A HYBRID MUSIC RETRIEVAL SYSTEM USING BELIEF NETWORKS TO INTEGRATE MULTIMODAL QUERIES AND CONTEXTUAL KNOWLEDGE

Björn Schuller, Martin Zobl, Gerhard Rigoll, and Manfred Lang

Institute for Human-Computer Communication
Technische Universität München
D-80290 München, Germany
(schuller | rigoll | lang)@ei.tum.de

ABSTRACT

Recently an increasing interest in music retrieval can be observed. Due to the growing amount of online and offline available music and a broadening user spectrum more efficient query methods are needed. We believe that only a parallel multimodal combination of different input modalities forms the most intuitive way to access desired media for any user. In this paper we introduce a query by humming, speaking, writing, and typing. The strengths of each modality are combined in a synergetic manner by a soft decision fusion. Songs can be referenced by their according melody, artist, title or other specific information. Further more the recognition of the actual user's emotion and external contextual knowledge helps to build an expectance of the intended song at a time. This constrains the hypothesis sphere of possible songs and leads to a more robust recognition or even a suggestive query. A combination of artificial neural networks, hidden Markov models and dynamic time warping integrated in a Bayesian belief network framework build the mathematical background of the chosen hybrid architecture. We address the implementation of a working system and results achieved by the introduced methods.

1. INTRODUCTION

In this work we aim to lead to an intuitive access to large databases of music. Due to actual development in audio compression technology and increases in the capacity of storage devices users are nowadays able to keep up to 10^4 or more tracks on their hard discs at home, in the car or on portable devices. In internet databases even higher numbers of songs may be available at a time. The user spectrum broadens as new products reach persons of all ages and expertise levels in using technical devices. These trends claim for a fast and intuitive possibility to access single tracks or construct play lists for every user in any scenario [1]. In general various challenges for a music retrieval system exist: incomplete information has to be complemented as people tend to know only fragments such as the artist, title or parts of the lyrics of a song. The information is often erroneous due to the multilingual nature of international titles or e.g. confusions of text fragments with the title. This results in wrong spelling or pronunciation challenging a typing or speech based query system. Recent approaches to query by humming form an interesting new chance for an intuitive search. However many users are shy, or cannot sing or hum too good. Even if a humming query is successful, several interpretations performed by different artists may exist belonging to the same melody and ask for a further specification by another modality. This clearly shows that humming alone can not lead to a final query result in large databases where different versions may exist. Therefore we believe that error-tolerant multimodal access is needed in advanced music retrieval. In the paper we present a system that interprets and combines user searches via humming, natural language, handwriting, and typing. For further improvement in interaction the system integrates contextual knowledge such as user preferences, the date and time, and the actual user emotion. The recognition of the multimodal input queries and their fusion in a hybrid architecture will be explained in detail throughout this work.

2. BELIEF NETWORKS AS FRAMEWORK

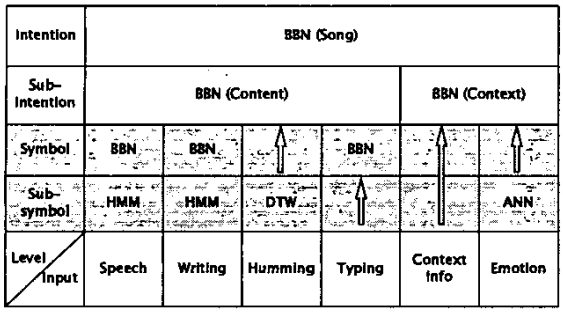
Due to the alluded requirements of handling incomplete and erroneous information [2] and integration in a hybrid architecture we decided for Bayesian belief networks (BBN) as mathematical background. We want to give just a short description of BBN and their capabilities in probabilistic reasoning. A BBN models state variables by nodes connected by directed edges that express statistical dependencies of nodes and their parents. It can be completely described in structure and conditional probabilities by its joint probability distribution. Knowledge of evidences of variables in the network allows for interfering states of variables. Since we want to allow the inheritance of probability scores provided by heterogeneous recognition instances we need to make probability based beliefs of evidences. Therefore the inference algorithm is extended in this work as described in [3].

3. SYSTEM OVERVIEW

The following figure gives an overview of the general architecture used. We divide the interpretation process into four levels: On the signal level called sub-symbol level throughout this paper, different task specific algorithms are used. The next level is the symbol level where BBNs combine these modality specific symbols to sub-intentions. On the sub-symbolic layer

continuous hidden Markov models with Gaussian mixtures (HMM), dynamic time warping (DTW), and a multi layer perceptron (MLP) are used for the recognition process. Their estimations are fused on the symbol layer by use of Bayesian belief networks (BBN) which leads to the sub-intention layer. On this layer a BBN units content and context based input.

