2010

Prominence Driven Character Animation

Charlie Cullen
Technological University Dublin, charlie.cullen@tudublin.ie

Paula McGloin
Technological University Dublin

Anna Deegan
Technological University Dublin

See next page for additional authors

PROMINENCE DRIVEN CHARACTER ANIMATION

Charlie Cullen[1], Paula McGloin[1], Anna Deegan[1], Evin McCarthy[1]

[1]Dublin Institute of Technology, Ireland, charlie.cullen@dmc.dit.ie

Abstract

This paper details the development of a fully automated system for character animation implemented in Autodesk Maya. The system uses prioritised speech events to algorithmically generate head, body, arms and leg movements alongside eyeblinks, eyebrow movements and lip-synching. In addition, gaze tracking is also generated automatically relative to the definition of focus objects-contextually important objects in the character's worldview. The plugin uses an animation profile to store the relevant controllers and movements for a specific character, allowing any character to run with the system. Once a profile has been created, an audio file can be loaded and animated with a single button click. The average time to animate is between 2-3 minutes for 1 minute of speech, and the plugin can be used either as a first pass system for high quality work or as part of a batch animation workflow for larger amounts of content as exemplified in television and online dissemination channels.

Keywords: animation, prominence speech analysis

1 Introduction

Existing work has considered both facial and full body animation from the perspective of Embodied Conversational Agents (ECAs) [1] which utilise script and behaviour annotation [2, 3]. In addition, natural language processing [4] or textual markup [5] methods have been proposed, though do not work with real speech. LivingActor™ [6] can utilise both real or synthesised speech (from input text) to produce automatic lip-synching for animated characters, but the inclusion of gestures is not possible. As there is arguably a prioritisation of movements and gestures in human communication [7], the proposed work contends that an animated character or avatar must perform much more than lip-synchronisation in order to communicate effectively.

Work has been performed linking vocal prosody to head movements [8] and facial expressions [9] and some systems do utilise real speech and gesture in the creation of avatars, notably in the real time synthesis of body language [10]. Although very useful work, the definition of speech prominence is based only on pitch and intensity (and thus is not hierarchical) and movements are defined as vector displacements rather than specific beat gestures or non-gestures. By distinction, we build on previous work in relation to reusable online character animation that developed a method of linking a prominence hierarchy of vowel events in the speech act [11] to movements and gestures for online avatars used in children's games [12]. This work led to discussions with animation houses that indicated interest in a more subtle and granular implementation of the system using industry standard animation tools (notably Autodesk Maya). The system proposed in this paper seeks to use a hierarchy of prominent events in the speech signal to control the allocation of beat movements and non-gestures. In so doing, the system aims for the production of avatars with higher quality real speech that possess a greater capability for gestural communication with the user.

2 Gesture Modelling

In general terms of gesture classification, McNeill [13] argues that all visible movements can be defined as either non-gestures (e.g. head scratching, object manipulation) or as one of a group of gesture types:

- **Iconics**- in which gestures relate closely to the semantic content of the speech, e.g. "…bends it way back" accompanied by gripping gesture moving back towards speaker's own shoulder.
- **Metaphorics**- a metaphoric gesture conveys an abstract concept consisting of a base and a referent.
- **Deictics**- are defined as pointing movements, mainly with the finger though other body parts (head, nose, chin) and extensible objects can be used.
- **Beats**- specify non-imaginistic movements that do not convey a discernable meaning.

At the lowest gesture level, beats are termed as biphasic (2 movement components) relating to movements of the hand. Beats are often considered to be small, of low energy and without a gesture space, thus occuring regardless of spatial position (e.g. finger tapping can occur on a table or on the thigh). Due to their lack of context, beat gestures can potentially be associated with acoustic features in a speech signal as a base level of movement. The animation system proposed in this paper focuses on the use of both non-gestures and beats to convey the rhythms of speech, in a similar manner to the work of Cassell that created beat gestures [14] with synthesised speech. The development of a prominence model for speech analysis allows beats and non-gestures to be automatically allocated to stressed events in the acoustic signal. In so doing, a character can be animated quickly to display the underlying movements that are core to human behaviour.

2.1 Acoustic Speech Analysis

The animation method described in this paper utilises existing work in the analysis of emotional speech, where a vowel stress tagging framework has been developed to define a speech asset by the prosody and rhythm of its vowel events.

