Continuous Control of Style and Style Transitions through Linear Interpolation in Hidden Markov Model Based Walk Synthesis

Joëlle Tilmanne and Thierry Dutoit

Numediart Institute / TCTS Lab.
University of Mons (UMons)
Mons, Belgium
joelle.tilmanne@umons.ac.be

Abstract. We present a Hidden Markov Model (HMM) based stylistic walk synthesizer, where the synthesized styles are combinations or exaggerations of the walk styles present in the training database. Our synthesizer is also capable of generating walk sequences with controlled style transitions. In a first stage, Hidden Markov Models of eleven different gait styles are trained, using a database of motion capture walk sequences. In a second stage, the probability density functions inside the stylistic models are interpolated or extrapolated in order to synthesize walks with styles or style intensities that were not present in the training database. A continuous model of the style parameter space is thus constructed around the eleven original walk styles. Qualitative user evaluation of the synthesized sequences showed that the naturalness of motions is preserved after linear interpolation between styles and that evaluators are sensitive to the interpolation factor.

Keywords: HMM, gait, style, control, synthesis, motion capture.

1 Introduction

The character animation field covers a lot of different domains, ranging from the coarse motions seen in video games to the precise humanlike motions of the last generation of 3D films, and including fields like virtual reality or character animation for human-computer interactions. In the framework of character animation, several approaches are available to produce realistic humanlike motion. There are currently three main kinds of techniques which are used for motion production in the animation field: traditional keyframe animation, procedural techniques in which a software based on a set of rules helps the animator for motion production, and motion capture based approaches where the animator uses real motions captured on an actor.

Producing natural looking animations of virtual humans is a very challenging task as human eyes are natural experts of human motion and naturalness. This is why motion capture, which consists in capturing human motion in the real world to transfer it to the virtual world under a mathematical form usable by

M.L. Gavrilova and C.J.K. Tan (Eds.): Trans. on Comput. Sci. XVI, LNCS 7380, pp. 34–54, 2012.

computers, is the only way to obtain truly realistic human motion [21]. Motion capture opens a huge field of study and of potential applications, and has received growing interest in the last years, especially since it is becoming more affordable and reliable. However, this technology suffers from several drawbacks. Motion capture data have a high dimensionality, the choice of the parameterization is not straightforward and motion is highly variable in general. Furthermore, motion capture is not very flexible. It is hard to reuse recorded motion capture segments and it is also very difficult to modify natural motion without loosing their naturalness, especially since the human eyes will be very sensitive to any inconsistency in motions. In this article, we address this last problem, by interpolating and extrapolating between and beyond the motion styles present in our motion capture training database. Finally, natural looking transitions are another challenge of motion capture sequences. For instance, generating a smooth transition between two separate walk styles is not straightforward if the transition has not been recorded. This issue is also addressed in this article, as our model is capable of handling controllable and natural looking transitions between distinct style models.

Two different approaches coexist about using motion capture data for producing animations: the “template-based” and the “model-based” approaches. In the “template-based” approach, a large database of motion sequences is built and algorithms are developed to retrieve, edit and blend together motion parts, to produce new sequences [8]. The “model-based” approach consists in training models based on motion capture data. The models can later be used to synthesize new motion sequences without resorting to the database initially used for training [9,1,31]. If the model is properly trained, the information contained in the database is summarized in the model parameters. This model gives then more freedom to the user for producing new plausible sequences, even if they were not present in the training database. This approach has been used for years in speech processing for example, first for recognition and more recently for synthesis [14]. Our work falls in the latter category, with the use of model-based techniques and more precisely of Hidden Markov Models (HMMs) [25], for the modeling and synthesis of humanlike motion. HMMs have been used for motion modeling and more especially motion recognition for a long time. The statistical nature of HMMs makes them perfect candidates for the modeling of spatio-temporal time series like human motion where both the tempo and the space trajectory can vary for several realizations of the same motion.

This approach is inscribed at the crossing between motion capture and procedural methods, as we use statistical learning techniques to automatically extract the underlying rules of human motion, without any prior knowledge, directly from training on 3D motion data. The position of our current work in the character animation field is illustrated in Figure 1.

In addition to template- or model-based approaches, other distinctions can be made between the methods applied to modify motion capture data. Parametric motion synthesis is a method that can be applied to both approaches and which consists in producing new motion by interpolating between motions that are

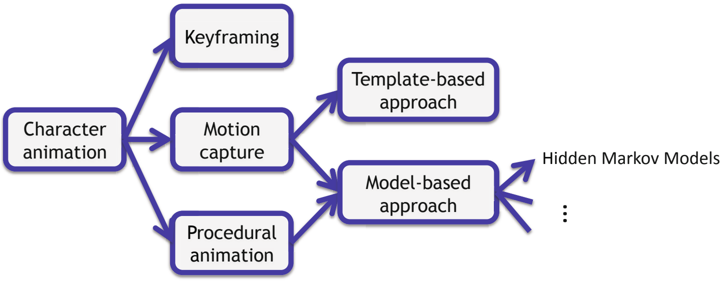

Fig. 1. Position of the HMM-based motion synthesis in the character animation hierarchy

visually similar and correspond to the same logical action [24]. In this article, our approach is parametric and model-based as we aim not only to synthesize a plausible human walk but also to take into account some kind of "style" component. "Style" can refer to emotions, but also to speed, gender (male or female), age (children or elderly,...), specific contexts, etc. The "style" studied in this paper consists of both emotions (like sadness for instance) or particular contexts (like "topmodel" walk). In this paper, style-specific models of walks are first trained over a training motion capture database. The parameters of these style-specific models are then used to produce new models corresponding to style combinations or style exaggerations not present in the training database. Furthermore, the parameters of the style-specific models can be used to build style-transition models and generate a sequence of walk presenting a smooth style transition, without jumps nor abrupt and disturbing changes.

The remainder of the paper is organized as follows. Section 2 presents related work. The motion capture training databases are presented in Section 3. Section 4 describes how the style-specific HMM models of walk were trained, and how these models can be used to synthesize new walk sequences. Section 5 explains how new style characteristics can be obtained by interpolation and extrapolation of the style-specific walk models. The results are presented in Section 6, and Section 7 describes the qualitative user evaluation. Conclusion and future works are addressed in Section 8.

