Virtual Character Performance From Speech

Stacy Marsella* Yuyu Xu† Margaux Lhommet‡ Andrew Feng§ Stefan Scherer¶ Ari Shapiro‖

Figure 1: *Our method can synthesize a virtual character performance from only an audio signal and a transcription of its word content. The character will perform semantically appropriate facial expressions and body movements that include gestures, lip synchronization to speech, head movements, saccadic eye movements, blinks and so forth. Our method can be used in various applications, such as previsualization tools, conversational agents, NPCs in video games, and avatars for interactive applications.*

Abstract

We demonstrate a method for generating a 3D virtual character performance from the audio signal by inferring the acoustic and semantic properties of the utterance. Through a prosodic analysis of the acoustic signal, we perform an analysis for stress and pitch, relate it to the spoken words and identify the agitation state. Our rule-based system performs a shallow analysis of the utterance text to determine its semantic, pragmatic and rhetorical content. Based on these analyses, the system generates facial expressions and behaviors including head movements, eye saccades, gestures, blinks and gazes. Our technique is able to synthesize the performance and generate novel gesture animations based on coarticulation with other closely scheduled animations. Because our method utilizes semantics in addition to prosody, we are able to generate virtual character performances that are more appropriate than methods that use only prosody. We perform a study that shows that our technique outperforms methods that use prosody alone.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation I.6.8 [Simulation and Modeling]: Types of Simulation—Animation;

Keywords: animation, gestures, behavior, conversational agent

1 Introduction

The flip of a hand, a raising of an eyebrow, a gaze shift: the physical, nonverbal behaviors that accompany speech convey a wide variety of information that powerfully influences face-to-face interactions. A nod can convey agreement, a gesture can emphasize a point and facial expressions can convey emotions. A speaker's aversion of gaze reflects they are thinking, in essence regulating cognitive load as they consider what to say next while also signaling they want to hold onto the dialog turn [Argyle and Cook 1976; Bavelas 1994]. Nonverbal behaviors are so pervasive in every moment of the dialog that their absence also signals information - that something is wrong, for example, about the physical health or mental state of the person.

Our interest in such behaviors lies in a desire to automate the selection and generation of nonverbal behavior for convincing, life-like virtual character performances. Specifically, in this paper we discuss the automatic generation of a character's nonverbal behavior from the audio of the dialog they must speak.

Whether or not through automatic means, the creation of a character's nonverbal behavior faces several challenges. Most fundamentally is the question of what behaviors to exhibit and when to exhibit them. The relation between nonverbal behavior and speech is complex. Nonverbals can stand in different, critical relations to the verbal content, providing information that embellishes, substitutes for and even contradicts the information provided verbally (e.g., [Ekman and Friesen 1969; Kendon 2000]). The form of these behaviors is often tied to physical metaphors; the rejection of an idea can be illustrated by a sideways flip of the hand that suggests discarding an object [Calbris 2011]. Nonverbal behaviors also serve a variety of rhetorical functions. Shifts in topic can be cued by shifts in posture or shifts in head pose. Comparison and contrasts between abstract ideas can be emphasized by abstract deictic (pointing) gestures that point at the opposing ideas as if they each had a distinct physical locus in space [McNeill 1992]. A wide range of mental states and character traits can be conveyed: gaze reveals thought processes, blushing suggests shyness and facial expressions intentionally or unintentionally convey emotions and attitudes. Finally, nonverbal behavior helps manage conversation, for example by signaling the desire to hold onto, get or hand over the dialog turn [Bavelas 1994].

Further the mapping between these communicative functions and the behaviors that realize them is many-to-many. Parts of the utterance can be emphasized using a hand gesture, a nod or eyebrow raise. On the other hand, a nod can be used for affirmation, emphasis or to hand over the dialog turn [Kendon 2002; McClave 2000]. The context in which the behavior occurs can transform the interpretation, as can even subtle changes in the dynamics of the behavior: head nods signaling affirmation versus emphasis typically have

*e-mail:marsella@ict.usc.edu
†yxu@ict.usc.edu
‡lhommet@ict.usc.edu
§feng@ict.usc.edu
¶scherer@ict.usc.edu
‖shapiro@ict.usc.edu

different dynamics. Moreover, behaviors can be composed with each other, further transforming their interpretation.

