PITCH BASED SOUND CLASSIFICATION

Andreas B. Nielsen, Lars K. Hansen

Intelligent Signal Processing,
IMM, DTU,
DK-2800 Lyngby

Ulrik Kjems

Oticon A/S,
Kongebakken 9,
DK-2765 Smørum

ABSTRACT

A sound classification model is presented that can classify signals into music, noise and speech. The model extracts the pitch of the signal using the harmonic product spectrum. Based on the pitch estimate and a pitch error measure, features are created and used in a probabilistic model with softmax output function. Both linear and quadratic inputs are used. The model is trained on 2 hours of sound and tested on publically available data. A test classification error below 0.05 with 1 s classification windows is achieved. Further more it is shown that linear input performs as well as a quadratic, and that even though classification gets marginally better, not much is achieved by increasing the window size beyond 1 s.

1. INTRODUCTION

The classification of sound is important for hearing aids because different sound environments, like music and speech, require different processing to get the optimal experience for the user. Hearing aids rely more and more on automatic program selection, which requires accurate sound environment classification.

The pitch is an interesting feature of sound and is used in various situations. One of the reasons pitch is interesting is that it is somewhat immune to frequency modulation like for example low pass filtering. If for example sound is transferred through a telephone, the pitch of the signal remains the same, and it is only the estimation of the pitch that is affected. Potentially this can make the classification more robust.

For a short survey of pitch estimation references look in [5]. In [2] the pitch is used for blind source separation, music information retrieval based on pitch is done in [7] and in [12] the pitch is used for language identification. The pitch has also been used for classification. In [4], [15] and [16] the pitch mean, variance and difference are used as features.

In this paper a classification system based solely on the pitch will be created. We are interested in three classes, broadly defined as; music, noise and speech. For such a system a pitch estimator, pitch features and a classification model is necessary. For running in a hearing aid it is important to have simple algorithms. The pitch estimation can be the most demanding part of the system, but in this paper the quite fast harmonic product spectrum will be used. Again to enhance efficiency, effort goes into finding features that separates the classes well, instead of using a complex classification model. A simple probabilistic model is used for the actual classification. To evaluate the system, it will be trained using a collection of sound that captures many of the aspects of both music and noise. A test set of publically available sounds has been gathered to facilitate comparisons.

In the second section the pitch estimation, the pitch features and the classification model will be presented. In section three the data used for both training and test is described. Section four contains the results followed by the discussion.

2. METHODS

In this section the pitch estimator is presented. A selection of the investigated features are described and finally the classification model is presented.

2.1. Pitch estimation

The harmonic product spectrum (HPS), see e.g. [6], exploits a simple feature of the frequency spectrum of a harmonic signal. When downsampling the spectrum the harmonics align with the fundamental frequency. When the original spectrum is multiplied with the downsampled spectra the HPS appears. The pitch is estimated by taking the maximum of the HPS. The value R defining the number of downsamplings is set to 5 in this paper.

In [13] a combined algorithm using the HPS was compared to a Bayesian pitch estimator [9] and HMUSIC [5]. The combined algorithm was superior for classification and the HPS shows similar performance on its own. The HPS has problems with doublings of the pitch, but this does not affect classification, because the pitch dynamics are not affected by doubling the pitch. The advantage of HPS is speed and lack of noise. Other algorithms have better accuracy, but tend to overlook frequencies if too small a search grid is used, which can be necessary to decrease computation time.

Apart from the pitch, a measure of the pitch error is found as well. This measure is called reliability, r. It is based on the relation between the energy in the maximum peak in HPS and the total energy in the HPS,

$$r = \frac{e_{pitch}}{e_{HPS}}. \tag{1}$$

This measure is close to zero when there is very little pitch in the signal and close to unity when the pitch is well defined.

2.2. Features

Features are build on top of the two signals, p and r. A total of 28 features are found, but only four will be described here. For the complete listing refer to [13]. An example from

each class is shown in Figures 1, 2 and 3. In the captions of the figures the characteristic properties of the classes are described and illustrated.

Reliable windows are used to divide pitch samples into smaller subsets. The pitch samples in a reliable window are characterized by abs-difference values smaller than a threshold. The reliable windows in music captures a musical note each and in speech a voiced region, see Figures 1 and 2. Some of the features use the reliable windows and some features use the complete classification window.

