough voice qualities.

[2] Interpretable in two ways: first degree of membership to different classes at the same time or as probabilities for single classes. For further information please refer to Sections 3 and 6.

[3] Crisp in this context means that annotations of samples or predictions of classifiers only support one class at a time, in contrast to fuzzy or soft annotations, which can support multiple classes at the same time to different degrees.

Table 1
Summary of features used as inputs to the F2SVM classification algorithm.

Feature	Measured from:
f_0	Speech signal
Ra, *Rk*, *Rg*. *EE*	Voice source estimate
Ra_f, Rk_f, Rg_f. EE_f	Voice source estimate
NAQ	Voice source estimate
$\Delta H_{1,2}$	Voice source estimate
OQG, GOG, SKG, RCG	Voice source estimate
Peak slope	Speech signal

1.1. Research hypothesis

We hypothesize that the extracted feature set representing the voice source (see Section 2) is suitable and robust enough for the classification of voice qualities. Additionally, we hypothesize that the sometimes ambiguous and mixed opinions provided by human expert annotators are indeed valuable and classification results can be improved by utilizing all the available information. We further show that the fuzzy classification approach presented in this paper provides a robust classifier that is even suitable for unseen speakers and novel recording conditions. This paper also compares the fuzzy predictions of the classifier with human opinions using a fuzzy distance measure and shows that the approach reaches human baseline performance.

These research hypothesis are targeted in multiple experiments including comparisons to standard classification approaches in cross validation, leave one speaker out, and cross corpus (i.e. training on one corpus testing on the other) experiments with crisp as well as fuzzy prediction evaluations.

1.2. Organization

The remainder of the paper is organized as follows: In Section 2 the utilized voice quality features for the classification experiments are introduced. Section 3 then derives fuzzy-input fuzzy-output support vector machines (F^2SVMs) from standard support vector machines (SVMs), explains the algorithm to find the separating hyperplanes, and introduces a fuzzy measure (D_1) to evaluate the results. Along with the introduction of the speech dataset used, Section 4 introduces the annotations by experts, which are later used as targets for the fuzzy classification experiments. In Section 5 the results for the manifold experiments are reported and discussed in Section 6. Finally, Sections 7 and 8 conclude and provide an outlook for future work.

2. Voice quality features

We selected voice quality features for the current study which have been previously shown to be useful for discriminating breathy, modal and tense voice qualities. The features described below in Sections 2.1–2.6 describe aspects of the voice source signal, which is derived using automatic iterative inverse filtering (see below). Automatic inverse filtering, however, can frequently produce estimates of the voice source signal that contain uncanceled formant oscillations that would clearly impact on the proceeding analysis (Walker and Murphy, 2007). Some voice quality features can be measured without inverse filtering, such as the peak slope feature described in Section 2.7. A summary of the features used as inputs to the classification method is shown in Table 1.

2.1. Pre-processing (f_0)

For the speech segments analyzed in the current study initial f_0 values are extracted using the f_0 tracker which is available in the ESPS/*waves+* software package. Glottal closure instants (GCIs) are then located using the method described in Goncharoff and Gries (1998), which identifies peaks in the filtered energy contour of the speech signal for identifying GCI candidates. A dynamic programming algorithm is then used to find the path of peaks that produce maximum energy. Note that for the methods used in the present work GCI detection is not critical. Using the GCIs

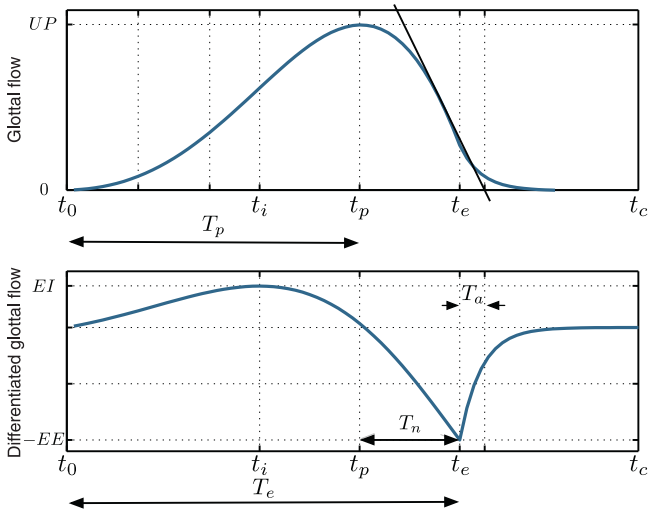

Fig. 1. A single cycle of an example synthetic LF model pulse of glottal flow (top) and differentiated glottal flow (bottom).

the speech signal is then automatically inverse filtered using the pitch synchronous iterative automatic inverse filtering (PSIAIF) method described in Alku et al. (1992).

This iterative inverse filtering method first applies a first-order high pass filter, estimated by LPC analysis, to roughly compensate for the roll-off in the voice source signal in order to help approximate the vocal tract filter response using LPC. The method then removes this vocal tract approximation from the speech segment yielding an approximation of the voice source signal. The speech signal is then inverse filtered using the filter coefficients obtained from LPC analysis of the residual.

This residual signal is then inverse filtered from the speech segment allowing improved estimation of the vocal tract frequency response. Finally, the speech signal is inverse filtered using the LPC coefficients derived from this vocal tract estimate giving the methods estimation of the voice source signal.

The features described in Sections 2.2–2.6 can then be measured on the output signal from this method.

2.2. LF model parameters from time domain estimation methods (Ra,Rk,Rg,EE)

The most commonly used acoustic voice source model is the Liljencrants–Fant (LF) model (Fant et al., 1985) (see Fig. 1). It is a five parameter (including f_0, and assuming $T_c = T_0$) model of differentiated glottal flow. The model has two segments. The first segment, the open phase, is a sinusoid function that increases exponentially:

$$U'_g(t) = E_0 e^{\alpha t} \sin(\omega_g t) \quad \text{for } t_o \le t \le t_e \tag{3}$$

where t_o denotes the starting point of the glottal cycle, t_e the point of the maximum excitation, $\omega_g = \pi/T_p$, with $T_p = t_p - t_0$ the time of glottal flow increase, α, which is responsible for the rate of amplitude increase, is solved implicitly, and

$$E_0 = -\frac{EE}{e^{\alpha T_e} \sin(\omega_g T_e)} \tag{4}$$

The second segment, which models the return phase, is an exponential function:

$$U'_g(t) = -\frac{EE}{\epsilon T_a}(e^{-\epsilon(t-t_e)} - e^{-\epsilon T_b}) \quad \text{for } t_e < t < t_c \tag{5}$$

where t_c is assumed to be $1/f_0$ (such a simplification is frequently used in the literature, see e.g., Gobl and Chasaide, 2003), the length of the glottal cycle, $T_b = t_c - t_e$, and where ϵ is solved iteratively following T_a. The Newton–Raphson method is used for solving ϵ, α and E_0 and a thorough description of the calculations is given in Gobl (2003).

The pulse shape of the LF model can be characterized using an amplitude parameter, *EE* (which is the negative amplitude corresponding to the main excitation), and three time based parameters *Ra*, *Rk* and *Rg* (see Eqs. (6)–(8)).

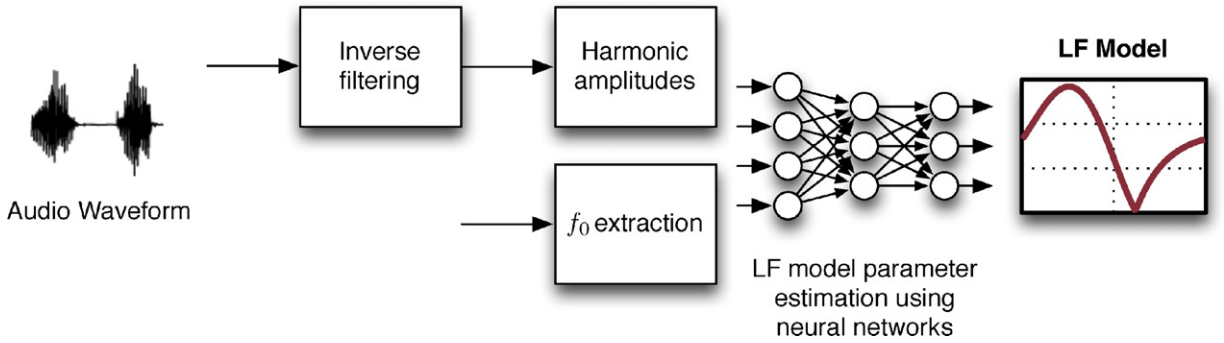

Fig. 2. Schematic overview of the algorithm for LF model parameter extraction in the frequency domain as described in Kane et al. (2010).

These parameters have been shown to be suitable for characterizing a range of voice qualities including breathiness and tenseness (Gobl, 1989):

$$Rg = \frac{1}{2T_p \cdot f_0} \tag{6}$$

$$Rk = \frac{T_e - T_p}{T_p} \tag{7}$$

$$Ra = T_a \cdot f_0 \tag{8}$$

Given an estimate of the voice source signal the model parameters can be derived by fitting the model to each of the voice source pulses. This can be done automatically by applying the algorithm used in Strik et al. (1993), however, other methods also exist for doing this in the time domain (e.g., Li et al., 2011). Though this approach is used in several software packages (e.g., Airas, 2008; Kreiman et al., 2006) the parameters have been shown to lack robustness in the presence of noise (Alku et al., 2002; Kane et al., 2010). The parameters are also sensitive to low frequency phase distortion which is commonly introduced even in high quality recording systems (Walker and Murphy, 2007).

2.3. LF model parameters from frequency domain methods (Ra_f,Rk_f,Rg_f,EE_f)

An alternative approach for mapping from frequency domain measurements to time domain LF model parameters has been developed which has built on previous work initially described in Kane et al. (2010). Fig. 2 shows a schematic overview of the approach.

