Accuracy assessment of thermoacoustic instability models using binary classication

Accuracy assessment of thermoacoustic instability models using binary classification

Maarten Hoeijmakers*[1], Ines Lopez Arteaga[1,2], Viktor Kornilov[1], Henk Nijmeijer[1], and Philip de Goey[1]

[1]Department of Mechanical Engineering, Eindhoven University of Technology, Netherlands
[2]KTH Royal Institute of Technology, Department of Aeronautical and Vehicle Engineering, Marcus Wallenberg Laboratory, Sweden

[Recieved 09/11/2012; Accepted 16/03/2013]

Abstract

We apply binary classification theory to assess the (in)stability prediction accuracy of thermoacoustic models. It is shown that by applying such methods to compare a large set of stability predictions and experiments it is possible to gain valuable qualitative insight in different aspects of prediction quality. The approach is illustrated with a 2-port model and a large experimental data set. The presented framework provides an unified and practical tool to answer questions such as (i) What is the chance that a stable prediction will be correct? and (ii) How conservative is the model? It is shown that the most suitable quality indicator is strongly dependent on the actual purpose of the model. The method provides a solid starting point for model comparison and optimization.

1 Introduction

In virtually any combustion device a feedback coupling between the system's acoustics and heat release can lead to a rich variety of resonance phenomena named thermoacoustic instabilities. In particular, the issue is frequently encountered during the development of low emission premixed combustion devices. Typical examples include high performance gas turbines, industrial ovens, and domestic heating systems [1–3]. Within the broad field of thermoacoustic instabilities there are many variations of the exact feedback mechanism and the type of noise produced. The heat release may be excited by velocity or mixture fluctuations for instance, whereas the noise can vary between anything from low-frequent humming or chugging, to high frequent pure tones and beating [4, 5]. Such instabilities generally limit the operation of the device because of unacceptable sound pressure levels or unstable flame anchoring. In extreme cases and in high power devices such as gas-turbines, this can even lead to structural failure [5]. Accordingly, the development of accurate thermoacoustic models is an actual and practically relevant problem.

Because of the coupled nature of the phenomena a thermoacoustic model always contains a description of the systems acoustics, and the coupling of the acoustics to the unsteady heat release. Within such a context many different variants are possible, ranging from lumped linear models using one dimensional acoustic approximations [6,7], to descriptions which are able to predict transient growth and complicated nonlinear phenomena [8, 9]. Irrespective of the specific model type however, the end goal is always to arrive at an accurate prediction of the observed phenomena. The most basic property which any thermoacoustic model should predict is the (in)stability of a certain system configuration. In case it will be unstable, the frequency of the instability, the limit cycle amplitude and the occurrence of possible mode switching become of interest. The next important question is then how to assess prediction quality. This issue forms the main subject of the current paper.

*Corresponding author: p.g.m.hoeijmakers@tue.nl

Currently, the assessment of model quality is often approached in an 'ad-hoc' sense, as the visual discrepancy of the model predictions versus the actual instabilities when some system parameter is varied. In [10, 11] for example, (in)stability experiments on a laboratory combustor were compared to a linear, and later a nonlinear model [9] by varying the length of the upstream duct. In all mentioned literature however [9–11], the model quality was inferred from comparing the visual discrepancy of the predictions versus the experimental points. Although such an approach gives a reasonable qualitative indication it remains difficult to answer important quantitative questions such as: 'What is the likelihood a stable prediction will be correct? and 'How conservative is my model'?

Furthermore, most of the aforementioned literature sources only consider a relatively small set of burner operating points, whereas in practice a combustor is often required to function over a wide regime of flow velocities and equivalence ratios. There is no guarantee that a model which is accurate in one operating window performs equally well in another. A similar statement can be made about the mutual differences between various burners.

To investigate these issues, large datasets are inevitable. The larger the dataset however, the more difficult it becomes to investigate and uncover trends using just visual inspection of the 'bare' results only. In fact, it is exactly this practical reason which motivated our search for usable quantitative performance indicators in the thermoacoustic context.

In essence a thermoacoustic model makes predictions of a binary nature, either stability or instability. The quantification of the performance of such a binary predictor is well developed problem in binary classification theory. Numerous examples can be found in the areas of material testing, medicine and weather forecasting [12]. In the current paper we apply binary classification theory to define quantitative measures of the (in)stability prediction quality of thermoacoustic models. Not only does such an approach provide us with a convenient quantitative quality assessment, it also gives valuable insight on different aspects of the prediction performance.

The remainder of the paper is organized as follows. In the next section the theory and methods used to arrive at a set of quantitative performance measures are described. The experimental setup and specifics of the model are treated in section 3. Then, in section 4 the theory is applied to a complete set of experiments and model simulations. Finally, section 5 is devoted to a discussion and summary of the results.

2 Model quality assessment

The goal of the current section is to present the methods and theory which is needed to perform a quantitative model quality assessment in the context of thermoacoustic stability prediction.

Before we start with the development of our approach it is worthwhile to linger on the exact meaning of 'model' and hence the interpretation of 'model quality'. In the wider context of the exact sciences, a model is usually understood to be a mathematical description of the physics occurring in the real world. To be more precise, one can distinguish between the actual mathematics and physics chosen to obtain the description, e.g. the governing equations and their simplifications, and the (physical) parameters, e.g. geometrical dimensions, which occur in the description. However, it is the combination of both which in the end constitutes the full model. To assess the model's quality, it is then common to vary one or more parameters consecutively in model and experiment and compare their results in relevant output parameters. In this context, the model quality is the ability of the model to capture the real world trends. It is clear that any discrepancy of model and experiment can be the result of two sources (i) wrongly chosen or measured parameter values, and (ii) errors in the physical description itself, e.g. oversimplification. It is important to realize that these sources cannot (without further investigation) be differentiated in the output of the model evaluation.

Any assessment of prediction capabilities of a thermoacoustic model starts with a set of model predictions and corresponding experiments. As mentioned before, the data usually takes the form of predictions and measurements of stability/instability when one or more of the system parameters is varied. Within this context, many variants of the dataset are possible, and different types of parameters and/or additional data may be gathered. Irrespective of the details however, the theory presented here is very general and the only prerequisite is that at least data on the presence and frequency of the instability is available.