Next, the four features that yielded the best performance in the forward selection procedure are described.
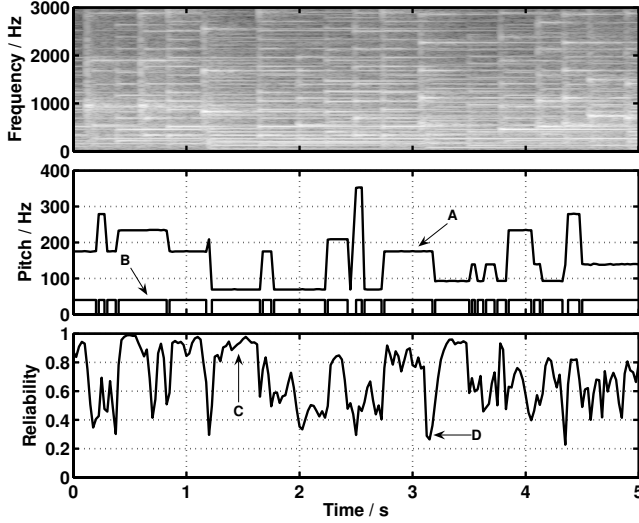
Fig. 1. Music. The pitch is confined in steps (A) which is caused by the musical notes. Note the very constant pitch in each step. In the bottom of the pitch plot is shown the reliable windows (B) (see text for explanation); each captures a note. The maximum of the reliability values is close to unity (C) and the minima are relatively high (D). This reflects the fact that pitch is dominant in music. Most dips occur between notes.

ReliabilityDev: This feature is the standard deviation of the reliability signal within the classification window,

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

where I is the number of pitch samples in the classification window. This feature shows good separation of all classes with speech having high values, noise low and music in between.

Difference1: Because it is especially the dynamic properties of the pitch that describe the classes well, the difference between subsequent pitch samples is a good measure. The feature is based on a histogram of the pitch abs-difference values and is the number of values in the interval [0;2[. It is good for separating music from speech and noise, because music has many constant pitch parts and therefore a high Difference1 value.

ToneDistance: This feature is good for separating the music from speech and noise. It is the distance from the pitch
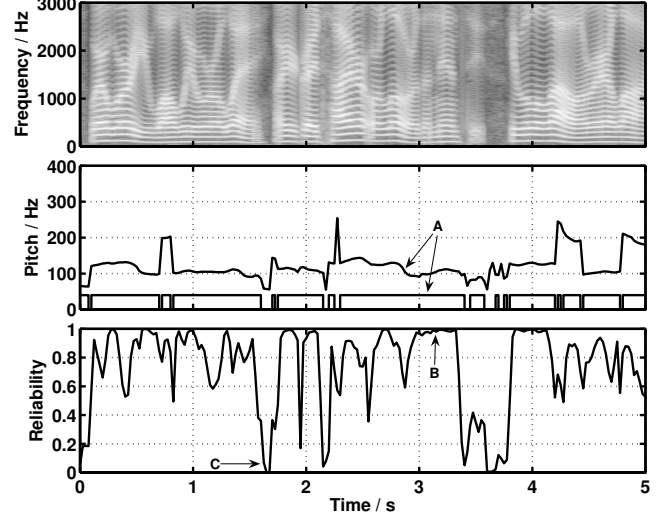
Fig. 2. Speech. Notice the long reliable windows and the changes of the pitch inside a window (A). The pitch is used to emphasize words in speech. Also notice the high maxima (B) and low minima (C) of the reliability. This reflects the differences between voiced and unvoiced regions, consisting of pitch and white noise respectively.

to a 12'th octave musical note. First the pitch is converted to the tone scale, t. This scale takes on integer values for notes and in between values for other pitches. The feature is the mean of the distances to the closest tones,

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

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

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,

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

$$f_{PitchChange} = \frac{1}{W} \sum_{w=1}^{W} d_w, \qquad (6)$$
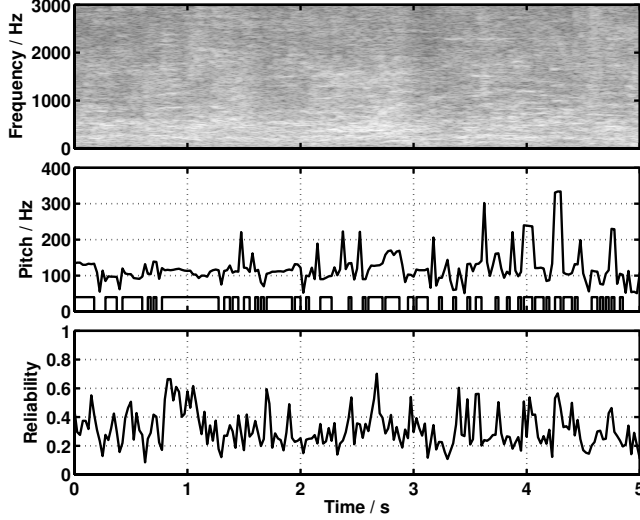
with W being the number of reliable windows, and $\boldsymbol{p}_w$ a vector of the pitch values in reliable window w.