2 Related Work

One of the main problems associated with the use of motion capture data to build animations is that using only the recorded motions can be very limitative. Even if the sequence of motion wanted by the animator can be found in the database, this sequence might not present the required style. Recording a database with all the style options for all the motions is impossible, and recording new motions each time a new animation has to be produced is very costly and often not materially possible. The goal is to find ways of parameterizing the style component of the motion independently from the functional part of the same motion, in order to give the user some kind of interactive control on the style of the output motion.

This approach is similar to approaches encountered in other aspects of human biometrical characteristics modeling or synthesis. Wecker et al. [42] for instance decompose iris images into basic iris properties and individual biometric characteristics, and use this decomposition in order to synthesize new plausible irises images not present in the training database. Lu et al. [20] start from a generic 3D face model and project 2D images of real faces on it. From one single image of each person, using their generic 3D model, their method generates new 2D images of the same person with different expressions or lighting conditions.

People have taken several approaches to address this problem of decomposition into a functional basis and variable characteristics. We will focus the present review of related work on research addressing 3D motion problems.

One approach is to use signal processing techniques to tackle the problem. Bruderlin and Williams [2] use a multiresolution filtering that decomposes the motion into several frequency bands whose amplitude can be modified in order to change the motion style. Unuma et al. [37] apply Fourier transform on the motion data and can modify the aspect of the motion by changing the weights of the Fourier transform in the frequency domain. Unfortunately, these approaches are not easy to use for style control, as changing the weights in the frequency domain is not an intuitive way of controlling the style of a motion.

The underlying principle of most works, including ours, is to use statistical learning techniques on a set of stylistic motion capture data. These techniques separate automatically, without any prior knowledge, the style component from the fundamental function of the motion. The statistical models can then give the user some kind of control on the style parameter, and synthesize motion according to the user's commands. If the principle is the same, several statistical approaches have been tested in the last years.

In [26], Rose et al. decompose motion into verb (fundamental of motion, like "walking") and adverb (which modifies the basic motion according to emotion, gender, "uphill" or "downhill", etc.). They propose a technique for real-time interpolation of motion sequences, based on radial basis functions and low order polynomials representation of the motion. Glardon et al. [9], Troje [36], Min et al. [22] and Tilmanne et al. [31] all use principal component analysis (PCA) not only for reducing the dimensionality of the problem, but to separate the influence of style from the functional part of the motion. In a similar way, Shapiro et al. [27] use Independent Component Analysis (ICA) for the same purpose. Min et al. [23] conduct a multilinear motion analysis to extract separately style and individuality variations, after time warping and PCA for dimensionality reduction.

Another interesting approach is to use Hidden Markov Models (HMMs) to synthesize motion, and to integrate a style variable into the HMM parameterization. One of the advantages of using Hidden Markov Models is that they exempt from using time warping procedures, needed in most approaches in order to align sequences prior to analyze them or extract the style component among them. HMMs integrate directly in their modeling the time variability of the motion. In their work, Wang et al. [41,40] present an HMM that can be trained as a parametric HMM incorporating a “style” parameter in the probability density functions (these densities are represented by SOMN (self organizing mixture networks) in [41] and by mix-SDTG (stylized decomposable triangulated graphs) in [40]). Brand and Hertzmann [1] include a style variable which is automatically extracted during the training process of the HMM and that can vary during the synthesis of a motion sequence. However, in their “Style Machine”, the style variable is not explicit, and changes some intrinsic style-related parameters which can make it hard to use as a style controller. In a work closely related to ours, Yamazaki et al. [46] model walk using a Hidden Semi-Markov Model (HSMM). Their model takes into account speed and stride length as a “style” variation using multiple regression. However, this method can only be use to model quantitative variations, and is thus not suited to model emotions or expressivity that can hardly be described by numerical values. In their approach, the whole training has to be done again if one wants to add a new style in the model.

3 Databases

In this work, two databases recorded with an inertial motion tracking system (IGS-190 from Animazoo [15]) were used. Our two databases, respectively called “eNTERFACE’08 3D” and “Mockey”, were recorded with the same motion capture suit but with different aims, subjects and settings.

The eNTERFACE’08 3D database is described in details in [33]. This database contains, among others, three sequences of straight “free” walk for 41 different subjects. In the “free” walk, subjects were invited to walk at their usual comfort speed. In the present work, the three free walk sequences of the 41 subjects were used to train our “neutral” walk model. In that database, the motion was captured at a frame rate of 60 frames per second (fps).

The Mockey Database aims to study the “expressivity” of walk [31]. All the different walks were performed by the same actor. He was given instructions about the “walking style” he had to act before each walk sequence recording. The eleven different acted styles were the following: proud, decided, sad, top-model, drunk, cool, afraid, tiptoeing, heavy, in a hurry and manly. Our “style” component consists thus in exaggerated variations that can be far from a plain walk. In this second database, motion was recorded at a frame rate of 30 fps. These eleven styles were arbitrarily chosen as they all have a recognizable influence on walk, as illustrated in Figure 2.

The walk sequences were manually segmented into left and right steps. The boundaries of the steps were arbitrarily defined as the moment the heel touches the ground. Depending on the style of walk performed and its corresponding step length, a different number of walk steps was recorded for each style. Each motion file contains two parts: the skeleton definition and the motion data. In the motion data part, the first three values of each frame give the 3D position of the

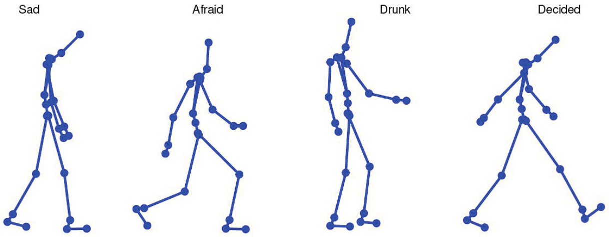

Fig. 2. Four example postures taken from the motion capture database (sad, afraid, drunk and decided walks)

root of the skeleton. They were discarded, as they depend on the displacement and orientation of the walk and can be recalculated given the foot contact with the ground and the leg segments lengths. The pose of the skeleton at each frame is then described by 18 tridimensional joint angles, which gives 54 values per frame to describe the motion.