In order to keep the theory as clear as possible, it is necessary to more precisely define the structure

of the data. First of all, let us restrict the dataset to compare the instability of only one frequency (or mode) per parameter setting, as is regularly done in literature, e.g. [10]. This means that in case multiple frequency components, or modes, are present in either the simulated or experimental datasets, one should select only one pair for the comparison and further quality assessment. Usually, this is achieved by only comparing the dominant frequency from the experimental dataset, to the most unstable mode, i.e. with the largest growth rate, from the simulations. This is also the method chosen in the current paper. Although it was recently shown that there exists a strong correlation between the (linear) growth rate and the limit cycle amplitude, it is not apparent that this is an universal law [13]. Clearly, any deviation from the assumed correlation will be detrimental to the apparent quality of the model, and this should be kept in mind when making statements about the model quality.

Now, we define the model simulation dataset as a collection of N data points $\mathcal{S} = \{(s_1, f_1^s), ...(s_N, f_N^s)\}$, where one element i out of the set corresponds to the (in)stability data at a specific value of a parameter, e.g. a duct length L. For any point (s_i, f_i^s), s_i is a boolean were 1 denotes instability, and 0 denotes stability. In case of instability, f_i^s corresponds to the predicted frequency of the instability. Likewise, the experimental dataset is given by N points, $\mathcal{E} = \{(e_1, f_1^e), ...(e_N, f_N^e)\}$.

2.1 Qualitative aspects of thermoacoustic models

Without loss of generality, consider the situation were the capabilities of a model are verified using a simple laboratory combustion geometry with (i) an inlet tube of length L_1, (2) burner/flame, and (3) exhaust tube of length L_2, as found in for example [14]. Then, it is typical to vary the upstream length L_1 in some way, e.g. by using a piston, and register the lengths at which it is unstable. A direct comparison between model and experiments can be made by plotting the length L_1 versus the unstable frequencies f_i^e and f_i^s. Two examples taken from the dataset described in section 3 are shown in figure 1a and b. As noted before, only the dominant frequency from the experimental data, and the most unstable mode from the simulations are shown.

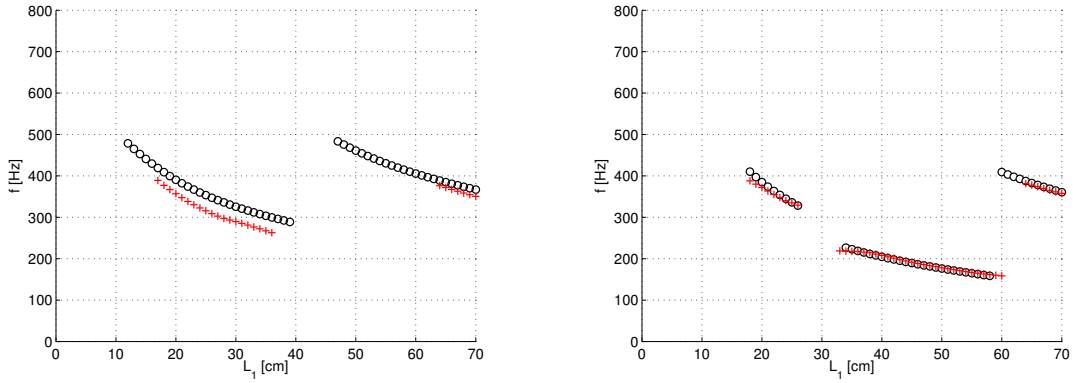

Figure 1: Comparison between experimental (+) and simulated instabilities (∘) as function of length L_1 of the upstream tube. (a) case 1 and (b) case 2.

A number of qualitative features about the ability of the model to capture the experimental results become immediately apparent from the figures shown. First and foremost is the correctness of the length ranges of the instability predictions. In case of perfect prediction, the transition points from stable to unstable and vice versa, are on exactly the same parameter values in both experiment and simulation. It is clear that the example cases show significant differences in this aspect of the (in)stability prediction.

Another feature is the correctness of the predicted frequencies, where it is obvious that large deviations are not desirable.

It is usually implicitly assumed that overprediction of the unstable range, as in the example shown, is not preferable. Although this seems a natural approach from a scientific modelling viewpoint, in practice it in fact might be desirable to have a model which is a bit on the conservative side when it comes to

predicting stability. As shall become apparent the ability to capture all unstable modes and a slight overprediction are intimately linked features of all except perfect models.

It is clear that in order to make quantitative statements about model conservatism, and the reliability for stable and unstable predictions, one has to approach the quality assessment from a more quantitative viewpoint.

The typical examples from figure 1 could be used as a first starting point to quantitatively define model quality. For example, one could define model quality as the ratio of the number of correctly predicted unstable points to the total number of unstable experiments. In other words, how many of the unstable observations were predicted as such? Alternatively, it is possible that one is more interested in the relative amount of correct unstable predictions to total unstable predictions. The difference in the questions is subtle, but important, whereas in the first number the wrong unstable predictions do not influence the score, in the second number they do. In the next sections it is shown how to structurally approach the problem using well developed methods from other branches of science.

2.2 Quantitative aspects of thermoacoustic models

In an abstract sense, the model poses a hypothesis, e.g. stable or unstable, which can be tested against the experimental results. The goal is then to rate the ability to make a correct hypothesis. Very similar problems occur in many other branches of science [12]. Usually however, the hypothesis is not the outcome of a model prediction but rather of a diagnosis or test. Typical examples include testing for the presence of disease in medicine, detection of material defects, weather forecasting, prediction of consumer risk in finance, or the prediction of protein structures in bioinformatics. Irrespective of the application however, a good accuracy measure is of crucial importance in order to assess the value of the diagnosis, test, or prediction. As a result, the methods to quantitatively assess the accuracy of a prediction, diagnosis, or test are well developed.

Similarly to the prediction of stability or instability in a thermoacoustics problem, many of these problems reduce to assigning a certain prediction instance to one out of two classes, commonly termed a binary classification problem. Notwithstanding the binary result, it is important to realize that the model, diagnosis or tests often have underlying data which is continuous. In such setting, the binary outcome is then based on whether or not a certain threshold value is exceeded. For example, the indicator for a disease test might be the concentration of certain hormone, and if the threshold value is exceeded the patient is considered sick. Again, this is very similar to the prediction of thermoacoustic instabilities, in which the underlying continuous variable is the growth rate of the mode, with zero being the threshold value. Nevertheless, the performance of the prediction, diagnosis, or test is usually assessed based on the ability to make a correct classification.

In the following, the standard terminology from binary classification shall be used on the thermoacoustic instability problem.

2.2.1 Binary classification and the contingency matrix

The model predictions and the experimental results can be categorized in two possible classes (1) positive, corresponding to unstable events and (2) negative, corresponding to stable events. In this setting, it is clear that the comparison between simulations and experiments is completely characterized by four possible outcomes: (i) a true positive - e.g. an unstable prediction is indeed unstable, (ii) a false positive - e.g. an unstable prediction is in fact stable, (iii) a false negative - e.g. a stable prediction is observed unstable, and (iv) a true negative, when a stable prediction is indeed stable. This is further summarized in table 1, using the abbreviations TP, FP, TN, and FN. Due to the binary nature of the problem, this is called binary classification, and the corresponding matrix is named a confusion or contingency matrix [15, 16].

