ESTIMATION OF SAXOPHONE CONTROL PARAMETERS BY CONVEX OPTIMIZATION

Cheng-i Wang[1], Tamara Smyth[1], Zachary C. Lipton[2]
[1]*Music Department, University of California, San Diego*
[2]*Computer Science and Engineering, University of California, San Diego*
Correspondence should be addressed to: chw160@ucsd.edu

Abstract: **In this work, an approach to jointly estimating the tone hole configuration (fingering) and reed model parameters of a saxophone is presented. The problem isn't one of merely estimating pitch as one applied fingering can be used to produce several different pitches by *bugling* or overblowing. Nor can a fingering be estimated solely by the spectral envelope of the produced sound (as it might for estimation of vocal tract shape in speech) since one fingering can produce markedly different spectral envelopes depending on the player's embouchure and control of the reed. The problem is therefore addressed by jointly estimating both the reed (source) parameters and the fingering (filter) of a saxophone model using convex optimization and 1) a bank of filter frequency responses derived from measurement of the saxophone configured with all possible fingerings and 2) sample recordings of notes produced using all possible fingerings, played with different overblowing, dynamics and timbre. The saxophone model couples one of several possible frequency response pairs (corresponding to the applied fingering), and a quasi-static reed model generating input pressure at the mouthpiece, with control parameters being blowing pressure and reed stiffness. Applied fingering and reed parameters are estimated for a given recording by formalizing a minimization problem, where the cost function is the error between the recording and the synthesized sound produced by the model having incremental parameter values for blowing pressure and reed stiffness. The minimization problem is nonlinear and not differentiable and is made solvable using convex optimization. The performance of the fingering identification is evaluated with better accuracy than previous reported value.**

1. INTRODUCTION

The problem of inverse modeling acoustic instruments is well addressed in Computer Music research, as solutions can lead to estimation of control parameters, and ultimately, provide information about a player's action during performance of the instrument. As human-computer interaction, and mapping between control and synthesis parameters, are important aspects of sound synthesis and live electronic music performance, the applications of inverse modeling abound. This work stems from [1], and builds upon previous work on modeling and system identification of reed-based wind instruments [2, 3, 4]. The ultimate aim is to extend the musical possibilities of the saxophone by estimating the player's control parameters, in real time, without hindering performance or technique. Though the work presented here does not run in real time, it provides valuable insights and offline results that could likey inform a real-time solution.

In this work, the focus is on the joint estimation of saxophone control parameters, that is, the tonehole configuration or *fingering* of the saxophone, as well as blowing pressure and stiffness (embouchure) of the reed. In [3], a pair of saxophone frequency responses, tapped at the mouthpiece and bell of a pure tenor without toneholes, are derrived from measurement. Later in [4], the measurement is extended and applied to the saxophone configured with every useable fingering, each measurement yielding a filter pair serving as the saxophone model with that applied fingering. This work was used to develop a real-time rule-based fingering identification system that operates on the spectral magnitude of the saxophone sound and estimated instrument frequency response pair comprising the saxophone model [1]. However, because the spectral magnitude of the saxophone's produced sound is substantially influenced by the player's control of the reed in addition to the fingering applied to the instrument, an alternative approach was suggested and partially explored in [1] whereby parameters of the reed model are jointly estimated with the applied fingering of the saxophone. Initial results in [1] suggest possible improved accuracy (though at the expense of computation), thus substantiating the work presented herein.

The influence of reed pulse on the resulting saxophone sound is similar to, and even more significant than, that of the glottal pulse in speech. A similar joint-estimation setup may be found in [5] where both the glottal waves and the all-pole filter coefficients of a source-filter speech model are jointly estimated using convex optimization. The coupling of the reed and saxophone differs from the source-filter model used in [5] in that vibration of the less massy saxophone reed (a source) is more effected by the internal state of traveling waves in the bore, creating more significant feedback between bore and reed than is typically seen between vocal tract and glottus. Nevertheless, in both systems, the spectrum of the produced sound is influenced by both source and filter, thus making the joint estimation of source and filter model parameters a reasonable approach. The main reason of forming and solving a convex optimization problem is that difficulties such as the aforementioned feedback dependency between source and filter signals, as well as the nonlinearity of the reed model, are made tractable and can be solved using off-shelf toolkits. Specifically, this parameter estimation problem is suited to convex optimization because: 1) the convolution operations between reed source and bore filters are linear and are thus naturally convex functions, 2) the nonlinear reed model and the signal feedback from the bore model can both be specified separately as different constraints in the optimization, and 3) since the estimation problem remains convex, so does the problem remain tractable. Tractability comes from three properties of convex optimization: 1) local optimum is global optimum; 2) feasibility of the constrained optimization problem can be determined unambiguously; 3) precise stopping criteria are available using *duality*, since the dual problem of the primal problem provides a lower bound for the primal problem [FIX: still unclear; be more specific]. [6, 7].