The generation of the nonverbal behaviors must additionally take into account that they are synchronized, often tightly, with the dialog and changes in this synchronization can lead to significant changes in what is conveyed to a listener. For instance, the stroke of a hand gesture, a nod or eyebrow raise performed individually or together are often used to emphasize the significance of a word or phrase in the speech. To achieve that emphasis the behavior must be closely synchronized with the utterance of the associated words being emphasized. Alteration of the timing will change what words are being emphasized and consequently change how the utterance is understood.

Such challenges make the pattern and timing of the behavior animations that accompany utterances unique to the utterance and the state of the character. Manual creation of the behaviors by hand animation and/or motion capture are consequently time consuming and costly, as well as requiring considerable expertise from the animator or the motion capture performer.

This has led us to research and develop an automatic method to generate expressive, life-like nonverbal behaviors. We have developed a technique that produces a character's speaking behavior through the use of acoustic, syntactic, semantic, pragmatic and rhetorical analyses of the utterance. These analyses seek to infer the communicative function of the utterance, including the communicative intent of the utterance as well as the underlying emotional and mental states of the speaker. The result of these analyses are then mapped to nonverbal behaviors, including head movements, facial expressions, gaze and gestures, that are composed and co-articulated into a final performance by an animation engine. By composing the overall performance for an utterance from behavioral components, we can get a wide range of unique, expressive performances, attuned to each utterance, while using a limited set of behavioral components. The quality of the result stems from a combination of several contributions that distinguish it from previous efforts. This includes the range and depths of analyses of the utterance text and audio, spanning prosodic, syntactic and semantic inferences, that have been be incorporated in a fully automated approach as well as an emphasis on a comprehensive approach that can generate the full range of synchronized nonverbal behaviors, spanning face, gesture, head movement, gaze and posture, needed to create a realistic performance.

In this paper, we discuss the approach we have taken and illustrate its effectiveness in a variety of applications: the use as an embodied conversational agent, the use inside a previsualization tool for film and television, the use for non-player characters (NPCs) in video games, and the use as an avatar for interactive applications. We also report on an evaluation study using human subjects that validate the success of our method.

Contributions A central contribution of this work is the deep and novel types of analysis incorporated in a comprehensive, automated approach that can generate a full range of nonverbal behaviors. Prosodic analysis is used to determine patterns of emphasis and overall emotional arousal that will drive when nonverbal behavior will happen, the style of the behavior and overall quantity of behavior. Syntactic and semantic analyses go on to undertake analysis, that for example detect the common use of metaphors in the language to drive selection of metaphoric gestures (e.g., "time as a moving object" or "abstract idea as a physical object" that allow them to have physical properties like "a big idea" that can be conveyed by gesture), the use of rhetorical structures like comparisons and contrasts that suggest abstract deictic gestures (e.g., this idea as opposed to that idea can be conveyed gesturing left than right), common ways of signifying affirmation, quantification and negation, to name a few of the analysis. All of these analysis can have nonverbal correlates.

Another important achievement of this work is the close coupling of the nonverbal behavior generation to a high quality virtual character animation system that can address key requirements for realistic nonverbal behavior generation such as the co-articulation of gestures and the use of gestural holds to provide emphasis.

2 Related Work

Previous work has differed in terms of degree of automation, depth of analysis or range of co-verbal and listening behavior generated. Our method generates a 3D virtual character performance based on an audio signal. Thus, our method shares similar goals to works that generates head movements [Busso et al. 2007] or gestures based on utterances or audio signals. [Levine et al. 2009] uses prosody-based features extracted from audio to train hidden Markov Models to generate appropriate gesture. Their later work [Levine et al. 2010] performed real time generation of gestures including word spotting (you, me). The major limitation of their work is that the synthesized gesture may not match the particular semantic content in the speech since the gesture generation is mainly based on prosody. Our method performs syntactic and semantic analysis to synthesize gestures that match the context of conversations. [Stone et al. 2004] used mocap segments that correspond to pre-recorded phrases and rearrange them to match the new sentences. Therefore their results are restricted to specific domain of the recorded phrases while our method can generate gestures for arbitrary input text. Similar works require some degree of manual annotation of the speaker's utterance long with manual additions to the system's knowledge to handle those utterances [DeCarlo et al. 2004].