No dimensionality reduction was conducted on the original set of 54 parameters in order to build a generic style interpolation procedure in which any new walk style could be added in the future without having anything to modify in the previously trained models. This would not have been the case if principal component analysis (PCA) had been used for dimensionality reduction for instance, since the PCA space would depend on the styles present in the training database. By keeping the 54 original motion observations, we avoid our modeling procedure from being dependent on the set of styles present in the training database.

In this paper, our 3D angles were converted from their original Euler angle representation into the exponential map angle parameterization which is locally linear and where singularities can be avoided [11,16]. Exponential maps represent each 3D rotation by 3 values.

4 Style Models Training

Our approach for stylistic model training is to start from a procedure originally developed for speaker adaptation in speech synthesis and to adapt it to our motion problem. This HMM-based procedure is presented in more details in [32] and is based on functions originally implemented for speech within the "HMM-based Speech Synthesis System" (HTS) framework, publicly available on the HTS website [14]. The dynamical aspect of the data is taken into account by integrating the first and second derivatives of our parameters both for reference and stylistic model training and for synthesis [35]. By adding these derivatives to our 54 original parameters, we obtain a 162 dimensional vector of observations to model. The time spent in each state of the HMM is explicitly modeled in duration probability density functions thanks to Hidden Semi-Markov Models

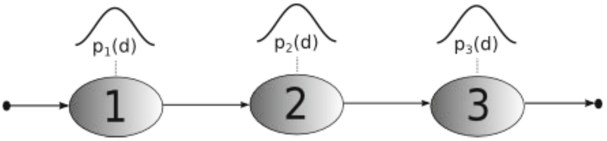

Fig. 3. A three-states HSMM with no skips (with $p_i(d)$ representing the density probability of the duration d of state i)

(HSMM) [47]. The schematic representation of the state duration modeling in a three-states HSMM is represented in Figure 3.

The first stage of the procedure consists in the training of a multi-walkers "reference" model, with the "neutral" walk sequences from the 41 subjects of the "eNTERFACE'08 3D" database. That reference HSMM model is trained using the HTS framework, with a procedure called SAT (for "speaker adaptive training" as it was designed for speech processing [44]) that removes the influence of the subject-specific variations in the final model.

Both steps (left or right) are modeled by separate left-right five-states HSMMs with no skips. Furthermore, contextual factors related to the position of the step in the whole walk sequence were taken into account during the training, thereby multiplying the number of models to train. We made the contextual distinction between five positions in the walk sequence for each step: the first, second, last, last-but-one steps of the sequence, and all the other steps. The training began thus with ten models to train (five for each step). These ten models can be clustered using decision trees. The decision tree is a binary tree, and in each of its nodes, a question splits contextual models into two groups. All possible contextual combinations can be found by traversing the tree. Once the decision tree is constructed, unseen contexts can be prepared and leave nodes containing little or very similar data can be merged (for more information on how the trees are built and used, please refer to [48]). After the decision tree clustering, six out of the ten originally possible contextual HSMMs remained to model the reference walk.

The reference model is used in the second stage of the training, as a basis from which the adaptive training is conducted in order to adapt the reference model to a specific style, using the stylistic walks from the Mockey database. The adaptive training is a method from the latest results of the HMM speech synthesis field, adapted to our motion synthesis problem [32]. Using this adaptation procedure for each style of the Mockey database, we obtain eleven style-specific HSMM models, with each stylistic walk model containing 6 contextual HSMMs.

Using these models, an HMM-based walk synthesis can be conducted and new sequences of walk can be calculated for each one of our eleven walk styles. First a model of the new sequence is obtained by concatenating HMMs corresponding to the succession of steps to be synthesized. This complete model is then used to calculate the corresponding optimal observation sequence, taking into account the dynamics of the synthesized data (thanks to the first and second derivatives of the parameters), which ensures the continuity of the synthesized sequence.

The model gives us joint angles and the cartesian coordinates of the root of the skeleton can then be computed. Using our knowledge of the boundaries of the synthesized walk cycles and calculating the height of each foot thanks to the known leg segment lengths, we determine which foot is in contact with the ground. From that fixed 3D position, we calculate the position of the whole body until the other foot becomes the reference, and so on for the whole sequence. This method enables us to ensure that no foot sliding effect can occur, as the displacement of the whole body is driven by the foot contact point with the ground.

5 Interpolation

Once our eleven style-dependent HSMM models are built, we can synthesize as many stylistic walk sequences as we want, for each one of the eleven styles. So far, each style is modeled separately, and in the synthesis step, the user's control on the output sequence is only the choice of one among the eleven possible styles. In this work, we want to go further in the style modeling and study how the styles can be controlled with more freedom and how the models behave outside of the styles trained from the database. In the model training stage presented in Section 4, each stylistic walk is modeled by six five-states HSMMs, and each HSMM model contains both state duration and 162-dimensional observation modeling. For each state of each HSMM, duration is modeled by one Gaussian probability density function (mean and variance) and observations are modeled by single Gaussian probability density functions (multidimensional Gaussian with diagonal covariance matrix). The set of parameters of one whole style model consists thus in 9780 parameters for 4890 probability density functions (pdfs), as each pdf consists in a Gaussian model (mean + variance) and $(\ 162\ (observation\ parameters)\ +\ 1\ (duration\ of\ one\ state)\ )$ $*\ 5\ (number\ of\ states)\ *\ 6\ (number\ of\ HSMMs) = 4890\ pdfs.$

Among these 4890 probability density functions, 4860 pdfs model the observations and 30 pdfs model the state durations.

In this work, we use these model parameters as a basis for obtaining new or exaggerated styles by extrapolation or interpolation. To do so, a simple procedure was applied. New models were obtained by linear combinations of the means of the probability density functions of each style-specific model. The 4890-dimensional model pdfs space is thus considered as a continuous space in which new styles can be produced by taking points in the space between two existing styles or slightly beyond these styles. Since the new style is obtained through a simple linear interpolation between two vectors, the high dimensionality of the model parameters is not an issue. The known styles are used as landmarks in the continuous model parameter space, around which new style possibilities can be produced by going further away or coming closer to another known style.