In [8], convex optimization techniques are used to inverse model a clarinet, with the aim of estimating reed parameters from synthesized clarinet sound. Their work has the advantage of knowing input pressure and volume flow signals at the mouthpiece when formulating the optimization problem. Here, saxophone reed control parameters, and the resulting volume flow and input pressure they generate, are optimization variables that are solved against sound recordings of an actual saxophone. The optimization is formed by minimizing the error between recorded saxophone sounds and sounds produces by the reed-saxophone model. The synthesis follows the convolutional synthesis method proposed in [9]. To evaluate the proposed methods, the precision of the estimated fingering is evaluated over a dataset consisting of fingering transfer functions (transfer functions derived from measurement of the saxophone with different tonehole configurations applied) and several example recordings of the saxophone played (by a professional saxophonist) with the corresponding fingering.

In Section 2, the details of the inverse modeling, including measurement, estimation and pre-processing of instrument responses, parametric reed model, optimization setup are provided. Experiments and results are described in Section 4 and conclusions and future works are discussed in Section 5 [FIX: once paper is finished, make sure this description is accurate].

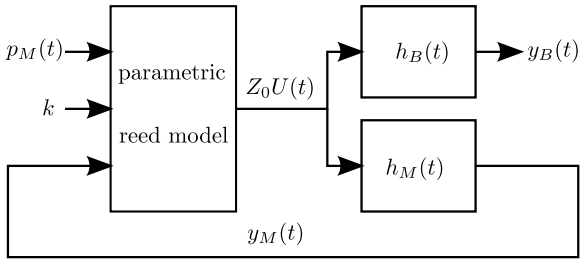

Figure 1: System diagram of convolutional synthesis

2. Saxophone Reed and Bore Models

The reed model used here is the quasi-static model describe in [10, 11, 2] among others, in which the reed vibrates in response to a pressure difference,

$$\triangle p(t) = p_M(t) - y_M(t), \tag{1}$$

across its surface. That is, introduction of blowing pressure $p_M(t)$ into the mouthpiece causes a pressure increase as compared to the valves's downstream (bore base) pressure $y_M(t)$, causing the reed to vibrate with a displacement given by

$$x(t) = \frac{\triangle p(t)}{k}, \tag{2}$$

where k is the stiffness of the reed. As the reed vibrates, it creates an aperture to the bore with time-varying cross-sectional area

$$A(t;x) = \lambda(h_0 - x(t)), \tag{3}$$

where λ is the effective jet width of the reed and h_0 is the reed rest opening. This results in a volume flow through the reed channel given by

$$U(t) = A(t;x)\sqrt{\frac{2\triangle p(t)}{\rho}} \tag{4}$$

for air density ρ, and ultimately the pressure input into the bore,

$$p_r(t) = Z_0 U(t), \tag{5}$$

where $Z_0 = \rho c/(\pi a^2)$ is the characteristic impedance of waves propagating in a bore with radius a. As shown in Figure 1, the saxophone signal produced at the bell $y_B(t)$ may be expressed in the time domain as the input pressure signal generated by the reed (5) convolved with the impulse response $h_B(t)$—the inverse of the frequency response $H_B(\omega)$ excited at the mouthpiece and tapped at the bell:

$$y_B(t) = (p_r * h_B)(t). \tag{6}$$

The signal at the mouthpiece $y_M(t)$ (base of the bore), on which the calculation of the reed model is dependent, is, in turn, given by the convolution of input pressure $p_r(t)$ with the impulse response $h_M(t)$—the inverse of the frequency response $H_M(\omega)$ excited and tapped at the mouthpiece, yielding

$$y_M(t) = (p_r * h_M)(t). \tag{7}$$

It is useful to note that the time-domain convolution given in (6) can be expressed in the frequency domain as

$$Y_B(\omega) = X_s(\omega)H_B(\omega), \tag{8}$$

where source $X_s(\omega)$ is the spectrum of the input pressure given in (5). Since applying filter

$$G(\omega) = \frac{1}{H_B(\omega)} \tag{9}$$

to $Y_B(\omega)$ yields $X_s(\omega)$, estimation of the source is often often referred to as an *inverse modeling* problem. Estimation of the fingering, and corresponding $H_{M,B}(\omega)$, is thus a prerequisite to tackling the inverse problem of source estimation.