The subscript f, in Ra_f,Rk_f,Rg_f,EE_f, is used to denote that these parameters are derived from frequency domain measures. The method involves using the amplitudes of the first eight harmonics ($H1, \ldots, H8$) from the voice source spectrum, as well as the local f_0 value, as inputs to a feed forward neural network (see Figs. 2 and 3), previously trained on a large volume of LF model configurations and their spectral information, in order to derive the four parameters Ra_f, Rk_f, Rg_f, and EE_f. Harmonic amplitudes are measured from the narrowband spectrum, obtained by fast Fourier transform (FFT) carried out on Hamming windowed, GCI centered frame, of length three times the local pitch period (note that this procedure is used for all narrowband spectral analysis carried out in this study). The three shaping parameters Ra_f, Rk_f, and Rg_f are learned jointly in one neural network, as they are interdependent. EE_f is trained separately in an additional neural network, as seen in Fig. 3.

For the ANN training and weight estimation, we first created a dataset, which involved generating a large number of LF model pulses and recording both the parameters used for generating each pulse as well as the corresponding spectral information. f_0 was used as the control parameter with $f_0 \in [50, 600]$. For each of the f_0 values 1000 LF model pulses were generated, each time randomly choosing values for each of the four other LF model parameters. Parameter value selection was also subject to certain restrictions stated in Gobl (2003) for biological and model plausibility.

After standard back-propagation training, the neural networks produce precise predictions of the parameters, with correlation coefficients R between predictions and target values (of the artificially generated parameters) > 0.9 and very low mean absolute errors (mae) (Ra_f, Rk_f, and Rg_f network: $R = 0.958$, $mae = 0.029$; EE_f network: $R = 0.981$, $mae = 0.012$).

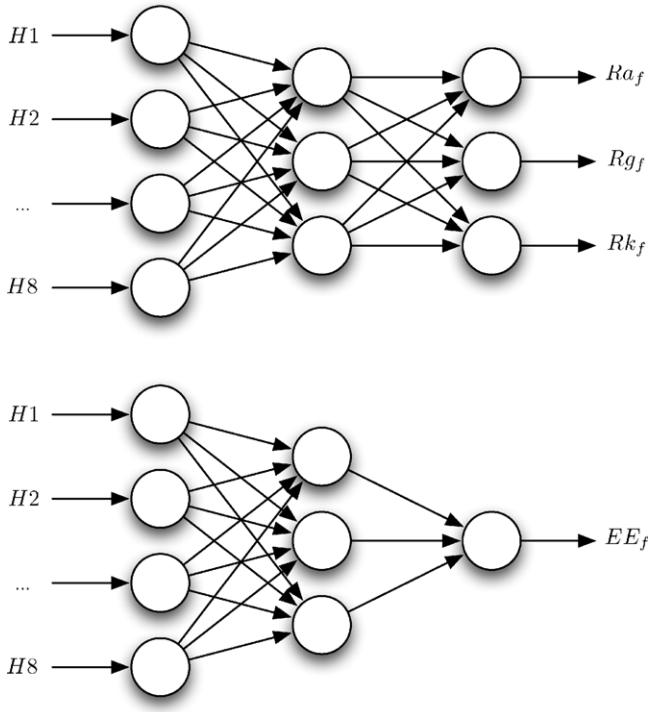

Fig. 3. The first eight harmonics [$H1, \ldots, H8$] are used as input for two separate networks approximating the four parameters Ra_f, Rk_f, Rg_f, and EE_f. As the parameters Ra_f, Rk_f, and Rg_f are not independent and influence each other a joint network is trained. EE_f is trained separately.

This approach was developed in order to improve the robustness of the extracted parameters to the presence of noise and phase distortion. A schematic overview of the processes involved in the extraction of the LF model parameters from the frequency domain is seen in Fig. 2.

2.4. Normalized amplitude quotient (NAQ)

The normalized amplitude quotient (NAQ) parameter was introduced as a global voice source parameter capable of differentiating breathy to tense voice qualities (Alku et al., 2002) and is closely related to the *Rd* parameter described in Fant et al. (1995). It is calculated using:

$$NAQ = \frac{f_{ac} \cdot f_0}{d_{peak}} \tag{9}$$

where f_{ac} is the maximum amplitude of the glottal flow, d_{peak} is the maximum negative amplitude of the differentiated glottal flow.

Although NAQ is closely related to *Rd*, it is subtly, but nevertheless significantly different. NAQ is a direct measure of the glottal flow and glottal flow derivative, whereas *Rd* is a measure derived from a fitted LF model pulse. As fitting an LF model pulse ultimately involves compromising modeling of some features of the voice source pulse more than others its value is not necessarily proportional to Rd. Indeed it is the amplitude features of the voice source pulse that are frequently poorly modeled when conducting model fitting.

NAQ, as an amplitude based parameter, was shown to be more robust to noise disturbances than parameters based on time instant measurements and has, as a result, been used in the analysis of conversational speech (Campbell and Mokhtari, 2003), which is frequently noisy. The parameter, however, may not be as effective as a voice quality indicator when a speaker is using a wide f_0 range (Gobl and Chasaide, 2003).

2.5. $\Delta H_{1,2}$

The difference in amplitude levels (in dB) between the first two harmonics of the narrowband voice source spectrum ($\Delta H_{1,2}$) is thought to be a rough correlate of the open quotient parameter (Henrich et al., 2001) and hence useful at discriminating breathy to tense voice qualities (Airas and Alku, 2007). The first two harmonics, however, are also affected by the asymmetry of the glottal pulse, but to a lesser degree. The narrowband spectrum is obtained by using three-pulse length sections, centered on a GCI and using a Hamming window.

2.6. Voice quality spectral gradients (OQG, GOG, SKG, RCG)

Lugger and Yang (2006) described a set of spectral gradient parameters for characterizing voice qualities from voice source signals which built on work presented in Stevens and Hanson (1994):

- Open quotient gradient (*OQG*):

$$OQG = \frac{\hat{H}_1 - \hat{H}_2}{f_0} \tag{10}$$

- Glottal opening gradient (*GOG*):

$$GOG = \frac{\hat{H}_1 - A_{1p}}{F_{1p} - f_0} \tag{11}$$

- Skewness gradient (*SKG*):

$$SKG = \frac{\hat{H}_1 - A_{2p}}{F_{2p} - f_0} \tag{12}$$

- Rate of closure gradient (*RCG*):

$$RCG = \frac{\hat{H}_1 - A_{3p}}{F_{3p} - f_0} \tag{13}$$

As in Lugger and Yang (2006), F_{1p}, F_{2p} and F_{3p} are frequencies of the harmonics nearest the first three formants. A_{1p}, A_{2p} and A_{3p} are the amplitudes at F_{1p}, F_{2p} and F_{3p}. Harmonic amplitudes $\hat{H}_1$ and $\hat{H}_2$ are measured from the voice source spectrum (note that the -symbol is used to denote that the harmonic measurements are made from the estimated voice source signal). Formants are derived using the *get_formants* script available in the SNACK toolkit.

The parameters were stated by the authors to be strongly correlated with typical glottal pulse shape parameters. The parameters have been shown to be useful in the classification of voice qualities, gender and emotion (Lugger and Yang, 2006). They have also been shown to be reasonably robust even with the presence of noise disturbances (Lugger et al., 2006).

2.7. Peak slope

The final parameter included in this work is based on features derived following wavelet based decomposition of the speech signal (Kane and Gobl, 2011). It was observed in Sturmel et al. (2009) that although only one or two of the waveforms derived using wavelet-based analysis, relating to high frequencies, were required for finding glottal closure instants (GCIs) in modal voice qualities. However, for breathier voice qualities the smoother GCIs require aspects of the decomposition were required. Based on these observations a measurement was designed to identify GCIs from glottal pulses with different glottal closure characteristics and hence differentiate between breathy and modal, and perhaps also tense voice qualities. The following equation is used for decomposing the speech signal:

$$g(t) = -\cos(2\pi f_n t) \cdot \exp\left(-\frac{t^2}{2\tau^2}\right), \tag{14}$$

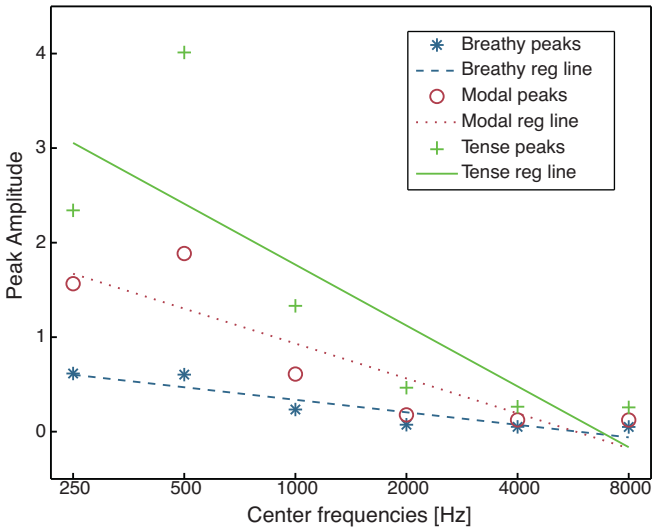

Fig. 4. Peak amplitudes, from signals with different center frequencies, with regression lines for an /o/ vowel produced by a male speaker in breathy, modal and tense voice qualities.

where the speech signal $s(t)$ is convolved with $g(t/s_i)$, and $s_i = 2^i$ and $i = 0, 1, 2, \ldots, 5$. This essentially is the application of an octave-band filter bank with the center 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 obtained from the decomposition and a regression line is fit to these peaks. In Fig. 4, it can be seen that for an /o/ vowel produced by a male speaker in breathy, modal and tense voice qualities the slope of the regression line is clearly different. Hence, the peak slope parameter is simply the slope coefficient of the regression line. In the original publication (Kane and Gobl, 2011) this was carried out on individual phone segments. In the current study it is carried out on the frame level following wavelet-based decomposition of the entire speech signal. A frame length of 32 ms and shift of 10 ms is then used, together with rectangular windowing.