In this continuous space, combinations of styles can be obtained, but the intensity of the eleven styles can also be controlled in a continuous manner, even though no control was available in the eleven style models themselves. An average walk model was calculated by taking the mean model parameters over the eleven style models, and used as a basis for controlling the intensity of each style separately. This average model corresponds to some kind of "neutral" style for the actor recorded in the stylistic database. Figure 4 illustrates the linear operations in the 4890-dimensional model space. In our previous work [32], eleven separate style models were obtained and could be used separately (Figure 4.A). In this work, we generalize the motion space by interpolating and extrapolating between two styles (Figure 4.B) or by continuously controlling the amplitude of a given style by interpolation and extrapolation with respect to the average style (mean of our eleven styles, (Figure 4.C)).

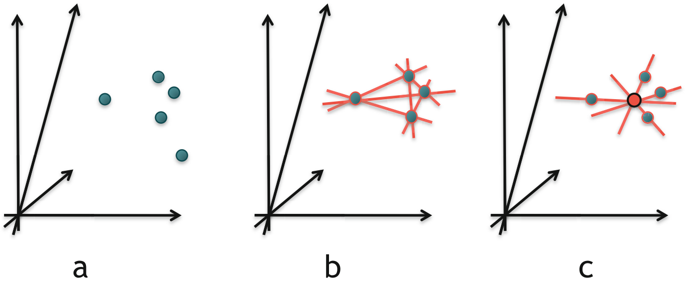

Fig. 4. Schematic representation of the interpolation/extrapolation procedure in the 4890-dimensional model space. Green dots represent the separate stylistic models, red lines the interpolation between models and the red dot is the averaged style model.

6 Results

Our first approach in building new models consists in linearly interpolating the whole set of 9780 parameters between two styles:

$$pdf_{new} = pdf_{style1} + interp * (pdf_{style2} - pdf_{style1}). \tag{1}$$

Where *pdf* is the 4890-dimensional vector of the probability density function means (i.e. the parameters of the model), and *interp* is a scalar value that gives the interpolation ratio between *style* 1 and *style* 2. The difference between *style* 2 and the reference *style* 1 is thus multiplied by a factor *interp* before being added to *style* 1. Figure 5 shows an example of interpolation between the original sad and decided walks, and illustrates that both postures and durations are interpolated. This approach can easily be generalized to a linear combination of any number of original styles.

The second approach aims at controlling the intensity of expression for each style separately. Style interpolation and exaggeration were obtained by using the averaged walk model as a reference (*style* 1 in equation 1), and the style whose

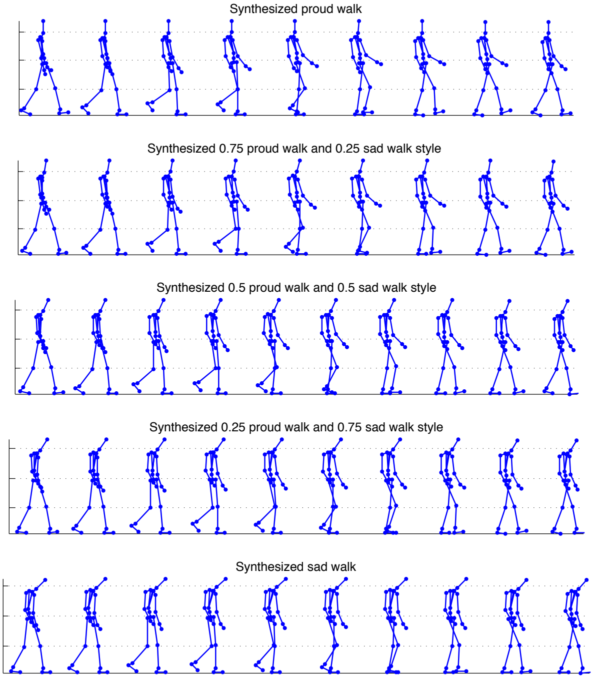

Fig. 5. Synthesized left step for original proud walk (first subfigure, $interp = 1$), 0.75 proud and 0.25 sad walk (second subfigure, $interp = 0.75$), half sad/half proud walk (third subfigure, $interp = 0.5$), 0.25 proud and 0.75 sad walk (fourth subfigure, $interp = 0.25$), original sad walk (last subfigure, $interp = 0$), . Synthesized skeleton poses are displayed every 0.1 second.

intensity we want to control (*style* 2) as an interpolation direction. The strength of the style can be diminished by decreasing the *interp* value from 1 to 0. For values of *interp* greater than 1, the style is exaggerated. Our synthesized walk sequences remained natural for values of *interp* lower than 2. For values around 2 and above, the quality of the synthesized walk depended on the original style as some exaggerated styles were affected by impossible movements (knees bending backwards or awkward bending of the spine for instance). These problems could be avoided in future studies by adding constraints to the joint angles so that they cannot take values beyond what is physically possible for a human being.

Another interesting result is obtained by giving negative values to the *interp* parameter. The difference between the controlled style and the average walk model is substracted to that average walk instead of being added. The obtained style then presents characteristics at the opposite of the controlled style. For instance the sad walk which is slower than the average model and where the pose of the character tends to bend inwards with respect to the average posture will give an opposite model where the character walks faster and looks much more "open" by its posture. We are thus able to synthesize styles that do not appear in our database but that show style characteristics that are opposite to the ones of recorded styles. A left step synthesized with the original sad walk model and with three new style intensities obtained by interpolation or extrapolation is illustrated in Figure 6. The figure shows that both poses and durations are affected by the interpolation process.

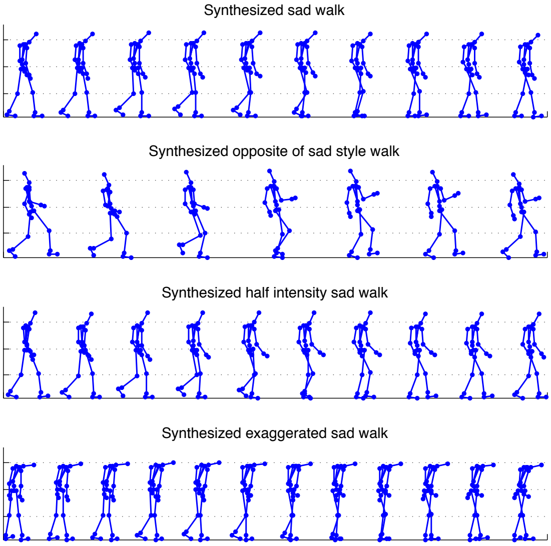

Fig. 6. Synthesized left step for original sad walk (first subfigure, $interp = 1$), interpolated opposite model (second subfigure, $interp =$ -1), half sadness intensity (third subfigure, $interp = 0.5$) and exaggerated sad walk (last subfigure, $interp = 2$). Synthesized skeleton poses are displayed every 0.1 second.