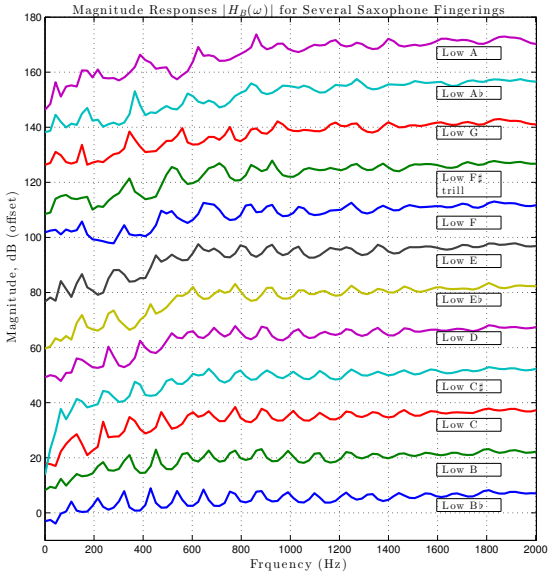

Figure 2: Magnitude responses at the bell for fingerings covering the lower register of a B-flat tenor saxophone.

3. Joint Reed-Bore Parameter Estimation

Estimation of the fingering amounts to determining which of several possible frequency response pairs $H_{M,B}(\omega)$, obtained apriori by measurement of the saxophone with all useable fingerings applied [1], is most likely to have produced the sound produced at the bell $Y_B(\omega)$. As the frequency response pairs $H_{M,B}(\omega)$ are estimated from measurement, they tend to be noisy below 100 Hz and above 5000 Hz. A denoising filter, linear in phase (with known delay), is thus applied to the inverse transformation of $H_{M,B}(\omega)$ to produce more suitable impulse responses $h_{M,B}(t)$ used in the subsequent discusstion. A subset of de-noised $H_B(\omega)$ magnitudes may be seen in Figure 2 for the lower register fingerings of the B-flat tenor saxophone.

Several recordings were made of a professional saxophonist playing a variety of notes using each tonehole configuration (including bugling/overblowing), yielding a sizeable dataset holding multiple examples of $y_B(t)$ for each possible fingering. The process of estimating reed and saxophone model parameters is done by first using convex optimization to find optimal reed parameters k and p_M for all instances of $h_{B,M}(t)$, then selecting which $h_{B,M}$ is most likely to have produced the target $y_B(t)$ (fingering estimation), and finally using corresponding values of k and p_M as the reed model estimates.

3.1. Convex Optimization

Formalizing as a convex optimization yields optimal values of k, p_m, and $U(t)$ for a given impulse response pair $h_{B,M}(t)$ corresponding to a particular fingering:

$$\begin{aligned} \underset{p_m,k,U(t)}{\text{minimize}} \quad & f_0(p_M,k,U(t)) \\ \text{subject to} \quad & (4), \\ & p_M,k \geq 0, \\ & A(t;x),U(t) \geq 0, \quad t = 1,\ldots,T, \end{aligned} \tag{10}$$

with the objective function

$$\begin{aligned} f_0(p_m,k,U(t)) &= \|y_B(t) - (p_r * h_B)(t)\|_2^2 \\ &= \|y_B(t) - (Z_0U * h_B)(t)\|_2^2, \end{aligned} \tag{11}$$

Parameter	Notation	Value
Air Density	ρ	$1.2\frac{kg}{m^3}$
Sound Velocity	c	$340\frac{m}{s}$
Bore Radius	a	$0.014m$
Characteristic Impedance	Z_0	$\frac{\rho c}{\pi a^2}$
Effective Jet Width	λ	$0.0055m$
Reed Rest Opening	h_0	$0.06m$

Table 1: Constants for parametric reed model

being the squared Euclidean distance between a recording at the bell $y_B(t)$ (having *unknown* fingering) and the signal synthesized according to (7) and (6), constrained by (4) and optimized over p_M, k and $U(t)$. The objective function here differs from the minimized negative cosine similarity used in [1] as the squared Euclidean distance avoids the need of a normalization factor to account for difference is units between model and target. Changing f_0 from negative cosine similarity to (11) improved the prediction accuracy from 60% to 93%.

