ABSTRACT
We present a framework for selecting best audio features for audio-visual analysis and synthesis of dance figures. Dance figures are performed synchronously with the musical rhythm. They can be analyzed through the audio spectra using spectral and rhythmic musical features. In the proposed audio feature evaluation system, dance figures are manually labeled over the video stream. The music segments, which correspond to labeled dance figures, are used to train hidden Markov model (HMM) structures to learn spectral audio patterns for the dance figure melodies. The melody recognition performance of the HMM models for various spectral feature sets are evaluated. Audio features, which are maximizing dance figure melody recognition performances, are selected as the best audio features for the analyzed audio-visual dance recordings. In our evaluations, mel-scale cepstral coefficients (MFCC) with their first and second derivatives, spectral centroid, spectral flux and spectral roll-off are used as candidate audio features. Selection of the best audio features can be used towards analysis and synthesis of audio-driven body animation.

1. INTRODUCTION
Installing human perception into a robot using audio-visual observations is one of the exciting problems in the humanoid robot studies [1]. Audio-visual locomotion (walking, running) using hidden Markov models is studied in [3]. Modeling human locomotion (walking, running) using hidden Markov models is studied in [3]. The resulting HMM is observed to be insufficient to recognize more complex motion such as dancing. In a recent study, we analyzed dance figures of a dancing person over an audio-visual database [2]. The 3D positions of body joints are extracted by tracking markers on the human body. HMM structures are used to model motion dynamics. Feature level audio-visual correlation analysis is performed using mel-scale cepstral coefficients (MFCC) for audio and 3D positions of body joints. Although we obtain encouraging results for basic dance sequences, an investigation and evaluation of candidate spectral audio features are needed to analyze complex dance figures. In [4], dance motion is detected through musical analysis under the assumption that dance motion primitives must be synchronized to the musical rhythm. Dance motion primitives and the melody, musical rhythm structure, is extracted. The experimental results confirmed that the primitive motions are in accordance to the musical rhythm for the given dance sequence. Estimation of dancing skills based on the same rhythmic factors is performed in [5]. In particular cases, rhythm information does not help to synthesize dance figures by itself. We need extra audio features to detect the changes of harmony of the audio to best describe different dance figures.
The short-time Fourier transform, defined as the center of gravity of the magnitude spectrum of the sound signal, is a fundamental tool in signal processing. This transform captures the dynamic spectral features of audio signals, which can be denoted as \( \tilde{F} \) on the magnitude spectrum of the short-time Fourier transform. These features are considered as dynamic spectral features and denoted as \( F \).

The first and second derivatives of the MFCC features are \( F_1 \) and \( F_2 \). We use 13 cepstral coefficients including the energy coefficient, \( F_m \). The first and second derivatives of the MFCC features are considered as dynamic spectral features, and denoted as \( F_{AM} \) and \( F_{AA} \).

Spectral centroid, flux and roll-off can be defined based on the magnitude spectrum of the short-time Fourier transform, which is denoted as \( U_{mj} \) for the \( j \)-th frequency bin at the \( m \)-th audio window. The spectral centroid, \( F_C \), is defined as the center of gravity of the magnitude spectrum of the short-time Fourier transform,

\[
F_C = \frac{\sum_j j \tilde{U}_{mj}}{\sum_j \tilde{U}_{mj}} \tag{1}
\]

The spectral flux, \( F_F \), is defined as the squared difference between the normalized magnitudes of successive spectral distributions to capture the local spectral change,

\[
F_F = \sum_j (\tilde{U}_{mj} - \tilde{U}_{(m-1)j})^2 \tag{2}
\]

The spectral roll-off is defined as the frequency \( F_R \) below which 85% of the spectral magnitude distribution is concentrated. It captures spectral shape information.

Various combinations of these features are used to train dance figures describing HMM structures and evaluated with respect to recognition performance.

### 2.1 Audio Feature Extraction

The act of dancing is the natural response of the body to the musical tempo and beat sequence. According to these responses, the movements of the body are shaped and dance figures are generated. Temporal features such as beat rate have been used to synthesize dance figures by [7].