Figure 1: Overview of the hybrid architecture

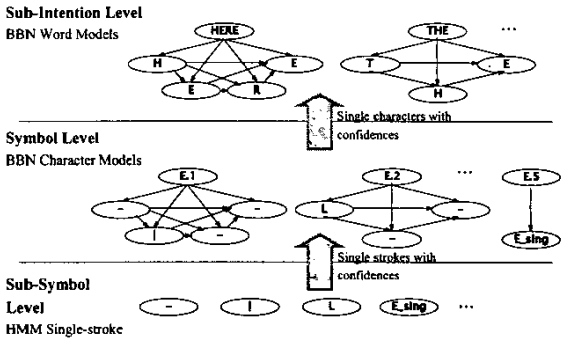


BBN	Bayesian Belief Network
HMM	Hidden Markov Model
ANN	Artificial Neural Network
DTW	Dynamic Time Warp

4. WRITTEN INPUT

As first input modality integration of written input will be described in detail as an example. Handwritten input can be entered via a touch screen. We decided to select printed letters. These are recognized by combined single strokes. The assumed characters build a word hypothesis by use of belief networks. An advantage is the capability of belief networks to cope with lacking data. The recognition process can be seen as a spotting-based approach. This allows handling of misspelling or correction of falsely recognized single strokes. The single strokes performed on the screen will be recognized on the sub-symbol level. As features the planar Cartesian x- and y-coordinates (plus δ and $\delta\delta$) are used. For invariance of the size and location they are normalized to their bounding box and the starting point of a stroke [4]. Each character is modeled by one continuous Bakis hidden Markov model [5]. The HMMs were trained using Baum Welch re-estimation with 10^5 iterations or an abruption criterion of a change in model parameters of $\varepsilon < 10^{-4}$. Up to four Gaussian mixtures and a variable optimized number of states were used. The HMM scores are fed forward to the BBN layer to avoid an early decision. A rule-based post-processing increased the recognition rate of finally 98.2% for an alphanumeric alphabet. By use of BBNs the single strokes are combined to characters on the symbol layer. Each character possesses a number of alternative stroke combinations modeled by a BBN. Each score is provided to the upper word-based BBN. On the word based sub-intention level different models basing on the underlying scores may achieve evidence. The words and their resulting scores are fed to the intention layer. The chosen architecture of the symbol-level networks models an increase in index, but allows for omissions at any index-point including the first. The believe in a symbol or sub-intention is increased by each further evidence occurring within the hypothesis.

Figure 2: Principle of BBN based spotting



5. TYPED INPUT

Typed input is processed analog to written input besides that no sub-symbolic recognition and symbolic fusion is needed. The BBNs on the sub-intention layer allow for correction of misspelling or completion of fragments.

6. SPOKEN INPUT

Speech is integrated in a natural way. A conventional HMM-based recognition engine is used on the sub-symbol level. We use hypothesis of semantic units and their scores on the sub-symbol level. A zero-gram is chosen as language model before the integration in the BBN on the symbol level. Like this only the raw single semantic units and their scores are provided. On the symbol level these units are matched to single phrases. We distinguish between facultative and obligatory phrases. These can be built by parameter words or normal words. While parameter words return a value such as a specific music genre, normal words add to the belief in a hypothesis [6]. Synonyms are clustered by the introduction of super words.

7. HUMMED INPUT

In this paper we aim to introduce the basic matching of hummed patterns by use of a Dynamic Time Warp (DTW) algorithm. A more complex preprocessing especially for matching polyphonic audio using frame length adaptation and stereophonic feature information will be described elsewhere. As features we use the spectral harmonic sum based on partial enhancement [7] calculated in the FFT of each frame. The spectrum is filtered by rectangular filters according to the distances of two musical notes in the western 12-tone scale [8]. Starting at the note D up to c''' we achieve a reduction to 47 note-bands. To enhance the melodic part in polyphonic audio we concentrate on the center panned parts where in general mostly the key melody can be found. The frame length can be adopted to the beats per minute of the audio track to increase the performance. In a databank

references for all songs in the database are stored [9]. The matching is done by use of a DTW with Euclidean distance metric and Itakura constraints. The endpoints constraints are loose to allow for beginning or ending at another note or bar within the intended song. Recognition rates reach from 46.5% matching polyphonic audio to 94.2% for the top one hit matching hummed data. Allowing the five best matches the rates increase significantly to 65.1% and 98.3% having 100 melodies in the databank. These values seem to be a reasonable trade-off between recognition rate and extra-effort for the users, which have to select between the five best titles by e.g. pointing on the touch-screen.