3. Fuzzy-input fuzzy-output support vector machines

Support vector machines (SVMs) have become one of the most popular classifiers in many different machine learning or pattern recognition applications (Bennett and Campbell, 2000). The principle idea is to find a hyperplane that maximizes the margin to the two linearly separable classes. In the case of non linearly separable classes the data are projected into a higher dimensional space using a kernel function. It is suspected that in a higher dimensional space a separating hyperplane is found easier than in a lower dimensional space (Schölkopf and Smola, 2001). The so-called support vectors are the nodes supporting the hyperplane. The main idea of a two dimensional SVM problem is seen in Fig. 5, where two classes are linearly separated by a hyperplane. The advantage over common multi layer perceptrons is that the SVM finds the hyperplane with the maximally wide margin as opposed to any hyperplane capable of separating the classes, which in turn is expected to have several advantages in particular generalization capabilities.

However, in many cases the task is to separate more than two classes. For these applications extended architectures like one-against-one SVM, one-against-all SVM or tree structured SVM (Schwenker, 2001) have been developed for the classification of crisp or hard labeled data (Kahsay et al., 2005). These extended architectures result in different computationally expensive decision and training algorithms.

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 SVM (FSVM) assigning memberships to several classes to single observations have been developed by (Lin and Wang, 2002) and (Huang and Liu, 2002). Though, the output of those FSVM is still crisp and no fuzzy output is generated. Therefore, so-called fuzzy-input fuzzy-output SVM (F^2SVM) capable of receiving soft labeled data and producing soft outputs with memberships assigned over multiple classes have been developed (Thiel et al., 2007; Borasca et al., 2006; Thiel, 2009). For the

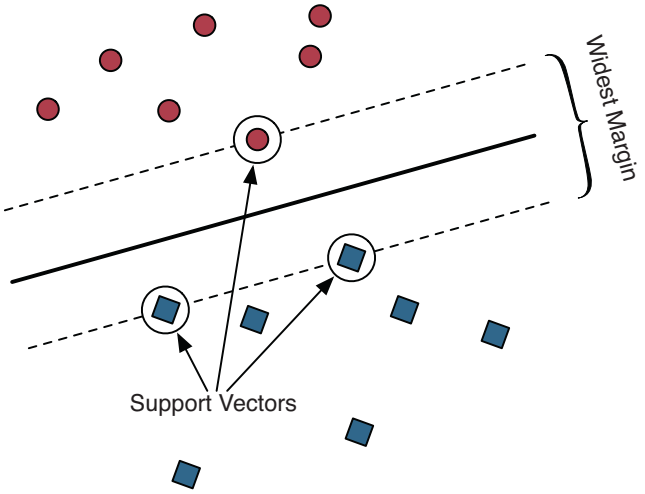

Fig. 5. A schematic view of a linearly separable two dimensional classification task. Class 1 is represented by shaded circles whereas class 2 is represented with shaded rectangles. The support vectors necessary for the representation of the hyperplane (solid line) are marked with a wider circle.

proper combination of the decisions of the single SVM in the case of a multi-class one-against-one SVM (three classes in the present study) a fuzzy output is more flexible than a binary output. Consider, for instance, that all three binary one-against-one SVM (i.e. in this study: breathy vs. modal; tense vs. modal; breathy vs. tense) disagree; a sound class estimate cannot be made for the given input based on these disagreeing predictions. In case of fuzzy classifier output, at least for real-valued output, such a tie-situation is unlikely. Further, as in the present study fuzzy outputs sometimes are a more suitable fit to the target application.

3.1. Training and retrieval

In order to emphasize on the differences and extensions in the F^{2}SVM approach the standard SVM training shall be introduced in the following.

3.1.1. Simple support vector machines

Starting from the typical two class problem a set M of training samples (x_μ, l_μ) may be defined as follows:

$$M = \{(x_\mu, l_\mu) | x_\mu \in \mathbb{R}^n, l_\mu \in \{-1, +1\}, \forall \mu = 1, \ldots, |M|\} \tag{15}$$

where x_μ denote the samples from the input space $\mathbb{R}^n$ and l_μ the target labels. Using these target labels as a category to divide set M into two subsets, the optimal separating hyperplane is determined by the vector $w \in \mathbb{R}^n$ and the bias $b \in \mathbb{R}$ fulfilling the separation constraints:

$$l_\mu (w^T x_\mu + b) \geq 1, \quad \forall \mu = 1, \ldots, |M|, \tag{16}$$

and the maximal margin condition (Bishop, 2006):

$$\theta(w) = \frac{||w||^2}{2} \rightarrow \min . \tag{17}$$

In order to solve this quadratic optimization problem so-called Lagrange multipliers $\alpha_\mu \geq 0$ are introduced, with one multiplier for each constraint, yielding the following Lagrangian function:

$$L(w, b, \alpha) = \frac{||w||^2}{2} - \sum_{\mu=1}^{|M|} \alpha_\mu \{l_\mu (w^T x_\mu + b) - 1\} \tag{18}$$

The dual form of this quadratic optimization problem defined through

$$W(\alpha) = \sum_{\mu=1}^{|M|}\alpha_\mu - \frac{1}{2}\sum_{\nu=1}^{|M|}\sum_{\mu=1}^{|M|}\alpha_\nu\alpha_\mu l_\nu l_\mu x_\nu^T x_\mu \tag{19}$$

subject to the following constraints:

$$\sum_{\mu=1}^{|M|}\alpha_\mu l_\mu = 0 \,\text{and}\, \alpha_\mu \geq 0, \quad \forall n = 1, \ldots, |M| \tag{20}$$

can be solved using standard optimization methods.

In real world problems a separating hyperplane might not be found at all times. Therefore, we can reformulate the optimization problem given in (17) by introducing so-called slack variables ξ_μ allowing a data point to be between the hyperplane and the margin ($0 \leq \xi_\mu < 1$) or even on the wrong side of the hyperplane ($\xi_\mu > 1$). The reformulated optimization problem is now (Bishop, 2006):

$$\theta(w, \xi) = \frac{||w||^2}{2} + C\sum_{\mu=1}^{|M|}\xi_\mu \rightarrow \min, \tag{21}$$

with the soft constraints:

$$l_\mu(w^T x_\mu + b) \geq 1 - \xi_\mu, \quad \xi_\mu \geq 0, \quad \forall\mu = 1, \ldots, |M|, \tag{22}$$

with a free parameter $C > 0$ regulating the amount of allowed errors by punishing them with an increased value of C.

3.1.2. Fuzzy-input fuzzy-output support vector machines

In the following the extensions necessary for the F^{2}SVM approach will be introduced. First of all it is necessary to define membership values m_μ^+ (the membership to the positive class) and m_μ^- (the membership to the negative class) for observation x_μ, which are now integrated to the optimization problem with slack variables:

$$\theta(w, \xi^+, \xi^-) = \frac{||w||^2}{2} + C\sum_{\mu=1}^{|M|}(\xi_\mu^+ m_\mu^+ + \xi_\mu^- m_\mu^-) \rightarrow \min \tag{23}$$

subject to the following constraints:

$$w^T x_\mu + b \geq 1 - \xi_\mu^+, \quad \text{with}\, \xi_\mu^+ \geq 0, \forall\mu = 1, \ldots, |M| \tag{24}$$

and

$$w^T x_\mu + b \geq -1 + \xi_\mu^-, \quad \text{with}\, \xi_\mu^- \geq 0, \forall\mu = 1, \ldots, |M| \tag{25}$$

This adaptation to the optimization problem doubles the storage complexity, as opposed to the standard SVM optimization, since each data point x_μ belongs to both classes (to a certain extent), whereas in the standard SVM case it can only belong to one, and therefore increases the runtime during training.

With the help of these four constraints we now can reformulate the optimization problem to the Lagrangian function $L(w, b, \xi^+, \xi^-, \alpha^+, \alpha^-, \beta^+, \beta^-)$, by introducing the Lagrangian multipliers $\alpha^+, \alpha^-, \beta^+, \beta^-$:

$$\begin{aligned} L(w, b, \xi^+, \xi^-, \alpha^+, \alpha^-, \beta^+, \beta^-) = {} & \frac{||w||^2}{2} + C\sum_{\mu=1}^{|M|}(\xi_\mu^+ m_\mu^+ + \xi_\mu^- m_\mu^-) - \sum_{\mu=1}^{|M|}\alpha_\mu^+((w^T x_\mu + b) - 1 + \xi_\mu^+) \\ & + \sum_{\mu=1}^{|M|}\alpha_\mu^-((w^T x_\mu + b) + 1 - \xi_\mu^-) - \sum_{\mu=1}^{|M|}\beta_\mu^+\xi_\mu^+ - \sum_{\mu=1}^{|M|}\beta_\mu^-\xi_\mu^- \end{aligned} \tag{26}$$

This Lagrangian function can now be transformed into the dual representation by eliminating w, b, ξ^+, and ξ^- by inserting the partial derivatives $\partial L/\partial w$, $\partial L/\partial b$, $\partial L/\partial \xi_\mu^+$, and $\partial L/\partial \xi_\mu^-$ and setting them equal to zero. Finally, leading to the Lagrangian requiring maximization:

$$W(\alpha) = \sum_{\mu=1}^{|M|} \alpha_\mu^+ + \sum_{\mu=1}^{|M|} \alpha_\mu^- - \frac{1}{2} \sum_{\nu=1}^{|M|} \sum_{\mu=1}^{|M|} (\alpha_\nu^+ - \alpha_\nu^-)(\alpha_\mu^+ - \alpha_\mu^-) x_\nu^T x_\mu \tag{27}$$

with $\alpha_\mu^+, \alpha_\mu^- \geq 0$, $\forall \mu = 1, \ldots, |M|$ and subject to

$$\sum_{\mu=1}^{|M|} (\alpha_\mu^+ - \alpha_\mu^-) = 0 \tag{28}$$

and

$$0 \leq \alpha_\mu^+ \leq C m_\mu^+, \quad 0 \leq \alpha_\mu^- \leq C m_\mu^- \tag{29}$$