We analyzed spectral audio features that would be helpful when used with temporal features that are proposed in [8, 9]. In our evaluations, mel-scale cepstral coefficients (MFCC) with their first and second derivatives, spectral centroid, spectral flux and spectral roll-off are used as candidate audio features. Short-time analysis of overlapping audio segments of size 25ms are performed for the extraction of audio features for each 10ms frame. Hamming window of size 23ms is applied to the analysis audio segment to remove edge effects. The MFCC feature vector is constructed with 13 cepstral coefficients including the energy coefficient, \( F_M \).

The first and second derivatives of the MFCC features are considered as dynamic spectral features, and denoted as \( F_{AM} \) and \( F_{AA} \).

Spectral centroid, flux and roll-off can be defined based on the magnitude spectrum of the short-time Fourier transform, which is denoted as \( U_{mj} \) for the \( j \)-th frequency bin at the \( m \)-th audio window. The spectral centroid, \( F_C \), is defined as the center of gravity of the magnitude spectrum of the short-time Fourier transform,

\[
F_C = \frac{\sum_j j \tilde{U}_{mj}}{\sum_j \tilde{U}_{mj}} \tag{1}
\]

The spectral flux, \( F_F \), is defined as the squared difference between the normalized magnitudes of successive spectral distributions to capture the local spectral change,

\[
F_F = \sum_j (\tilde{U}_{mj} - \tilde{U}_{(m-1)j})^2 \tag{2}
\]

The spectral roll-off is defined as the frequency \( F_R \) below which 85% of the spectral magnitude distribution is concentrated. It captures spectral shape information.

Various combinations of these features are used to train dance figure describing HMM structures and evaluated with respect to recognition performance.

### 2.2 Evaluation of Audio Features

We assume, spectrum and energy flux of audio tracks contain repeated patterns; these repeated patterns are correlated with separate dance figures in a dance sequence and spectral features can be used to extract these audio patterns. Our purpose is to select audio features that best describe temporal patterns associated with dance figures.

We analyze four candidate audio feature sets that are combinations of extracted spectral features, i.e., \( F_C \), \( F_S \), \( F_R \). We applied HMMs, \( \lambda_{nm} \), to model these candidate features regarding of manually labeled dance figures recurring in the dance performance. The \( \lambda_{nm} \)'s are trained supervised for each dance figure, where \( n \) and \( m \) are figure and feature indices respectively. We perform a left-to-right HMM structure to model each dance figure conceding the assumption that each dance figure is composed of a sequence of well-defined motion primitive.

Furthermore, we extract figures for given audio and \( \lambda_{nm} \) by using Viterbi decoding as a result of supervised segmentation. These figures are compared with the corresponding visually labeled dance figure boundaries. The comparison results are calculated according to the number of correct labels, \( D_{nm} \), the number of deletions, \( D_{nm} \), the number of insertions, \( D_{nm} \), out of the total number of labels, \( T \), in the defining transcription file. These parameters are function of figures, candidate feature sets and state number that are notated as \( n \), \( m \) and \( s \) respectively.

Each \( \lambda_{nm} \) model has one important parameter to set which is the number of states, \( S \). In order to determine the optimal number of states for each of the \( \lambda_{nm} \), we train each \( \lambda_{nm} \)
with different number of states and apply multiple objective optimization (MOO) method [10], which defines a so-called best compromise solution that is closest to the defined utopia point. We have three optimization criteria, that are recognition correctness, $RC_{n,m,s}$, recognition accuracy, $RA_{n,m,s}$, and recognition missed-matches, $RM_{n,m,s}$, that are calculated in percentages as given in the equations below.

$$RC_{n,m,s} = \frac{H_{n,m,s}}{T} \times 100$$ (3)

$$RA_{n,m,s} = \frac{H_{n,m,s} - I_{n,m,s}}{T} \times 100$$ (4)

$$RM_{n,m,s} = \frac{I_{n,m,s} + S_{n,m,s} + D_{n,m,s}}{T} \times 100$$ (5)

In our scenario, our aim is to find the solution point that is closest to the infeasible utopia point, (1,1,0). The utopia point is the one where all optimization criteria are satisfied, in other words, where best matched audio features would be recognized correctly and accurately with no substitutions, deletions or insertions. To find the closest feasible point, we use Euclidian distance, and calculate the distance for each different audio feature set and we select the set of audio features with minimum distance as the MOO solution of our system.

