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.