SINGER-DEPENDENT FALSETTO DETECTION FOR LIVE VOCAL PROCESSING BASED ON SUPPORT VECTOR CLASSIFICATION

Gautham J. Mysore, Ryan J. Cassidy, Julius O. Smith III

Center for Computer Research in Music and Acoustics (CCRMA)
Stanford University
Stanford, CA 94305

ABSTRACT

We present and analyze a machine learning technique to determine from an input sung vocal waveform if falsetto (also known as the *head voice*) is being used. Such a system may be used to tune signal processing parameters, ideally in real-time, for such applications as intelligibility enhancement of high-pitched sung notes, and other musical systems which tune signal processing parameters according to detected performance parameters. Our falsetto detector uses a support vector classifier trained on Mel-Frequency Cepstral Coefficients (MFCCs) computed from a newly collected database of anechoic sung notes. It is shown to give correct classification with better than 95% accuracy.

1. INTRODUCTION

An important problem in musical signal processing is the adaptation of signal processing parameters to detected (and usually hidden) musical parameters. In one recent product [1], the parameters of an effects device (e.g., an audio equalizer used for aesthetic sound modifications) are modified in sync with a pre-programmed tempo. In a related application [2], a system automatically generates cues and musical accompaniment from detected performance changes dictated by a musical score. We attempt to solve another related problem: a vocalist wishes to change the processing applied to his voice when he sings in his falsetto, or the upper portion of his vocal range, versus when he sings in his modal voice, or the lower portion of his range. One example of such a processing change involves the recently reviewed fact that the intelligibility of soprano-sung vowels is reduced when singing in the head voice[1] [3], and thus it would be advantageous to provide certain enhancement processing when such notes are sung. A block diagram of such a system is shown in Figure 1.

This work focuses on the following sub-problem: given a sung input waveform, determine whether or not the vocalist in question is employing his falsetto. In this way, the detection system should yield a discrete variable $\hat{f}(m)$ equal to 1 to indicate falsetto, and equal to -1 otherwise, for each frame[2] m.

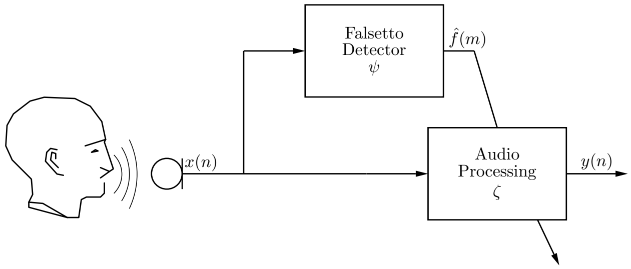

Fig. 1. Main block diagram for falsetto-driven audio processing. This work focuses on the design of the falsetto detector.

2. FALSETTO ACOUSTICS

Sundberg [4] defines a *register* as "a phonation frequency range in which all tones are perceived as being produced in a similar way and which possess a similar voice timbre." Rossing [5] further defines two such registers which correspond to different modes of vocal fold vibration: *modal voice* (or *chest voice* [4]), and *head voice* (or *falsetto* for male singers). While the former corresponds to lower phonation frequencies, the latter corresponds to high frequencies. Unfortunately, there is, as of yet, no precise physical or spectral definition as to what separates the two modes, and so we must rely on a trained subject's knowledge as to when he/she is singing in which mode (this is the method adopted in various prior scientific studies (see, e.g., [4]).

What is known physically, however, is that the male falsetto arises from the so-called "light" vibration of the vocal folds [5], in which there is relatively little contact between the folds for the duration of the pitch cycle they produce. By contrast, there is a distinct opening and closing of the folds for each pitch period produced in the "heavy" mechanism of the folds, which is associated with the (lower-pitched) male "chest voice." The reduced contact of the folds during a falsetto pitch cycle results from the greater tension applied to the folds by the glottal muscles, which is coincident with the falsetto's higher pitch.