Further, we get the Karush–Kuhn–Tucker conditions ($\forall \mu = 1, \ldots, |M|$) (Thiel et al., 2007):

$$\alpha_\mu^+ ((w^T x_\mu + b) - 1 + \xi_\mu^+) = 0 \tag{30}$$

and

$$\alpha_\mu^- ((w^T x_\mu + b) + 1 - \xi_\mu^-) = 0 \tag{31}$$

Therefore, the samples x_μ with $(\alpha_\mu^+ - \alpha_\mu^-) \neq 0$ are the support vectors defining the decision function:

$$y(x) = sign \left(\sum_{\mu=1}^{|M|} (\alpha_\mu^+ - \alpha_\mu^-) x^T x_\mu + b \right). \tag{32}$$

In order to extend this fuzzy approach to a multi-class approach in which we do have a set of samples:

$$M = \left\{ (x_\mu, l_\mu) | x_\mu \in \mathbb{R}^n, \quad l_\mu \in \mathbb{R}^k, \quad \text{with} \sum_{j=1}^{k} l_{\mu,j} = 1, \quad \forall \mu = 1, \ldots, |M| \right\} \tag{33}$$

where $l_{\mu,j}$ indicates the grade of membership of sample x_μ to class j and k is the total number of classes. Note that a crisp classification problem is simply a special case of the fuzzy one with all $l_{\mu,j} = 0$ except for one $l_{\mu,j^*} = 1$, where j^* denotes the target class. In order to solve the fuzzy multi-class problem we train $k(k-1)/2$ SVM $S_{i,j}$ in the one-against-one paradigm for each possible pair of classes i and j. The sample set for each of the SVM to be trained is constructed as follows (Thiel et al., 2007), analogously the standard SVM can be extended to solve a multi-class problem:

$$M_{i,j} = \{(x_\mu, m_{\mu,i}^+) | m_{\mu,i}^+ = l_{\mu,i}\} \cup \{(x_\mu, m_{\mu,j}^-) | m_{\mu,j}^- = l_{\mu,j}\} \tag{34}$$

Now suppose we have trained all of the one-vs.-one SVM we can construct a fuzzy output vector as follows:

1. Given a previously unseen sample $z \in \mathbb{R}^n$ and trained SVM $S_{i,j}$, $\forall i, j = 1, \ldots, k$.
2. Compute distances $d_{i,j}(z) \in \mathbb{R}$ for each class using all $S_{i,j}$ to the respective hyperplanes.
3. Transform those distances $d_{i,j}(z)$ using a logistic function $f(d_{i,j}) = 1/1 + \exp(-A d_{i,j})$, with $A \in \mathbb{R}$ and subject to optimization if required. A common fixed value is $A = 2$.
4. Finally a pairwise coupling as in Thiel et al. (2009) and Thiel (2009) is employed to receive the final fuzzy labels $\tilde{y}(z) \in \mathbb{R}^k$ one dimension for each of the k classes and with $\sum_{i=1}^{k} \tilde{y}_i(z) = 1$.

3.2. Fuzzy distance measure

In order to be able to compare the combined annotator opinion[4] $o_z \in \mathbb{R}^k$, with $\sum_{i=1}^{k} o_{z,i} = 1$, where $o_{z,i}$ denotes the annotator assigned membership to class i for any sample z and the fuzzy output of the F^{2}SVM $\tilde{y}(z)$, it is important to utilize a measure capturing the similarity or distance between o_z and $\tilde{y}(z)$. To achieve this, we employed the so-called S_1 measure, commonly used for fuzzy classifier fusion (Kuncheva et al., 2001; Kuncheva, 2001):

$$S_1(x, y) = \frac{\sum_{i=1}^{L} \min(x_i, y_i)}{\sum_{i=1}^{L} \max(x_i, y_i)} \in [0, 1], \tag{35}$$

capturing the similarity between x and y in this study, or the D_1 distance defined as

$$D_1(x, y) = 1 - S_1(x, y) \in [0, 1], \tag{36}$$

which is a pseudo metric sufficing the following properties:

- Non-negativity: $D_1(x, y) \geq 0, \forall x, y$.
- Identity of indiscernibles: $D_1(x, y) = 0 \Leftrightarrow x = y$.
- Symmetry: $D_1(x, y) = D_1(y, x)$.

Furthermore, the following properties hold:

- Maximal distance: $D_1(x, y) = 1 \Leftrightarrow \forall i : x_i = 0 \vee y_i = 0$.
- Minimal distance: $D_1(x, y) = 0 \Leftrightarrow x = y$.

In order to get a feel for the distance D_1 consider the following examples:

- $x = [1, 0, 0]$, $y = [1, 0, 0]$:

$$D_1(x, y) = 1 - \frac{1 + 0 + 0}{1 + 0 + 0} = 0$$

- $x = [1, 0, 0]$, $y = [0, 1, 0]$:

$$D_1(x, y) = 1 - \frac{0 + 0 + 0}{1 + 1 + 0} = 1$$

- $x = [0.5, 0.5, 0]$, $y = [0, 0.5, 0.5]$:

$$D_1(x, y) = 1 - \frac{0 + 0.5 + 0}{0.5 + 0.5 + 0.5} = \frac{2}{3}$$

- $x = [0.1, 0.9, 0]$, $y = [0, 0.9, 0.1]$:

$$D_1(x, y) = 1 - \frac{0 + 0.9 + 0}{0.1 + 0.9 + 0.1} = \frac{2}{11}$$

[4] The combined annotator opinion is the fuzzy target. In this study it has been derived following the algorithm in Fig. 7.

- $x = [0.4, 0.2, 0.4]$, $y = [0.3, 0.1, 0.6]$:

$$D_1(x, y) = 1 - \frac{0.3 + 0.1 + 0.4}{0.4 + 0.2 + 0.6} = \frac{1}{3}$$

As it might become obvious from the above examples crisp decisions as in the first two examples might result in the optimum (i.e. zero) or worst value (i.e. one) for the distance. Whereas, the distribution of membership values over all the classes might not result in the optimum but in measures above the minimum. As it should be shown in the last examples fitting trends in membership assignment result in values well above zero, but also far from one. Therefore, we would expect values of around 0.33 to be an indication of a close overall match between the fuzzy prediction and the target values in the results of the classification experiments using this fuzzy measure.

4. Evaluation

4.1. Speech data

There is a distinctive lack of widely available speech data with voice quality annotation. Further, as voice quality annotation schemes differ and as the annotator's interpretation of voice quality labels may not be consistent, it is difficult to combine multiple smaller data collections to form larger resources for evaluation and analysis.

4.1.1. Finnish vowel dataset

The speech data for this study comes from the recordings used in Airas and Alku (2007) and was provided to us by the authors. 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 voice quality 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. The impulse response of the recording system was obtained using a maximum length sequences (MLS) method (Rife and Vanderkooy, 1989). Analysis of the impulse response reveals a very flat amplitude response down to the very low frequencies, e.g., frequencies 60–100 Hz show a 0.026 dB variance in the amplitude response and frequencies 20–60 Hz show a variance of 0.768 dB. The phase response of the recording system is linear down to well below the lowest f_0 values. However, slight phase non-linearity in the very low frequencies was compensated for by convolving recorded signals with this impulse response time reversed.

4.1.2. Sentence dataset

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 high quality recording equipment (a B&K 4191 free-field microphone and a B&K 7749 pre-amplifier). Audio was digitized at 44.1 kHz (using a Lynx-two sound card) and then downsampled to 16 kHz. Sentences were then manually annotated on the phoneme level (446 phoneme segments in total).

The speaker is active in voice quality research and produced the sentences repeatedly until it was deemed that the voice quality had been successfully maintained throughout the entire sentence. This process required several iterations of re-recording. The sentences were then listened to independently by another researcher also experienced in voice quality research as well as with Laver's labeling scheme. She confirmed that the sentences were produced consistently using three discrete voice qualities and that these voice qualities corresponded to Laver's description of breathy, modal and tense. She did comment, however, that the perceptual distance of the breathy sentences compared to the modal sentences was greater than the tense sentences compared to the modal sentences. Following this process we are confident

that the sentences do indeed correspond to the voice quality labels and that they are suitable for use in the current study.[5]

4.2. Expert voice quality label assessment

For the purpose of the current study we required three independent sets of voice qualities to be used for the analysis. The initial labeling of the speech samples (i.e. breathy, normal and tense) was deemed inappropriate as a person's 'normal' voice quality could, for instance, be intrinsically breathy or tense and, hence, this would result in three overlapping sets of voice qualities. Instead we opted to use a voice quality labeling system based on Laver's framework (Laver, 1980). The label 'modal', which replaced 'normal', under Laver's framework has particular physiological and acoustic attributes which means that not every speaker's 'normal' voice quality would be considered 'modal'.

In order to decide on the new labeling, listening tests were conducted with three participants. All participants were experienced in voice quality research and were also familiar with Laver's labeling framework (Laver, 1980). The participants were required to rate the speech samples on a forced choice five point scale (see Table 6). The scale gave the option of choosing breathy/modal and modal/tense as well as the three individual labels. This allowed participants to indicate their uncertainty over the voice quality label if elements of the two voice qualities were perceived. Say, for instance, if breathiness was only weakly perceived (perhaps a weak sensation of aspiration noise) the option breathy/modal could be chosen. Note that there is an assumption here that the voice qualities breathy, modal and tense lie on a continuum. Such an assumption has also be made in previous studies (e.g., Gobl and Ní Chasaide, 2003). Other studies (e.g., Edmondson et al., 2001) have reported on the possibility for the voice quality compound *breathy-tense*, however, such a voice quality was deemed not to be contained within the current datasets.

Samples were presented to the participants in a randomized order, resulting in an inter-rater agreement of $\kappa = 0.526$. A criteria was defined based on the numerical values of the participants ratings in order to determine whether samples were to be considered 'included' or 'excluded'. If the mean rating for the given sample was more than 0.75 (in numerical scores) away from its original voice quality label then the sample was considered 'excluded'. Samples were also excluded if the standard deviation of its rating were more than 1, as this demonstrated disagreement on the part of the participants on the labeling to be used. This resulted in 314 of the 792 total samples being considered 'excluded', with 478 being allocated the class 'included'. For the subset 'included', an inter-rater agreement of $\kappa = 0.717$ is reached. This subset contains no samples for which the maximum annotator opinion, i.e. the voice quality that received the most support from the experts, diverges from the actual intended label, rendering a fair comparison between the crisp and fuzzy-input approaches. To be precise, the distribution of classes in the reduced dataset included 170 breathy (i.e. 35.5%), 170 modal (i.e. 35.5%), and 138 tense (i.e. 29.0%) samples respectively, which is an almost even distribution of samples. In the second part of the experiments described in Section 5.3 the whole set of 792 samples was used.