Every single point resulting from comparing the simulation and experimental set can now be assigned (classified) to one of the four possible cells. The goal is then to count the number of occurrences in each cell. Note that the row and column totals S_u, S_s denote the total number of stable and unstable simulations, where E_u and E_s are the totals for the experiments. However, because the data contains information about both stability *and* frequency, one has to be careful in the exact definitions of the various categories. In case both experiment and simulations are unstable ($e_i = s_i = 1$), the difference in predicted and observed frequency becomes of interest. Here, we only consider the instability prediction

	Exp. Unstable	Exp. Stable	Total
Sim. unstable	TP	FP	S_u
Sim. stable	FN	TN	S_s
Total	E_u	E_s	

Table 1: The contingency table. TP: True positive. FP: False positive. FN: False negative. TN: True negative.

as correct (TP) if $|f_i^e - f_i^s|/f_i^e < 0.2$. Thus in essence a prediction is only considered satisfactory if it lies within a 20% margin of the observed frequency. In practice, any predicted frequency which has an error larger than 20% results from the model considering the wrong mode number as most unstable. In that case, we will consider it as two separate points, where one is classified as a false positive and the other as a false negative. In this way, it only decreases model performance, but does not bias the error measures introduced in the next section. The other possible situations are more straightforward and are shown in table 2.

Condition	Category
$e_i = s_i = 1, \|f_i^e - f_i^s\|/f_i^e < 0.2$	TP
$e_i = s_i = 1, \|f_i^e - f_i^s\|/f_i^e > 0.2$	FN,FP
$e_i = s_i = 0$	TN
$e_i = 0, s_i = 1$	FP
$e_i = 1, s_i = 0$	FN

Table 2: Classification rules.

Given a set of experiments and simulations, a total overview of the comparison is obtained by counting the number of instances in each of the four possible matrix entries. Table 3 summarizes the results for the two example cases of figure 1. Note that the contingency matrix basically shows if the row variables (the simulations) are *contingent* on the column variables (the experiments). As is clear from the example tables, a good case tends to have the largest numbers on the diagonal terms of the matrix.

(a)

	E_u	E_s
S_u	27	25
S_s	0	11

(b)

	E_u	E_s
S_u	41	3
S_s	3	16

Table 3: The contingency matrix for (a) Case 1, and (b) Case 2, corresponding to figure 1a-b.

Although the contingency matrix gives a complete overview of the comparison between simulations and experiments, it is not immediately clear how good the predictions are.

In the following, we will present several different performance indicators and investigate their meaning and usefulness in the current context.

2.2.2 Performance metrics

Starting from the contingency matrix, one may derive several common performance metrics.

Let us begin with considering the portion of experimental unstable observations which were predicted as such. This measure is usually called the true positive rate, sensitivity, hit rate, or recall:

$$TPR = 100\frac{TP}{TP + FN}. \tag{1}$$

A similar number can be defined for the percentage of stable experiments which were predicted correctly,

the true negative rate, specificity, or false alarm rate:

$$TNR = 100\frac{TN}{TN + FP}. \tag{2}$$

Both indicators have in common that the maximal score is achieved if all the stable, or unstable observations are predicted correctly, irrespective of how many false predictions from the opposite class are made. For example, in the extreme limit one can simply achieve a high TPR by predicting all instances as unstable. Note however, that this would in turn lead to a very low TNR as the number of FP would increase dramatically. As shall become clear in the next section, it is therefore the combination of TPR and TNR which can reveal interesting information about the model quality.

Instead of basing the metrics on correctness one may also characterize the performance in terms of the relative error. This leads to the false positive and false negative rates (FPR, FNR):

$$FPR = 100\frac{FP}{TN + FP} = 100 - TNR \quad \text{and} \quad FNR = 100\frac{FN}{TN + FP} = 100 - TPR. \tag{3}$$

Naturally, the error measures are simply complementary to the measures based on correctness. Given the TPR and TNR from a data set, it becomes straightforward to answer questions regarding the likelihood that an unstable or stable measured point would have been predicted as such. In this setting, one has to ascertain that the size of the dataset is sufficiently large to obtain statistically relevant results.

Alternatively, one can wonder about the likelihood that an unstable or stable prediction will indeed be correct. This question might be much more appropriate when assessing the actual usefulness of model predictions. In this context, one may define the positive and negative predictive values:

$$PPV = 100\frac{TP}{TP + FP} \quad \text{and} \quad NPV = 100\frac{TN}{FN + TN}. \tag{4}$$

Clearly, the PPV and NPV simply give the ratio of correct predictions from each class to total number of predictions of the given class. Note that the difference between the true and negative rates on one hand, and the positive and negative predictive values on the other, consists of using the column totals (experiments) or row totals (simulations) of the contingency matrix as reference for a score. It should be clear that a combination of TPR and TNR gives essentially the same information as combining the PPV with the NPV. The reason for this is simply that all four matrix entries appear in both combinations.

The foregoing indicators are useful to quantify the prediction performance of a single class only, independent on the other class. Nevertheless, the total performance irrespective of the specific class can also be of interest. This is often termed the accuracy, and defined as the ratio of total correct predictions, to total predictions and observations as

$$ACC = 100\frac{TP + TN}{TP + FP + FN + TN}. \tag{5}$$

It is clear that a high accuracy can only be achieved when a model does not over- or underpredict the (in)stability ranges. A short overview of the various measures is presented in table 4.

Metric	Description
TPR	Likelihood that unstable observations will be predicted correctly.
TNR	Likelihood that stable observations will be predicted correctly.
FPR	Likelihood that stable observations will be incorrectly predicted.
FNR	Likelihood that unstable observations will be incorrectly predicted.
PPV	Likelihood that unstable predictions will be correct.
NPV	Likelihood that stable predictions will be correct.
ACC	Overall accuracy of model irrespective of the specific class of the event.

Table 4: Overview of performance metrics and their meaning.

Notwithstanding the intuitiveness of the presented indicators one has to be aware of implicit subtleties. The most important one relates to the fact that the measures are based on *relative* fractions. Because

of this, their sensitivity to variations depend on the bias between the positive (stable) and negative (unstable) instances (either predictions or simulations) in the data set. For example, the less observed unstable points, the more sensitive the TPR is to wrong stable predictions (FN). At the same time the sensitivity of the NPV for FN type errors decreases due to the additional FN's. One subtle consequence is that the NPV can be high when the dataset has a high bias towards stability, even though the model strongly overpredicts the stable ranges.

