Pitch Based Sound Classification

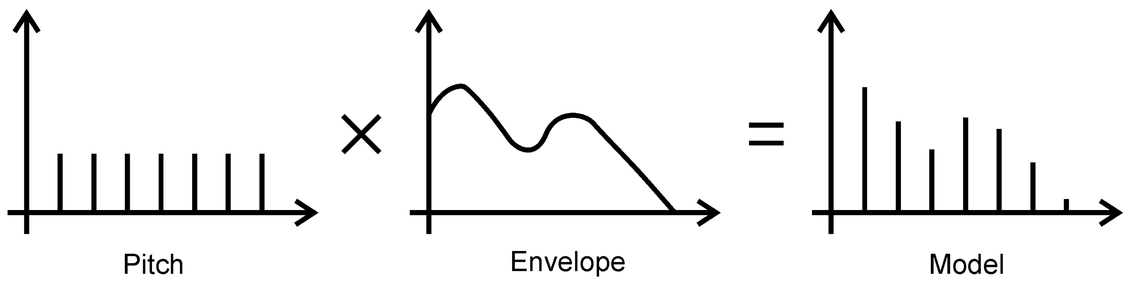
Andreas Brinch Nielsen, Lars Kai Hansen
Technical University of Denmark
abn,lkh@imm.dtu.dk

Ulrik Kjems
Oticon A/S,
uk@oticon.dk

IMM/ISP

Introduction

Ideally you can divide a harmonic signal into two parts. The pitch and the envelope.



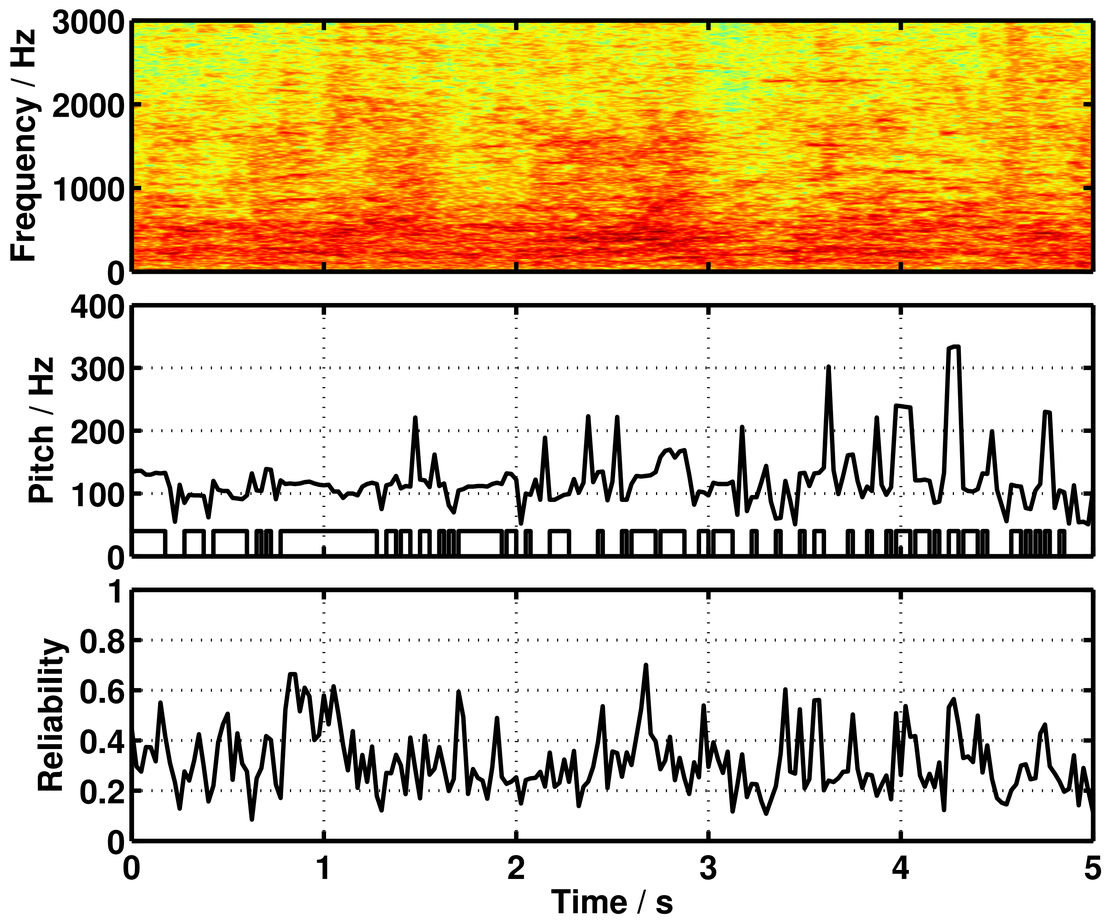
When filtering a signal, like for example on a phone line, only the envelope is affected. A classifier based on the pitch alone should therefore be very robust to this kind of filtering. As stated, the pitch is not affected but the estimation of the pitch is, so this is not a magic trick, but might still give some possibilities.

