PROSODY-DRIVEN HEAD-GESTURE ANIMATION

*M.E. Sargın, E. Erzin, Y. Yemez, A.M. Tekalp**

Multimedia, Vision and Graphics Laboratory
Koç University, Istanbul, Turkey
{msargin,eerzin,yyemez,mtekalp}@ku.edu.tr

A.T. Erdem, C. Erdem, M. Özkan†

R&D, Momentum A.S.
TUBITAK TEKSEB A-205, Gebze, Turkey
terdem@momentum-dmt.com

ABSTRACT

We present a new framework for joint analysis of head gesture and speech prosody patterns of a speaker towards automatic realistic synthesis of head gestures from speech prosody. The proposed two-stage analysis aims to "learn" both elementary prosody and head gesture patterns for a particular speaker, as well as the correlations between these head gesture and prosody patterns from a training video sequence. The resulting audio-visual mapping model is then employed to synthesize natural head gestures from arbitrary input test speech given a head model for the speaker. Objective and subjective evaluations indicate that the proposed synthesis by analysis scheme provides natural looking head gestures for the speaker with any input test speech.

Index Terms— Man-machine systems, multimedia systems, gesture and prosody analysis, gesture synthesis

1. INTRODUCTION

State of the art visual speaker animation methods are capable of generating synchronized lip movements automatically from speech content; however, they lack automatic synthesis of speaker gestures from speech. Head and face gestures are usually added manually by artists, which is costly and often look unrealistic. Hence, learning the correlation between gesture and speech patterns of a speaker towards automatic realistic synthesis of speaker gestures from speech remains as a challenging research problem.

There exists significant literature on speaker lip animation, that is, rendering lip movements synchronized with the speech signal [1, 2]. Despite exhibiting variations from person to person and in time, head and body gestures are also correlated with speech. For example, it has been observed that manual gestures are correlated with prosody [3] and verbal content of the speech [4], whereas head gestures are mostly correlated with the prosody [5, 3]. In [6], we presented a preliminary demonstration of natural looking head and arm gesture synthesis from speech using a manually determined audio-visual mapping from speech to head and arm motions.

In this paper, we present a framework for joint analysis of head gesture and speech prosody patterns towards automatic generation of the audio-visual mapping from speech prosody to head gestures. There are some open challenges involved in the joint analysis of head gestures and prosody towards prosody-driven head gesture synthesis: First, there does not exist a well-established set of elementary prosody and gesture patterns for gesture synthesis, unlike phonemes and visemes in speech articulation. Second, prosody and gesture patterns are speaker dependent, and may exhibit variations in time even for the same speaker. Third, synchronicity of gesture and prosody patterns may exhibit variations. We address these challenges by first processing the head gesture and prosody features separately by a parallel HMM structure to learn and model the gestural and prosodic elements (elementary patterns), respectively, over training data for a particular speaker. We then employ a multi-stream parallel HMM structure to find the jointly recurring gesture-prosody patterns and the corresponding audio-to-visual mapping.

*This work has been supported by the European FP6 Network of Excellence SIMILAR and by the European FP6 Network of Excellence 3DTV-511568.

†This work has been supported by TUBITAK under project TIDEB-3050135 and by the European FP6 Network of Excellence 3DTV-511568.

2. OVERVIEW OF THE PROPOSED SYSTEM AND FEATURE EXTRACTION

A block diagram of the proposed system for prosody-driven head gesture animation, which consists of analysis and synthesis parts, is depicted in Fig. 1. The analysis part includes two feature extraction modules and two-stages of analysis. Feature extraction modules compute the head gesture features $\boldsymbol{f}^g$ and speech prosody features $\boldsymbol{f}^p$, respectively, from training stereo video sequences of a speaker. At the first stage analysis, individual feature streams are used to train separate parallel HMM structures, which provide probabilistic models for temporal recurrent patterns in the corresponding modalities, respectively. The segments corresponding to these patterns are detected and labeled over the training video streams, where pattern labels for prosody and gesture are denoted by $\boldsymbol{l}^p$ and $\boldsymbol{l}^g$, respectively. At the second stage, the labels of temporally segmented gesture and prosody streams are used together to train a discrete multi-stream parallel HMM to identify jointly recurring patterns. The resulting joint HMM structure models the correlation between speech prosody and head gestures. The synthesis part makes use of the joint HMM to predict the gesture labels from the prosody labels computed for a test input speech using the prosody HMM obtained by the first stage analysis. The corresponding gesture features, i.e., head motion parameters, are synthesized using the gesture HMM obtained at the first stage analysis and finally animated on a 3D head model.