ACE [Kopp et al. 2003] is a rule-based system that focuses on deictic and iconic gestures. The virtual human reads the text input and looks for specific words in order to display associated gestures in right-time using prosody analysis. [Kipp et al. 2007b] introduced a system that generates gestures in the style of the input motion based on probabilistic reproduction of captured subject, and an extension that inclusions dynamics [Neff et al. 2008]. They annotated the training video of a specific subject using an annotation tool called ANVIL [Kipp et al. 2007a] and learned a probabilistic model to recreate gestures for new utterances. Therefore, their result can only represent the emotional state for that particular speaker while our rule-based method can adapt to different emotional states and characters.

Alternatively, there is work on overall nonverbal behavior generation but using limited forms of analyses such as detecting individual keywords [Bergmann and Kopp 2009; Lee and Marsella 2010] or using textual analysis to derive factors like rheme (or focus) and theme (or topic) to determine emphasis [Cassell et al. 2001]. Extending this framework, the Nonverbal Behavior Generator (NVBG) [Lee and Marsella 2006] is a rule-based system that uses the communicative intent embedded in the surface text as well as information on the agent's cognitive processing, such as its internal goals and emotional state to generate a range of nonverbal behaviors.

Expressive facial animation driven by audio has been explored in various works. Cassell [Cassell et al. 1994] developed an automatic rule-based system to animate faces. [Brand 1999] reconstructed video images with facial expressions from audio. [Deng et al. 2006] learned models for speech and expression then blended them to synthesize an expressive face. [Cao et al. 2003] extracted expressive components with ICA then synthesized a facial performance with a novel utterance. [Albrecht et al. 2002] extracted prosodic features

from the speech signal to generate facial expressions. [Li and Shum 2006] uses HMM models to learn mappings from audio. [Chuang and Bregler 2005] learns speech and emotional content using a bilinear model from video, and generates head movements from audio pitch. [Ju and Lee 2008] generates expressive facial motions from motion capture data using stochastic methods. Our method is primarily concerned with concurrent generation of facial expressions, gestures and head movements, rather than in the generation high-quality facial movements. However, the incorporation of some of the above methods could be used in combination with our technique to improve the facial performance.

Our animation system [Shapiro 2011] translates Behavior Markup Language (BML) instructions into animated performances, similar to other BML realizers [Niewiadomski et al. 2009; van Welbergen et al. 2009; Thiebaux et al. 2008; Heloir and Kipp 2009].

3 System Overview

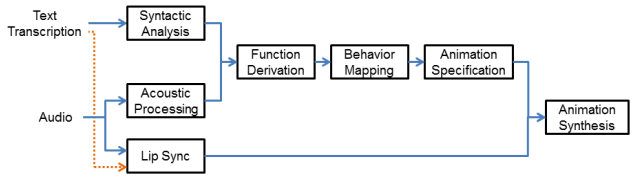

Figure 2: *Overview of the our method. Our system accepts an audio signal and a transcription of the audio as input and generates an animated performance as output.*

An overview of our system can be seen in Figure 2. To generate the animated performance, the following process occurs:

1. **Acoustic Processing:** The sentence audio is acoustically analyzed to derive information on what words are being stressed and the overall agitation of the speaker which is then interpreted as sad, neutral, agitated/excited (see 3.1).
2. **Syntactic Analysis:** The sentence text is then parsed to derive the syntactic structure (see 3.2).
3. **Function Derivation:** The function analysis phase infers aspects of the utterance's communicative functions using forward-chaining inference rules to build up a hierarchically structured lexical, semantic and pragmatic analysis. Examples of these communicative functions include affirmation and emphasis (see 3.3).
4. **Behavior Mapping:** Our method then goes through a behavior analysis stage, in which a set of nonverbal behavior rules map from communicative functions to classes of nonverbal behaviors (see 3.4).
5. **Animation Specification:** The animation specification phase then maps those behavior classes to specific behaviors. Character specific mappings can be designed to support individual differences including personality, culture, gender and body types. The final result is a schedule of behaviors (see 3.5).
6. **Animation Synthesis:** The animation engine processes the schedule of behaviors and synthesizes the performance. The animation system might drop, shorten, or coarticulate various movements based on the given constraints (see 3.8).

The acoustic and syntactic results of Steps 1 and 2 initiate a knowledge base used in subsequent steps. Starting from that initial knowledge base, Steps 3 - 5 use a knowledge based inference engine that employs pattern-based invocation of rules to build an increasingly deeper analysis of the communicative function of the utterance and then transforms those communicative functions into actual behaviors that will effectively convey these functions. Conflict resolution occurs at several phases in the overall process. For example, if there are two or more rules overlapping with each other causing conflict, our method resolves the conflict by filtering out the rule with lower priority. The priority value of rules has been set through a study of human behaviors using video corpora.