5. Experiments and results

In the following we have listed the results of the manifold recognition experiments (11 overall) that we conducted, Table 2 lists all the conducted experiments, the used data and targets. A schematic overview of the experimental setup is shown in Fig. 6. Basically, there is one major distinction separating the experiments into two groups, namely the crisp classification experiments, for benchmarking and comparison to standard methods, and the fuzzy output experiments, with the target to match the mixed votes of the annotators. For the crisp experiments we chose the reduced dataset,[6] by excluding the samples for which the annotators did not clearly choose the targeted voice quality as described in Section 4.2. The standard methods of choice for comparison were naive Bayes classifier (NB), giving a rough baseline, and standard crisp support vector machines (SVMs) utilizing the same radial basis function (RBF) kernel as the fuzzy-input fuzzy-output support vector machines (F^2SVMs). Both SVM types utilize the one-against-one multi-class paradigm. For all the experiments we conduct a z-score transformation of the features based on the data within the respective training sets before classification.

[5] The sound files of the 30 sentences used in the current study are included in the paper submission.

[6] Set only includes samples for which the intended voice quality coincides with the maximum membership assignment by the annotators.

Table 2
List of the 11 experiments conducted with their respective training and test dataset, as well as the targets, and methods.

Experiment	Train data	Test data	Train targets	Test targets	Method
Crisp classification experiments: Sections 5.1 and 5.2					
Cross validation (90 %/10 % split)	Vowels	Vowels	Actual label	Actual label	Naive Bayes Standard SVM F^{2}SVM
Leave one speaker out (10/1 speaker split)	Vowels	Vowels	Actual label	Actual label	Naive Bayes Standard SVM F^{2}SVM
Cross corpus	Vowels	Sentence data	Actual label	Actual label	Naive Bayes Standard SVM F^{2}SVM
Fuzzy classification experiments: Section 5.3					
Cross validation	Vowels	Vowels	Membership	Membership	F^{2}SVM
Leave one speaker out	Vowels	Vowels	Membership	Membership	F^{2}SVM

Table 3
Error (in %) comparison of Naive Bayes (NB), standard SVM and crisp F^{2}SVM outputs for cross validation (10-fold) and leave one speaker out experiments.

	Cross validation		Leave out one speaker	
	Err. (%)	Std.	Err. (%)	Std.
Naive Bayes	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**

The error (Err.) and standard deviation (Std.) are calculated by comparing to the true label. The experiments were conducted on the reduced dataset.
* Significant differences to the best performing approach F^{2}SVM in paired t-tests with $p < 0.05$.
** Significant differences to the best performing approach F^{2}SVM in paired t-tests with $p < 0.01$.

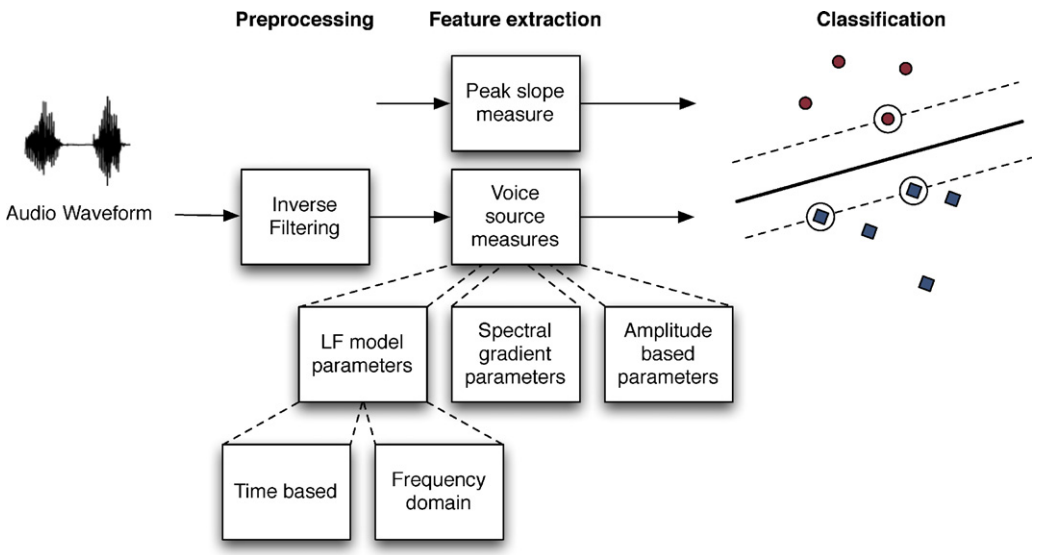

Fig. 6. Schematic system overview, including data preprocessing (i.e. inverse filtering), feature extraction, and classification.

Furthermore, for both groups an additional experiment partition criterium exists: 10-fold cross validation experiments as well as leave one speaker out experiments were conducted. In the former we randomized the dataset and split it into 10 training and test sets, plus one validation fold for parameter optimization. The leave one speaker out experiment, in order to verify generalization abilities of the approaches over different speakers within the Finnish

```
Algorithm 5.1: membershipAssignment()

comment: Assign memberships m_x

for each x ∈ data
  m_x ← {0, 0, 0}
  for each e ∈ experts
    l_x ← e(x)
    if l_x is 1
    m_x(breathy) ← m_x(breathy) + 1
    if l_x is 2
    m_x(breathy) ← m_x(breathy) + 0.5
    m_x(modal) ← m_x(modal) + 0.5
    if l_x is 3
    m_x(modal) ← m_x(modal) + 1
    if l_x is 4
    m_x(modal) ← m_x(modal) + 0.5
    m_x(tense) ← m_x(tense) + 0.5
    if l_x is 5
    m_x(tense) ← m_x(tense) + 1
  m_x ← m_x/n_exp
  return (m_x)
```

Fig. 7. Pseudocode snippet explaining the assignment of membership values to the recordings. x denotes a recording, m_x the membership assignment for x, e an expert, and $e(x)$ the opinion of expert e on sample x. n_{exp} is the number of available experts. The final membership assignment with $\sum m_x = 1$ is returned.

vowel set data, was executed as follows: for each fold one of the eleven speakers was left out of the training set and was solely used for testing.

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 introduced in Section 4.1. In the cross corpus experiment we train the classifier utilizing the data from one dataset and test it in an entirely new setting with data from another corpus. This type of experiment poses a considerable challenge to the classifier as strong generalization capabilities are required.

For the F^{2}SVM experiments it was necessary to generate fuzzy inputs resembling the degree of membership of each sample towards the three voice qualities. For each of the recordings these membership values were calculated using the labels mentioned in Table 6, as indicated by all the experts, and following the pseudo algorithm shown in Fig. 7. These newly calculated values were then used as the teacher signal for the F^{2}SVM in the experiments. In Fig. 8 the assigned memberships for each of the voice qualities as calculated are visualized using box plots.

5.1. Crisp classification experiments

In Table 3 the error rates of all of the crisp classification experiments are listed. It is clearly visible that in all the experiments the F^{2}SVM outperforms the other baseline approaches. For the cross validation 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; $\sigma = 3.89$). The standard SVM, receiving the actual label as target in training, produced 16.09% error ($\sigma = 4.59$) in the cross validation and 18.33% ($\sigma = 6.99$) in leave one speaker out. Both times the F^{2}SVM outperforms the standard SVM statistically significantly in paired t-tests (cross validation $p = 0.02$; leave one speaker out $p = 0.04$). The baseline performance of the naive Bayes classifier results in errors slightly over

Table 4
Comparison of confusion matrices using Naive Bayes, standard SVM and F^{2}SVM approaches for crisp cross validation (10-fold) experiments with all speakers.

	Naive Bayes			SVM			F^{2}SVM		
	Breathy	Modal	Tense	Breathy	Modal	Tense	Breathy	Modal	Tense
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**

Numbers are hit rates and lines sum up to one for each confusion matrix modulo rounding errors.

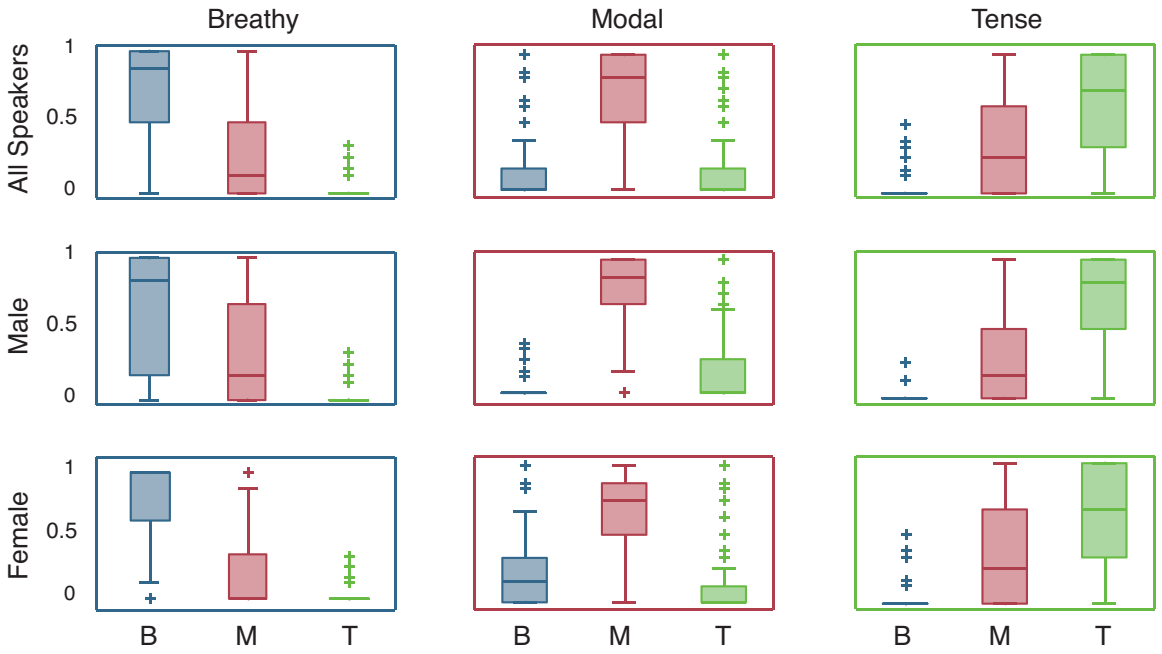

Fig. 8. Distribution of the memberships for the reduced vowel dataset as assigned by the annotators following the membership assignment algorithm in Fig. 7. The top row resembles the membership assignments for all speakers, the middle row for male speakers and the bottom row for female speakers. The letters denote the voice quality by the starting letter, i.e. B . . .breathy, M . . .modal, and T . . .tense. The center line denotes the median and the box is limited by the third and first quartile. Whiskers include the furthest outlying points that are not yet outliers, i.e. more than about 2.5 times the standard deviation, away from the median. Outliers marked as crosses are further away from the median.