2.1. Extraction of Head Gesture Features

We define the head gesture feature vector, $\boldsymbol{f}_k^g$, for frame k to include the Euler angles associated with the 3D head rotation and their first differences,

$$\boldsymbol{f}_k^g = [\theta_k, \phi_k, \psi_k, \Delta\theta_k, \Delta\phi_k, \Delta\psi_k]^T \quad (1)$$

where θ_k, ϕ_k and ψ_k are the Euler angles of rotation, with respect to a reference frame k_r, around the x, y and z axes, respectively, and $\Delta\theta_k$, $\Delta\phi_k$, $\Delta\psi_k$ denote their respective first differences. The

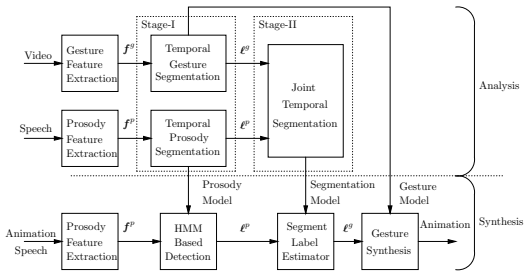

Fig. 1. Overview of the proposed synthesis-by-analysis system.

reference frame k_r can be selected as the first frame in which the subject's head is assumed to be at neutral position.

The set-up and algorithm for extraction of the feature vectors at each frame can be summarized as follows: We use a rectified stereo camera system with two identical cameras, and assume that the intrinsic camera parameters are known *a priori*. We first locate an ellipse for the head region in the reference frame. For each frame k, the 2D optical flow vectors are computed within the head region with respect to the reference frame k_r. Then, the 3D world coordinates of the 2D image points within the head region are calculated using disparity estimation and triangulation. Next, the rigid rotation and translation matrices are computed based on the resulting 3D point correspondences. Finally the Euler angles are extracted from the rotation matrix. We also employ a Kalman filter for post-smoothing of the estimated Euler angles.

2.2. Extraction of Prosody Features

The prosodic speech events can be described by the temporal variations of loudness/intensity and pitch as well as pauses between phrases, phoneme durations, timing, and rhythm. Among these, the most expressive one is the pitch, which is the rate of vocal-fold cycling. In this study, pitch frequency, P, and speech intensity, I, are considered as prosody features.

The pitch contour is extracted at a rate of 100 Hz from the speech signal using the autocorrelation method. The mean of the pitch contour is removed over the active utterances to emphasize local variations and later it is low pass filtered to reduce discontinuities. The regions between utterances without a valid pitch are filled with zero mean unit variance Gaussian noise. The intensity features are also extracted over the active utterances. The squared sound intensities are weighted with a 32 ms Kaiser-20 window, and the speech signal intensity is calculated as the sum of these weighted samples. The 32 ms window is shifted by 10 ms for each frame to extract intensity values at 100 Hz frame rate. The intensity features are also mean removed over active utterances and between-utterance regions are filled with zero mean unit variance Gaussian noise. Finally, the pitch frequency, its derivative and the intensity are concatenated to form the 3 dimensional prosody feature vector $\boldsymbol{f}_k^p$ at frame k:

$$\boldsymbol{f}_k^p = [P_k \ \Delta P_k \ I_k]^T \tag{2}$$

3. HEAD GESTURE-PROSODY PATTERN ANALYSIS

We propose a two stage HMM-based unsupervised analysis framework, where the first stage aims to separately extract elementary gesture and prosody patterns for a speaker, and the second stage determines a correlation model between these head gesture and prosody patterns.