In the first two approaches presented, a new model is calculated by interpolation of the original style models. The interpolated model represents a single new style and is used to synthesize a sequence of walk with a style not present in the original database. The HMM model takes into account the dynamics of the data, and the continuity between the steps of the walk sequence is hence ensured. But it also enables us to synthesize a sequence presenting style transitions. During the synthesis process, step models are concatenated in order to form a complete model of the entire walk sequence. If the concatenated step models represent different walks, the resulting synthesized sequence will display the corresponding style change. As the synthesis process takes into account the dynamics of the data to calculate the optimal parameter sequence corresponding to the concatenated walk model, the style transition will occur without abrupt jerks. This result is similar to the blending procedures applied to smooth the transition between two different mocap files over a few frames as it is usually done in pure mocap approaches. However, even if no jerks are present in the angle data and that the motion continues smoothly when the character goes from a step with style A to a step with style B, the transition will not appear as "natural" for all style transitions. If they are not surprised by something, humans do not abruptly change the style of their walk from one step to the next one. The change will rather occur continuously over a few steps.

Rather than concatenating directly a sequence of "style 1" step models to "style 2" step models, our interpolation procedure is able to produce step models corresponding to styles between style 1 and style 2. The style transition can thus be distributed over as many steps as the user wants. The steps concatenated in the whole sequence correspond to gradual interpolations, starting from style 1 and going to style 2. The walk sequence produced this way is thus a smooth and gradual transition between style 1 and style 2, transiting by in-between styles that were not present in the original database. The resulting transition looks more natural than a brutal style change occurring between one step and the next one, as illustrated in Figure 7.

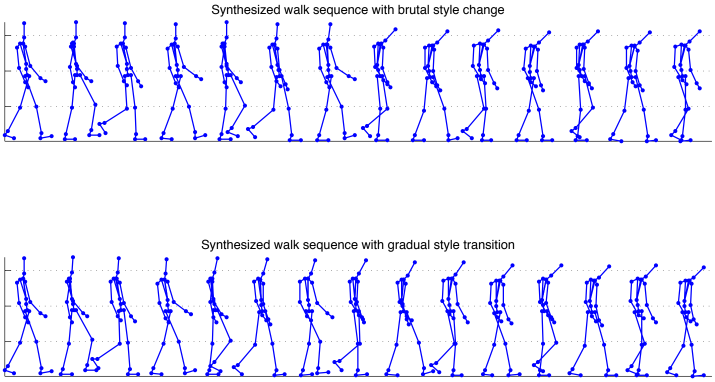

Fig. 7. Synthesized walk sequence with abrupt proud to sad style transition (first subfigure), and synthesized walk sequence with gradual proud to sad style transition (second subfigure). Synthesized skeleton poses are displayed every 0.5 second.

7 User Evaluation

Evaluating the quality of motion sequences is an open problem common to the whole character animation field, independently from the studied approach. Results are often illustrated with a video displaying some motion examples, but most works are presented without any kind of subjective evaluation by the user.

In [32], we proposed three tests to assess the quality of the synthesis results. These tests evaluated the naturalness of the synthesized motion, the style recognition of our eleven original styles and the comparison between original motion capture and synthesized examples. In one of the tests, users were asked to classify the original walk styles (without any modification) among 11 possibilities, and the results showed that the styles were correctly perceived most of the time, even if some close styles were sometimes switched. In this paper, a similar evaluation approach was chosen. Our evaluation consisted in two tests, evaluating respectively the quality of the interpolation between two styles (Section 7.1) and the style intensity control (Section 7.2).

Participants accessed to the evaluation tests through a web browser. They were presented one video at a time and asked to evaluate its content. Once their answer was selected and saved, they could not come back to previous videos. If they did not complete the test thoroughly, they could come back later, but the participant's results were saved even if the two tests were not completely finished. They had to start the video themselves by clicking on it, and could watch it as many times as they wanted. In the video sequences, motion was performed by a basic blue stick-figure character as shown in the previous figures.

Fifty-two naive evaluators took part in the evaluation, 22 females and 30 males, from 16 to 66 years old, with an average and standard deviation of 32 and 11 years, respectively. Every evaluator was presented a set of 10 videos or couples of videos for each test. Those videos were randomly picked by the evaluation program, and the evaluated set was thus different for each evaluator. The result of each test for each evaluator were taken into account only if the user completed all the evaluations of the given test. The final number of evaluators taken into account is not the same for both test as some users dropped the evaluation after taking the first test.

7.1 Style Interpolation Evaluation

In the first test, the evaluator was presented one video displaying three walk sequences with different styles, in a row. In the video, the first sequence was a walk with style A, the second one was a walk with style B, and the third sequence was an interpolation between styles A and B. The evaluation sequences corresponded to five possible pairs of A and B styles, and to five interpolation factors (0, 0.25, 0.5, 0.75 or 1) for each A and B pair. The set of evaluation videos for this first test thus contained 25 videos from which ten were randomly picked for each user to evaluate.

The user was asked to position the interpolated walk between style A and style B, by choosing between five possible answers:

- identical to style A (100% A + 0% B)
- close to style A (75% A + 25% B)
- in the middle between style A and style B (50% A + 50% B)
- close to style B (25% A + 75% B)
- identical to style B (0% A + 100% B)

Table 1. Confusion matrix of interpolation factor recognition test. The recognition rate is expressed in percents of the actual interpolation factor sequences presented to the evaluators, rounded to the unit.

		Evaluators classification (%) of interpolation factor				
		0	0.25	0.5	0.75	1
Actual interpolation factor	0	31	47	17	3	2
	0.25	3	48	39	8	2
	0.5	2	13	41	39	5
	0.75	3	5	22	46	24
	1	1	1	8	43	47

He was also asked to assess the naturalness of the interpolated walk by choosing if it seemed “Real”, “Synthetic”, or “I don’t know”.

Fifty-two participants completed this first test. Table 1 presents the confusion matrix of the interpolation factor recognition.