3. DATA COLLECTION

A significant database of sung musical notes has been collected in an anechoic environment. In total, 3–6 takes of 35 pitches spanning the range of an individual male subject[3] were recorded. For each note, the subject sung the vowel "Aah" exclusively. The subject was recorded using a Bruel & Kjaer 4133 microphone spaced 1–2 m from the singer's mouth, and powered by a Bruel & Kjaer

[1] Here *head voice* is the gender-neutral term for sounds produced by male and female singers when they sing in the upper portion of their range; *falsetto* is a term often applied to the male head voice.

[2] Assume a frame rate of 50 Hz.

[3] Each semitone between G2 and F♯5 (35 semitones in total) was recorded.

2690 Conditioning Amplifier. The recording was performed in an anechoic chamber with a volume of approximately 1000 cubic feet. A laptop-based audio recording device was used to record the pre-amplified vocal signal.

4. MEL-FREQUENCY CEPSTRAL COEFFICIENTS

A wide assortment of features computed from the input waveform may be considered. For the purposes of this work, we have selected a conventional feature, the Mel-Frequency Cepstral Coefficients (MFCCs) [6, 7].

MFCCs [6] have been a popular feature in speech recognition for several decades. They are obtained by computing the Discrete-Cosine Transform (DCT) of the log-energy non-uniform filter bank outputs:

$$\mathit{MFCC}_c = \sum_{b=1}^{N_b} \mathit{LE}(b) \cos\left(c(b-1/2)\frac{\pi}{N_b}\right), c = 1, 2, \ldots, N_c, \tag{1}$$

where N_c is the number of coefficients, N_b is the number of bands in the non-uniform filter bank, b is the band index, and c is the coefficient index. The log-energy outputs $LE(b)$ for each band b are computed as

$$\mathit{LE}(b) = \log_{10}\left(\sum_{k=0}^{N-1} |X_m(k)|\, H_b(k)\right), b = 1, 2, \ldots, N_b, \tag{2}$$

where $X_m(k)$ is the Discrete Fourier Transform (DFT) of the input $x(n)$ evaluated at frame m as

$$X_m(k) = \sum_{n=0}^{N-1} x(n+mR) \exp\left(\frac{-j2\pi nk}{N}\right), k = 0, 1, \ldots, N-1, \tag{3}$$

and the non-uniform filter bank responses $H_b(k)$ are typically set to be approximately triangular in shape.

The filter-bank details of different MFCC implementations vary from one implementation to the next [8]. We have adopted the MFCC implementation of Slaney [7], which uses 40 equal-area triangular filters, with uniform band spacing up to 1000 Hz, and logarithmic spacing above 1000 Hz. For each frame[4], $N_c = 13$ cepstral coefficients are generated.

5. SUPPORT VECTOR CLASSIFICATION

The support vector machine (SVM) [9] is a supervised machine learning algorithm that has been used for our classification task of distinguishing MFCCs obtained from falsetto waveforms from those obtained from modal waveforms. The SVM can find an optimal classification boundary (hyperplane) by the solving the following primal convex optimization problem:

$$\begin{array}{ll} \min_{\xi,w,b} & \frac{1}{2}||w||^2 + C\sum_{i=1}^m \xi_i \\ \text{s. t.} & y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i, \quad i = 1, \ldots, m \\ & \xi_i \geq 0, i = 1, \ldots, m. \end{array} \tag{4}$$

where $x^{(i)}$ is feature vector i and $y^{(i)}$ is the corresponding label (1 or -1) $\forall i$. In our application, a feature vector $x^{(i)}$ corresponds to a single frame of MFCC data, with the corresponding $y^{(i)}$ equal to 1 to indicate a falsetto frame, and equal to -1 to indicate a modal frame.

If the data is linearly separable, the solution to this problem yields the maximum margin hyperplane. However, if the data is not linearly separable, some degree of misclassification of training samples is allowed. A competing objective (to having a large margin) is to reduce this degree of misclassification. The parameter C controls the trade-off between these 2 competing objectives.