978-0-7695-4268-3/10 $26.00 © 2010 IEEE
DOI 10.1109/CVMP.2010.28

In the plugin, vowel onset detection [15] is performed by using a Hann bandpass filter centred around 750Hz to accentuate vowels in the signal for both male and female speakers (whilst removing the higher fricative and plosive energy) and then taking a 1st order derivative curve of the intensity contour to detect vowel energy, Equation (1):

$$\frac{d}{dt}v(t,s_c,s_t) = v(t,0,0) * \left[\frac{d}{dt}h(s_t)\right] * g(0,s_c) \quad (1)$$

This equation defines the formula for obtaining the intensity derivative curve of a Hann bandpass filtered signal, adapted for a single channel from [15]. The initial intensity is denoted by v(t,0,0) at time t where $h(s_t)$ is a low-pass filter and $g(0,s_c)$ is a Gaussian function with zero mean and standard deviation s_c. In calculating the intensity contour, the Gaussian analysis window is set to a long value (80ms) to avoid shimmer [16]. The subsequent derivative curve is then analysed to define the positive maxima (points of greatest rise in the curve) and negative minima (steepest falls in the curve) that indicate candidate vowel onset and offsets. Using these points to define a vowel, each vowel is then queried to produce a stress tagging framework that defines various parameters in the speech signal including: pitch data [17] and contour; intensity data [18] and contour; voice quality [19] and emotional dimensions [20] (Figure 1):

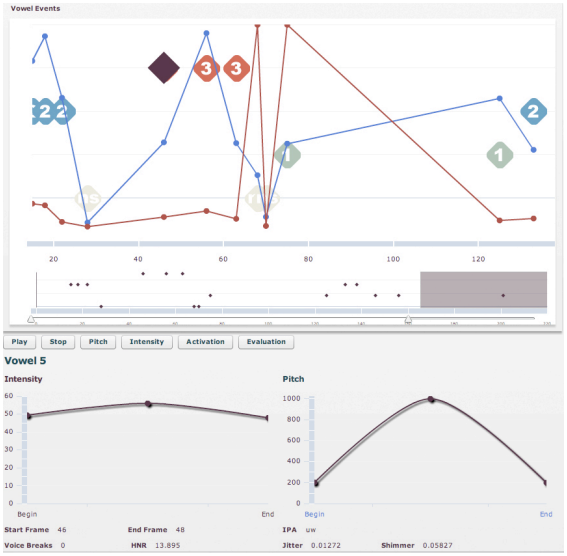


Figure 1: Vowel stress tagging framework speech analysis, taken from Cullen [21]. The screenshot shows vowels defined by prominence, pitch and intensity contours, with micro-prosody and voice quality listed below.

The first chart shows vowel events in a speech asset defined in terms of their prominence level (0-3). The pitch and intensity contours for the overall asset are also shown, while individual vowel contours (pitch, intensity) and voice quality information are displayed in the graphs below. The prominence (stress) level of each vowel is then calculated using a combination of these features, applying a simple promoter system based around deviation from the mean [22].

The acoustic analysis data from this speech analysis framework is then used in the authoring tool to allocate gestures, facial expressions and lip-synchronisation animation by linking it to an established hierarchy of gesture types as detailed below.

2.2 Kinesic Structure

McNeill proposes a hierarchy of gesture types, but this does not directly consider the grouping of gestures within the time domain. By extension of this concept, Kendon [23] defines a temporal hierarchy of gesture movements as follows:

- **Arm use and body posture**- where the speaker adopts different body postures and arm usage patterns. Changes in these patters are considered to be kinesic units or paragraphs.
- **Head movements**- Within a kinesic unit, the same head movements often take place (e.g. head moves right several times)
- **Gesture unit** (G-unit)- defined as the time from beginning of limb movement to end
- **Gesture phrase** (G-phrase)- subset of a G-unit, contains a set of optional phases (e.g. preparation, retraction) based around a stroke (peak gesture effort)

Of particular interest to the work in this paper is the linkage between this kinesic hierarchy to the phonological model of speech:

Kinesic	Phonological
arm use/body posture	Locution cluster
head movement	Locution group
1 G-unit	Locution
1 G-phrase	Tone unit
1 Stroke	Most prominent syllable

Table 1: Mapping of kinesic structure to the phonological speech model, taken from Kendon [23].

This phonological model has direct correlation with data in the stress tagging framework, whereby a stressed (prominent) vowel would be commensurate with a prominent syllable- and hence a stroke. At higher levels, tone units are similar to pitch and intensity contours, while locutions, groups and clusters broadly represent the various grammatical structures (phrase, sentence, paragraph) used in human speech. The stress tagging framework is currently implemented at syllable and tone unit level (where frequency of movement is defined relative to groups of vowels in a tone unit). Future work will investigate the role of locutions and locution groups more fully, as provision has been made in the framework for hierarchical vowel grouping and cluster analysis.