In the following sections, we discuss these processing steps in more detail.

3.1 Acoustic Processing

A key component for driving the behavior generation is the detection and assessment of the overall agitation level (high/mid/low) for a given utterance. We analyze the voice quality, which refers to the coloring or timbre of the voice [Laver 1980], on a tense to lax dimension. We associate tense speech to high agitation, modal speech to mid level agitation, and lax speech to low agitation. For the recognition of the voice quality we employ fuzzy-input fuzzy-output support vector machines (F^2SVM) as in [Scherer et al. 2013; Kane et al. 2013] and standard Mel frequency cepstral coefficients ($mfcc$).

Further, to inform the virtual human's behavior generation about which parts of the speech and the analyzed utterances are emphasized, we employ a simple algorithm to detect word prominence and stress in speech, similarly to the more complex algorithm introduced in [Mishra et al. 2012]. The algorithm is based on the fundamental frequency (f_0) of the voiced parts and the audio signal's intensity.

3.1.1 Mel frequency cepstral coefficients ($mfcc$)

We extract 12 coefficients capturing the overall spectral information from the speech signal using Hanning windowed 32 ms frames with a 10 ms shift (i.e. 100Hz sample rate)[Davis and Mermelstein 1980].

3.1.2 Energy in dB (e_{dB}):

The energy of each speech frame is calculated on 32 ms windows with a shift of 10 ms. This speech window $w(t)$ is filtered with a hamming window and the energy

$$e(t) = \sum_{i=1}^{|w(t)|} w_i(t)^2 \quad (1)$$

is calculated and converted to the dB-scale

$$e_{dB}(t) = 10 \cdot \log_{10}(e(t)). \quad (2)$$

3.1.3 Fundamental frequency (f_0)

We use the method in [Drugman and Abeer 2011] for f_0 tracking based on residual harmonics, which is especially suitable in noisy conditions. The residual signal $r(t)$ is calculated from the speech signal $s(t)$ for each frame using inverse filtering. This process removes strong influences of noise and vocal tract resonances. For each $r(t)$ the amplitude spectrum $E(f)$ is computed, showing peaks for the harmonics of f_0, the fundamental frequency. Then,

the summation of residual harmonics (SRH) is computed as follows:

$$SRH(f) = E(f) + \sum_{k=2}^{N_{harm}} [E(k \cdot f) - E((k - \frac{1}{2}) \cdot f)], \quad (3)$$

for $f \in [f_{0,\min}, f_{0,\max}]$, with $f_{0,min} = 20$ and $f_{0,max} = 500$. The frequency f for which $SRH(f)$ is maximal is considered the fundamental frequency of this frame. By using a simple threshold, the unvoiced frames are discarded as in [Drugman and Abeer 2011].

3.1.4 Agitation level detection

For the agitation level detection, we utilize F^{2}SVM, which have previously shown to robustly detect voice quality on a tense to lax dimension [Scherer et al. 2013], where accuracies of up to about 90% are observed. The F^{2}SVM produce soft outputs with predicted memberships m_i assigned over multiple classes for each presented sample. The F^{2}SVM are trained using a speech corpus recorded by CereProc for their speech synthesis product CereVoice. The CereVoice speech synthesis system uses sub-corpora of neutral (modal), tense and lax voice quality data in order to produce subtle changes in emotion [Aylett and Pidcock 2007; Aylett et al. 2013]. These sub-corpora have been recorded over a five year period across several languages, and covering different accents of English. We compute the median $mfcc$ over each available utterance (about 4000) and train the F^{2}SVM.

In order to detect the agitation level, we compute the median $mfcc$ value for an unseen utterance and choose the class (i.e. high/mid/low) with maximal membership value assigned $\arg\max_i(m_i)$ by the trained F^{2}SVM as the observed agitation level.

3.1.5 Word stress recognition

In order to identify the parts of speech that are stressed we set thresholds $\theta(f_0)$ and $\theta(e_{dB})$ as the 90th percentile of both the observed signal intensity (e_{dB}) and the fundamental frequency f_0 for each utterance individually. For each analyzed feature frame we check if it is larger than one of the thresholds and consider this 10 ms frame as stressed. Based on the word timings for the analyzed utterances we calculate a *stressed ratio* $\in [0, 1]$ for each word (i.e. the amount of stressed frames divided by the total length of the word). This ratio is then passed to the behavior generation system for further processing.