Let us now illustrate the various indicators for the two example problems. Figure 2 gives an overview of the metrics for each case. Because for both examples virtually all unstable points were predicted unstable the TPR scores are high. The TNR score however is much higher in the second case. This is a result of the trend to overpredict the instability ranges in the first case, which leads to many false positives. As shown, such overprediction also affects the PPV in a negative and the NPV in a positive way. Finally, the accuracy scores then confirm the intuition that the second case is the most accurate. It should be clear that the outlined method provides only a convenient tool to analyse the model quality in a systematic way, but will not provide the user with answers about the exact reason why a model performs good or bad.

In summary, the proposed quality indicators can be very useful in quantifying the model performance in a precise way. In particular, the TPR, TNR, FPR, FNR, and ACC highlight different aspects of model quality and provide convenient tools to assess the performance of a thermoacoustic model.

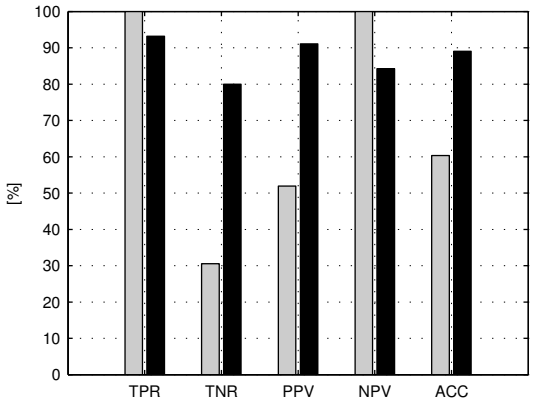

Figure 2: The performance of the example cases, Grey: case 1. Black: case 2.

3 Experimental setup and model

In this section an overview is given of the experimental setup, the measurement procedure, and the modeling approach.

The thermoacoustic configuration of interest is shown in figure 3a. The setup consists of a variable length upstream duct L_1 of 50 mm diameter, a burner deck, and a downstream duct $L_2 = 0.17$ m with a diameter of 90 mm and open end.

The (in)stability of the setup will in general depend on the burner, flow velocity, equivalence ratio, area change, and the up- and downstream duct length. In our case, the (in)stability of the setup is characterized for four different burners, three different burner flow velocities (velocity through the burner area), and three different equivalence ratio's. In total, this results in $4 \times 3 \times 3 = 36$ different cases. The details are given in table 5.

All burners used are of perforated plate type, with a thickness of 1 mm, and a hexagonal pattern of holes in the central 50 mm diameter of the burner. The complete burner is then characterized by the diameter of the holes D and the pitch between them P in mm. The total open areas A_b of the burners are 267.04 (D2-P5), 342.43 (D2-P4.5), 431.18 (D3-P6), and 512.08 mm^2 (D2-P3.5). Further information about the burners and transfer functions can be found in [17].

Additional data from the experiments is obtained by various sensors located in the setup. Sound recordings of the instabilities are made possible by a pressure transducer connected to the data acquisition

Table 5: Overview of case parameters.

Parameter	Range	Unit
Burner	D2-P5, D2-P4.5, D3-P6, D2-P3.5	D=Hole diameter, P=Pitch [mm]
Equivalence ratio	0.75, 0.85, 0.95	
Velocity	80, 100, 120	[cm/s]

system. Furthermore, the burner deck temperature is measured by a N-type thermocouple while the downstream temperature was measured by a R-type couple.

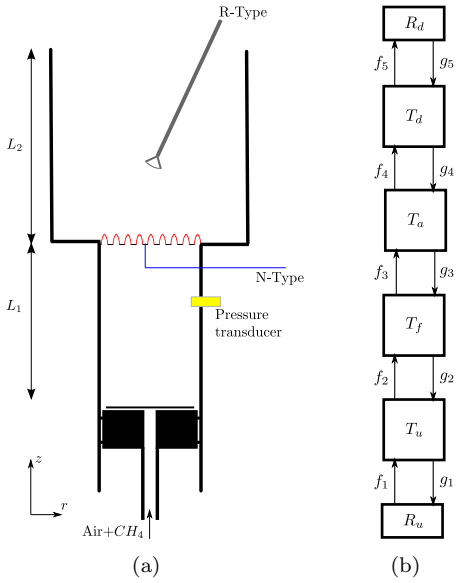

Figure 3: (a) Experimental setup. (b) Acoustic network model.

3.1 Instability experiments

The basic instability measurement for each case consists of changing the tube length (L_1) between [8 70] cm by moving the piston with 63 steps of 1 cm. This measurement is made both in the upward, and downward movement direction. In case an instability occurs, it is recorded for a 10 s period for subsequent post processing. Following the actual experiments, the data is post processed in two steps.

First, the common instability peaks are selected from the up- and down measurements. Since hysteresis behaviour due to the nonlinearity of the flame can play an important role, the up- and down-stream unstable points are not necessarily the same. In some cases a discrepancy of more than 10 points was observed. This highlights an important question: which points to select as input for the model comparison. Since the aim of the current paper is to define and illustrate a number of prediction quality measures in a linear framework we only consider the common unstable points from the up- and down sets for the experimental-model comparison. Note however, that the described methods may be very well applied to do a more extensive comparison of the up- and down- sets using linear and especially non-linear models.

The second post-processing step consists of extracting the strongest frequency component from each unstable recording. Finally, the results are stored as described earlier. Note that the total experimental data set consists out $4 \times 3 \times 3 \times 63 = 2268$ points.

3.2 Model

Various modelling approaches can be used to predict thermoacoustic instabilities. In general, one can distinguish between lumped and non-lumped approaches and linear and nonlinear variants. In the current

paper, we use a lumped linear model. To be more precise, when a system has relatively simple geometry and the system's streamwise dimensions are much larger than the transversal ones it can be assumed that only one-dimensional wave propagation is present. In that case, so called acoustical 2-port models [18] may be used to describe the network of interconnections of the various acoustic parts such as ducts, flames and area changes.