In this paper the pitch is used for the classification of sound into three classes; music, speech and noise. Features have been found mostly on the dynamic features of the pitch and the pitchness of a signal. These features are used in a linear network using the softmax output function. The features are used directly, but also quadratic combinations of the features equal to having diagonal and full covariance are used, giving three complexities of the model (Linear, QuadDiag and QuadComp).

A window size of 100 ms with 75 % overlap is used when estimating the pitch. Then, pitch estimates are integrated over a window of 1s - 5s and various features are found. Examples of the extracted pitch in the three classes are shown on this poster and some examples of features are shown and the other poster.

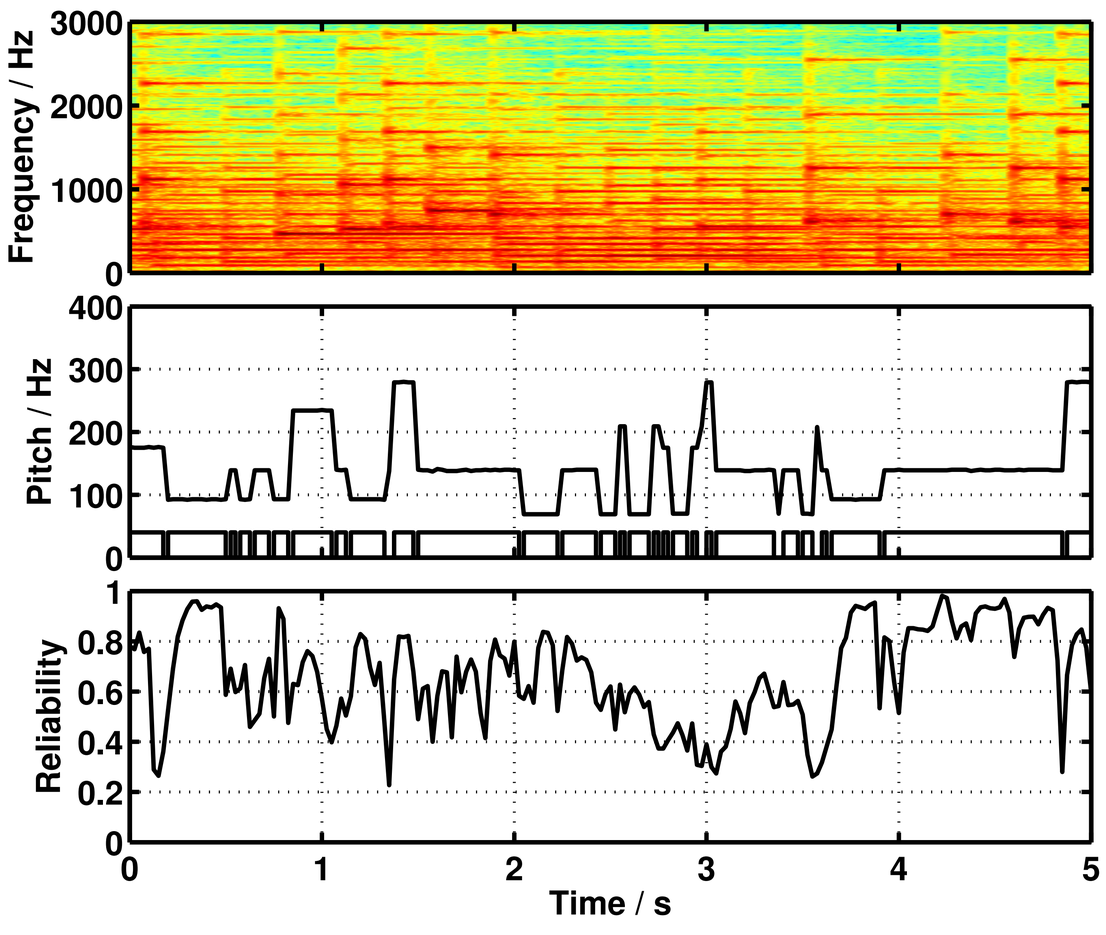
The signal is divided into reliable windows which are shown in the bottom of the pitch plots.