3.2 Syntactic Analysis

After the audio processing, sentence text is parsed to derive its syntactic structure and that information is added to the knowledge base. This serves two purposes. Most importantly, it facilitates subsequent analysis of the communicative function at the level of the syntactic substructures. For example, an entire noun phrase can be identified in subsequent steps as requiring emphasis and a behavior such as nodding can be synchronized to span that phrase. Another example of this is to identify the functional relation between substructures, such as connective phrases like "as opposed to" that can relate two clauses. Some behaviors, like small head movements, tend to be correlated to syntactic structures [Lee and Marsella 2010]. One of the challenges of parsing utterances is that most parsers openly available assume complete, grammatically correct sentences which is typically not the case for natural spoken dialog. We have explored the use of several different parsing technologies (e.g., [Klein and Manning 2003; Lafferty et al. 1992; Charniak 2000]) designed for text but we continue to explore alternatives. The results reported here use the parser developed by Eugene Charniak [Charniak 2000].

3.3 Function Derivation

The function derivation phase derives the meaning for the utterance's elements relevant to nonverbal behavior. Specifically, the goal is targeted to identifying lexical, syntactic and rhetorical structures closely tied to nonverbal behavior as opposed to attempting to derive the full semantic analysis of the utterance. The basic classes of functional analyses currently performed by the system are listed in Table 1. As noted above, the function derivation process relies on pattern-matching rule invocation that uses forward chaining to build up hierarchical interpretations starting from the text, the syntactic structure and word stress patterns stored in the knowledge base during steps 1 and 2.

For example the analysis of a phrase such as "a lot more important" is built up incrementally from the basic functional classes in Table refFcnTable. First, lexical analysis rules identify the word pattern of "a lot" as an instance of positive *Quantification* and "more important" as a positive instance of the *Comparison*, resulting in that knowledge being added to the knowledge base. The addition of that knowledge would in turn allow the match of the rule that identifies the *strong positive comparative*, asserting that knowledge. Figure 3 shows this rule (simplified somewhat to facilitate exposition). Variables are denoted by dollar signs ($). The statements in the "foreach" part of the rule (aka the left hand side) matches knowledge in the knowledge base while the "assert" part adds the knowledge resulting from the unification of the variables.

```
semantics_strong_positive_comparative
 foreach
   fcn(quantifier,positive,$start1,$end1,$priority1)
   fcn(comparative,positive,$end1,$end2,$priority2)
   $newPri = increase_priority($priority1,$priority2)
 assert
   fcn(comparative,strong_positive,$start1,$end2,$newPri)
```

Figure 3: *Strong comparative detection rule. If all the foreach statements are true, i.e., match existing facts in the knowledge base, the assert statements will be added to the knowledge base. Dollar signs identify match variables. The increase_priority is a function that will increase the priority of strong_positive_comparative over its components.*

There are several benefits to this pattern-based, forward chaining of rules. First, it reduces the combinatorics. For example, consider the later example. There are multiple forms of quantifiers and multiple ways of specifying comparatives. We do not want to explicitly enumerate all possible combinations of quantifiers and comparatives that express strong comparatives but rather rely on rule-based forward-chaining to avoid such explicit enumeration. Further because the pattern matching supports variable unification and can match any knowledge in the knowledge base, it is straightforward to tie together knowledge from different sources in general ways. For example, an emphasis rule increases the priority of any communicative function that begins or ends with a word that the audio processing has determined as being stressed.

Another functional class concerns identifying relations that are based in physical metaphors. For example, words like "further" versus "a lot" can be used in similar ways, for example in comparisons, but also suggest different physical gestures, marking distance as opposed to cardinality. Specifically a gesture that sweeps away from the speaker conveys "further" effectively. As a final example, consider an example from the Mental State class, dysfluency. Dysfluency is detected, for example, if the audio contains filled pauses

such as “um..” or “er..” or repeated words. This indicates that the character is deep in thought, which in turn triggers a rule that requests the character reduce cognitive load by gaze aversions using saccadic eye motion..