Table 5
Comparison of confusion matrices using Naive Bayes, standard SVM and F^2SVM approaches for crisp leave one speaker out experiments with all speakers (eleven speakers).

	Naive Bayes			SVM			F^2SVM		
	Breathy	Modal	Tense	Breathy	Modal	Tense	Breathy	Modal	Tense
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**

Numbers are hit rates and lines sum up to one for each confusion matrix modulo rounding errors.

20% for both the cross validation and the leave one speaker out experiment. Both times the naive Bayes classifier is strongly outperformed by the F^2SVM with significant differences (cross validation $p < 0.001$; leave one speaker out $p = 0.008$). No statistically significant difference between the standard SVM and the naive Bayes classifier was found.

The confusion matrices of these experiments can be seen in Table 4 (cross validation experiment) and Table 5 (leave one speaker out experiment). All approaches result in very similar confusion matrices where almost no confusion

Table 6
Possible voice quality labels to be given to speech samples by expert judges (left column) and their numerical value (right column).

Voice quality	Number given
Breathy	1
Breathy/modal	2
Modal	3
Modal/tense	4
Tense	5

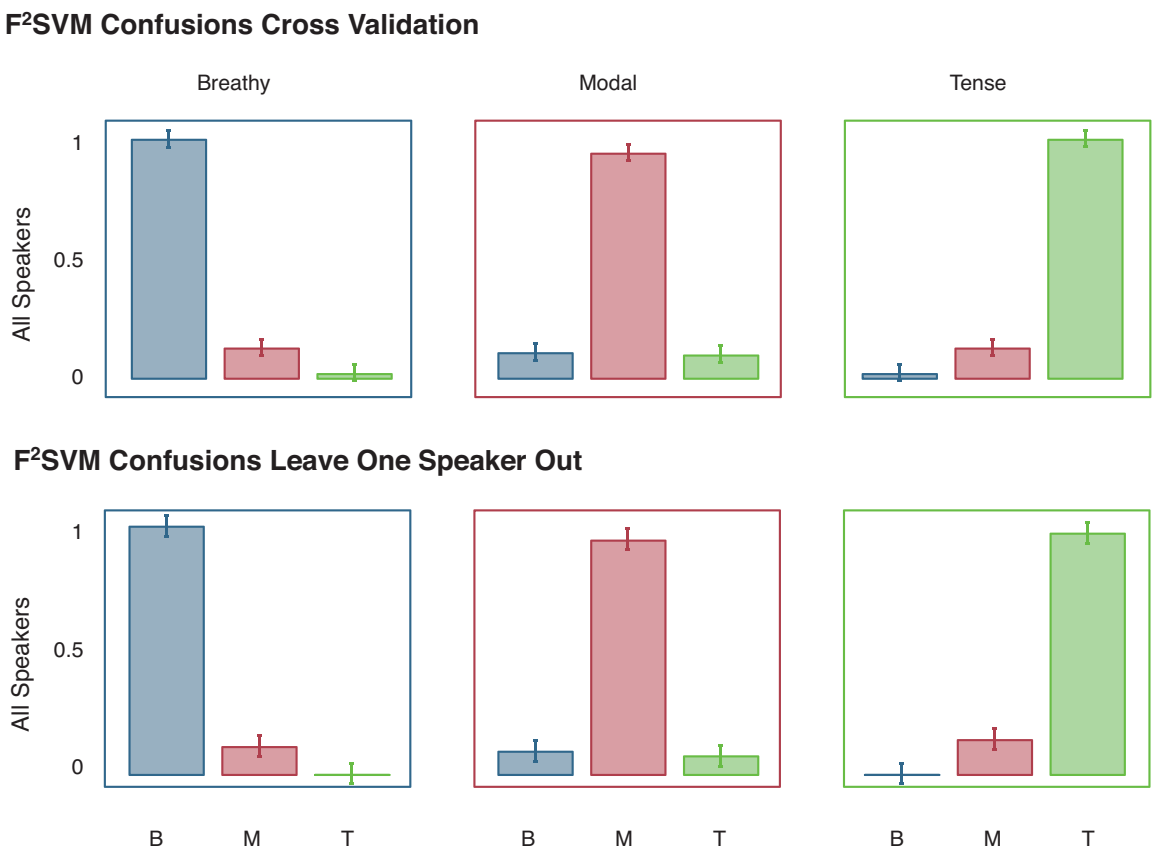

Fig. 9. Distribution of the confusions for the reduced vowel dataset as assigned by the F^{2}SVM for the crisp cross validation (top row) and leave one speaker out experiments (bottom row). The letters denote the voice quality by the starting letter, i.e. B . . .breathy, M . . .modal, and T . . .tense. Values are in accuracy rate and error bars are given over the single folds of the validation.

between breathy and tense voice qualities are present. For the F^{2}SVM and the naive Bayes classifier these errors are not reported in the cross validation experiments, further, in the leave one speaker out experiment they do not exceed 1%. In the standard SVM case breathy is confused with tense in 6% of the cases for the leave one speaker out experiment (only 3% in the cross validation experiment). The errors of the naive Bayes between neighboring voice qualities are, however, more frequent than in the other approaches. The confusions of the F^{2}SVM are further visualized in Fig. 9 and show strong resemblance of the human performance and distribution of labels shown in Fig. 8.

5.2. *Cross corpus experiments*

In order to further check the generalization ability of the approach a cross corpus experiment was conducted. All the mentioned methods, i.e. naive Bayes, 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 Table 7 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 voice quality. The temporally integrated decisions of the classifiers are computed by aggregating the frame-wise decisions over the full length of the sentence. 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.

The confusion matrices for the approaches on the frame level are shown in Table 8. It is seen that the reference approaches perform poorly for breathy and modal voice qualities, whereas they hardly ever confuse the tense quality.

Table 7
Error (in %) comparison of Naive Bayes, 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.

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

The classifiers are trained on the Finnish vowel set data and tested on the sentence data (compare Section 4.1). The error is calculated by comparing the classifiers' output to the target label of the sentences.

Table 8
Comparison of confusion matrices using Naive Bayes, standard SVM and F^{2}SVM approaches for cross corpus experiments. The classifiers are trained on the Finnish vowel set data and tested on the sentence data (compare Section 4.1). Numbers are hit rates and lines sum up to one for each confusion matrix modulo rounding errors.

	Naive Bayes			SVM			F^{2}SVM		
	Breathy	Modal	Tense	Breathy	Modal	Tense	Breathy	Modal	Tense
Breathy	0.65	0.35	0.00	0.56	0.43	0.01	**0.96**	0.04	0.00
Modal	0.03	0.57	0.40	0.02	0.54	0.44	0.12	**0.73**	0.15
Tense	0.00	0.11	0.89	0.00	0.10	**0.90**	0.01	0.18	0.80

In contrast to that the F^{2}SVM performs best with respect to the breathy voice quality. Confusions of breathy and tense are again very rare.

5.3. Fuzzy classification experiments

We conducted a second type of experiments in which we did not target high accuracy rates as in the experiments reported in Sections 5.1 and 5.2 and, but sufficiently low D_1 measures as described in Section 3.2. The D_1 distance measure indicates if the classifier's fuzzy output is close to the fuzzy annotation as assigned by the expert annotators. For the fuzzy classification experiments we utilized the full dataset available and the targets were the assigned memberships of the annotators to each of the samples as explained in Fig. 7. Since the standard SVM and naive Bayes approaches do not yield fuzzy outputs that are comparable to the F^{2}SVM results we decided to utilize the actual crisp label of a recording as well as the combined annotator opinion as a benchmark.

If the actual label of a recording x was breathy the fuzzy memberships $m_x = \{1, 0, 0\}$. This is then compared to the classification output as well as the annotator opinions using the distance measure D_1 introduced in Eq. 36. In Table 9 the full set of mean D_1 measures for the experiments is listed. The top line compares the perception of the annotators to the classification of the F^{2}SVM. The next one compares the classification with the actual target label of the recording. Finally the perception of the annotators is compared to the target labels, as a baseline. It is seen that the baseline outperforms the F^{2}SVM in all the cases, as expected. However, the mean D_1 measures for the classification are quite

Table 9
D_1 distance measures and standard deviations (Std.) for fuzzy F^{2}SVM outputs for cross validation (10-fold; X-Val) and leave one speaker out (LOSO) experiments and comparative values as baseline.

	X-Val		LOSO	
	D_1	Std.	$\mathbf{D}_1$	Std.
Classification vs. perception	0.3884	0.0288	0.3877	0.0260
Classification vs. target	0.5022	0.0320	0.5026	0.0325
Target vs. perception	0.3709	0.0256	0.3709	0.0496

No sample from the dataset was excluded for the experiments. Classification denotes the output of the F^{2}SVM, whereas perception denotes the memberships assigned according to the annotator opinions. Further, target stands for the actual label of the recording.