For the experiments described herein, the optimization problem is evaluated over a frame size of $T = 2048$ samples, during which time p_M and k are assumed constant. The optimization problem may be translated into a conic programming problem then solved via general optimization toolkits (MOSEK [12] is used here). The other model constants used for the optimization are listed in Table 1. The detail setup of the conic programming is provided in appendix A.

3.2. Fingering Estimation

The optimization (10) for a given target $y_B(t)$ is done over all 27 possible $h_{B,M}(t)$ pairs corresponding to all possible fingerings. The synthesized signal that is most similar to target $y_B(t)$ is then selected among these 27 possibilities, and the final parameter estimation is the $h_{B_M}(t)$ pair (fingering) and the corresponding optimal values of p_M and k used to generate the selected synthesis.

Though it's reasonable to assume that the convex optimization (10) yielding satisfactory optimized parameters $\{p_M, k, U(t)\}$ would also give an optimal $h_{M,B}$ pair, it was found that the objective function f_0 was not an accurate indicator of the fingering, likely due to the energy term of $(p_r * h_B)(t)$ in (**??**) causing a selection bias toward a lower energy synthesis [FIX: still not clear... is this what you mean?] To remedy, the cosine similarity

$$g_0(p_m, k, U(t)) = \frac{{y_B}^T (p_r * h_B)(t)}{\|y_B(t)\| \|(p_r * h_B)(t)\|} \quad t = 1, \ldots, T \qquad (12)$$

which provides a similarity measure based more on the shape of the waveform, is used to estimate the fingering. It should be noted that, after f_0 is expanded and the two squared terms $y_B(t)^2$ and $(p_r * h_B)(t)^2$, corresponding to the energy of $y_B(t)$ and the synthesized signal, respectively, are discarded, the remaining $-{y_B}^T (p_r * h_B)(t)$ is equal to the negative numerator of g_0, which is minimized (maximized as g_0).

4. Results

Figure 3 illustrates a frame size of a target recording $y_B(t)$ (black), a middle F note played with low B-flat fingering (thus overblown the fifth), along with three synthesis attempts (blue, green and red) using (7) and (6) with three sets of optimized variables $\{p_M, k, U(t)\}$ and impulse responses $h_B(t)$ and $h_M(t)$ corresponding to 1) the *correct* low B-flat fingering (blue), 2) the *feasible but incorrect* low F fingering (green), and 3) the *infeasible and incorrect* low B fingering. When the impulse responses corresponding to the correct fingering are used, the optimization yields parameter values for $\{p_M, k, U(t)\}$ that produce a synthesized signal having reasonable likeness to target $y_B(t)$ (see Figure 3, blue). When impulse responses corresponding to *incorrect* fingerings are used however, the optimization produces parameters values yielding synthesis more dissimilar from $y_B(t)$ than the *correct* one, and *infeasible* one (red) is farther than *feasible* one (green).

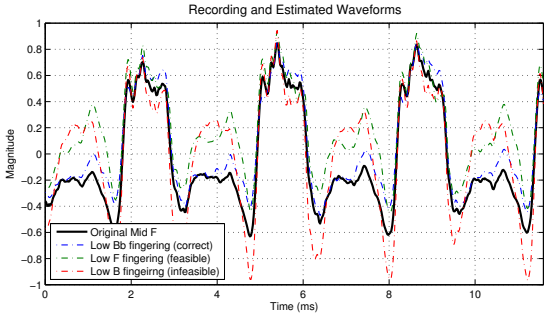

Figure 3: Synthesis of (6) using optimized variables p_M, k and $U(t)$ and impulse responses corresponding to *correct* fingering low B-flat (top) and *incorrect* fingering C-sharp (bottom), for target sound middle B-flat with low B-flat fingering.

used \ **tested**	Low B♭	Low F
Low B♭	**0.6185**	0.334
Low F	0.5398	**0.544**
used \ **tested**	Low D	Low A
Low D	**0.7958**	0.75
Low A	0.77	**0.78**

Table 2: Fingering **used** in target versus fingerings **tested** in optimization for $y_B(t)$ having pitch middle F (top) and middle A (bottom).

