Discrimination of Speech and Monophonic Singing in Continuous Audio Streams Applying Multi-Layer Support Vector Machines

Björn Schuller, Gerhard Rigoll, and Manfred Lang
Institute for Human-Machine Communication
Technische Universität München
D-80290 München, Germany
(schuller | rigoll | lang)@ei.tum.de

Abstract

In this paper we present a novel approach to the discrimination of speech and monophonic singing for the use in Music Information Retrieval applications. A working prototype is introduced applying Multi-Layer Support Vector Machines for the discrimination, and static high-level features derived of the pitch and energy contours of an acoustic signal. The feature set for the discrimination is presented and ranked according to a Linear Discriminant Analysis. For the automatic segmentation within an input signal stream a further feature set is used for the discrimination of signal and noise. A corpus for training and evaluation comprising speech and monophonic singing data of nine performers is described in detail. The data has been labeled according to the judgments of another set of probands. A recognition rate of correct assignments of 99.2 % could be reached, and demonstrates the high performance of the proposed methods.

1. Introduction

Considering the rapidly growing sizes of music databases novel retrieval methods are commonly regarded as obligatory for efficient and fun access. Furthermore increasingly smaller portable music devices are able to store very large databases of in general compressed digital music. On these devices very limited display space and only a few hard-keys are provided. A search for a musical piece can therefore often not be performed by a keyword search. Considering furthermore special situations as automotive environments, it may not be desired or possible to select a title within a large list by typing or scrolling through such a list in view of distraction. In these described cases speech as an input modality seems very reasonable, as speech leaves the hands and eyes free and requires only a microphone for the input recording. An especially promising approach is content-based retrieval encouraging the user to hum or sing the key-melody of a desired musical piece. However, melody-based retrieval has its limits as the key melody does not allow for a final identification of the correct artist considering cover interpretations of songs, or the exact version having live or studio recordings, remixes or edits in different languages in mind. Therefore melody based music retrieval cannot lead to a final result in general, but can only help to dramatically constrain the number of titles in question. A further constraint or even final selection of titles can be fulfilled by speech, which seems very natural as the query was initiated by means of vocal articulation. By a natural speech utterance the user can either provide more information as the year, genre, type of recording and, depending on the power and size of the retrieval engine and database, even more exact information as the artist, title or parts of the lyrics. Further more a referencing selection between presented alternatives can be intuitively fulfilled by speech as well.

In view of this background an automatic discrimination between speech and monophonic singing input is desirable [1] preventing to ask the user for explicit labeling of his vocal input type. While the use in music retrieval is our goal [2], a number of further application scenarios exist. Such comprise automatic labeling of speech and singing data e.g. in entertainment shows or radio interviews.

In this paper we therefore evaluate an optimal feature set for the discrimination of speech and monophonic a-cappella singing strongly differing from the far spread discrimination of polyphonic music and speech.

0-7803-8603-5/04/$20.00

2. Selected Features

In order to be able to discriminate speech and monophonic singing we aim to find features carrying the desired information of the nature of the underlying acoustic signal. Furthermore the features should not depend on the actual spoken or sung content itself too strongly. Finally they should fit the chosen modeling by means of classification algorithms. As there is a number of features suited for this task it seems to be important to find the optimal feature set in view of a maximum performance and great generalization capability in order of speaker and content independence. We chose the analysis of the contours of pitch and energy for their well-known capability to carry a large amount of information considering the perceptual difference of spoken and sung signals. In comparable works [1] the use of pitch information is also propagated. However, further features, among them energy based duration of silences, and durations of voiced sounds are introduced in this paper. The selected contours rely rather on broad classes of sounds while spectral characteristics in general seem to depend too strongly on phonemes and therefore on the phonetic content. In order to calculate the contours, frames of the signal are analyzed every 10ms using a 20ms Hamming window function. The values of energy are calculated by the logarithmic mean energy within a frame. The pitch contour is achieved by use of the average magnitude difference function (AMDF) as can be seen in the equation, where $F_{0,i}$ represents the fundamental frequency in the frame i, $s(k)$ the signal at a discrete time instant k, N stands for the last sample in the frame and f_s is the sampling frequency. The order of the propagated AMDF-function is represented by j.

$$F_{0,i} = \left(\frac{\arg\min_k \frac{1}{N} \sum_{k=0}^{N-1} \left| s_i(k) - s_i(k+k) \right|^j}{f_s} \right)^{-1} ; N = T_w \cdot f_s$$

The AMDF provides a faster alternative to the calculation of the autocorrelation function of a frame. The precondition however is that it is calculated in first order which results in additions instead of multiplications compared to the related auto correlation function, and claims the search of the minimum instead of the maximum to achieve the instantaneous pitch value. As all estimation methods for pitch contour, this technique also underlies deviations from the original contour, which could only be measured by glottal measurement. AMDF-based pitch calculation proved robust against noise but susceptible to dominant formants.