For each of the elements the output variables, namely pressure p' and velocity u' are linked to the input variables by linear equations. For signals with a harmonic time dependence in the form $e^{i\omega t}$ with (complex) frequency $\omega = \omega_r + i\omega_i$ these relations can be represented in a transfer matrix (TM) form

$$\begin{bmatrix} f' \\ g' \end{bmatrix}_{out} = \begin{bmatrix} T_{11}(\omega) & T_{12}(\omega) \\ T_{21}(\omega) & T_{22}(\omega) \end{bmatrix} \begin{bmatrix} f' \\ g' \end{bmatrix}_{in}, \tag{6}$$

where the up- and down-stream propagating waves $f' = \frac{1}{2}(\frac{p'}{\rho c} + u')$ and $g' = \frac{1}{2}(\frac{p'}{\rho c} - u')$ are a linear combination of p' and u' . Note that f', g', p' and u' are in general complex amplitudes of the acoustic waves. Here ρ and c are the mean gas density and speed of sound, respectively.

An overview of the network model of the setup is given in figure 3b. For many acoustic elements the analytical expressions for the acoustic transfer matrices are known [18]. For elements like the flame/burner combination however, the transfer matrix should be measured by experiments. In addition, in order to model the downstream wave propagation correctly it is desirable to have a good estimate of the corresponding temperature distribution. These two items are treated separately in the next subsections.

3.2.1 Flame element.

Although it is possible to directly measure the TM of the flame burner combination it is a technically complicated and challenging task [19]. An alternative method to obtain the TM is by using the linearized Rankine-Hugionot equations in combination with a measured thermoacoustic transfer function $\mathcal{F}$. This $\mathcal{F}$ characterizes the flame heat release rate response q' in W to acoustic velocity excitation u' in m/s

$$\mathcal{F}(\omega) = \frac{q(\omega)'}{u(\omega)'} \frac{\overline{u}}{\overline{q}}, \tag{7}$$

where $\overline{q}$ and $\overline{u}$ denote the mean of the respective variables.

The measurement of the flame frequency response $\mathcal{F}(\omega_r)$ is a relatively straightforward procedure, and shall not be treated here. For further information we refer to [20]. Note that the transfer functions were measured for each individual burner and flow setting separately, resulting in 36 measured cases. Following the frequency response measurement a rational transfer function is fitted to the measurement data and incorporated in the TM linking the cold (2) and hot (3) sides

$$\begin{bmatrix} f \\ g \end{bmatrix}_3 = \frac{1}{2} \begin{bmatrix} \epsilon + 1 + \theta\mathcal{F} & \epsilon - 1 - \theta\mathcal{F} \\ \epsilon - 1 - \theta\mathcal{F} & \epsilon + 1 + \theta\mathcal{F} \end{bmatrix} \begin{bmatrix} f \\ g \end{bmatrix}_2, \tag{8}$$

where $\theta = \frac{T_3}{T_2} - 1$ the temperature ratio, and $\epsilon = \frac{\rho_2 c_2}{\rho_3 c_3}$, is the jump in specific acoustic impedance across the flame.

3.2.2 Downstream temperature

The downstream temperature distribution due to the cooling of the flue gas in the outflow tube can play an important role in the total thermoacoustic feedback loop because it directly influences the speed of sound and the effective length of the tube. In order to be able to provide the model with the best input data possible, the downstream temperatures are measured using a R-type thermocouple at three radial positions: $r = [0, 2.5, 4]$ cm, and six axial positions $z = [2, 3, 6, 9, 12, 15]$ cm, with z the distance above the burner deck. Furthermore, to exclude effects from short time flow fluctuations, every temperature is obtained from averaging the thermocouple readings in a 10 s window ($F_s = 1000$ Hz). Finally, the approximate axial temperature profile was constructed from the radial averages at each axial position. Because the flow velocity, particular burnerdeck pattern, and equivalence ratio have a direct effect on the downstream temperature the whole procedure was done for every case individually.

In the model the downstream temperature profile is included by discretization of the duct in $n = 10$ isothermal ducts and corresponding temperature jumps, with section i defined as

$$T_d^k = \frac{1}{2}\begin{bmatrix}(\epsilon_k + 1)e^{-i\omega L_k/c_k} & \epsilon_k - 1 \\ \epsilon_k - 1 & (\epsilon_k + 1)e^{i\omega L_k/c_k}\end{bmatrix}, \tag{9}$$

with $k = 1, .., n$, $\epsilon_k = \frac{\rho_{k-1}c_{k-1}}{\rho_k c_k}$, where $\rho_0 c_0$ are taken from the element preceding the variable temperature duct. Here, the temperatures of the isothermal ducts were defined as the measured temperature at the location of the duct center, where linear interpolation was used in case this center was not an exact measurement point. The total variable temperature profile can then simply be included by multiplying the individual sections as $T_d = \prod_{k=0}^{k=(n-1)} T_d^{n-k}$.

3.2.3 Complete system

A complete system description can be obtained by matching the subsequent in- and outputs of the elements. This leads to an overall system transfer matrix $T_s = T_d T_f T_a T_u$, with T_f the flame element, T_a the area change, and T_u the upstream duct. Finally, the total set of system equations is obtained by combining T_s with the up- and downstream reflection coefficients R_u and R_d,

$$\begin{bmatrix} 1 & -R_u & 0 & 0 \\ 0 & 0 & -R_d & 1 \\ T_s(1,1) & T_s(1,2) & -1 & 0 \\ T_s(2,1) & T_s(2,2) & 0 & -1 \end{bmatrix} \begin{bmatrix} f_1 \\ g_1 \\ f_5 \\ g_5 \end{bmatrix} = 0. \tag{10}$$

Here the upstream reflection was determined via impedance measurements to be $R_u \approx 0.96$ over the frequency range of interest, while for the downstream reflection R_d the usual Levine-Schwinger relations were used [21].

In order to determine the eigenfrequencies the non-trivial solutions of equation (10) have to be found, which are given by the specific ω at which the determinant of the matrix (10) equals zero. The solutions can be found numerically by applying the Nelder-Mead root finding algorithm on the resulting equation [22]. The resulting $Re(\omega)$ denotes the eigenfrequency and $Im(\omega)$ the stability, with $Im(\omega) > 0$ being exponentially stable.

For each individual combination of upstream length, burner, and flow setting, the foregoing procedure yields a number of eigenfrequencies and corresponding eigenvectors. However, it is not uncommon that more than one eigenfrequency is unstable. In this situation, only the mode with the largest $|\omega_i|$ is selected. The data is stored as described before.

4 Results

In the following two sections we illustrate the proposed quantitative measures of section 2.2.2 by using them to (i) compare different burner cases using a single model, and (ii) to compare different models for a single burner case. Thus two slightly different applications of the quality measures are considered.