As a proof of concept, the confusion matrix for two examples of binary classification are shown in Tab. 2, one for target sound having pitch middle F (top) and one having pitch middle A (bottom). For each pitch, two target recordings $y_B(t)$ are considered, each produced using one of two possible fingerings. The results of the cosine similarity between a $y_B(t)$ and synthesis using impulse responses corresponding to two possible fingerings (g_0) is shown, with the higher value, in bold, being the estimated fingering.

4.1. Saxophone Fingering Identification

To further investigate the saxophone fingering identification problem, an experiment is conducted on a larger subset of the recorded saxophone notes covering pitches producible from low B♭, B, C, C♯ and D fingering, with results shown in Tab. 3. For each fingering there are three recordings each having a different pitch (15 recordings total). In this experiment, the correct fingering was estimated from five possible candidates 93% of the time, substantially above a baseline of 20% for a random guessing. An experiment on the full dataset, 57 recordings against all 27 possible fingerings, is conducted as well later with an accuracy rate of 19.3%, which is better than a random baseline of approximately 4%, but definitely not ready for practical use [FIX: can we omit this last part? It sounds bad in that it devalues the work...].

5. Conclusion and Future Work

In this work, convex optimization techniques are deployed to identify the fingering used for producing a recorded saxophone sound. The optimization is formed based on a coupled saxophone model consists of a parametric reed model and estimated impulse responses of the saxophone. The accuracy of the identification is assessed, and the performance (93%) on the subset of the dataset in comparison to that of the full dataset (19.3%) indicates that the proposed formalization indeed has the potential to be further developed and investigated to have practical uses. Also the accuracy

Pitch, fingering \ Test against	low B♭	low B	low C	low C♯	low D
mid B♭, low B♭	**0.8**	0.7	0.66	0.67	0.61
mid F, low B♭	**0.74**	0.47	0.61	0.66	0.65
high B♭, low B♭	*0.62*	0.57	0.60	0.71	**0.75**
mid B, low B	0.77	**0.84**	0.75	0.61	0.69
mid F♯, low B	0.77	**0.86**	0.60	0.67	0.67
high B, low B	0.64	**0.85**	0.73	0.7	0.66
mid C, low C	0.65	0.77	**0.82**	0.78	0.63
mid G, low C	0.7	0.59	**0.88**	0.63	0.68
high C, low C	0.67	0.72	**0.83**	0.72	0.69
mid C♯, low C♯	0.67	0.61	0.70	**0.76**	0.74
mid G♯, low C♯	0.70	0.75	0.62	**0.84**	0.71
high C♯, low C♯	0.62	0.63	0.65	**0.72**	0.69
mid D, low D	0.67	0.69	0.66	0.70	**0.76**
mid A, low D	0.6	0.65	0.68	0.75	**0.8**
high D, low D	0.66	0.7	0.7	0.75	**0.77**

Table 3: Prediction of fingerings using cosine similarity from between target pitch with corresponding fingering and synthesis (Eqn. (6)) using impulse responses for low B♭, B, C, C♯ and D fingering. Gray boxes indicate correct fingerings (fingering used in target). Bold fonts indicate identified fingerings based on maximum cosine similarity. Slanted fonts are used if the runner-up in identification is the correct fingering.

on the subset of the dataset (93%) is significantly improved from previous reported result (60%) [1] by using squared euclidean distance as objective function.

For future works, the first step is to further boiled down the optimization formalization into small building blocks to have better understanding on how parameters interact with each and to improve computational efficiency towards real-time application. Also, more saxophone recordings for each fingering should be gathered to have a better validation on how the formalization could generalize. The other interesting part that was not explored in this work is the investigation of player's mouthpiece control, mainly input mouthpiece pressure and reed stiffness. The challenge facing such investigation is that when comparing $y_B(t)$ to the synthesis by Eqn. (6), there is a scaling problem since $y_B(t)$ is represented as real numbers between $-1 \sim 1$ while the synthesis outputs waveforms in units of pressure (Pascals). Due to such scaling disparity, the value of the estimated $\{p_M, k, U(t)\}$ does not reflect real world measurements. To tackle such problem, the studies done in [11] about the characteristics of single-reed instrument could be leveraged to choose an appropriate scaling factor into the optimization setup. Optimization in the frequency domain should also be considered since fitting the waveform shapes in time domain is an overkill when what matters is the spectral characteristics.