Considering our related works in speech emotion recognition [3], derived static features show more independence of the content compared with direct analysis of the raw contours using e.g. Hidden Markov Models. Therefore we decided to derive higher-level static features out of the mentioned contours. Initially a set of 33 features was analyzed. The durations of pauses are calculated according to an energy level within a frame. The durations of voiced sounds are estimated by the pitch contour. As unvoiced sounds have no harmonic character only frames with a clearly determinable pitch value are considered as voiced frames. For the integration of spectral information in a very universal way also spectral energy below 250Hz and 650 Hz was computed following the computation of a Fast Fourier Transform. However, this information did not show significant correlation with singing or speech. The final feature vector consists of the 10 features that showed the greatest potential in our tests. A simple feature-wise ranking according to a Linear Discriminant Analysis (LDA) of these overall 10 features is shown in table 1.

Table 1. LDA-Ranking Speech / Singing Discrimination features

Feature	Performance (LDA) in %
Rate of voiced sounds	90.4
Duration of voiced sounds standard deviation	88.8
Mean silence durations	85.8
Mean duration of voiced sounds	85.4
Pitch relative absolute area	72.3
Pitch relative minimum	72.3
Pitch range	63.8
Pitch maximum gradient	62.3
Pitch relative maximum	42.7
Pitch standard deviation	33.7

Some pitch information as the mean pitch is dependent of the speaker, and even in speaker-dependent recognition do not show different behavior in speech and singing. The relative positions of the maximum and minimum pitch, which are often used in speech emotion recognition [3] as well, did expectedly not help for the discrimination. The mean gradient of pitch, the distance between reversal points, and their standard deviation also showed only

insignificant changes between speech and singing in our evaluation.

As for energy the derived features besides the mean silence durations did not show a significant potential. Among those were: standard deviation, mean value, mean and median of fall- and rise-time, the maximum value and its relative position, the maximum gradient, the mean distance between reversal points and its standard deviation, and the median of silence durations.

Derived features of the signal-contour itself were also neglected in the final feature vector. Those comprise the number of zero-crossings, the median of sample values and the mean value.

3. Support Vector classification

In this paper we can provide only a very brief introduction to the theory of Support Vector Machines (SVM). The basic classification separates two classes [4]. The learning problem can be seen as finite training set

$$\Psi = \left\{ \left(x^m, d^m \right) / m = 1,...M \right\} \text{ with}$$
$$x^m \in \Re^n \text{ and } d^m \in \{+1,-1\}.$$

The sample vectors of a class x^m with $d^m=+1$ shall be considered as positive instances, while the vectors of the other class x^m shall be seen as negative instances with $d^m=-1$. To be able to separate these two classes we span the following hyper-plane by the vector $w \in \Re^n$ and a bias $b \in \Re$ with the given condition:

$$\mathrm{H} = \mathrm{H}(w,b) = \left\{ x \in \Re^n / w^T x + b = 0 \right\}$$
$$\mathrm{d^m} = \pm 1 \quad \Rightarrow \quad \mathrm{w^T x_m} + \mathrm{b} = \pm 1$$

In order to achieve a discriminant separation this hyper-plane is placed optimally between the two classes providing the maximum distance to each class according to a Lagrange optimization. The so-called support vectors then span the plane. As only these vectors are needed a great reduction of references is achieved by this approach compared to other distance-based classifiers. In the recognition phase the distance of a test-sample to the separation hyper plane forms the basis of decision. As the problem we consider is not of a linear characteristic we use a transform $\Phi : \Re^n \to \Re^N$ into a higher dimensional feature space by a kernel mapping function $K^{\Phi}(x,y) = \Phi(x)^T \Phi(y)$. The aim is linear separability in the new feature space. In our problem a radial basis function kernel showed optimal performance.

After training of the SVM the class with the minimum distance is selected throughout the recognition process. The fact that SVM tend to show high generalization potential due to their structural risk minimization is especially useful as the samples of speech and singing show weak intra-class correlation considering independence of the content.

4. Data corpus

Due to the lack of a public corpus nine probands were invited to sing and speak a corpus for test and evaluation purposes. Two of them were female. A set of 50 selected songs of contemporary music had to be sung twice by each performer in two cycles. Additionally a phase of free singing in five different musical genres took place. The same test-persons had to read out loud ten sentences and talk freely about their opinion considering Music Information Retrieval in order to collect speech samples under the same conditions. The samples were recorded in an acoustically isolated room by use of an AKG-1000s MK II condenser microphone. The sung samples were reclassified by three further probands in order to assure their perceptual difference to spoken utterances. In total 1035 sung samples and 10.5h of speech were collected.