3 Authoring Plugin Operation

The character rigs used in prototyping and testing of the current system are simple biped models that are given character sets with authored movements for each body area within the gesture space (section 0). A more generic system

implementation is planned, whereby a skeleton biped containing the requisite movement libraries is mapped onto a specific rig during the profile construction phase, but at time of writing no definite skeleton based system has been finished. The current Maya plugin is written in python (Figure 2):

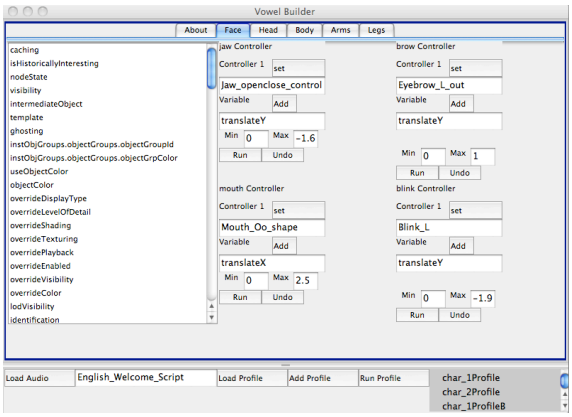


Figure 2: Screenshot of the Maya animation plugin interface. Controllers and variables are allocated to movements, and this profile can be stored for reuse.

Standard Maya interface classes were used to build a simple GUI for setting up the animation profile for a given character. The plugin allows animation to be authored with a single click once a character profile has been defined. To define a profile, the animator selects controllers and variables that relate the features animated by the system. For example, to control the jaw movements of the character, selecting the jaw controller will provide a list of the available variables, setting this variable also requires providing minimum and maximum bounds and the plugin can then control this aspect of the character. In the case of gestures, the Maya clip editor is used to list the available animation clips contained in each character set that can then be added to the animation list. In this manner, as the number of clips increases (for a more complex character) the options for different types of animation output are possible.

Future revisions of the plugin interface will provide more extensive authoring functions, notably provision for emotional modelling of the characters based on the acoustic features of the input speech signal [24]. Consideration will also be given to semi-automated operation [12], where available vowels can be used as authoring points for gesture allocation.

3.1 Production Review Group

To facilitate ease of prototyping and development, different elements of plugin functionality were previewed by creating a series of short playblast animations and then informally testing these animations with a small viewer group. This production review group (made up of researchers) was organised in a similar manner to production review meetings where changes and variations can be discussed as part of a working design brief. These initial tests were carried out during the design stage of the system, and thus are distinguished from formal user tests (section 6).

4 Gesture Library Design

In building a library of gestures for use with the system, the notion of gesture space was used to build character sets relating to different areas within the space (Figure 3):

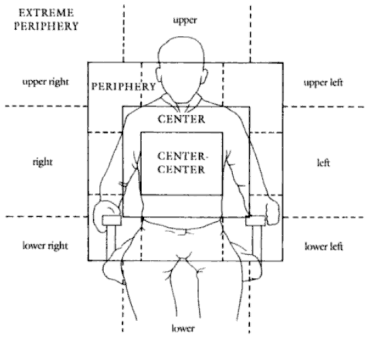


Figure 3: Defined areas within the gesture space, Adapted from McNeill [13].

In this paper, upper, centre, left/right and lower are used to define character sets for head, body, arms and legs respectively. Statistical grouping of gesture types to areas of the gesture space [13] leads to distinct differentiation between iconics (centre), metaphorics (lower centre), deictics (periphery) and beats (speaker defined cluster). In considering beats, the notion that an individual speaker has a particular beat reference point (or area) is used by the work of this paper to create a set of arm and hand gestures relating to a specific area within the gesture space. Levine et al [10] define 3 areas (head, torso/arms and legs) of body movement, which we extend to include a separation of the torso and arms. Although limited in the current implementation, this distinction is intended to facilitate implementation of locution cluster and grouping analysis in future work relating to posture and body weight. A separate character set of gestures is created for the upper (head), centre (body), left/right (arms) and lower left/right (legs) areas of the gesture space. Each of these character sets is then processed separately by the plugin to allocate gestures to prominent events in the speech signal. As a result, the movement of the head can be controlled by certain vowels in specific correlation to the prominence of the signal, whilst less important movements of the feet (for example) can be allocated statistically in relation to all vowel events.