Noise



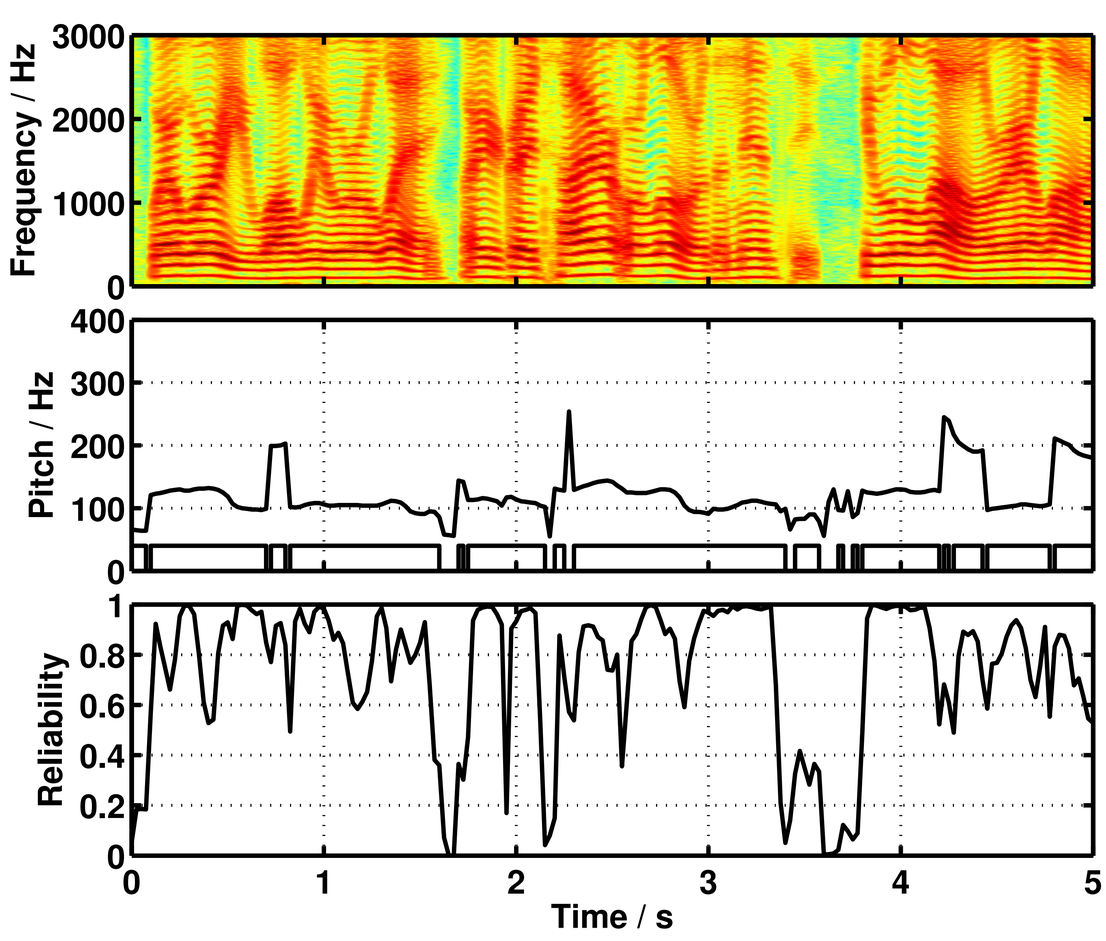
The pitch estimate in noise is often random in nature, but can also be constant at times, depending on the type of noise. The reliable windows are generally short with little change of the pitch and do not hit musical notes. The reliability of noise is often low and with low variation compared to music and speech.