The dual problem of the above primal convex optimization problem is as follows:

$$\begin{array}{ll} \max_\alpha & W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i,j=1}^m y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)}, x^{(j)}\rangle \\ \text{s. t.} & 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m \\ & \sum_{i=1}^m \alpha_i y^{(i)} = 0. \end{array} \tag{5}$$

As the Karush-Kuhn-Tucker (KKT) conditions are satisfied in this formulation of the dual problem, there is zero duality gap. We can therefore solve this dual problem in lieu of solving the primal problem.

Since every occurrence of a training sample in the dual problem is in the form of an inner product with another training sample, we can replace each inner product $\langle x^{(i)}, x^{(j)}\rangle$ with $\langle\phi(x^{(i)}), \phi(x^{(j)})\rangle$. This is effectively transforming each feature vector to a higher dimensional space using the feature mapping ϕ. It is more likely that the data set will be linearly separable in a higher dimensional space than in the original feature space.

The inner products in the higher dimensional space can be specified using a kernel:

$$K(x^{(i)}, x^{(j)}) = \langle\phi(x^{(i)}), \phi(x^{(j)})\rangle \tag{6}$$

These inner products do not have to be explicitly computed. We can instead specify a function (known as a Mercer kernel) which implicitly corresponds to transforming the feature vectors to a higher dimensional space and then taking their inner products.

In this work, we consider a linear kernel and a Gaussian (or RBF) kernel. The linear kernel (no transformation) is given by:

$$K(x^{(i)}, x^{(j)}) = \langle x^{(i)}, x^{(j)}\rangle \tag{7}$$

The Gaussian kernel is given by:

$$K_G(x^{(i)}, x^{(j)}) = \exp\left(-\frac{||x^{(i)} - x^{(j)}||^2}{2\sigma^2}\right). \tag{8}$$

When using the Gaussian kernel, we have to specify the parameter σ in addition to the original parameter C in the optimization problem. We consider a number of different values of the parameters C and σ and choose the parameter values that yield the lowest estimated generalization error.

These algorithms have been implemented using Spider [10] (an object oriented environment for machine learning in Matlab).

6. RESULTS

7-fold cross validation [11] was used to determine the optimal values of the parameters (C and σ). The cross validation was performed on a training set that consists of 21 modal notes and 21 falsetto notes, and therefore split into folds of 6 notes each (each

[4] All input waveforms have a sampling rate of 44.1 kHz; the frame rate is 50 Hz, and the frame size is 256 samples (thus, the frame expansion is undersampled). If lower latency is desired, a higher frame rate may be used, with a corresponding increase in computational requirements.

notes corresponds to 24 feature vectors). For each of the kernels (linear and Gaussian), we chose the parameters that yielded the lowest estimated generalization error. For the linear kernel, we used 100 values of C logarithmically spaced between 10^{-3} and 10. For the Gaussian kernel, we used the same range of values of C and 100 values of σ logarithmically spaced between 10^{-2} and 100. We therefore used 10000 possible combinations of parameters for the Gaussian kernel.

Once we chose the optimal value of C for the linear kernel, we retrained the SVM on the entire training set. We then tested on a test set that consists of 7 modal notes and 7 falsetto notes, obtaining the test error. We repeated this procedure with the optimal values of C and σ for the Gaussian kernel, obtaining the test error for this kernel.

The results of our study are shown in Table 1. Appealingly, both errors are less than 5%, with the Gaussian kernel outperforming the linear kernel. Plots of the error performance of each kernel as a function of C and σ (Gaussian kernel only) are shown in Figure 2 and Figure 3.

Kernel	Test Error	Optimal C	Optimal σ
Linear	4.8%	0.0413	N/A
Gaussian	1.5%	2.057	2.42

Table 1. Table showing results of SVM classification with linear and Gaussian kernels.