A. Conic Programming Setup

To transform the optimization problem in Eqn. 10 into convex optimization form, slack variables and conic programming techniques are introduced as described here. An n-dimensional rotated quadratic cone is convex and defined as

$$Q_r^n = \{x \in \mathbb{R}^n | 2x_1x_2 \geq x_3^2 + \cdots + x_n^2, x_1, x_2 \geq 0\}. \quad (13)$$

First the objective function f_0 is rewritten as a rotated quadratic cone and the optimization given by (10) is rewritten accordingly with other inequalities staying the same (omitted for brevity)[FIX] and a slack variable γ introduced as

$$\begin{aligned} &\text{minimize} && \gamma \\ &\text{subject to} && (\frac{1}{2}, \gamma, \triangle y) \in Q_r^{2+T}, \quad (14) \\ & && \triangle y = y_B(t) - (p_r * h_B)(t). \quad (15) \end{aligned}$$

Equation (4) also has to be rewritten as two rotated quadratic cones. Let $\triangle p = p_M(t) - y_M(t)$, $k' = k^2$ and introduce slack variables η, α, β and ζ, the equality constraint (4) is transformed to

$$U(t) - \lambda\sqrt{\frac{2}{\rho}}(h_0\alpha - \beta) - \eta = 0, \quad (16)$$

$$(\triangle p, \frac{1}{2}, \alpha) \in Q_r^{2+T}, \quad (17)$$

$$(\zeta, \beta, \triangle p)(t), (\frac{1}{8} \triangle p, k', \zeta)(t) \in Q_r^3,\ t = 1, \ldots, T. \quad (18)$$

Finally the full convex optimization is written as

$$\begin{aligned} &\text{minimize} && \gamma + \eta \\ &\text{subject to} && (14), (15), (16), (17), (18), \\ & && p_M, k', \gamma, \eta \geq 0, \\ & && A(t;x), U(t) \geq 0,\ t = 1, \ldots, T \end{aligned}$$

with k recovered by calculating square root of the optimized k'.

References

[1] T. Smyth and C.-i. Wang: *Toward Real-Time Estimation of Tonehole Configuration*. In *40th International Computer Music Conference joint with the 11th Sound and Music Computing conference*. 2014.

[2] T. Smyth and J. S. Abel: *Toward an estimation of the clarinet reed pulse from instrument performance*. In *The Journal of the Acoustical Society of America*, volume 131(6):4799–4810, 2012.

[3] T. Smyth and S. Cherla: *The Saxophone by Model and Measurement*. In *Proceedings of the 9th Sound and Music Computing Conference, Copenhagen, Denmark*, page 6. 2012.

[4] T. Smyth and M. Rouhipour: *Saxophone modelling and system identification*. In *Proceedings of Meetings on Acoustics*, volume 19, page 035010. Acoustical Society of America, 2013.

[5] H.-L. Lu and J. O. Smith: *Joint estimation of vocal tract filter and glottal source waveform via convex optimization*. In *Applications of Signal Processing to Audio and Acoustics, 1999 IEEE Workshop on*, pages 79–82. IEEE, 1999.

[6] J.-B. Hiriart-Urruty and C. Lemaréchal: *Fundamentals of convex analysis*. Springer, 2001.

[7] S. Boyd and L. Vandenberghe: *Convex optimization*. Cambridge university press, 2009.

[8] V. Chatziioannou and M. van Walstijn: *Estimation of clarinet reed parameters by inverse modelling*. In *Acta Acustica united with Acustica*, volume 98(4):629–639, 2012.

[9] T. Smyth and J. S. Abel: *Convolutional synthesis of wind instruments*. In *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, pages 219–222. IEEE, 2007.

[10] N. H. Fletcher and T. D. Rossing: *The physics of musical instruments*. Springer, 1998.

[11] J.-P. Dalmont, J. Gilbert, and S. Ollivier: *Nonlinear characteristics of single-reed instruments: Quasistatic volume flow and reed opening measurements*. In *The Journal of the Acoustical Society of America*, volume 114(4):2253–2262, 2003.

[12] *The MOSEK optimization software*.