4.1 Head movements

Head movements are used in human communication and interaction [25, 26] to indicate contextual importance in speaking and attention [27] when listening. In the plugin, level 3 prominence vowels are used to allocate movements clips from a created head character set (Figure 4):

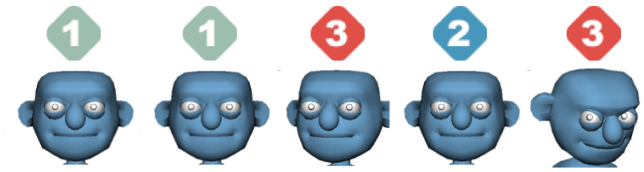


Figure 4: Allocation of head movement clips based on speech prominence.

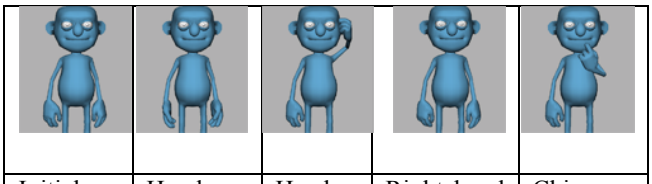
In the above example, a level 3 stress denotes a head movement whereas lower stress levels do not change the position of the head. Discussions by the production review group indicated that movements on both level 2 and 3 stress vowels were too verbose and distracting, often suggesting that the character was disturbed in some manner. For this reason, the current version of the plugin defaults to level 3 movements only. Having said this, a more robust statistical approach is planned based on vowel clustering wherein movements occur relative to locutions and locution groups as defined by Kendon.

4.2 Body Movements

Although the current system functions at the kinesic stroke level (corresponding to beats) body movements are arguably defined more broadly in terms of shifts in weight and posture that would better correspond to kinesic units. To facilitate this, shifts in body weight and posture are currently allocated relative to the position of all vowels in the speech asset. This action is performed in the plugin by allocating a value to a variable slider in the profile section of the application. The slider defines a frequency of body movement from 0 (no vowels) to 1 (all vowels) and allows the user to specify the value as part of a particular profile for a character.

4.3 Arm Movements

The prototype character contains a set of arm movements that are used during the authoring process (Figure 5):



Initial pose	Hands lift	Head scratch	Right hand close	Chin scratch

Figure 5: Example character set gestures.

The movements shown are examples of those that can be performed by the avatar. Further extension of the movement library will consider the development of metaphorics and iconics in separate character sets. Each gesture is made available at authoring, and selection is made from the list for allocation to a specific vowel as a beat (or non-gesture). The plugin allows subsets of gestures to be stored in a specific profile, such that a character may be predefined to only perform certain gestures. This facilitates scenarios where the character is not intended to move the hand to either head or mouth, or when the head is not intended to move in certain ways. Although limited in its current application, this will be extended in future work to operate in animation processes where other head tracking algorithms may be employed to prevent movements clashing (e.g. head scratch while nodding).

4.4 Leg Movements

Leg movements are focussed in the lower areas of the gesture space, they are more likely to be defined relative to metaphorics or perhaps even deictics (periphery). In addition, much of the work on gestural behaviour during speech has prioritised arm and hand movements over those relating to the rest of the body (rather than body language studies that focus on posture and whole body movement). As a result, leg and foot movements are allocated using the same frequency slider approach employed for the body character set.

5 Facial Animation

In addition to gesture allocation, prominent speech events are also used to drive various elements of the characters face. Although simple lip synchronisation based on amplitude is provided (section 5.4), this is based on input from the production review group that suggested lack of mouth movement was distracting from other aspects of the animation. Having said this, the lip synch provided is not considered to be a comprehensive attempt at proper mouth animation and thus is not included in user testing (section 6).

5.1 Eyebrow movements

In the current implementation of the plugin, a simple linear mapping between higher prominence vowels (level 2 and 3) and eyebrow movement is performed wherein the brows move the same amount for every vowel. The production review group indicated that a level 1 (and thus 2 and 3) mapping was too frequent, while a level 3 mapping was not frequent enough to suggest facial expression and so level 2 (2 and 3) was chosen. Although this simple mapping provides useful additional expression, a more complex mapping involving a correlate of pitch (and perhaps intensity) to both eyebrow position and direction (producing frowns) will be considered in future revisions of the plugin. In addition, more complex characters may also use prominence to define cheek movements in conjunction with eyebrows- while non-human rigs often have facility for ears/atenna or other virtual mandibles to be moved in this manner.