3.1. Stage-I: Extraction of Elementary Head Gesture and Prosody Patterns

The first stage analysis defines recurrent elementary head gesture and prosody patterns separately using unsupervised temporal clustering over individual feature streams. The gesture and prosody feature streams $\boldsymbol{F}^g$ and $\boldsymbol{F}^p$ are separately used to train two HMM structures Λ_g and Λ_p, which capture recurrent head gesture segments $\boldsymbol{\varepsilon}^g$ and prosody segments $\boldsymbol{\varepsilon}^p$. For ease of notation, we use a generic notation to represent the HMM structure which is identical for the gesture and prosody streams. The HMM structure Λ, which is used for unsupervised temporal segmentation, has M parallel branches and N states as shown in Fig. 2. The states labeled as s_s and s_e are non emitting start and end states of the parallel HMM structure. The parallel HMM Λ is composed of M parallel left-to-right HMMs, $\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$, where each λ_m is composed of N states, $\{s_{m,1}, s_{m,2}, \ldots, s_{m,N}\}$. The feature stream is a sequence of feature vectors, $\boldsymbol{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_T\}$, where $\boldsymbol{f}_t$ denotes the feature vector at frame t. Unsupervised temporal segmentation using HMM model Λ yields L number of segments $\boldsymbol{\varepsilon} = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L\}$. The l-th temporal segment is associated with the following sequence of feature vectors,

$$\varepsilon_l = \{\boldsymbol{f}_{t_l}, \boldsymbol{f}_{t_l+1}, \ldots, \boldsymbol{f}_{t_{l+1}-1}\} \quad l = 1, 2, \ldots, L \tag{3}$$

where $\boldsymbol{f}_{t_1}$ is the first feature vector $\boldsymbol{f}_1$ and $\boldsymbol{f}_{t_{L+1}-1}$ is the last feature vector $\boldsymbol{f}_T$.

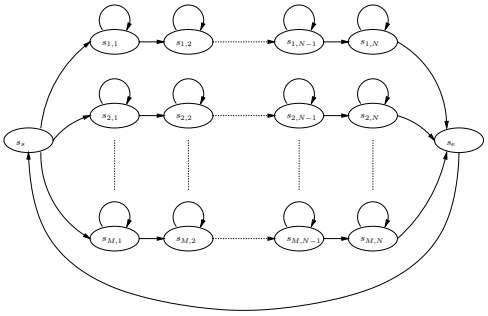

Fig. 2. Parallel HMM structure

The segmentation of the feature stream is performed using Viterbi decoding to maximize the probability of model match, which is the probability of feature sequence given the trained parallel HMM,

$$\mathrm{P}(\boldsymbol{F}|\Lambda) \quad = \quad \max_{\varepsilon_l, m_l} \prod_{l=1}^{L} \mathrm{P}(\varepsilon_l|\lambda_{m_l}) \tag{4}$$

where ε_l is the l-th temporal segment, which is modeled by the m_l-th branch of the parallel HMM Λ. One can show that λ_{m_l} is the best match for the feature sequence ε_l, that is,

$$m_l = \underset{m}{\operatorname{argmax}} \, \mathrm{P}(\varepsilon_l|\lambda_m) \tag{5}$$

Since, the temporal segment ε_l from frame t_l to $(t_{l+1} - 1)$ is associated with segment label m_l, we define the sequence of frame labels based on this association as,

$$\ell_t = m_l \qquad \text{for } \ t = t_l, t_l + 1, \ldots, t_{l+1} - 1 \tag{6}$$

where ℓ_t is the label of the t-th frame and we have a label sequence $\boldsymbol{\ell} = \{\ell_1, \ell_2, \ldots, \ell_T\}$ corresponding to the feature sequence $\boldsymbol{F}$. The

first stage analysis extracts the frame label sequences $\boldsymbol{\ell}^g$ and $\boldsymbol{\ell}^p$ given the head gesture and prosody feature streams $\boldsymbol{F}^g$ and $\boldsymbol{F}^p$. While mapping the gesture and prosody features to discrete frame labels, the mismatch between the frame rates of gesture and prosody is eliminated by downsampling the frame rate of prosody label stream to the rate of gesture label stream.

3.2. Stage-II: Joint Modeling of Prosody-Gesture Patterns

In the second stage, unsupervised segmentation of the joint gesture-prosody label stream is performed to detect recurrent joint label patterns. Note that, this task is similar to the task of stage I, except in the second stage we have a multi-stream discrete observation sequence. For this task, the parallel HMM structure in Fig. 2 is used with discrete multi-stream HMM branches. In multi-stream HMMs, all streams share the same state transition structure however emission probabilities are determined independently for each stream.