Table 10
Comparison of confusion matrices using the F^{2}SVM approach for cross validation (10-fold) and leave one speaker out experiments with all speakers.

	Cross validation			Leave out one speaker		
	Breathy	Modal	Tense	Breathy	Modal	Tense
Breathy	0.87	0.10	0.03	0.85	0.12	0.03
Modal	0.21	0.60	0.19	0.14	0.66	0.19
Tense	0.04	0.21	0.75	0.03	0.21	0.77

Numbers are hit rates and lines sum up to one for each confusion matrix modulo rounding errors. The maximum likelihood of the **fuzzy output** of the F^{2}SVM is compared to the maximum likelihood of the merged **annotator opinion**. Further, no sample has been removed from the dataset.

close to the human performance and there is no statistically significant difference found utilizing a t-test over the 10 folds of the cross validation and the leave one out speaker folds, respectively.

As an additional benchmark, we included the confusion matrices of the experiments by hardening the output of the classifier as well as the annotator opinion, i.e. the maximum of the assigned memberships resembles the crisp label $\max \tilde{y}(z)$ (compare $\tilde{y}(z)$ in Section 3.1.2). These matrices are found in Table 10. Since all of the samples available in the dataset were taken into consideration in these experiments the results show less accuracy, but the error types stay the same: Confusions involving modal are still responsible for the majority of the errors.

6. Discussion of statistical evaluation

The most striking result drawn 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 crisp classification experiments shown in Table 3. Therefore, it seems quite obvious that there is some 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 Section 4.2, allowing the annotator to assign a label between breathy and modal (the value 2 in Table 6) and a value between modal and tense (the value 4 in Table 6) is supported by the classification results shown in Section 5. This conclusion can be drawn since all the classifiers, comprising naive Bayes, standard SVM, and F^{2}SVM, confuse neighboring classes more often than the two extreme classes, breathy and tense. This statement is underlined by almost all the confusion matrices found in the paper. Furthermore, human perception results as shown in Fig. 8 show similar confusions due to the increased assignment of memberships to modal in the breathy and tense cases, as seen in the results shown in Fig. 9 for the F^{2}SVM.

Overall, the approach is apparently 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 Table 3).

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^{2}SVM clearly outperformed the reference approaches, with an overall accuracy of around 82% for the frame-wise decisions and for the single voice qualities breathy (96% accuracy) and modal (73% accuracy). The standard approaches recognized the voice quality tense with a higher accuracy of about 90%, however, the accuracies for the other two voice qualities were much lower (see Table 8). Further, after integrating the decisions over the whole sentences the accuracy rose up to more than 95%, meaning that only one out of the thirty sentences was confused. Whereas this result is encouraging, it needs to be stated that in natural recordings it is not expected that the voice quality of a speaker is constant over a full sentence as in the investigated corpus, nor are the boundaries of the voice quality known beforehand.

The interpretation of the fuzzy experimental results as reported in Section 5.3 is not as straightforward as for the crisp classification results. The output $\tilde{y}(x) \in \mathbb{R}$ (with $\sum_{i=1}^{k} \tilde{y}_i(x) = 1$) of the F^{2}SVM can be interpreted basically in two ways: the values $\tilde{y}_i(x)$ may be interpreted as degree of membership to class $i \in \{1, \ldots, k\}$, allowing for mixed

states of two or more voice qualities at the same time. Further, they can be interpreted in a probabilistic manner, where the values $\tilde{y}_i(x)$ denote the probability of x belonging to class i. And therefore we can interpret $i^* = arg\max_i \tilde{y}_i(x)$ as the crisp and most probable decision of the F^{2}SVM. In Section 5 we list results for both interpretations, we however believe that the degree of membership interpretation is more suitable for the current study, but it is harder to interpret in terms of accuracy and therefore we included both.

First for the degree of membership interpretation, it can be observed, that human performance is not quite reached. While comparing the distance of the actual class labels towards the expert annotations, with the one for the automatic classification towards the expert annotations, the former always outperforms the latter with respect to the D_1 measure. However, the values are not that much higher as seen in Table 9, and no significant difference was found between the observed D_1 measures.

By interpreting the results as probabilities for one class, we can gain another set of confusion matrices, that makes it a bit easier to quantify the error, however, it does not support the idea of mixed voice qualities being present to different degrees. The confusion matrices found in Table 10 show similar results as in the crisp classification experiments, but less accuracy, which is of course due to the additional samples used in the experiment.[7] The matrices are filled by comparing the most probable F^{2}SVM output to the most probable expert opinion. An accuracy of around 74% was achieved in the cross validation experiment. For the leave one speaker out experiments (all speakers) 76% accuracy was achieved.

7. Conclusion

In the present study we investigated the capability of F^{2}SVM to classify voice quality samples from a vowel corpus, as well as in a cross corpus study using data taken from full sentences. The results in Section 5 show relatively high accuracy rates for the proposed F^{2}SVM approach in comparison to the standard approaches in many of the experiments including cross validation and leave one speaker out validation conditions as well as when comparing crisp and fuzzy measures. Additionally, it has shown strong generalization capabilities in cross corpus analysis and leave one speaker out experiments. The proposed method outperformed its competitors (standard SVM, and naive Bayes) in crisp classification experiments clearly, by only utilizing the information present in fuzzy labels during training. This is a very encouraging result supporting the importance of fuzzy treatment of voice quality data and annotations. The results are very promising for future work including the extension of the approach to running speech and more naturalistic data.

Additionally the fuzzy outputs are advantageous, if the classifier is used in a larger architecture, where voice quality classification is only one of many predictions that are spatiotemporally integrated, such as it is proposed in Scherer et al. (2012). The information that is passed from layer to layer in the overall architecture is often error prone and early crisp decisions can often not be compensated for in later stages, whereas in fuzzy predictions all the information retains. It could be shown that classifiers relying on fuzzy outputs significantly outperform those, that combine early crisp decisions (Scherer et al., 2012).

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

When considering realistic data recorded outside ideal laboratory conditions the voice quality ground truth will again not be known. Therefore, expert ratings, that lead to fuzzy labels and mixed categories, are required to assess the quality of an automatic and robust voice quality classifier. Realistic data further requires alternative measures for the assessment of accuracy, which is why we introduced the D_1 measure to determine the distance of a fuzzy result towards a fuzzy target.

Unfortunately, the results here using the D_1 measure are not easily comparable with other studies or indeed with other classifiers in the present study. We chose the D_1 distance measure as it reflects the full distance between two fuzzy labels (i.e. the prediction and the target) and as it is an adaptation of the widely used S_1 similarity measure (Kuncheva,

[7] No sample of the 792 was excluded according to the criteria in Section 4.2 for all the fuzzy experiments.

2004). We hope that future research involving the classification of data with fuzzy ground truths (e.g., voice qualities and affective states) will utilize a measure of this nature which would facilitate the comparability of similar results.

8. Future work

Evidence has been provided in the present study, that the utilized features and classification methods are suitable for voice quality analysis and classification. One specific direction of our research is to define an optimal voice quality feature set (initially for breathy to tense voice qualities) following formal robustness testing as well as automatic feature selection experiments. Further, we wish to develop a separate feature set suited to a different dimension of voice quality (i.e. harshness, creakiness) and apply a similar classification approach. We then wish to migrate this approach (both in terms of acoustic features and classification method) to the analysis of more authentic productions of voice qualities recorded in naturalistic settings. A longer term goal would be to apply voice quality characterization to speech technology input and output applications in order to be able to adapt more naturalistically to the interacting users and their affective states.

Acknowledgments

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). This work was further supported as part of the FASTNET project – Focus on Action in Social Talk: Network Enabling Technology funded by Science Foundation Ireland (SFI) 09/IN.1/I2631 and by the Science Foundation Ireland (Grant 07/CE/I 1142) as part of the Centre for Next Generation Localisation (www.cngl.ie). We are very grateful to Dr. Matti Airas and Prof. Paavo Alku for providing us with the vowel dataset used in this study.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.csl.2012.06.001.

References

Airas, M., 2008. TKK aparat: an environment for voice inverse filtering and parameterization. Logopedics Phoniatrics Vocology 33 (1), 49–64.

Airas, M., Alku, P.,2007. Comparison of multiple voice source parameters in different phonation types. In: Proceedings of Interspeech 2007. ISCA, pp. 1410–1413.

Alku, P., Bäckström, T., Vilkman, E., 1992. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Communication 11 (2–3), 109–118.

Alku, P., Bäckström, T., Vilkman, E., 2002. Normalized amplitude quotient for parameterization of the glottal flow. Journal of the Acoustical Society of America 112 (2), 701–710.

Alku, P., Magi, C., Yrttiaho, S., Bäckström, T., Story, B., 2009. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. Journal of the Acoustical Society of America 125 (5), 3289–3305.

Bennett, K.P., Campbell, C., 2000. Support vector machines: hype or hallelujah? ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations Newsletter 2 (2), 1–13.

Bishop, C.M., October 2006. Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edition. Springer.

Blomgren, M., Chen, Y., Ng, M.L., Gilbert, H.R., 1998. Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. Journal of the Acoustical Society of America 103 (5), 2649–2658.

Borasca, B., Bruzzone, L., Carlin, L., Zusi, M.,2006. A fuzzy-input fuzzy-output SVM technique for classification of hyperspectral remote sensing images. In: Proceedings of the 7th Nordic Signal Processing Symposium, 2006 (NORSIG 2006). IEEE, pp. 2–5.

Cabral, J., Renals, S., Richmond, K., Yamagishi, J.,2008. Glottal spectral separation for parametric speech synthesis. In: Proceedings of Interspeech 2008. ISCA, pp. 1829–1832.

Campbell, N., 2004. Specifying affect and emotion for expressive speech synthesis. In: Gelbukh, A. (Ed.), Computational linguistics and intelligent text processing. Vol. 2945 of Lecture Notes in Computer Science. Springer, pp. 395–406.