5.2 Eyeblinks

Blinking is considered to be a natural and necessary function [28], protecting the eye from foreign bodies and maintaining correct corneal moisture. Blink rate is linked to health [29], eye sensitivity [30] and reaction to alerts [31]. Doughty [32] defines spontaneous eyeblink rate (SEBR) as a common standard, though indicates that large variance of SEBR values are reported between medical studies. Variance aside, some consensus is given to an increase in SEBR from reading to

primary gaze, and increasing again during conversation (Table 1):

Task	Reading		Primary		Conversation	
	min	max	min	max	min	max
SEBR (mean)	1.4	14.4	8.0	21.0	10.5	32.5

Table 1: Overall SEBR value ranges, adapted from [32]. All values are in blinks/min.

In the above table, increases in SEBR are observed from reading (lowest) through normal gaze to conversation (highest) that would correlate broadly with expected physical behaviour of the eye for each activity. In addition, Doughty [33] distinguishes inter-eyeblink interval (IEBI) between normal (regular interval between blinks), I-type (random) and J-Type (mainly short with occasional longer intervals) and provides suggested values for all quantities (Table 2):

Characteristic	I-type	J-type	normal
SEBR(mean/std. dev)	7.5 ±2.7	10.7 ±2.5	12.3 ±2.4
IEBI (mean/std. dev)	8.7 ±2.5	5.8 ±1.6	5.1 ±1.3

Table 2: SEBR and IEBI values for different blink characteristics, adapted from [33]. All values are in seconds.

As Table 2 shows, a broad inverse correlation exists between SEBR and IEBI (as would be expected). Although Doughty indicates no absolute values can be taken from this study, a broad indication of SEBR increase depending on task (reading/gazing/conversation) and inverse variation between SEBR and IEBI depending on interval type (I-type/J-type/normal) can be made. Having said this, no temporal hierarchy to distinguish between blink interval type currently exists in the literature and thus variation of interval type requires further investigation. Indeed, the IEBI tests were performed using a single focus point in an otherwise empty room, rather than measurement while speaking, performing a task or more commonly in dialogue. As a result, the system detailed in this paper defaults to a median SEBR value for conversation (it is assumed the avatar will be delivering presentation or monologue) of 21.5 blinks/min. Although IEBI has been coded into the system, production review group testing highlighted that both normal and J-type blinks appeared repetitive after short periods of time (J-type often producing clusters of blinks). For this reason, a simple randomized I-type blink is implemented in the current system. For a static blink time (full open/close cycle) of 0.5 seconds, this specifies a randomized range of IEBI between one blink length and SEBR (0.5-2.4 seconds). Future revisions of the system will consider how best to vary blink length, SEBR and IEBI to produce greater variation and naturalness, ideally linked to the emotions and characteristics (e.g. health) of a given character.

5.3 Gaze Tracking

Various models of gaze tracking have been discussed in the literature, including statistical models [34] and hierarchical state machines [35]. Fukayama et al. [36] propose a gaze behaviour model defined by gaze ammount, duration and gaze points that correlates with patterns found in human interactions. The model relates to subjective scales for friendliness and dominance based on a 50% gaze ammount for 'like' (friendliness), upward gaze points linked to strong (dominance) and downward gaze points linked to warm (friendliness). Subjective scales for emotion and mood make direct parameterisation difficult when seeking to combine these terms with other descriptors such as the stress tagging framework delineation of the emotional dimensions of activation and evaluation [20]. Having said this, the basic model of gaze amount, duration and gaze points can still be implemented without direct recourse to these terms. The system described in this paper thus employs a root gaze point (which defaults to 50% gaze ammount) and a series of user defined gaze points that are then rendered by the system at random intervals relative to vowels in the speech act. Future revisions will consider linkage between emotional modelling and features such as gaze tracking, while the definition of gaze points may also be considered in relation to the gesture space. In addition, a correlation between eyeblink phase [37] and gaze position may also be considered both in terms of general behaviour and also as an indication of health [38] (which is correlated to mood and emotion).

5.4 Lip synchronisation

Initial lip-synchronisation was driven directly by the stress tagging framework, but missed vowels (due to the size of the detector window, dipthongs and lack of formant tracking) were immediately noticed by all member of the production review group. In practice, no untrained vowel detector for real unconstrained speech exists which performs with 100% accuracy. For this reason, jaw movements are driven directly by the input audio file in the current revision of the plugin. It is considered a preferable result to have jaw and lip movement at all times when speech is detected (even though sometimes noisy) than to run the risk of missing a vowel, which will be instantly noticed by the viewer. As no dedicated audio float is present in Maya (unlike 3DSMax), the python audioop classes have been used to create an audio float that drives jaw keys based on the input amplitude of the speech signal. To counteract noise, the jaw animation curves are smoothed using an Euler filter (Maya API filterCurve function) to minimise jittery mouth movements. Future revisions of the plug will also investigate how best to reduce the number of keyed frames without missing signal peaks (as would occur with a simple linear frame skip) so that animation time may be reduced.