The joint gesture-prosody frame label stream, denoted by $\boldsymbol{\ell}^{gp}$, is defined such that for every frame k, $\ell_k^{gp} = [\ell_k^g,\ \ell_k^p]^T$. We represent the discrete multi-stream parallel HMM structure with Γ_{gp} and its m-th branch with γ_m^{gp}. The discrete HMM Γ_{gp} is trained over the joint gesture-prosody frame label stream. Each branch γ_m^{gp}, associated with a joint gesture-prosody temporal label pattern, is then described with a state transition matrix $\boldsymbol{A}_{\gamma_m^{gp}}$, a discrete observation probability distribution $\boldsymbol{B}_{\gamma_m^{gp}}$ and an initial state probability matrix $\boldsymbol{\Pi}_{\gamma_m^{gp}}$,

$$\gamma_m^{gp} = (\boldsymbol{A}_{\gamma_m^{gp}}, \boldsymbol{B}_{\gamma_m^{gp}}, \boldsymbol{\Pi}_{\gamma_m^{gp}}) \tag{7}$$

The discrete observation probability distribution $\boldsymbol{B}_{\gamma_m^{gp}}$ defines the probability of observing a gesture-prosody frame label at state s and frame k,

$$P(\ell_k^{gp}|s) = P(\ell_k^g|s)^{\kappa_g} P(\ell_k^p|s)^{\kappa_p} \tag{8}$$

where the exponents, κ_g and κ_p, are the stream weights and they are selected equal to each other as 1.

4. PROSODY-DRIVEN GESTURE SYNTHESIS

The proposed prosody-driven gesture synthesis system takes speech as input and produces a sequence of head gesture features, i.e., Euler angle vectors, which are naturally correlated with the input speech. The detailed flow of the synthesis is described in the following.

i. The prosody features, $\boldsymbol{F}^p$, are extracted from the input speech signal.

ii. Temporal segmentation of prosody feature sequence $\boldsymbol{F}^p$ is performed using the HMM model Λ_p, which is trained in the stage I analysis to extract the temporal prosody segment sequence, $\boldsymbol{\varepsilon}^p$, and the sequence of prosody frame labels, $\boldsymbol{\ell}^p$.

iii. The aim of this step is to predict the sequence of gesture frame labels, $\boldsymbol{\ell}^g$, given the prosody frame labels $\boldsymbol{\ell}^p$. To this effect, temporal segmentation of the prosody frame labels, $\boldsymbol{\ell}^p$ is performed using the HMM model Γ_p, which is extracted by splitting the jointly trained gesture-prosody HMM model Γ_{gp}. As a result of this temporal prosody label segmentation, a state sequence $\boldsymbol{s}^p = \{s_1^p, s_2^p, \ldots, s_K^p\}$ associated with $\boldsymbol{\ell}^p = \{\ell_1^p, \ell_2^p, \ldots, \ell_K^p\}$ is extracted. Then, the gesture frame label sequence $\boldsymbol{\ell}^g$ is predicted by maximizing the probability of observing gesture label on the state sequence path $\boldsymbol{s}^p$ over the gesture HMM model Γ_g.

iv. This step computes the gesture segment sequence $\boldsymbol{\varepsilon}^g$, consisting of the Euler angle features, given the gesture frame label sequence $\boldsymbol{\ell}^g$. First, we find the segment frame boundaries, $\{t_l\}_{l=1}^L$, by merging the same gesture frame labels in the sequence $\boldsymbol{\ell}^g$. Then, the Euler angle features for the l-th segment, $\varepsilon_l^g = \{\boldsymbol{f}_{t_l}^g, \boldsymbol{f}_{t_l+1}^g, \ldots, \boldsymbol{f}_{t_{l+1}-1}^g\}$, are generated from the HMM $\lambda_{\ell_{t_l}}^g$, which is the ℓ_{t_l}-th branch of the parallel HMM model Λ_g (computed in stage I).

v. As the final step of the gesture synthesis, the Euler angles are smoothed using median filtering followed by a Gaussian low pass filter to remove motion jerkiness. The median filtering is performed over 11 visual frames and the Gaussian smoothing is performed over 15 visual frames.