4.1 Quality as function of burner

Here the goal is to apply the various measures to assess the different aspects of model quality, and discuss their relevance. In order to keep the amount of information manageable, the results are evaluated on a per burner basis. Consequently, all the cases per burner are assigned to the same contingency matrix, and yield only one value for each of the presented performance metrics. The results per burner are shown in figure 4-6 in order of increasing burner porosity. The example cases shown in figure 1a correspond to (1) D2-P5, at $\Phi = 0.85$ and $V = 100\ cm/s$ and (2) D3-P6, at $\Phi = 0.85$ and $V = 100\ cm/s$ respectively. Naturally, these cases form only a subset of the total data for the respective burners.

Observing the data for the the TPR and TNR in figure 4 it becomes clear that there are large differences between the four burners. Apparently the ability of the model to predict all observed instabilities varies from 86% to only 54%. On the other hand, the percentage of experimentally stable cases which were predicted stable varies from 38% to 97%. Comparing the TPR and TNR directly, it is clear that

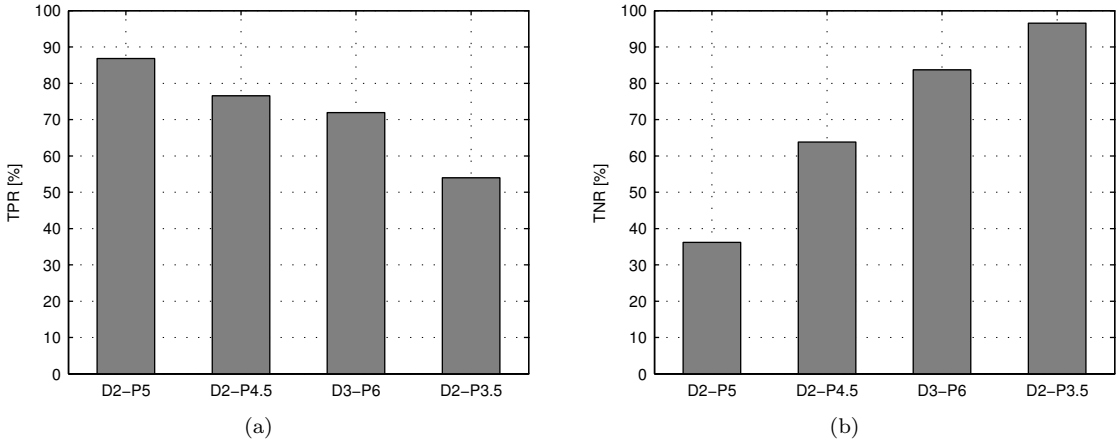

Figure 4: The (a) true positive rate (TPR) and (b) true negative rate (TNR) per burner.

an increase in the TNR seems to be directly related to a decrease in TPR. This in turn suggests that the ability to correctly predict as many as possible unstable cases from the experimental set is in conflict with doing the same for the stable cases.

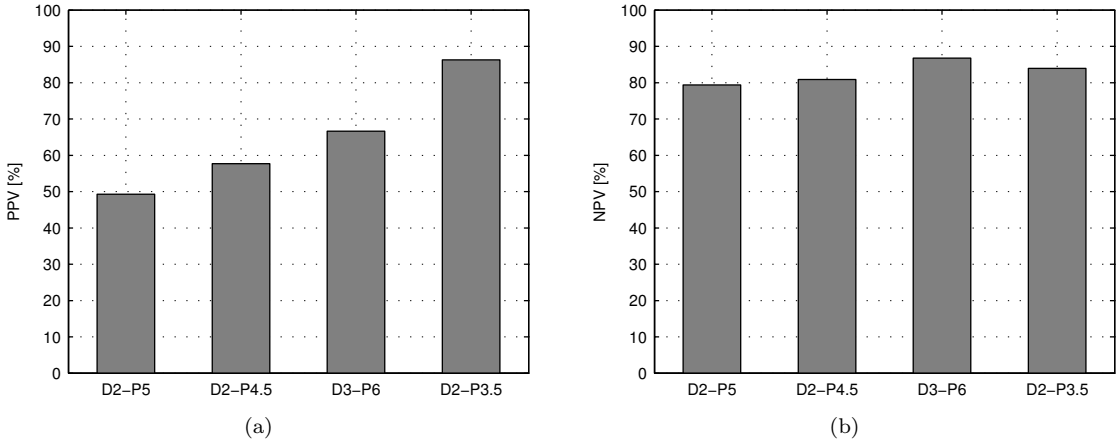

Figure 5: The (a) positive predictive value (PPV) and (b) negative predictive value (NPV), per burner.

Considering the data for the PPV and NPV in figure 5a and b respectively, a slightly different picture emerges. Clearly, the PPV also varies significantly across the cases, and increases from 50% to 85%. In contrast, the NPV fluctuates only lightly around 85%. Thus, it seems that the confidence one can have in stable model predictions is roughly constant across the cases. In contrast, the certainty of an unstable prediction being correct varies.

Finally, the accuracy measure shown in figure 6 indicates the combined prediction quality of unstable and stable predictions. As shown, the accuracy increases for increasing porosity. Note that in essence this reflects the combined data from the TPR, TNR, PPV and NPV plots.

In relation to the correctness of the predicted frequencies, it has to be noted that 79% of all correct unstable predictions were within 8% of the observed frequency. This confirms that the upper bound on the frequency error of 20% is only violated when the wrong mode is considered in which case it should not be regarded as a correct unstable prediction.

Let us now relate the results for PPV, TNR and TPR and NPV to each other. Clearly, the low porosity burners of the set have a high TPR coupled to a low TNR and PPV but with a relatively high NPV. This can only be the result of over-predicting the instability ranges. In that case most of the experimentally observed unstable ranges are captured well, giving a high TPR, but with many false

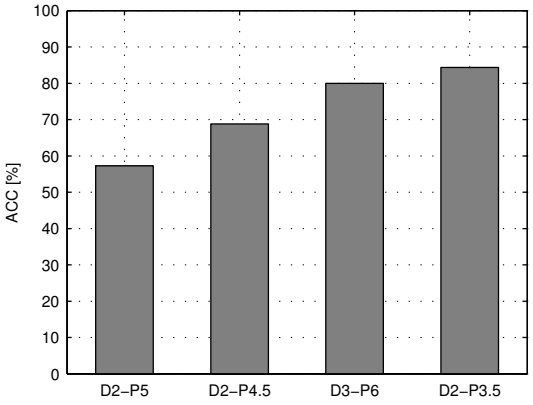

Figure 6: The accuracy ACC per burner.

unstable predictions (overpredicting instability), leading to low TNR and PPV numbers. The NPV is only affected little since compared to the total number of stable points relatively few false negatives are made. A totally different situation is the high porosity burner D2-P3.5. In this case, a low TPR goes together with a high TNR and PPV, which suggest over-prediction of stable cases. Contrary to what one expects however, the NPV is equally high as the other cases. This is an example of the aforementioned effect of difference in bias in the underlying datasets. Because there is a relatively large number of stable compared to unstable cases, the NPV is not significantly affected by a few wrong stable predictions, whereas the TPR is strongly influenced.