Once jaw movements have been created, overlaid mouth shapes are then added at each vowel stress tag, using prominence to accentuate certain vowels of importance in the signal. This feature is implemented with cognisance of planned additions to plugin functionality that will use vowel identification (rather than merely detection) to select specific mouth shapes [39] for blending. At time of writing, the default glottal stop which relates to the most basic human

vocal utterance of the schwa [40] is mapped to a simple U shape for the character. Although lacking in detailed expression, the movement of mouth in both up/down (jaw) and lips (schwa) provides a useable first pass of lip synchronisation that is arguably sufficient for automated content. Having said this, higher production values dictate that a trained animator will ideally post-process the lip movements in commercial scenarios.

6 User Testing

Testing was performed to examine the performance of the gesture and facial animation concepts described in this paper. In the case of gesture, 2 sets of animations were produced: one set containing containing randomised gestures and another allocated based on information from the vowel stress tagging framework. To assess facial animation, a second set of animations was created for both random and stress tagged gesture sets containing additional eyebrow movements, blinks and gaze tracking (Table 3):

Gesture Type	**Random**		**Stress Tagged**	
Facial expressions	No	Yes	No	Yes

Table 3: Test design for comparison of gesture allocation and facial expressions to control conditions.

The test group contained 20 participants, all of whom had no involvement in this research. For each animation, participants were asked to evaluate 4 statements using a Likert scale:

1. The animation movements were timed appropriately
2. The animation movements were consistent with speech
3. The facial expressions were timed appropriately
4. The facial expressions were consistent with speech

The results of testing were as follows (Figure 5):

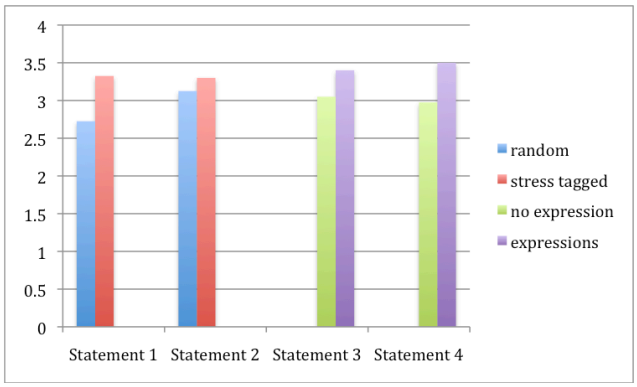


Figure 6: Average test results for each statement. Results are given for each statement organised by gesture type (questions 1&2) and facial expression (questions 3&4).

Post-test analysis of the results using T-tests gave significant differences at the $\alpha<0.01$ level for statements 1 & 4, with statement 3 being significant at the $\alpha<0.05$ level. Statement 2 (that movements were consistent with speech) was not validated with a result of $\alpha>0.25$. These results indicate that the allocation of beat and non-gesture gestures based on vowel stress levels is perceived to have greater rhythmic consonance with the acoustic patterns of speech, as is the similar addition of facial expressions and gaze tracking. Having said this, the movements were not considered to be consistent with speech due to the emphasis on vowel stress (statement 2), and indeed the lack of variance in the results suggests that test participants did not consider that the movements themselves reflected the spoken content. Indeed, all average values for responses ranged between 2.725 to 3.5, which does not suggest a particularly high performance of the animation in general. Further, the less significant variance for statement 3 (that facial expressions were timed appropriately) is also possibly due to perceived inaccuracies in the lip synchronisation performed. This finding is also borne out by feedback from professional animators [41-43] who suggest that the lack of proper lipsynch detracts from the overall gestural behaviour of the character.

7 Conclusions

The character animation system detailed in this paper provides a means of automatically assigning non-gestures and beat gestures based on the acoustic prominence of vowels in speech. The system defines character sets for each area of the gesture space, and then allocates gestures from these sets in time with the speech act. In addition, the basic facial expressions of the character including eyebrow movements, eyeblinks, gaze tracking and lip synchronisation are also added in correspondance to acoustic events. Testing of the system to determine the effect of prominence driven gesture allocation (in comparison to random gestures) and the effect of adding facial expressions demonstrated significant improvement, though overall results were not particularly high. The movements used by the character were not considered significantly consistent with speech, though the lack of iconic and metaphoric gestures may be a factor in this result.