There are two main advantages of using HMMs for gesture synthesis. The first is the random variations in the synthesized gesture patterns for each segment. This variation yields more natural looking synthesis results than using a fixed gesture dictionary, since humans produce slightly varying gestures at different occasions for the same semantics. The second advantage is generating gestures with varying durations in accordance with prosody of the speaker.

5. EVALUATION AND RESULTS

We have conducted experiments using the MVGL-MASAL gesture-speech database. The database includes four recordings of a single subject telling stories in Turkish. Each story is approximately 7 minutes long and the total duration of the database is 27 min and 45 seconds. The audio-visual data is synchronously captured from the stereo camera and sound card. The stereo video includes only upper body gestures with 30 frames per second whereas the audio is recorded with 16 kHz sampling rate and 16 bits per sample. The database is partitioned into two parts such that three stories are used for training of the models and one story is used for testing. For objective evaluation of the synthesis, the Euler angles extracted from the test sequence are considered as the ground truth for the synthesized head motion.

The head gesture and prosody correlation analysis includes unsupervised temporal segmentation of the individual feature streams as well as the joint gesture-prosody label stream. The parallel HMM Λ_g is trained with features extracted from the training video using Expectation-Maximization (EM) algorithm. The resulting HMM structure provides a probabilistic cluster model for unsupervised segmentation of head gestures into recurring elementary patterns. We select the number of states in each branch of the head gesture HMM Λ_g as $N_{\Lambda_g} = 10$, corresponding to the minimum gesture pattern duration of 10 frames ($\frac{1}{3}$ sec assuming 30 video frames/sec). We set the number of gesture patterns M_{Λ_g} to 5.

The speech prosody feature sequence is extracted from the audio part of the training database. As defined in stage I, the HMM model Λ_p is trained with prosodic features to obtain unsupervised temporal segmentation of the prosody stream. The prosody patterns are expected to follow smooth pitch frequency movements over several syllables. Considering the average syllable durations and smoothness of the pitch contours, we set $N_{\Lambda_p} = 5$ in each branch of the prosody HMM model Λ_p. The number of prosody patterns M_{Λ_p} is set to 5.

In the second stage, the discrete multi-stream HMM structure Γ_{gp} is trained using EM over the joint gesture-prosody pattern label stream to perform unsupervised segmentation. The number of states for each branch of Γ_{gp} is selected as $N_{\Gamma_{gp}} = 4$ to model possible label pair transitions. These four states model four different gesture-prosody label pair combinations within a joint gesture-prosody label pattern. The number of joint label patterns is set to $M_{\Gamma_{gp}} = 6$.

Table 1. The distance measures between the original and the two sets of synthesized Euler angles: from the proposed Γ_{gp} and IOHMM models

Model	Γ_{gp}	IOHMM
ϵ_e	12.518	13.287
ϵ_m	1.798	1.857

Table 2. The Subjective A-B Comparison Results

A-B pair	Preference Score
Original - Γ_{gp}	-0.23
Original - IOHMM	-0.83
Γ_{gp} - IOHMM	-0.56
Identical pairs	0.04

5.1. Synthesis Results

In this section, we present objective and subjective evaluations of the prosody-driven head gesture synthesis process. The objective evaluations compare the difference between original and synthesized Euler angles. Furthermore, A-B comparison type subjective evaluations are performed using the talking head avatar of *Momentum Inc.*, where the Euler angles that we deliver are used to drive head gestures/motion of the speech-driven talking head animation.

We have also considered using an Input-Output Hidden Markov Model (IOHMM) structure [1] for joint analysis of head gestures and prosody. In that case, the IOHMM structure replaces the HMM Γ_{gp} and builds gesture-prosody segment label mapping to predict the gesture segment labels from the prosody. The states in the IOHMM are fully connected and the number of states is selected to be same to the number of states in the Γ_{gp} model, which is 24.

In our evaluations, the distance between the original and synthesized Euler angles are measured using the Euclidian distance ϵ_e and the Mahalanobis distance ϵ_m. The original Euler angles from the visual part of the test database are extracted to be used as the ground truth in the objective evaluations. Two sets of synthesized Euler angles are generated using the audio part of the test database. The first set is the proposed head gesture synthesis system based on the Γ_{gp} model. The second set is generated with the same head gesture synthesis system as defined in Section 4 by replacing the second stage joint gesture-prosody correlation model Γ_{gp} by IOHMM. The average distances are given in Table 1. Note that, all the three distance measures yield better distances for the proposed joint gesture-prosody correlation model Γ_{gp}.