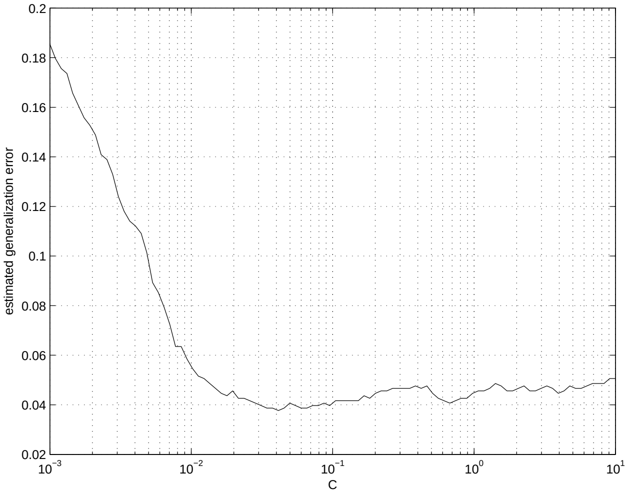

Fig. 2. Graph showing estimated generalization error for the linear kernel versus the value of the parameter C. For optimal C, the SVM achieves better than 95% correct classification.

7. CONCLUSION

We have demonstrated techniques that determine from an input waveform whether a vocalist is singing in his falsetto, with an accuracy greater than 95% on recorded data. Future possible enhancements include the use of alternative features (e.g. relative amplitude of partials in the sung note).

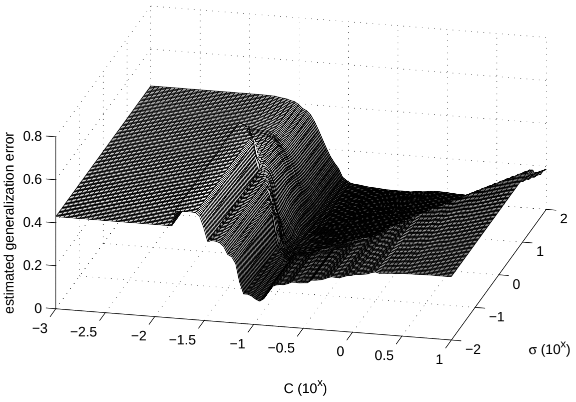

Fig. 3. Graph showing estimated generalization error for the Gaussian kernel versus the value of the parameters C and σ. For optimal values, the SVM achieves better than 98% correct classification.

8. REFERENCES

[1] Paul White, “Roger Linn AdrenaLinn II: Filter, effects, amp modeller & drum box,” *Sound On Sound*, pp. 128–130, Mar. 2004.

[2] M. Puckette and A. C. Lippe, “Score following in practice,” in *Proceedings of the 1992 International Computer Music Conference, San Jose*. 1992, Computer Music Association.

[3] Elodie Joliveau, John Smith, and Joe Wolfe, “Vocal tract resonances in singing: The soprano voice,” *Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2434–2439, Oct. 2004.

[4] Johan Sundberg, *The Science of the Singing Voice*, Northern Illinois University Press, Dekalb, IL, 1987.

[5] Thomas D. Rossing, F. Richard Moore, and Paul A. Wheeler, *The Science of Sound*, Addison Wesley, San Francisco, 2002.

[6] Steven B. Davis and Paul Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[7] Malcolm Slaney, “Auditory toolbox. version 2,” Tech. Rep. 1998-010, Interval Research Corporation, 1998, available online at http://rvl4.ecn.purdue.edu/˜malcolm/interval/1998-010/.

[8] Todor Ganchev, Nikos Fakotakis, and George Kokkinakis, “Comparative evaluation of various MFCC implementations on the speaker verification task,” in *Proceedings of the 10th International Conference “Speech and Computer” (SPECOM 2005), Moscow, Russia*. 2005, pp. 191–194, ELSNET, available online at http://www.wcl.ee.upatras.gr/ai/papers/ganchev17.pdf.

[9] Christopher J. C. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[10] Jason Weston, Andre Elisseeff, Gokhan Bakir, and Fabian Sinz, *The Spider*, 2006, available online at http://www.kyb.tuebingen.mpg.de/bs/people/spider/main.html.

[11] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, Berlin, 2003.