7.1 Future Work

Future work will consider further extensions to the hierarchy of the vowel stress tagging framework to include locutions and locution groups, notably in relation to clustering for head movements and kinesic units for body posture. Head tracking algorithms will also be employed to better consider the role of clip blending between head and arm movements, notably when moving towards the inclusion of iconics and metaphorics. Eyebrow movements will also be defined in terms of pitch (to produce frowns) alongside prominence, and these variations will also be considered alongside eyeblinks in terms of the emotion and health of the character.

This inclusion of emotional modelling will also consider how best to implement more complex gaze tracking, which will also be more closely defined in terms of the gesture space. Lip synchronisation will also be addressed (considered the use of vowel classifiers to shape the mouth) while the functionality of the plugin itself will be extended to investigate the possibilities of both emotional modelling and semi-automated animation authoring.

Acknowledgements

The research leading to this paper was partially supported by the European Commission under contract IST-FP6-027122 "SALERO" and by the Enterprise Ireland Proof of Concept PC-2008-0335 "AniMA".

References

[1] J. Cassell, "Embodied conversational interface agents," *Communications of the ACM,* vol. 43, pp. 70-78, 2000.

[2] S. Kopp and I. Wachsmuth, "Synthesizing multimodal utterances for conversational agents: Research articles," *Computer Animation and Virtual Worlds,* vol. 15, pp. 39-52, 2004.

[3] M. Kipp, M. Neff, and I. Albrecht, "An annotation scheme for conversational gestures: how to economically capture timing and form," *Language Resources and Evaluation,* vol. 41, pp. 325-339, 2007.

[4] J. Cassell, H. H. Vilhjálmsson, and T. W. Bickmore, "BEAT: the Behavior Expression Animation Toolkit," *SIGGRAPH,* pp. 477-486, 2001.

[5] M. Neff, M. Kipp, I. Albrecht, and H. Seidel, "Gesture modeling and animation based on a probabilistic recreation of speaker style," *ACM Transactions on Graphics 27,* vol. 1, pp. 1-24, 2008.

[6] B. Morel, "Living Actor Avatars arrive on the scene," 2009.

[7] K. Atzwanger, K. Grammer, K. Schäfer, and A. Schmitt, *New Aspects of Human Ethology (Recent Advances in Phytochemistry)*: Springer, 1997.

[8] E. Chuang and C. Bregler, "Mood swings: expressive speech animation," *ACM Transactions on Graphics,* vol. 24, pp. 331-347, 2005.

[9] E. Ju and J. Lee, "Expressive facial gestures from motion capture data," *Computer Graphics Forum,* vol. 27, pp. 381-388, 2008.

[10] S. Levine, C. Theobalt, and V. Koltun, "Real-Time Prosody-Driven Synthesis of Body Language," in *ACM TOG (Proc. of SIGGRAPH Asia)*, Yokohama, Japan, 2009.

[11] C. Cullen, B. Vaughan, and S. Kousidis, "Emotional speech corpus construction, annotation and distribution," in *The sixth international conference on Language Resources and Evaluation, LREC 2008* Marrakech, Morocco, 2008.

[12] C. Cullen, P. McGloin, A. Deegan, E. McCarthy, and C. Goodman, "Reusable, Interactive, Multilingual Online Avatars," in *The 6th European Conference on Visual Media Production (CVMP)* London, England, 2009.

[13] D. McNeill, *Hand and mind: what gestures reveal about thought*. Chicago: University of Chicago Press, 1996.

[14] J. Cassell, H. H. Vilhjalmsson, and T. Bickmore, "BEAT: the Behavior Expression Animation Toolkit," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*: ACM, 2001, pp. 477-486.

[15] G. Hu and D. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, pp. 396-405, 2007.

[16] P. Boersma, "Should jitter be measured by peak picking or by waveform matching?," *INTERNational Journal of Phoniatrics, Speech Therapy and Communication Pathology,* vol. 61, pp. 305-308, 2009.

[17] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *Journal of the Acoustic Society of America,* vol. 93, pp. 1097-1108, 1993.

[18] M. Schröder and J. Trouvain, "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching," *International Journal of Speech Technology,* pp. 365-377, 2003.

[19] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities," *Acta Otolargngologica,* vol. 90, pp. 441-451, 1980.

[20] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication Special Issue on Speech and Emotion,* vol. 40, pp. 5-32, 2003.