Campbell, N., Mokhtari, P.,2003. Voice quality: the 4th prosodic dimension. In: Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003). ICPhS, pp. 2417–2420.

Campbell, W.N., 2007. On the Use of Non Verbal Speech Sounds in Human Communication. Vol. 4775 of Lecture Notes in Computer Science. Springer, pp. 117–128.

Childers, D.G., Lee, C.K., 1991. Voice quality factors: analysis, synthesis and perception. Journal of the Acoustical Society of America 90 (5), 2394–2410.

Drugman, T., Bozkurt, B., Dutoit, T.,2009. Complex cepstrum-based decomposition of speech for glottal source estimation. In: Proceedings of Interspeech 2009. ISCA, pp. 116–119.

Dubois, D., Prade, H., 1980. Fuzzy Sets and Systems: Theory and Applications. Academic Press, New York.

Edmondson, J., Esling, J., Harris, J., Shaoni, L., Ziwo, L., 2001. The aryepiglottic folds and voice quality in the yi and bai languages: laryngoscopic case studies. Mon-Khmer Studies: A Journal of Southeast Asian Linguistics and Languages 31, 83–100.

Fant, G., 1960. The Acoustic Theory of Speech Production, 2nd edition. Mouton De Gruyter, The Hague.

Fant, G., Liljencrants, J., Lin, Q., 1985. A four parameter model of glottal flow. KTH, Speech Transmission Laboratory, Quarterly Report 4, 1–13.

Fant, G., Liljencrants, J., Lin, Q., 1995. The LF-model revisited. transformations and frequency domain analysis. KTH, Speech Transmission Laboratory, Quarterly Report 2–3, 119–156.

Gobl, C., 1989. A preliminary study of acoustic voice quality correlates. KTH, Speech Transmission Laboratory, Quarterly Report 4, 9–21.

Gobl, C., 2003. The voice source in speech communication. Ph.D. Thesis, KTH Speech Music and Hearing, Stockholm.

Gobl, C., Chasaide, A.N.,2003. Amplitude-based source parameters for measuring voice quality. In: Proceedings of ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis (VOQUAL 2003). ISCA, pp. 151–156.

Gobl, C., Ní Chasaide, A., 1992. Acoustic characteristics of voice quality. Speech Communication 11, 481–490.

Gobl, C., Ní Chasaide, A., 2003. The role of voice quality in communicating emotion, mood and attitude. Speech Communication 40, 189–212.

Goncharoff, V., Gries, P.,1998. An algorithm for accurately marking pitch pulses in speech signals. In: Proceedings of IASTED International Conference on Signal and Image Processing (SIP 1998). IASTED/ACTA Press, Nevada, USA, pp. 281–284.

Henrich, N., d'Alessandro, C., Doval, B., 2001. Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data. In: Proceedings of EUROSPEECH, Scandanavia, pp. 47–50.

Hillenbrand, J., Cleveland, R., Erickson, R., 1994. Acoustic correlates of breathy vocal quality. Journal of Speech and Hearing Research 37, 769–778.

Hillenbrand, J., Houde, R., 1996. Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. Journal of Speech and Hearing Research 39, 311–321.

Huang, H.P., Liu, Y.H., 2002. Fuzzy support vector machines for pattern recognition and data mining. International Journal of Fuzzy Systems (4), 826–835.

Ishi, C.T., Sakakibara, K.I., Ishiguro, H., Hagita, N., 2008. A method for automatic detection of vocal fry. IEEE Transactions on Speech and Audio Processing 16 (1), 47–56.

Ito, M.,2004. Politeness and voice quality—the alternative method to measure aspiration noise. In: Proceedings of Speech Prosody 2004. ISCA, Nara, Japan, pp. 213–216.

Kahsay, L., Schwenker, F., Palm, G., 2005. Comparison of multiclass svm decomposition schemes for visual object recognition. In: Kropatsch, W.G., Sablatnig, R., Hanbury, A. (Eds.), Pattern Recognition. Vol. 3663 of Lecture Notes in Computer Science. Springer, pp. 334–341.

Kane, J., Gobl, C., 2011. Identifying regions of non-modal phonation using features of the wavelet transform. In: Proceedings of Interspeech, Florence, Italy, pp. 177–180.

Kane, J., Kane, M., Gobl, C.,2010. A spectral LF model based approach to voice source parameterisation. In: Proceedings of Interspeech 2010. ISCA, pp. 2606–2609.

Kreiman, J., Gerratt, B.R., Antonanzas-Barroso, N., 2006. Analysis and Synthesis of Pathological Voice Quality (Software Manual). The Regents of the University of California.

Krishnamurthy, A.K., Childers, D.G., 1986. Two-channel speech analysis. IEEE Transactions on Acoustics, Speech and Signal Processing 34 (8), 730–743.

Kuncheva, L.I., 2001. Using measures of similarity and inclusion for multiple classifier fusion by decision templates. Fuzzy Sets and Systems 122 (3), 401–407.

Kuncheva, L.I., 2004. Combining Pattern Classifiers: Methods and Algorithms. Wiley.

Kuncheva, L.I., Bezdek, J.C., Duin, R.P.W., 2001. Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognition 34 (2), 299–314.

Ladefoged, P., Maddieson, I., 1996. The Sounds of the World's Languages. Blackwell.

Laver, J., 1979. The description of voice quality in general phonetic theory. Edinburgh University Department of Linguistics Work in Progress 12, 30–52.

Laver, J., 1980. The Phonetic Description of Voice Quality. Cambridge University Press.

Li, H., Scaife, R., O' Brien, D., 2011. LF model based glottal source parameter estimation by extended Kalman filtering. In: Proceedings of the Irish Signals and Systems Conference (ISSC 2011).

Lin, C.F., Wang, S.D., 2002. Fuzzy support vector machines. IEEE Transactions on Neural Networks (13), 464–471.

Lugger, M., Stimm, F., Yang, B.,2008. Extracting voice quality contours using discrete hidden Markov models. In: Proceedings of Speech Prosody 2008. ISCA, Campinas, Brazil, pp. 29–32.

Lugger, M., Yang, B.,2006. Classification of different speaking groups by means of voice quality parameters. In: Proceedings of ITG-Fachtagung Sprach-Kommunikation. VDE.

Lugger, M., Yang, B.,2008. Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008). IEEE, pp. 4945–4948.

Lugger, M., Yang, B., Wokurek, W.,2006. Robust estimation of voice quality parameters under real world disturbances. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006). IEEE, pp. 1110–1197.

Mackenzie Beck, J., 2005. Perceptual analysis of voice quality: the place of vocal profile analysis. In: Laver, J., Hardcastle, W., Beck, J.M. (Eds.), A Figure of Speech: A Festschrift for John Laver. , pp. 285–322 (Chapter 12).

Markel, J., Gray, A., 1982. Linear Prediction of Speech. Springer, New York.

Ogden, R., 2001. Turn transition, creak and glottal stop in Finnish talk-in-interaction. Journal of the International Phonetic Association 31 (1), 139–152.

Raito, T., Suni, A., Pulakka, H., Vainio, M., Alku, P.,2008. HMM-based Finnish text-to-speech system utilizing glottal inverse filtering. In: Proceedings of Interspeech 2008. ISCA, pp. 1881–1884.

Rife, D.D., Vanderkooy, J., 1989. Transfer-function measurement with maximum-length sequences. Journal of the Audio Engineering Society 37, 102–113.

Scherer, S., Glodek, M., Layher, G., Schels, M., Schmidt, M., Brosch, T., Tschechne, S., Schwenker, F., Neumann, H., Palm, G., 2012. A generic framework for the inference of user states in human computer interaction: how patterns of low level communicational cues support complex affective states. Journal on Multimodal User Interfaces, special issue on: Conceptual Frameworks for Multimodal Social Signal Processing, 1–25.

Schölkopf, B., Smola, A.J., 2001. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA.

Schwenker, F., 2001. Solving multi-class pattern recognition problems with tree-structured support vector machines. In: Radig, B., Florczyk, S. (Eds.), DAGM-Symposium. Vol. 2191 of Lecture Notes in Computer Science. Springer, pp. 283–290.

Stevens, K., Hanson, H., 1994. Classification of glottal vibration from acoustic measurements. Vocal Fold Physiology: Voice Quality Control, 147–170.

Strik, H., Cranen, B., Boves, L.,1993. Fitting a LF-model to inverse filter signals. In: Proceedings of the European Conference on Speech Communication and Technology (Eurospeech). ISCA, Berlin, Germany, pp. 103–106.

Sturmel, N., d'Alessandro, C., Rigaud, F.,2009. Glottal closure instant detection using lines of maximum amplitudes (LOMA) of the wavelet transform. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009). IEEE, pp. 4517–4520.

Thiel, C., 2009. Multiple classifier systems incorporating uncertainty. Ph.D. Thesis, Ulm University.

Thiel, C., Giacco, F., Schwenker, F., Palm, G.,2009. Comparison of neural classification algorithms applied to land cover mapping. In: Proceeding of the 2009 Conference on New Directions in Neural Networks. IOS Press, Amsterdam, The Netherlands, pp. 254–263.

Thiel, C., Scherer, S., Schwenker, F.,2007. Fuzzy-input fuzzy-output one-against-all support vector machines. In: 11th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES 2007). Vol. 3 of Lecture Notes in Artificial Intelligence. Springer, pp. 156–165.

Walker, J., Murphy, P., 2007. A review of glottal waveform analysis. In: Stylianou, Y., Faundez-Zanuy, M., Esposito, A. (Eds.), Progress in Nonlinear Speech Processing. Springer, pp. 1–21.

Wester, M.,1998. Automatic classification of voice quality: comparing regression models and hidden Markov models. In: Proceedings of the Symposium on Databases in Voice Quality Research and Education (VOICEDATA 1998). Utrecht Institute of Linguistics OTS, Utrecht, pp. 92–97.

Yanushevskaya, I., Gobl, C., Ní Chasaide, A.,2005. Voice quality and f_0 cues for affect expression. In: Proceedings of Interspeech 2005. ISCA, pp. 1849–1852.