The interpolation factor was properly recognized by the evaluator in 42.7% of the cases, much higher than the 20% of mere chance. And in 46% of the cases, the interpolation factor was misidentified for one of its direct neighbors (for instance 0.5 or 0 instead of 0.25). In 60.6% of the cases, the evaluator was not capable of recognizing that the same file was displayed twice (when $interp = 0 \ or \ 1$), which demonstrates that even if the evaluators were capable of recognizing the trend (more like A or more like B), he is not very good at assessing small style variations. This poor recognition of the original styles can also be explained by the fact that even if the answers “identical to style A (or B)” were proposed to the evaluators, our formulation of the question asked to position the interpolated walk “between” styles A and B which might have mislead some people.

The mean value of the interpolation factor evaluated by the user for each one of the actual interpolation factor is presented in Figure 8, along with a 95% confidence interval. It can be observed that the confidence interval of the different interpolation factors do not overlap. Even if the exact interpolation factor was not always recognized, especially at the extremities ($interp = 0 \ or \ 1$), the global trend of the interpolation is very clear to the user who can make the distinction between each one of the interpolation factors even if he does not evaluate them at their exact value.

Figure 9 presents the results of the naturalness evaluation of the interpolated sequence. It can be observed that the value of the interpolation factor did not influence the naturalness of the sequence as perceived by the user, and our interpolated style models were seen in the same way as models of the original style ($interp = 0 \ or \ 1$).

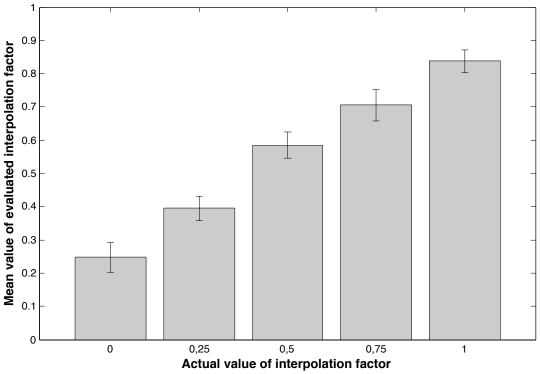

Fig. 8. Mean value of the evaluated interpolation factor (with .95 confidence intervals) for each one of the actual interpolation factor

7.2 Style Intensity Control Evaluation

In this second test, the displayed video contained two walk sequences in a row. The first sequence was a walk with style A, and the second sequence was a walk presenting a variation of style A. The evaluation sequences corresponded to five possible "A" styles, and to five interpolation factors between the style and the averaged style (-1, 0, 0.5, 1 or 2). The set of evaluation videos contained thus 25 files, from which ten were randomly picked for each user. The evaluator was asked to position the gradation of the style intensity of the interpolated walk in comparison to the original style A. The five possible answers he had to choose from were:

- Opposite of style A
- Neutral style
- Half intensity of style A
- Identical to style A
- Exaggeration of style A

The participant was also asked to assess the naturalness of the style intensity variation sequence, in the same manner as in the first test.

Forty-one users completed this second test. Table 2 presents the confusion matrix of the style intensity factor recognition.

The style intensity factor was properly recognized by the evaluator in 69% of the cases, which is much higher than the 20% of mere chance and even better than the result obtained in the first evaluation test. These results show that it was easier for the evaluator to quantify the intensity of a given single style than to evaluate the interpolation factor between two styles that might have been an improbable mix. In 27% of the cases, the interpolation factor was misidentified for one of its direct neighbors (for instance 0.5 or -1 instead of 0). Figure 10 presents the average intensity factor evaluated by the participants for each one of

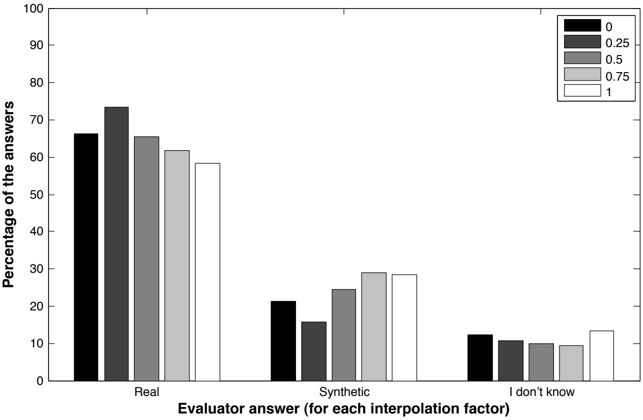

Fig. 9. Results of the naturalness test comparing the perception (real, synthetic or I dont know) of interpolated style sequences with each interpolation factor

Table 2. Confusion matrix of style intensity factor recognition test. The recognition rate is expressed in percents of the actual intensity factor sequences presented to the evaluators, rounded to the unit.

		Evaluators classification (%) of interpolation factor				
		-1	0	0.5	1	2
Actual interpolation factor	-1	92	6	0	0	2
	0	16	31	47	4	2
	0.5	0	4	52	45	0
	1	0	0	5	89	6
	2	11	0	3	4	82

the actual style intensities. It can be noticed that even if the exact interpolation factor was not always recognized, the global trend of the interpolation is very clear to the user, as it was also observed in the first test.

Figure 11 presents the results of the naturalness evaluation of the style intensity control. It can be observed that as long as the style intensity factor stayed in the 1 to 0 range, it did not influence the naturalness of the sequence as perceived by the user. Our averaged sequence (corresponding to $interp = 0$) seems thus as natural to the user as styles from the original database. However, the perceived naturalness decreases dramatically when we exaggerate the style or when we take its opposite. This can easily be explained as, by taking styles outside of the range of value of the walk present in our database (which were already exaggerated styles performed by an actor), the synthesis gives angles values that are outside of the range of possible humans movements and completely ruin the

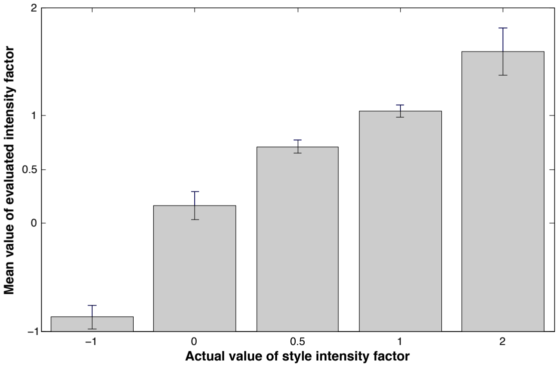

Fig. 10. Mean value of the evaluated style intensity factor (with .95 confidence intervals) for each one of the actual style intensity factors