8. CONTEXT INTEGRATION

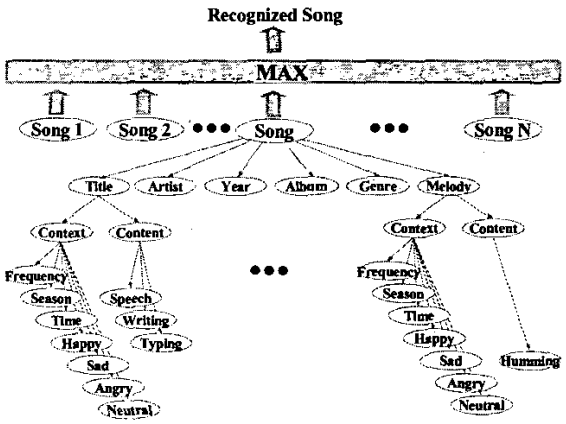
The frequency of listening to a song by the user, the actual season (winter, summer) and time (morning, afternoon, evening, night) and actual user emotion (happy, sad, angry, neutral) are chosen as contextual variables. In our opinion the selection of a musical style, or even a concrete song, can depend on these aspects. For example one might expect the wish of a ballad on a winter night if the user is sad. On the other hand a joyful user might select a Latin-dance track on a summer morning. However the system learns the typical behavior patterns of the user, if any can be observed. This can even lead to a suggestive query.

Emotion recognition is based on prosodic features in the speech signal and on the touch interaction on the touch screen [10]. In speech analysis we use a feature vector of 25 derived global features of the pitch and energy contour, spectral aspects, and durations and pauses. For the touch recognition the x-,y-, and z-axis coordinates form the basis of the calculation for the touch-energy and a transformation into cylinder coordinates. In cylinder coordinates the phase information alludes e.g. tremors. The touch screen used is based on absorption of acoustic surface waves. The z-coordinate can be measured by the degree of wave absorption. The features are freed of their means and normalized to their standard deviation. They are classified after each user utterance or touch interaction by an artificial neural network. We chose a multi layer perceptron with 10^3 hidden neurons in one hidden layer. The model was trained using back propagation at 10^3 iterations using a gain of $5*10^{-5}$. The system has been trained and tested with a corpus of 3000 acted emotion samples and touch-patterns. The achieved recognition rate of 85.2% in average resembles the rate of human classifiers that had to classify the same data. However recognition seems to be quiet user dependent. A loss of 20% in performance can be observed, if a training disjunctive person uses the system.

9. MULTIMODAL SOFT DECISION FUSION

Each song is modeled in a BBN. In the recognition phase the maximum likelihood model is selected as the recognized song. We chose a singly connected net architecture without parallel paths on the intention layer. Singly connected nets can be calculated very fast, while BBNs in general tend to demand enormous computation effort. The primary attributes of a song are its title, artist, year of publication, album, genre and melody in accordance to the ID3 tags of the MP3-audio compression standard. Only the attribute melody has been added. They allow for a fast play-list creation by demanding music of a specific year, genre, etc. Each attribute consists of a content and a context related variable. Each evidence in a content related attribute raises the belief in the song. On the other hand the actual context provides an a-priori belief in a song as stated earlier. The fact that the content state depends on every modality allows parallel and complementary use of the modalities. For example a user can write a title. After a first query result showing several interprets she can select by speech. On the other hand more conventional modalities less prone to errors as typing can be used as fall-back solution. If e.g. a query by humming fails, the user can write down information related to the song. The next figure 3 provides a graphical overview of the principle of the intention layer BBNs.

Figure 3: Song models and maximum likelihood decision



In each final node the connection to the sub-intention layer BBNs takes place.

10. RESULTS

In the following the results obtained with a working implementation of the suggested methods are introduced and explained in detail. The results are not related with a complete successful search. They rather refer to the sub-intention based performance.

Figure 4: Table of recognition results

%	Top 1	Top 2	Top 3	Top 5	Top 10
Speech	93.1	94.2	96.1	98.4	98.9
Humming	94.2	95.0	97.1	98.3	100.0
Writing	96.3	98.2	99.2	99.8	100.0
Typing	98.2	98.8	99.8	100.0	100.0

However, these results strongly depend on the knowledge of the attributes of the desired song. Therefore we regard a user study as more significant to evaluate the idea of multimodal music retrieval. A prototype has been realized and designed especially for the automotive environment. In a study ten users who did not previously know the system had to search 40 specific titles. In a next step they were allowed to choose 10 titles of the actual charts. The users rated the system as very comfortable and made

broad use of the ability to select by different modalities. The users were of different ages between 25 and 68 years with an average of 38.2 years. They possessed different technical background and experience in music retrieval e.g. in the net. The total success rate was 100% meaning that each user succeeded in the selection of each desired song. In the following table the relative amount of selection via a specific modality is shown. However, these numbers are strongly dependent on the users and the songs. A user of course cannot hum an unknown song. Anyways the numbers clearly show that several modalities were used. The fact that typing was not used too much bases in our opinion on the high performance of the hand-writing recognition.