The results shown suggest that a single model may perform quite differently depending on which specific burner is considered. More specifically, the overall accuracy of the model increases with increasing porosity. At this stage it is not completely clear what the reason is for such dependence. Two possibilities come to mind, namely (i) bias of input parameters, and (ii) relative sensitivity to parameter errors.

First of all, and as mentioned in section 2, there might be an inherent bias error in the parameters used in the model. In this case, it is especially the flame transfer function and downstream temperature measurement which could introduce errors. For example, a decreasing bias in the flame transfer function measurement for increasing burner deck porosity might explain the observed trend. A similar statement can be made for the temperature distribution. However, there is no obvious reason as to why such trend in the bias might occur. Furthermore, it is the actual sensitivity of the model to the bias error in the input parameter that determines how it will appear in the output quality. This is precisely the subject of the second point.

A possible cause for the observed trend could be the difference in the relative role of model inaccuracies compared to the accuracy probing parameter L_1. To be more precise, let us give a fictive but illustrative example. Suppose the 'real-world' dependence of the stability indicator ω_i can be written (locally) around $\omega_i = 0$, as $\omega_i(L_1, T_2, R_2) = a(T_2)L_1 - R_2$, where $a(T_2)$ is the sensitivity of ω_i to L_1, and R_2 is the downstream end reflection coefficient. Note that it seems reasonable that $a(T_2)$ is a function of the downstream temperature, since it determines the relative role of L_1 compared to the total travelling time in the system. Furthermore, decreasing or increasing the losses through reflection coefficient R_2 will only offset the ω_i value. Next, we assume that in the model we underestimate the actual losses $\overline{R}_2$ by ΔR_2 as $R_2 = \overline{R}_2 - \Delta R_2$. In such a setting, it can be easily checked that the range of L_1 were the stability is wrongly predicted will scale with $a(T_2)^{-1}$. And thus the presented quality measure will depend on T_2. In practice, such interactions occur between all of the model's parameters. As a result, the accuracy measure will tend to score higher when the sensitivity of stability to L_1 is large compared to the model inaccuracies. In this case, the increase in accuracy is not a direct result of higher porosity, but of the implicit degree to which the different burners influence the sensitivity of the model to changes in L_1.

Returning to the performance indicators themself, it was discussed that it is the combination of the various measures which gives the most information about the quality properties of the model. In this light, it is worthwhile to consider the TPR versus the FPR. This leads to points in the so called *receiver operating characteristic* (ROC) plane, plotted in figure 7 for all four cases [16]. This plot further illustrates the cost versus benefit trade-off that the combination of the FPR and TPR represent. For

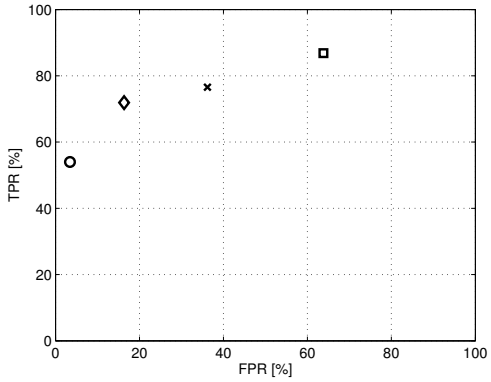

Figure 7: The TPR vs FPR. ($\circ$) D2-P3.5, ($\diamondsuit$) D3-P6, ($\times$): D2-P4.5, ($\square$) D2-P5.

example, the cost of having the high true positive rate in the D2-P5 case clearly is making a lot of false positives, thus wrong unstable predictions. In the following section we shall explore this concept further to illustrate model conservatism and how the ROC plane can be used to optimize a certain parameter in the model.

4.2 Quality as function of model parameter

The *ROC* plane can provide a valuable tool in case one is interested in the model performance as function of one of the models parameters. In particular, it is useful to assess the conservatism of a model in predicting stable behaviour. It is the most intuitive to consider a model conservative if it is 'careful' in assigning stability. From this viewpoint, a conservative model will score in the top right corner of the ROC plane, with both a high FPR and high TPR rate. A model with liberal stability predictions on the other hand, will score closer to the $(FPR, TPR) = (0, 0)$ point, indicating a low TPR and FPR rate. A perfectly accurate model will score near the top left corner, the $(0, 100)$ point. In most practical cases however, a model is not perfect, and creates points somewhere in the triangle spanned by the $(0, 0)$, $(0, 100)$ and $(100, 100)$ points. Thus, the closer the (FPR, TPR) points are to the $(0, 100)$ point the more accurate a model tends to be. In this setting, the most optimal choice of model parameters will minimize the distance of the (FPR, TPR) score to the point $(0, 100)$, representing the best possible trade-off between having a high true positive rate combined with a low false positive rate.

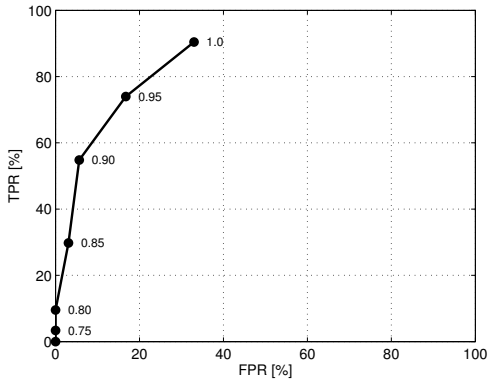

Figure 8: The TPR versus the FPR for different values of the upstream reflection coefficient R_2.

In the current context we provide a simple example by investigating the effect of a varying amount of acoustic radiation loss on the results for the D3-P6 burner. In particular, the upstream reflection coefficient R_u is varied in the range between $R_u = 0.5$ and 1. The results in figure 8 clearly show the drastic change which an increase in R_u causes. As can be expected, the model changes from very liberal

for $R_u < 0.85$ to rather conservative for $R_u = 1$. Note a value of $R_u \approx 0.95$ provides a good balance between predicting all unstable points and over-predicting the unstable ranges. This value is remarkably close to $R_u = 0.96$ which was measured using impedance measurements of the slitted piston and supports the usefulness of using the *ROC* plot to further optimize a model based on experimental validation.