A description of the remaining features can be found in [13].

2.3. Classification model

In this paper a probabilistic model is used based on the softmax output function [3]. The model is used because of its simplicity,

$$z_c = \sum_{i=0}^{F} a_{i,c} f_i, \qquad (7)$$

Fig. 3. Noise from a café. The pitch estimate in noise is often random in nature, but is also constant at times, depending on the type of noise. The reliable windows are generally short with little change of the pitch and does not hit musical notes. The reliability of noise is often low and with low variation compared to music and speech.

where F is the number of features, f_i are the feature values, $f_0 = 1$ is the bias and $a_{i,c}$ are the variables of the model. In order to get the desired output interval of 0 to 1, the softmax output function is used,

$$y_c = \frac{\exp(z_c)}{\sum_{c'}^{C} \exp(z_{c'})}, \qquad (8)$$

for C classes. This output function also assures that the outputs sum to 1 and they can be interpreted as probabilities. The model is trained using maximum likelihood.

Three variations of the input are used. The linear with the features used directly (Linear). A quadratic including the squares of the features, thus doubling the input size (Quad-Diag), and last a quadratic where all the covariance combinations are used making a total of $i(i + 3)/2$ inputs (excluding
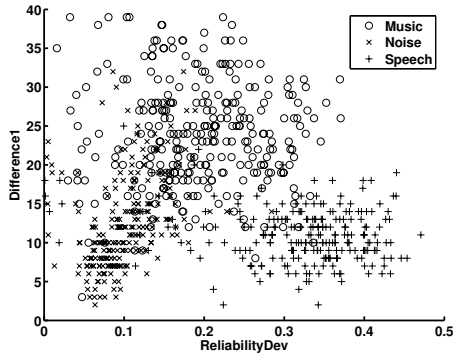
Fig. 4. Scatter plot of the first two features described. The clustering is apparent. Especially speech separates well from the other two classes.

the bias) to the model (QuadComp).

3. DATA

The database used for training the model consists of the clean classes of music, noise and speech. The music has been taken from various recordings from a variety of genres totalling, 50 minutes. The noise consists of traffic, factory noise, a helicopter, many people talking and various other sources, totalling 40 minutes. The speech was taken partly from the Keele [11] and KED Timit [10] clean speech databases and from other clean speech sources in different languages, totalling 42 minutes. A complete description can be found in [13]. For the test set publically available sounds were used. Eight music files totalling 38 minutes, 23 minutes of noise, and 35 minutes of speech were included. The specific files can be found here [14].

The pitch estimator uses windows of 100 ms and overlap of 75 ms. This makes approximately 40 pitch samples per second. The classification windows of above 1 s uses overlap to obtain a classification every second. With these settings the training set size is approximately 7000 samples and the test set is approximately 5500 samples.

4. RESULTS

In total 28 features were found. Forward selection is used to select between them. Results with 1 s classification windows is shown in Figures 5 and 6.
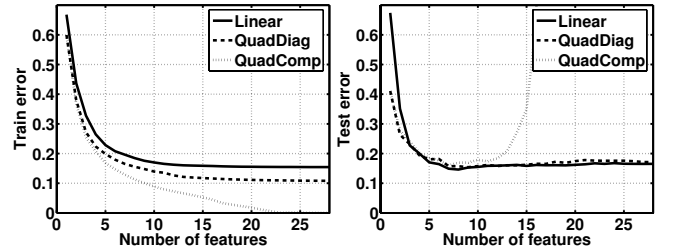

Fig. 5. Negative log likelihoods. The test error shows a minimum at 7 features depending. Somewhat surprising the linear model shows superior performance, which might be explained by overfitting of the larger models.

The three plots of the test errors of Figure 5 shows no improvement when using more than 7 features. The more complex models show better training error, but when it comes to test error not much is gained, and from five features the linear model performs better. This is verified in Figure 6 where the three models show very close performance.

All three models reach a minimum classification error of approximately 0.05. If the linear model with seven features is chosen, only 24 parameters $(8 \cdot 3)$ is necessary to achieve a classification error of close to 0.05.