It sometimes occurs that functions conflict with each other. For example, the functional analysis phase may infer a comparative, as well as that the same word should be heavily emphasized. So as part of the function derivation phase, each function inferred is assigned with a priority based on the class of the function and whether the words it spans are stressed. Then these priorities are used to resolve conflicts between overlapping functions with lower priorities being dropped.

Currently, the Function Derivation phase employs 91 rules along with a dictionary of 170 words and phrases employed specifically by the lexical analysis rules. Increasing the size of this dictionary is the simplest, most straightforward way to extend the capabilities of the system. For example, we could replace this dictionary with a link to dictionary systems like Wordnet[Miller 1995] to identify synonymous words and phrases.

Table 1: *Function Derivation Classes*

Function Class	Description	Examples
Affirmation	Agree or accept	okay, yes
Negation	Negate or reject	not, no
Interrogative	Direct or indirect questions	what, where, when
Contrast	Phrase level contrasts	aside from, other than
Process	Process denote & state	stop, continue, done
Quantification	Set cardinality	a lot, none, everything
Comparison	Comparatives & superlatives	better, best
Spatial relation	Actual or metaphor	beyond, further
Physical relation	Actual or metaphor	longer, larger
Mental state	Cognitive & emotional states	uncertainty
Performance factor	Cognitive load & processes	dysfluency, word search
Deixis	Abstract or concrete pointing	those people
Modal	Likelihood, ability, permission and obligation	have to, must, should
Emphasis	Rhythmic/prosodic	stressed words/phrases
Intensifier	Adverbs of degree	very, extremely
Valence	Emotion/valence words	awesome, awful

3.4 Behavior Mapping

The behavior mapping process takes each function derived by the previous phase and maps it to a set of alternative sequences of behavioral types. The alternatives allow variability in the character’s behavior from one utterance to the next, as well as specialization by character. Allowing for temporally synchronized sequences of behaviors to realize a communicative function permits a schedule of multiple behaviors to realize a function. For example, a strong positive comparative function might be realized by a synchronized beat gesture, nodding of the head and an eyebrow lift as well as any subset of those behaviors. There are currently 97 function-to-behavior mapping rules, each of which can suggest multiple alternative behavior sequences to realize the function - providing for a rich behavioral space. An example of this mapping between functional classes and behaviors is shown in Table 2. Many behaviors can be mapped to the same functional class. For example, a functional class may be realized either by a gesture, a head nod or a gaze aversion or combination thereof. Our system chooses one of the mapped behaviors randomly to provide variation in the performance. In addition, heuristics are implemented in order to prevent inordinate amounts of hand switching while gesturing, or overly repetitive activations of the same behavior.

The agitation state determined from the audio signal analysis affects the rules in the behavior mapping phase. For example, characters in the low agitation state, which can be correlated to sad or lethargic, tend to move their heads from side to side instead of front to back. Characters in the high agitation state, which can be correlated to angry or highly energetic, would tend to emphasize words with a beat, rather than the subtler eyebrow raise.

Table 2: *Behavior Mapping*

Function Class	Example Behaviors
Affirmation	big nod, tilt left nod, tilt right nod
Negation	gesture negation, shake, small shake
Interogative	gesture question, brow raise
Contrast	gesture contrast, tilt right, tilt left
Comparison	gesture comparative, gesture comparative longer, gesture comparative bigger
Mental state	tilt half nod left, shrug
Performance factor	gaze aversion upright, shortblink
Deixis	gesture point left, gesture point right, gaze left, gaze right
Emphasis	gesture rhythm, small nod, beat gesture
Intensifier	brow frown, half nod

3.5 Animation Specification

Behaviors are mapped to the BML language, a high-level XML language that describes a behavior and an execution schedule [Kopp et al. 2006]. Behaviors that can be specified include: head movements along each X, Y, or Z axes, gazing at objects in the virtual scene or angular offsets from them, blinking, saccadic eye movements, gesturing including deictics (pointing), facial expressions and speech. Behaviors are specified with start and end times such that they correspond to start or endings of words or of other behaviors. The purpose of the BML layer is to provide an abstraction to the animation system. Thus, a behavior would be specified like “gesture to the left starting at .34 and ending at 1.2” or “nod your head slightly starting at 3.4 and ending at 3.8”. The animation synthesis system must then interpret these high-level instructions to synthesize the final motion. Such details are explained in Section 3.8. There are currently 101 default behavior-to-BML rules but in addition there may be character specific ones added to override those defaults.