5 Conclusions

In this paper a suitable quality indicator for thermoacoustic (in)stability prediction is proposed. Such indicator is desirable both from a practical, to assess large datasets, but also from a more fundamental viewpoint to investigate the reliability of (in)stability predictions. We have shown that using binary classification theory it is possible to define not only one, but a number of intuitive indicators to more precisely quantify model performance.

More precisely, the TPR and FPR give information about the amount of experimentally unstable and stable points which are predicted correctly. The PPV and NPV on the other hand can be interpreted as the likelihood an unstable or stable prediction is correct. Finally, the ACC gives an accuracy measure which does not distinguish in prediction quality between the two classes. Further insight can be gained by plotting the indicators in an *ROC* plot, which provides an excellent tool to analyse the optimality of the model, and investigate the trade-off between capturing all unstable points and making false unstable predictions.

The methods are applied to analyse the prediction quality of linear network model for a large experimental dataset. The data revealed large differences between the various burner cases, and disclosed a trend of increasing accuracy with increasing porosity. The reasons for this are not clear yet, but it is likely that parameter uncertainties relative to the probing parameter L_1 play in important role. The results indicate that a single model may perform quite differently for different burners and operating conditions.

The presented results clearly highlight the requirement to carefully think about the particular purpose of the thermoacoustic model. Depending on the exact modelling goal, different choices for the quality indicator can be preferable. For example, if the target of the model is to guarantee stable operation the NPV alone can be an appropriate indicator. Nevertheless, a high NPV can simply be obtained by assigning all but a few points as unstable. Clearly, a high NPV rating alone has only limited applied value.

From a practical viewpoint, it is important to realize that a conservative model which over-predicts the unstable ranges will limit the set of allowable parameters for stable operation. In the context of the current experiments, this means that the length range L_1 for which stable operation is obtained is in reality larger than predicted. Thus from a design perspective, a more logical modelling goal would be to achieve very trustworthy stable predictions but without limiting the practical design freedom of the parameter of interest too greatly. This demand is equivalent to a high NPV combined with a high TNR, as it means that the certainty of stable prediction is high, without limiting the range of allowable design parameters significantly. Note that this is similar to requiring a high TNR with a low FNR, in which case the *ROC* plot of the TNR versus the FNR could be used.

From the perspective of the data analysis process, it should be remembered that the method provides only a tool to assess quality and condense large amounts of data in a more suitable form for analysis. It does not immediately provide the user with possible (deterministic) reasons for these trends. Hence it should be viewed as an addition to current methods of analysis.

Finally, the approach is well suited to investigate the prediction quality of other, possibly more advanced, thermoacoustic models. For instance, with a few modifications to the input data, it is possible to compare the performance of a non-linear to a linear model. Given the amount of extra measurements, e.g. flame describing functions, and hence time and work, needed to perform non-linear predictions, it is worthwhile to investigate the relative improvement achieved in the predictions. The presented method provides the tools to investigate this quantitatively.

6 Acknowledgements

The authors wish to thank the referees for the constructive suggestions. This research was supported by the Dutch technology foundation STW .

References

[1] J.J. Keller. Thermoacoustic oscillations in combustion chambers of gas turbines. *AIAA Journal*, 1995, 33, 2280-2287.

[2] A.A. Putnam. *Combustion driven oscillations in industry.* American Elsevier, 1971.

[3] C.M. Coats, Z. Chang, and P.D. Williams. Excitation of thermoacoustic oscillations by premixing domestic gas burners. *Combustion Science and Technology*, 2008, 180, 314-342.

[4] S. Candel. Combustion dynamics and control: Progress and challenges. *Proceedings of the Combustion Institute*, 2002, 29, 1-28.

[5] T.C. Lieuwen. Modeling premixed combustion-acoustic wave interaction: A review. *Journal of Propulsion and Power*, 2003, 19, 765-781.

[6] A. P. Dowling and S.R. Stow. Acoustic analysis of gas turbine combustors. *Journal of Propulsion and Power*, 2003, 19, 751-764.

[7] F.E.C. Culick. Spatial averaging combined with a perturbation/iteration procedure. *International Journal of spray and combustion dynamics*, 2012, 4, 185-216.

[8] S. Nagaraja, K. Kedia, and R.I. Sujith. Characterizing energy growth during combustion instabilities: Singularvalues or eigenvalues? *Proceedings of the Combustion Institute*, 2009, 32, 2933-2940.

[9] N. Noiray, D. Durox, T. Schuller, and S. Candel. A unified framework for nonlinear combustion instability analysis based on the flame describing function. *Journal of Fluid Mechanics*, 2008, 615, 139-167.

[10] N. Noiray, D. Durox, T. Schuller, and S. Candel. Self-induced instabilities of premixed flames in a multiple injection configuration. *Combustion and Flame*, 2006, 145, 435-446.

[11] R. Kaess, T. Poinsot, and W. Polifke. Determination of the stability map of a premix burner based on flame transfer function computed with transient CFD. *European Combustion Conference*, 2009

[12] J.A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 1988, 240, 1285-1289.

[13] W. Krebs, T. Poinsot, H. Krediet, E. Portillo, S. Hermeth, C. Paschereit, and S. Schimek. Comparison of nonlinear to linear thermoacoustic stability analysis of a gas turbine combustion system. In *Proceedings of ASME Turbo Expo 2012*, 2012.

[14] F. Boudy, D. Durox, T. Schuller, G. Jomaas, and S. Candel. Describing function analysis of limit cycle in a multiple flame combustor. *Proceedings of the gas turbine technical congress and exposition*, 2010.

[15] N.A. Macmillan and C. Douglas Creelman. *Detection Theory, A User's Guide.* Lawrence Erlbaum Associates, second edition, 2005.

[16] J.P. Egan. *Signal Detection Theory and ROC Analysis.* Academic Press, 1975.

[17] Manohar. *Thermo-acoustics of bunsen type premixed flames.* PhD thesis, Technical University of Eindhoven, 2011.

[18] M.L. Munjal. *Acoustics of ducts and mufflers.* John Wiley & Sons, 1987.

[19] K. Truffin and T. Poinsot. Comparison and extension of methods for acoustic identification of burners. *Combustion and Flame*, 2005, 142, 388-400.

[20] V.N. Kornilov, K.R.A.M. Schreel, and L.P.H. de Goey. Experimental assessment of the acoustic response of laminar premixed bunsen flames. *Proceedings of the Combustion Institute*, 2007 ,31 , 1239-1246.

[21] H. Levine and J. Schwinger. On the radiation of sound from an unflanged circular pipe. *Phys. Rev.*, 1948, 73, 383-406.

[22] J.C. Lagarias, J.A. Reeds, M.H. Wright, and P.E. Wright. Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on Optimization*, 1998, 9, 112-147.