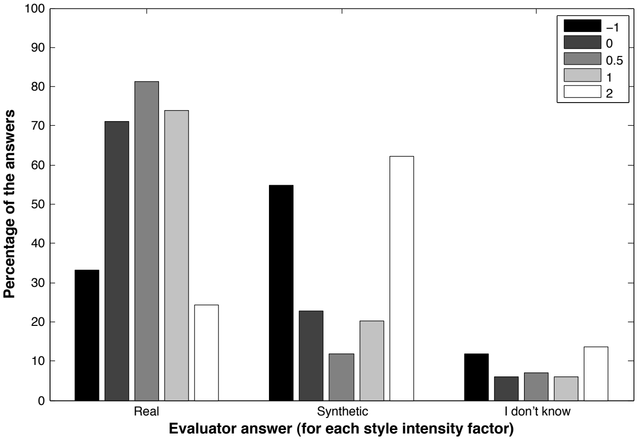

Fig. 11. Results of the naturalness test comparing the perception (real, synthetic or I dont know) of interpolated style sequences with each style intensity factor

naturalness of the motion. This issue will have to be further investigated in the future, and adding rules that constrain the angle values to stay in a plausible human range might be a way of dealing with this problem.

8 Conclusion

In this work, a set of eleven stylistic walk models based on Hidden Semi Markov Models was used as a basis for style interpolation and extrapolation, giving the user continuous control on the style of the synthesized motion while preserving its naturalness. New walk styles not present in the original database could be synthesized by interpolating the model parameters between different original styles. An average style model was also calculated by computing the mean of the parameters of the eleven style models. By taking this average model as a reference and changing the values of the interpolation factor, we were able to control the intensity of the style expression (values of *interp* between zero and one), to exaggerate the controlled style (values of *interp* greater than one), and to obtain new styles with characteristics at the opposite of the controlled style (values of *interp* lower than zero). Our model also enabled us to synthesize smooth and natural looking transitions between two different styles by progressive interpolation. Some examples of walk sequences synthesized with our method can be found at `http://tcts.fpms.ac.be/~tilmanne`. Qualitative user evaluation assessed that the trend of the interpolation factor was perceived by the user and that the naturalness of the motion was preserved for styles between the original styles.

In future works, constraints on the range of variation of the angles should also be added for style extrapolation, so that the synthesized styles remain physically plausible. The interpolation and extrapolation was conducted on 4890 parameters without any feature selection, but the influence of these parameters on the stylistic variations could be investigated, as some of them might be of lesser or greater influence than others on the perceived style. Our next step will be to implement our continuous style control and HMM synthesis procedure into a real-time framework, giving the user the possibility to control the synthesized walk in real time. We will also study how the style characteristics could be added directly on plain motion capture walk sequences.

Acknowledgment. This project was partly funded by the Ministry of Région Wallonne under the Numediart research program (grant N0716631). Joëlle Tilmanne was supported by the "Fonds pour la formation à la recherche dans l'industrie et l'agriculture" (FRIA) during part of this work. We gratefully acknowledge Alexis Moinet for his help in designing the online evaluation procedure and all our evaluators for their participation.

References

1. Brand, M., Hertzmann, A.: Style machines. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, pp. 183–192 (2000)
2. Bruderlin, A., Williams, L.: Motion signal processing. In: SIGGRAPH 1995 Proceedings, pp. 97–104 (1995)
3. Calinon, S., Guenter, F., Billard, A.: On Learning, Representing, and Generalizing a Task in a Humanoid Robot. IEEE Transactions on Systems, Man and Cybernetics 37(2), 286–298 (2007)
4. Chiu, C., Marsella, S.: A style controller for generating virtual human behaviors. In: Proceedings of AAMAS 2011, The 10th International Conference on Autonomous Agents and Multiagent Systems, vol. 3, pp. 1023–1030 (2011)
5. Elgammal, A., Lee, C.S.: The Role of Manifold Learning in Human Motion Analysis Human Motion Understanding, Modeling, Capture and Animation, pp. 1–29 (2008)

6. Forbes, K., Fiume, E.: An efficient search algorithm for motion data using weighted PCA. In: Proceedings of the 2005 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp. 67–76 (2005)
7. Forsyth, D.A., Arikan, O., Ikemoto, L., O'Brien, J., Ramanan, D.: Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis. Foundations and Trends in Computer Graphics and Vision 1(2/3) (2006)
8. Geng, W., Yu, G.: Reuse of Motion Capture Data in Animation: A Review. In: Kumar, V., Gavrilova, M.L., Tan, C.J.K., L'Ecuyer, P. (eds.) ICCSA 2003. LNCS, vol. 2669, pp. 620–629. Springer, Heidelberg (2003)
9. Glardon, P., Boulic, R., Thalmann, D.: PCA-based walking engine using motion capture data. In: Computer Graphics International, pp. 292–298 (2004)
10. Glardon, P., Boulic, R., Thalmann, D.: A Coherent Locomotion Engine Extrapolating Beyond Experimental Data. In: Proceedings of Computer Animation and Social Agent (CASA), Geneva, Switzerland, pp. 73–84 (2004)
11. Grassia, F.S.: Practical parameterization of rotations using the exponential map. Journal of Graphics Tools 3, 29–48 (1998)
12. Grudzinski, T.: Exploiting Quaternion PCA in Virtual Character Motion Analysis. In: Bolc, L., Kulikowski, J.L., Wojciechowski, K. (eds.) ICCVG 2008. LNCS, vol. 5337, pp. 420–429. Springer, Heidelberg (2009)
13. Hsu, E., Pulli, K., Popovic, J.: Style Translation for Human Motion. In: SIGGRAPH 2005 Proceedings, pp. 1082–1089 (2005)
14. HTS working group: The HMM-based speech synthesis system (HTS) Version 2.1, `http://hts.sp.nitech.ac.jp/` (accessed 2010)
15. IGS-190. Animazoo (2010), `http://www.animazoo.com`
16. Johnson, M.P.: Exploiting quaternions to support expressive interactive character motion. PhD Thesis (2002)
17. Jolliffe, I.T.: Principal Component Analysis, 2nd edn. Springer Series in Statistic. Springer, New York (2002)
18. Lau, M., Bar-Joseph, Z., Kuffner, J.: Modeling Spatial and Temporal Variation in Motion Data. ACM Transactions on Graphics (SIGGRAPH ASIA) 28(5), 171 (2009)
19. Li, Y., Wang, T., Shum, H.: Motion texture: a two-level statistical model for character motion synthesis. In: Proc. of SIGGRAPH 2002, New York, NY, USA, pp. 465–472 (2002)
20. Lu, X., Hsu, R.-L., Jain, A.K., Kamgar-Parsi, B., Kamgar-Parsi, B.: Face Recognition with 3D Model-Based Synthesis. In: Zhang, D., Jain, A.K. (eds.) ICBA 2004. LNCS, vol. 3072, pp. 139–146. Springer, Heidelberg (2004)
21. Menache, A.: Understanding motion Capture for Computer Animation and Video Games. Morgan Kauffman Publishers Inc., San Francisco (1999)
22. Min, J., Chan, Y., Chai, J.: Interactive generation of human animation with deformable motion models. ACM Trans. Graph. 29(1), 9:1–9:12 (2009)
23. Min, J., Liu, H., Chai, J.: Synthesis and editing of personalized stylistic human motion. In: Proceedings of SI3D, pp. 39–46 (2010)
24. Pejsa, T., Pandzic, I.S.: State of the Art in Example-Based Motion Synthesis for Virtual Characters in Interactive Applications. Computer Graphics Forum 29(1), 202–226 (2010)
25. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proc. of IEEE 77(2), 257–286 (1989)
26. Rose, C., Cohen, M.F., Bodenheimer, B.: Verbs and Adverbs: Multidimensional Motion Interpolation. IEEE Comput. Graph. Appl. 18(5), 32–40 (1998)