[21] C. Cullen, B. Vaughan, S. Kousidis, E. McCarthy, and J. McAuley, " CorpVis: An Online Emotional Speech Corpora Visualisation Interface," in *4th International Conference on Semantic and Digital Media Technologies (SAMT)*, Graz, Austria, 2009.

[22] C. Cullen, B. Vaughan, S. Kousidis, and F. Reilly, "A vowel-stress emotional speech analysis method," in *CITSA* Genoa, Italy, 2008.

[23] A. Kendon, "Human gesture," in *Tools, Language and Cognition in Human Evolution*, K. R. Gibson and T. Ingold, Eds. Cambridge: Cambridge University Press, 1993, pp. 43-62.

[24] C. Cullen, B. Vaughan, and S. Kousidis, "Emotional speech corpus construction, annotation and distribution," in *The sixth international conference on Language Resources and Evaluation, LREC*, Marrakech, Morocco, 2008.

[25] J. K. Lee and C. Breazeal, "Human social response toward humanoid robot's head and facial features," in *Proceedings of the 28th International conference on Human factors in computing systems*, Atlanta, Georgia, USA, 2010, pp. 4237-4242.

[26] C. Goodwin, "Action and embodiment within situated human interaction," *Journal of Pragmatics,* vol. 32, pp. 1489-1522, 2000.

[27] U. Hadar, T. J. Steiner, E. C. Grant, and F. C. Rose, "Kinematics of head movements accompanying speech during conversation," *Human Movement Science,* vol. 2, pp. 35-46, 1983.

[28] E. Ponder and W. Kennedy, " On the act of blinking," *Quarterly Journal Experimental Physiology,* pp. 89-110, 1927.

[29] L. Colzato, W. v. d. Wildenberg, and B. Hommel, "Reduced Spontaneous Eye Blink Rates in Recreational Cocaine Users: Evidence for Dopaminergic Hypoactivity," *Public Library of Science (PLoS ONE),* vol. 3, 2008.

[30] M. Collins, R. Seeto, L. Campbell, and M. Ros, "Blinking and corneal sensitivity," *Acta Ophthalmologica,* vol. 67, pp. 525-531, 2009.

[31] A. Schulza, J. Lass-Hennemanna, S. Richtera, S. Römera, T. D. Blumenthalb, and H. Schächinger, "Lateralization effects on the cardiac modulation of acoustic startle eye blink," *Biological Psychology,* vol. 80, pp. 287-291, 2009.

[32] M. Doughty, "Consideration of Three Types of Spontaneous Eyeblink Activity in Normal Humans: during Reading and Video Display Terminal Use, in Primary Gaze, and while in Conversation," *Optometry and Vision Science,* vol. 78, pp. 712-725, 2001.

[33] M. Doughty, "Further Assessment of Gender- and Blink Pattern-Related Differences in the Spontaneous Eyeblink Activity in Primary Gaze in Young Adult Humans," *Optometry and Vision Science,* vol. 79, pp. 439-447, 2002.

[34] C. Pelachaud and M. Bilvi, "Modelling gaze behavior for conversational agents," in *Proceedings of Intelligent Virtual Agents (IVA)*, Kloster Irsee, Germany, 2003.

[35] A. Colburn, M. Cohen, and S. Drucker, "The Role of Eye Gaze in Avatar Mediated Conversational Interfaces," 2000.

[36] A. Fukayama, T. Ohno, N. Mukawa, M. Sawaki, and N. Hagita, "Messages embedded in gaze of interface agents - impression management with agent's gaze," in *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*, Minneapolis, Minnesota, USA, 2002, pp. 41-48.

[37] F. VanderWerf, P. Brassinga, D. Reits, M. Aramideh, and B. O. d. Visser, "Eyelid Movements: Behavioral Studies of Blinking in Humans Under Different Stimulus Conditions," *Journal of Neurophysiology,* vol. 89, pp. 2784-2796, 2003.

[38] H. A. Slagtera, R. J. Davidson, and R. Tomer, "Eye-blink rate predicts individual differences in pseudoneglect," *Neuropsychologia,* vol. 48, pp. 1265-1268, 2010.

[39] R. Williams, *The Animator's Survival Kit*. London: Faber & Faber, 2002.

[40] J. Pickett, *The Acoustics of Speech Communication; Fundamentals, Speech Perception Theory, and Technology*. Boston, MA: Allyn and Bacon, 1999.

[41] N. Marsden, "Automated character animation prototype," London: Hibbert Ralph, 2010.

[42] C. Goodman, "Vowel stress driven character animation," London: Pepper's Ghost Productions, 2010.

[43] S. McDonnell, "Online character animation tools," Dublin: Jam Media, 2010.