Subjective A-B comparisons are performed using the speech-driven talking head animations to measure opinions on the naturalness of the synthesized head gestures. The subjects are asked to evaluate the naturalness of the speech-driven synthesized head gestures for an A-B test pair on a scale of $(-2,\ -1,\ 0,\ 1,\ 2)$, where the scale corresponds to (A much better, A better, no preference, B better, B much better).

The whole test database is manually partitioned into meaningful 15 segments, where each segment is approximately 12 seconds. For each evaluation 8 segments out of 15 are randomly selected. Three sets of A-B comparison pairs, each including these 8 segments, are considered for the speech-driven talking head animations using the original and two sets of synthesized Euler angles. Furthermore, three random startup A-B test pairs and another three test pairs with identical synthesis algorithms are also included to the subjective test set. Hence, the total number of A-B pairs in a test is 30. Apart from the three random start-up A-B pairs, all the pairs are randomized across conditions and pairwise. The subjective tests are performed over 15 subjects. The average preference scores for the three comparison sets are presented in Table 2. The subjective A-B comparisons, as expected, indicate a preference for the talking head animations using original Euler angles. The animations that are derived with the proposed joint gesture-prosody correlation model Γ_{gp} are preferred over the animations using IOHMM correlation model with an average preference score of -0.63. Also note that, the preference of the animations using the original Euler angles is stronger for the IOHMM driven animations than the proposed Γ_{gp} driven animations. Samples of the audio-visual sequences for the prosody-driven talking head animations are available online *http://mvgl.ku.edu.tr*.

6. CONCLUSIONS

We proposed a new two-stage joint head gesture and speech prosody analysis framework to drive automatic realistic synthesis of head gestures from speech prosody. The proposed two-stage analysis framework offers the following advantages: i) Meaningful elementary gesture and prosody patterns are defined for a speaker in the first stage. ii) A mapping between these elementary prosody and head gesture patterns is obtained with the unsupervised segmentation of joint gesture-prosody label stream. iii) The HMM-based analysis and synthesis yields flexibility in modeling structural and durational variations within gestural and prosodic patterns. iv) Automatic generation of the elementary gesture patterns produces natural looking prosody-driven head gesture synthesis.

7. REFERENCES

[1] Y. Li and H.-Y. Shum, "Learning dynamic audio-visual mapping with inputoutput hidden markov models," *IEEE Trans. on Multimedia*, vol. 8, no. 3, pp. 542–549, 2006.

[2] J. Xue, J. Borgstrom, J. Jiang, L.E. Bernstein, and A. Alwan, "Acoustically-driven talking face synthesis using dynamic bayesian networks," in *Proc. of the Int. Conf. on Multimedia and Expo 2006 (ICME 2006)*, 2006, pp. 1165–1168.

[3] K.G. Munhall, Jeffery A. Jones, Daniel E. Callan, Takaaki Kuratate, and Eric Vatikiotis-Bateson, "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," in *PSYCHOLOGICAL SCIENCE*, 2004, vol. 15, pp. 133–137.

[4] F. Quek, D. McNeill, R. Ansari, X. Ma, R. Bryll, S. Duncan, and K.E. McCullough, "Gesture cues for conversational interaction in monocular video," *ICCV99 Wksp on RATFGRTS*, pp. 64–69, 1999.

[5] Takaaki Kuratate, Kevin G. Munhall, Philip E. Rubin, Eric Vatikiotis-Bateson, and Hani Yehia, "Audio-visual synthesis of talking faces from speech production correlates," in *Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99)*, 1999, pp. 1279–1282.

[6] M. E. Sargin, F. Ofli, Y. Yasinnik, O. Aran, A. Karpov, S. Wilson, E. Erzin, Y. Yemez, and A. M. Tekalp, "Gesture-speech correlation analysis and speech driven gesture synthesis," in *Proc. of the Int. Conf. on Multimedia and Expo 2006 (ICME 2006)*, 2006.