27. Shapiro, A., Cao, Y., Faloutsos, P.: Style components. In: Proceedings of Graphics Interface, Quebec, Canada, pp. 33–39 (2006)
28. Shoemake, K.: Animating Rotations with Quaternion Curves. In: Proc. of SIGGRAPH 2005, San Francisco, vol. 19(3), pp. 245–254 (1985)
29. Tanco, L.M., Hilton, A.: Realistic synthesis of novel human movements from a database of motion capture examples. In: Proc. of the Workshop on Human Motion (HUMO 2000), Washington DC, USA, pp. 137–142 (2000)
30. Taylor, G.W., Hinton, G.E.: Factored conditional restricted Boltzmann Machines for modeling motion style. In: ICML 2009 Proceedings of the 26th Annual International Conference on Machine Learning, pp. 1025–1032 (2009)
31. Tilmanne, J., Dutoit, T.: Expressive Gait Synthesis Using PCA and Gaussian Modeling. In: Boulic, R., Chrysanthou, Y., Komura, T. (eds.) MIG 2010. LNCS, vol. 6459, pp. 363–374. Springer, Heidelberg (2010)
32. Tilmanne, J., Moinet, A., Dutoit, T.: Stylistic gait synthesis based on hidden Markov models. EURASIP Journal on Advances in Signal Processing, 72 (2012)
33. Tilmanne, J., Sebbe, R., Dutoit, T.: A Database for Stylistic Human Gait Modeling and Synthesis. In: Proceedings of the eNTERFACE 2008 Workshop on Multimodal Interfaces, Paris, France, pp. 91–94 (2008)
34. Toda, T., Tokuda, K.: A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis. IEICE-Transactions on Information and Systems 90(5), 816–824 (2007)
35. Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: Speech parameter generation algorithms for HMM-based speech synthesis. In: Proc. of ICASSP (June 2000)
36. Troje, N.F.: Retrieving information from human movement patterns. In: Understanding Events: How Humans See, Represent, and Act on Events, pp. 308–334. Oxford University Press (2008)
37. Unuma, M., Anjyo, K., Takeuchi, R.: Fourier principles for emotion-based human figure animation. In: SIGGRAPH 1995 Proceedings, pp. 91–96 (1995)
38. Urtasun, R., Glardon, P., Boulic, R., Thalmann, D., Fua, P.: Style-based Motion Synthesis. Computer Graphics Forum 23(4), 799–812 (2004)
39. Wang, Y., Liu, Z., Zhou, L.: Automatic 3D Motion Synthesis with Time-Striding Hidden Markov Model. In: Yeung, D.S., Liu, Z.-Q., Wang, X.-Z., Yan, H. (eds.) ICMLC 2005. LNCS (LNAI), vol. 3930, pp. 558–567. Springer, Heidelberg (2006)
40. Wang, Y., Liu, Z., Zhou, L.: Learning Style-directed Dynamics of Human Motion for Automatic Motion Synthesis. In: IEEE Conference on Systems, Man, and Cybernetics 2006, Taiwan, pp. 4428–4433 (2006)
41. Wang, Y., Xie, L., Liu, Z., Zhou, L.: The SOMN-HMM Model and Its Application to Automatic Synthesis of 3D Character Animation. In: IEEE Conference on Systems, Man, and Cybernetics 2006, Taiwan, pp. 4948–4952 (2006)
42. Wecker, L., Samavati, F., Gavrilova, M.: A multiresolution approach to iris synthesis. Computers & Graphics 34(4), 468–478 (2010)
43. Yamagishi, J., Nose, T., Zen, H., Ling, Z.H., Toda, T., Tokuda, K., King, S., Renals, S.: Robust speaker-adaptive HMM-based text-to-speech synthesis. IEEE Transactions on Audio, Speech, and Language Processing 17(6), 1208–1230 (2009)
44. Yamagishi, J., Kobayashi, T.: Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. IEICE TRANSACTIONS on Information and Systems 90(2), 533–543 (2007)

45. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J.: Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Transactions on Audio, Speech, and Language Processing 17(1), 66–83 (2009)
46. Yamazaki, T., Niwase, N., Yamagishi, J., Kobayashi, T.: HumanWalking Motion Synthesis Based on Multiple Regression Hidden Semi-Markov Model. In: 2005 International Conference on Cyberworlds (CW 2005), pp. 445–452 (2005)
47. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.: Duration modeling for HMM-based speech synthesis. In: Fifth International Conference on Spoken Language Processing (ICSLP), pp. 29–32 (1998)
48. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book (for HTK Version 3.4) (2009)