Music



The pitch is confined in steps which are caused by the musical notes. Note the very constant pitch within each step. Each reliable window captures a note. The maximum reliability values are close to unity and the minima are relatively high. This reflects the fact that pitch is dominant in music. Most dips occur between notes.

Speech



Notice the long reliable windows and the changes of the pitch inside a window. The pitch is used to emphasize words in speech. Also notice the high maxima and low minima of the reliability. This reflects the differences between voiced and unvoiced regions, consisting of pitch and white noise respectively.

Pitch Based Sound Classification

Andreas Brinch Nielsen, Lars Kai Hansen
Technical University of Denmark
abn,lkh@imm.dtu.dk

Ulrik Kjems
Oticon A/S,
uk@oticon.dk

IMM/ISP

Features

This is the four features that performed the best.

ReliabilityDev: The standard deviation of the reliability signal (r_i) within the classification window,

$$f_{ReliabilityDev} = \sqrt{\frac{1}{I-1}\sum_{i=1}^{I}(r_i - \mu_r)^2}, \quad (1)$$

where I is the number of pitch samples in the classification window.

Difference1: The number of pitch (p_i) abs-difference values in the interval [0;2[,

$$f_{Difference1} = \sum_{i=2}^{I}\left(|p_i - p_{i-1}| < 2\right), \quad (2)$$

ToneDistance: The average distance from the estimated pitch to a 12'th octave musical note is found,