Figure 5: Table of modality preferences

Speech	Humming	Writing	Typing
25.8%	48.1%	23.2%	2.9%

The next table shows the preference of attributes chosen for the selection. The attribute melody directly corresponds to the modality humming since it can be entered only by this modality. The fact that users did select songs by their titles with a surprisingly low occurrence shows that melody-based access seems more intuitive. However, this again depends on the test setup.

Figure 6: Table of distribution of the attributes

Title	Artist	Year	Album	Genre	Melody
10.2%	34.6%	3.4%	0.5%	3.2%	48.1%

The average number of modalities used for the selection of one title was 2.27. This may change if a user uses the system more often. However, the system provides the chance for each user to choose a favorite modality for personal access. The contextual knowledge integration needs longtime testing to allow the system to learn and profile the user and his habits. It has therefore yet been tested only with three probands. The functionality could be verified and the choice of the contextual variables seems reasonable. More testing will be needed to verify the results.

11. DISCUSSION

In this paper we presented a novel approach to multimodal music retrieval. Users were provided the ability to select titles by humming, typing, naturally speaking or writing or selecting on a touch screen. The chance to use different modalities found broad acceptance among ten probands. The test subjects used more than one modality throughout our tests. Humming proved as the first choice modality, but all other modalities helped as well to access a desired song. The use of Bayesian belief networks for the multimodal integration allowed for a robust parallel and complementary analysis of different user inputs for different attributes. As contextual variables the emotion of a user leads to a suggestive query providing "the right song at the right time". In our future research we aim to combine emotion recognition and music retrieval to achieve an estimation of the transmitted emotion within a song. Like this the system can itself find convenient tracks for an actual user emotion.

12. ACKNOWLEDGEMENTS

The work presented in this paper has been supported by the FERMUS project, a cooperation of BMW Group, DaimlerChrysler, SiemensVDO and the institute of Human-Machine Communication at the Munich University of Technology. The project stands for error-robust multimodal speech dialogues. The contents discussed in this paper largely benefits from the collaboration with the student assistants Ting-Yap Tong, Elmar Sommer, and Florian Hörger.

13. REFERENCES

[1] J. Reiss, M. Sandler: *"Benchmarking Music Information Retrieval Systems,"* presented at the JCDL Workshop on the Creation of Standardized Test Collections, Tasks, and Metrics for Music Information Retrieval (MIR) and Music Digital Library (MDL) Evaluation, Portland, Oregon, July 18th, 2002.

[2] F. V. Jensen, *"An Introduction to Bayesian Networks,"* UCL Press, 1996

[3] M. Hofmann, M. Lang: *"Intention-based Probabilistic Phrase Spotting for Speech Understanding,"* Proc. of the Int. Symp. on Intelligent Multimedia, Video and Speech Processing, ISIMP 2001, Hong Kong, China, 2.5-4.5.2001. Ed.: IEEE Hong Kong Chapter of Signal Processing, S. 99-102.

[4] B. Schuller, M. Lang: *"Integrative rapid-prototyping for multimodal user interfaces,"* USEWARE 2002, Darmstadt, 11.-12.06.2002. Düsseldorf, VDI-Report No. 1678, pp. 279-284.

[5] L. Rabiner: *"A tutorial on Hidden Markov Models and selected applications in Speech Recognition,"* Proceedings IEEE, pp. 257-284, February 1989.

[6] B. Schuller, M. Lang, G. Rigoll: *"Automatic Emotion Recognition by the Speech Signal,"* SCI 2002, 6th World Multiconference on Systemics, Cybernetics and Informatics, Orlando, Florida, USA, Proceedings Vol. IX, "Image, Acoustic, Speech and Signal Processing II", IIIS, pp. 367-372.

[7] J. Song, S. Bae, K. Yoon: *"Query by humming: Matching humming query to polyphonic audio,"* ICME 2002, 3rd International Conference on Multimedia and Expo, IEEE, Lausanne, Switzerland, CD-Rom Proceedings.

[8] H. Nagano, K. Kashino, H. Murase: *"Fast Music Retrieval using polyphonic binary feature vectors,"* ICME 2002, 3rd International Conference on Multimedia and Expo, IEEE, Lausanne, Switzerland, CD-Rom Proceedings.

[9] J. Pickens: *"A Survey of Feature Selection Techniques for Music Information Retrieval,"* Technical report, Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, MA, 2001.

[10] B. Schuller, M. Lang, G. Rigoll: *"Multimodal Emotion Recognition in Audiovisual Communication,"* ICME 2002, 3rd International Conference on Multimedia and Expo, IEEE, Lausanne, Switzerland, CD-Rom Proceedings.