5. Segmentation

As speech and singing shall be recognized in a continuous stream, an automatic segmentation is needed prior to the final discrimination. A first segmentation is achieved by use of a conventional energy threshold, which has to be exceeded, or respectively under-run for a set time interval to indicate the start or end of a signal of interest.

Secondly a Support Vector classification is fulfilled applying Mel Frequency Cepstral Coefficients (MFCC) and δMFCCs for the discrimination of singing or speech and noise. Noise is considered here as ambient noise and comprises background polyphonic music, as one target scenario is user selection of musical titles via singing while expectedly already listening to music. MFCC have proven highly effective in the field of automatic speech recognition as they model the subjective pitch and frequency content of audio signals. They are computed from the FFT power coefficients. These are filtered by a triangular band pass filter bank, which consists of 17 filters. In Mel-frequency their interval is constant. The total frequency ranges from 0 Hz to 22050 Hz.

The overall discrimination between the three classes noise, speech and singing is fulfilled layer-wisely using Multi-Layer SVMs for two-class separation on each layer as can be seen in figure 1. Multi-Layer SVMs form one of several approaches to multi-class classification applying SVMs. While no confidence measure can be provided for any class in this solution, an advantage is that different feature sets can be used on each layer. In our case this helps to respect the different requirements in discrimination of voice-like signals in general, and secondly between speech and singing. Multi-Layer SVMs have also been successfully applied in musical genre recognition [5].

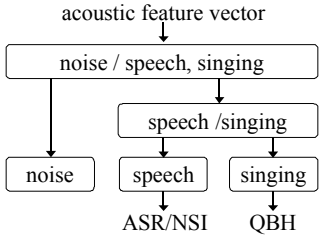

Figure 1. Multi-Layer SVM Discrimination

If the decision is made for noise no further action takes place. In the case of speech an Automatic Speech Recognition (*ASR*) and Natural Speech Interpretation (*NSI*) module can be used for further processing. In a similar manner the input can be forwarded to a Query by Humming (*QBH*) module in the case of singing.

6. Results

In five cycles using $4/5^{th}$ for training and $1/5^{th}$ of training disjoint data for the evaluation an overall performance evaluation was fulfilled. The following table shows the confusion matrices of the average performance. In the first table on the left-hand side the confusion of noise and voice is shown in respect to the confusions on the first layer of the multi-layer SVM as seen in figure 1. In the second table on the right-hand side the confusions of speech and singing can be found.

Table 2. Confusion matrices on the SVM layers

	Noise	Voice
Noise	**99.5%**	0.5%
Voice	0.1%	**99.9%**

	Speech	Singing
Speech	**99.2%**	0.8%
Singing	0.7%	**99.3%**

In the tables the recognized class is found to the right, and the actual class downwards. The overall recognition rate for the discrimination of noise and voice resembles 99.72±0.1% throughout the test-cycles. The mean performance for the discrimination of speech and singing was 99.22±0.7%.

7. Conclusion

We believe that the achieved average recognition performance of 99.2% clearly demonstrates the robustness of the presented novel feature set for the discrimination of speech and monophonic singing. The pitch-derived features showed the greatest contribution for the discrimination of speech and singing. While energy-derived features contain rhythmic information in singing, they tend to show no differences in spoken or sung acoustic signals. The highest gain in discrimination performance could be obtained by integration of information of durations of voiced sounds and silences which are not considered in the works of Gerhard [1]. By use of an energy threshold and Mel Frequency Cepstral Coefficients the signal type could be determined out of a continuous stream by means of Multi-Layer Support Vector Machines. The proposed methods have been successfully integrated into a multimodal Music Information Retrieval system [2], which is controlled among others via natural speech and singing queries.

8. References

[1] D. Gerhard: "Pitch-based acoustic feature analysis for the discrimination of speech and monophonic singing," *Journal of the Canadian Acoustical Association 30:3*, pp. 152-153, 2002.

[2] B. Schuller, et al.: "A Hybrid Music Retrieval System using Belief Networks to Integrate Queries and Contextual Knowledge," *Proc. of the ICME 2003*: *Vol. I*, pp. 57-60, Baltimore, MD, USA, 2003.

[3] B. Schuller, G. Rigoll, M. Lang: "Hidden Markov Model-Based Speech Emotion Recognition," *Proc. of the ICASSP 2003 Vol. II*, pp. 1-4, Hong Kong, China, 2003.

[4] N. Cristianini, J. Shawe-Taylor: *"An Introduction to Support Vector Machines and other kernel-based learning methods,"* Cambridge University Press, 2000.

[5] C. Xu, N. Maddage, X. Shao, F. Cao, Q. Tian: "Musical Genre Classification Using Support Vector Machines," Proc. of the ICASSP 2003 Vol. V, pp. 429-432, Hong Kong, China, 2003.