Especially interesting is the classification of speech. For seven features and the linear model the speech classification error is 0.01 and the false alarm rate is 0.07. With 5 s windows no speech samples are misclassified and false alarm rate is 0.03. This can be interesting especially for hearing aids where speech is an important class to recognize.

If longer windows are used better classification can be obtained. Figure 7 shows the relation between window size and test classification error.
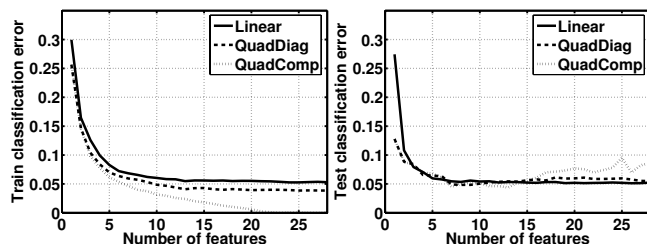
Fig. 6. Classification error for both training and test data with 1 s windows. Only a small improvement is achieved with the larger models. A test classification error of just below 0.05 is achieved.
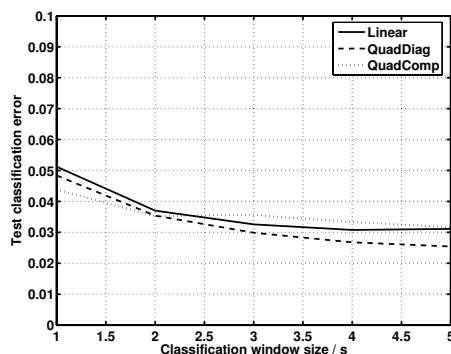
Fig. 7. Test classification error with different window sizes. The minimum classification error is chosen for each model and window size.

5. DISCUSSION

In this paper, a new way of using the pitch for classification is presented. It shows that the pitch can discriminate between the three classes; music, noise and speech. Results of 0.05 general miss rate and 0.01 miss rate of speech were obtained with a 1 s windows. Hence, we find classification results that are significantly better than in earlier systems [1].

The HPS pitch estimator and the linear model are both very efficient algorithms. It might be feasible to incorporate some of the functionality in hearing aids, and this could possibly increase the classification functionality.

6. ACKNOWLEDGEMENT

7. REFERENCES

[1] S. Allegro, M. C. Büchler and S. Launer, Automatic Sound Classification Inspired by Auditory Scene Analysis, *Eurospeech*, Aalborg, Denmark, 2001.

[2] F. R. Bach and M. I. Jordan, Blind one-microphone speech separation, Neural Information Processing Systems 17, pp. 65-72, 2005.

[3] C. M. Bishop, *Neural networks for pattern recognition*, Oxford University Press, 2004.

[4] M. C. Büchler, *Algorithms for sound classification in hearing instruments*, Swiss Federal Institute of Technology, 2002.

[5] M. G. Christensen, S. H. Jensen, S. V. Andersen and A. Jakobsson, Subspace-based Fundamental Frequency Estimation, *Proc. 12th European Signal Processing Conference*, pp. 637-640, Vienna, 2002.

[6] P. de la Cuadra, A. Master and C. Sapp, "Efficient pitch detection techniques for interactive music", *Proc. Int. Computer Music Conference*, Havana, 2001.

[7] A. Durey, M. Clements, "Direct estimation of musical pitch contour from audio data", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, 2004.

[8] D. Ellis, www.ee.columbia.edu/~dpwe/.

[9] L. K. Hansen, F. Å. Nielsen and J. Larsen, "Exploring fMRI Data for Periodic Signal Components", *Artificial Intelligence in Medicine*, vol. 25, pp. 25-44, 2002.

[10] Festvox project, www.festvox.org/dbs/.

[11] C. F. Meyer, www.liv.ac.uk/Psychology/hmp/projects/pitch.html.

[12] C.-Y. Lin and H.-C. Wang, "Language Identification Using Pitch Contour Information", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, 2005.

[13] A. B. Nielsen, *Pitch based sound classification*, IMM, DTU, 2005.

[14] A. B. Nielsen, www.imm.dtu.dk/~abn.

[15] S. Pfeiffer, S. Fischer and W. Effelsberg, "Automatic Audio Content Analysis", *ACM Multimedia*, pp. 21-30, Boston, 1996

[16] E. Wold, T. Blum, D. Keislar and J. Wheaton, "Content-Based Classification, Search, and Retrieval of Audio", *IEEE MultiMedia*, vol. 3, no. 3, pp. 27-36, 1996.