$$t_i = 12\log_2\frac{p_i}{440}, \quad (3)$$

$$f_{ToneDistance} = \frac{1}{I}\sum_{i=1}^{I}|t_i - round(t_i)|. \quad (4)$$

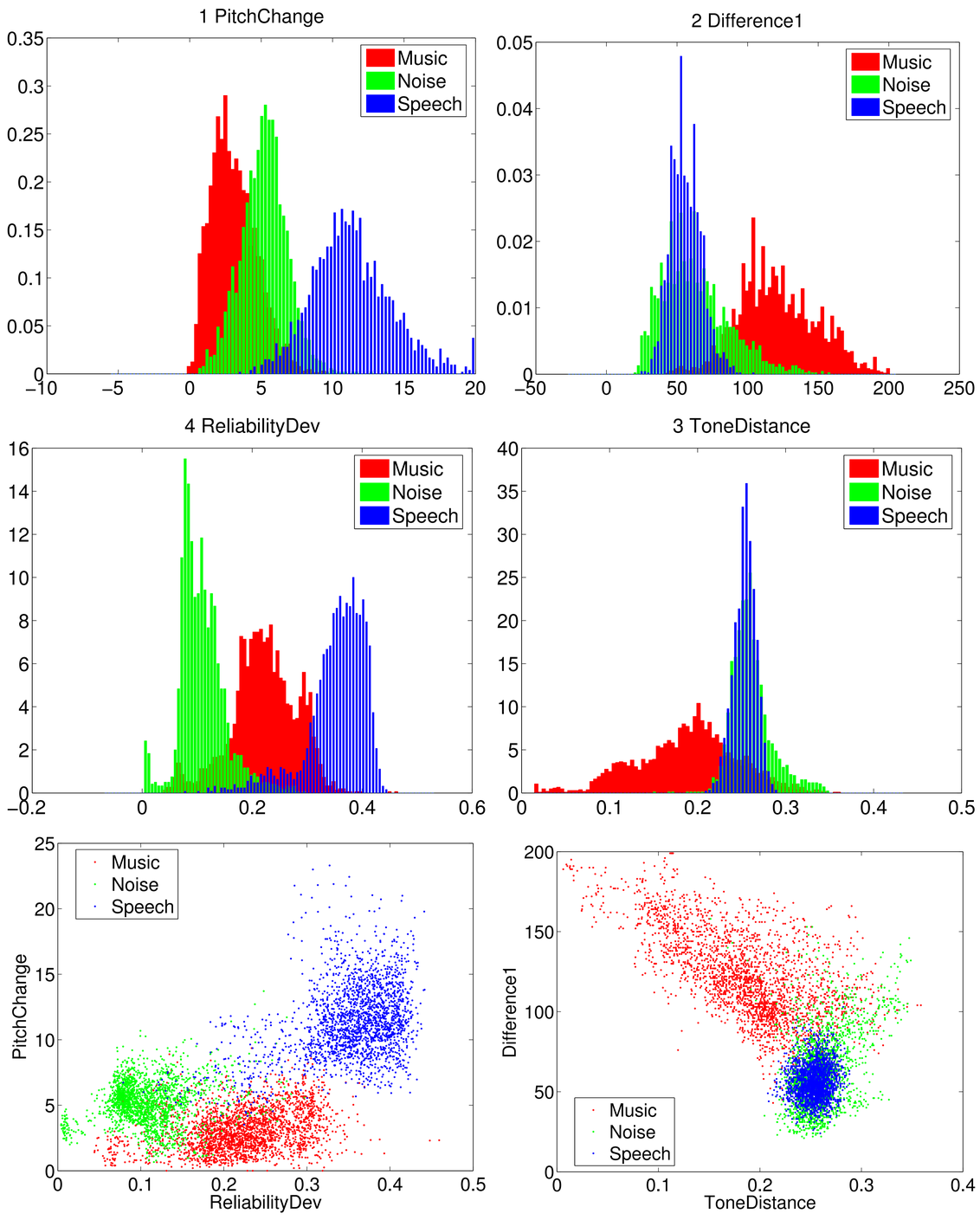
PitchChange: The PitchChange feature measures the difference between the highest and the lowest pitch in a reliable window and calculates the mean over a classification window,