We set the state number of each $\lambda_{n,m}$, to the state number minimizing the euclidian distance in MOO which is calculated for different number of states. Hence number of states for each $\lambda_{n,m}$ is set to a different value according to each training feature combination. Then each $\lambda_{n,m}$, which is fixed to optimum state number, with corresponding candidate feature set is evaluated by employing MOO method regarding three optimization criteria. Euclidian distance from the utopia point (1,1,0) is calculated for each feature set’s recognition correctness, accuracy and missed-matched values. Audio features with smallest Euclidian distance is selected where this feature set best describes temporal patterns associated with dance figures.

3. RESULTS

Our training dataset includes multiview video recordings of a dance performance for the traditional dance, zeybekiko, with a duration of approximately 5 minutes. The signals are analyzed over a 23 ms Hamming window at every 10 ms. Each video recording consists of one single dance figure repeated successively during the whole performance.

For audio analysis, we manually label the start and end frames of each dance figure in the video file. The label file is used for supervised temporal training of $\lambda_{n,m}$’s for each different dance figure and candidate feature set. Audio file is divided into four equal pieces and each piece of audio is used for testing the other three pieces which served as the training data for the HMMs. After training and testing different combinations of these four audio pieces, the audio is completely segmented for each different feature combination. The four candidate feature sets are $F_S$, $F_C$, $F_F$ and $F_R$.

In order to determine the optimal number of states for each of the $\lambda_{n,m}$, each $\lambda_{n,m}$ is trained with different number of states from 2 to 40. By computing the euclidian distance to the utopia point of each of the model match for each state number, we examine the progression of the learning process and the accuracy of the trained model. The evolution of this parameter for the four of total HMM structures is displayed in Fig. 3. $S$ is set to the value where Euclidian distance is minimized in MOO solution for each different feature combination. These values are $S = 11$ for $F_S$, $S = 20$ for $F_C$, $S = 22$ for $F_F$ and $S = 24$ for $F_R$.

As a result of supervised segmentation we investigated the correlation between dance figures and audio patterns. We defined the correlation with euclidian distance measures that is calculated by using three optimization criteria for each $\lambda_{n,m}$, with optimum number of state, of each feature set. The calculated euclidian distance measures, recognition correctness, accuracy and missed-matched values are given in Table 1. The best correlated feature set is determined as $F_C$ according to our algorithm.

<table>
<thead>
<tr>
<th>$Euc.\text{Dist.}$</th>
<th>$RC$</th>
<th>$RA$</th>
<th>$RM$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_S$</td>
<td>0.688</td>
<td>1</td>
<td>0.2252</td>
</tr>
<tr>
<td>$F_{SC}$</td>
<td>0.5021</td>
<td>0.9099</td>
<td>0.5676</td>
</tr>
<tr>
<td>$F_{SF}$</td>
<td>0.7007</td>
<td>0.991</td>
<td>0.4775</td>
</tr>
<tr>
<td>$F_{MR}$</td>
<td>0.688</td>
<td>1</td>
<td>0.4505</td>
</tr>
</tbody>
</table>

Figure 3: Results for selection of $S$ for each feature combination: (a) $F_S$, (b) $F_C$, (c) $F_F$, (d) $F_R$. Table 1: Calculated Euclidian distance measures, recognition correctness, accuracy and missed-matches for each feature set.
4. CONCLUSIONS AND FUTURE WORK

Evaluation of audio features serve as a basis for video mark up and realistic avatar synthesis applications. In this paper, we investigate correlation of spectral features with dance figures that would improve automatic dancing avatar synthesis and video mark up performances.

Calculated distance measures indicates that combination of MFCCs and spectral centroid yields small distance measure. These extracted temporal patterns of audio is useful for synthetic agents and/or robots to learn dance figures from audio and able the humanoid perceiving in a way of humans. For the future work, the variety of features would be increased and temporal features would be added to these analysis, which will provide more robustness in both analysis and synthesis of dance figures and the results would be tested on different dance sequences.

5. ACKNOWLEDGEMENTS

This work has been supported by the European FP6 Network of Excellence SIMILAR (http://www.similar.cc) and by TUBITAK under project EEEAG-106E201 and COST2102 action.

REFERENCES