$$d_w = \max(\boldsymbol{p}_w) - \min(\boldsymbol{p}_w), \quad (5)$$

$$f_{PitchChange} = \frac{1}{W}\sum_{w=1}^{W} d_w, \quad (6)$$

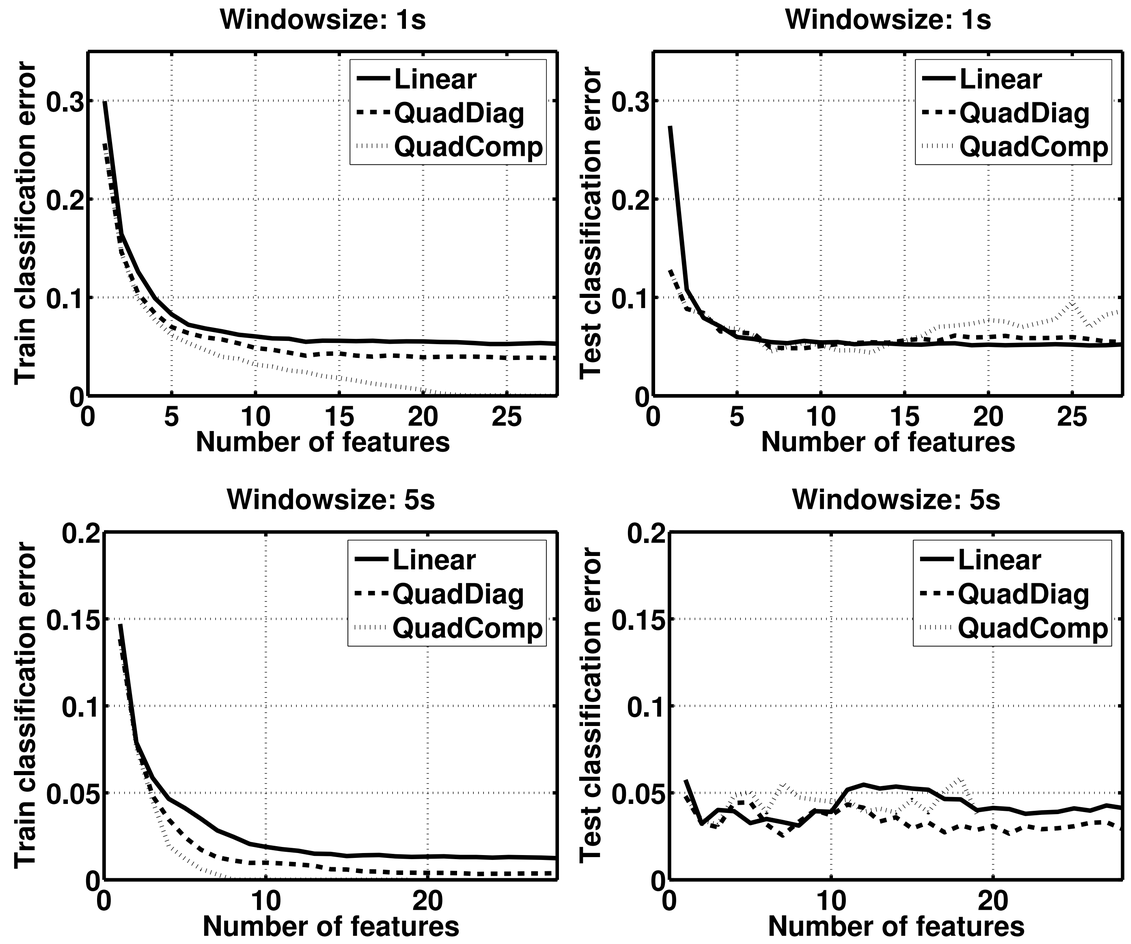
with W being the number of reliable windows, and $\boldsymbol{p}_w$ a vector of the pitch values in reliable window w.

Feature comparison

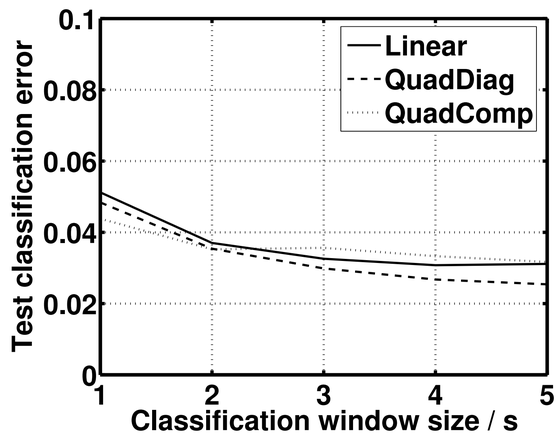


Results

For testing the model a test set is used. The training and test log likelihoods for increasing number of features is shown below. The features have been ranked using forward selection and are used in that sequence. Results for both windows of 1s and 5s are shown. To the right the best value is plotted for each model and window size.



Here the minimum for each model and for each window size is plotted, to show the influence of the window size on the classification rate.



Conclusion

The results show that the longer window size the better. An error rate of 3 % is achieved for 5s. The shorter window of 1s is not that much worse with an error rate of 5 %.

The results also show that not much is gained when increasing the model complexity and actually the full covariance matrix seems to perform worse than the linear. It would be expected though that with a bigger training set which would avoid overfitting this should at least perform the same.

A five dimensional feature vector and a linear network achieves an error rate of only 6 %.