As each function gets mapped to its corresponding behavior, additional conflicts appear due to timing and implementation. Such conflicts are not detectable during the function derivation phase, and become clearer when concrete behaviors are derived from their original abstraction. Here conflicts involve either overlapping of behaviors or behaviors that are too close in time for the resulting animation to be realized smoothly. Thus, to reduce conflicts, behaviors with higher priorities are retained while lower priority ones are removed.

3.6 Listener Feedback

In addition to handling the behaviors of the speaker, our system implements a listener feedback pipeline. Currently, a listening character performs mirroring of head movement behaviors of the speaking character with a .5 second delay. As such, our method handles

a limited form of generic feedback driven by speaker's nonverbal behaviors. This could easily be extended to other forms of generic feedback such as nodding in response to the pauses in the speaker's utterance. It does not handle specific feedback which is feedback driven by the characters unfolding interpretation of, and reaction to, the speaker's utterance. Specific feedback would require a deeper understanding of the utterance than our method handles with its shallow parse of the utterance. Nonetheless, we find simple listener feedback rules are a positive addition to the performance when more than one character is present in the scene. See [Wang et al. 2013] for an approach to handling specific feedback.

3.7 Knowledge Encoding

The knowledge encoded in the system represent a multi-year effort, exploring several approaches to encoding the knowledge used in the function derivation and behavior mapping rule sets. Initially, an extensive literature review of the research on nonverbal behavior was undertaken. This seeded the development of rules encoding the function derivation and behavior mapping rules. Also, videos of real human face-to-face interactions have been annotated and analyzed to verify the rule knowledge, embellish knowledge with dynamic information about behaviors and develop the priority system used to resolve conflicts between behavior suggestions. This annotation and analysis was critical because existing literature says little about dynamics of behaviors and further conflict resolution was needed to resolve potential conflicts both between the behaviors suggested by the rules as well as differences across literature sources. One can characterize this approach as a *expert knowledge plus semi-automated analysis* approach. More recently, pure data-driven machine learning techniques have been used as a way to validate the features used in the rules, including Hidden Markov Models and Latent-Dynamic Conditional Random Fields to learn the mapping between features of an utterance and nonverbal behaviors, using annotated human face-to-face interactions. See [Anonymous].

3.8 Animation Synthesis

The scheduled behaviors that are derived from the Behavior Mapping stage (see 3.4) are then interpreted and processed by our animation synthesis system into an animated performance for the entire body of a virtual character. The animation system synthesizes gesture motions, lip-synced mouth animations, and other auxiliary motions such as head movements, eye darts, gazes and emotive facial expressions. These motions are processed in stages by a series of controllers which synthesize motion over sets of joints in the character's skeletal hierarchy. It uses a control hierarchy which progresses from gross movements that involve the entire body, to smaller movements that only override a small number of joints, such as the eye joints. In the following order, our animation component evaluates the underlying idle movement, gestures, gazing, head movements, facial expressions, and finally eye and eyelid motion.

Gestures

Gestures animations are derived from motion data either hand-constructed by an animator, or generated from motion capture. We have identified approximately 29 types of gestures, both single and double handed for each agitation state. The single handed versions were mirrored to generate an equivalent gesture on the other hand. Our male and female characters each have different gesture sets to reflect different styles of movement. The amount of gesture data required for our method is similar in number to other gesture synthesis implementations (eg. [Neff et al. 2008]).

Each agitation state (low, medium or high as described in Section 3.1.4) utilizes a different gesture clip for each gesture category. For example, a beat gesture in the high agitation state is faster and is performed with more energy than the beat gesture from the medium agitation state.

A typical gesture animation has seven time markers: start time t_s, ready time t_{rd}, stroke start time t_{ss}, stroke time t_{st}, stroke end time t_{se}, relax time t_{rl} and end time t_e which are manually annotated by a digital artist or gesture expert. The markers are organized into five functional phases as shown in Table 3:

Table 3: *Gesture Phases*

Gesture Phase	Description
t_s to t_{rd}	preparation of a gesture
t_{rd} to t_{ss}	pre-stroke hold period
t_{ss} to t_{se}	stroke phase, t_{st} indicates emphasis point of the gesture
t_{se} to t_{rl}	post-stroke hold period
t_{rl} to t_e	retraction phase

Each gesture behavior is specified in BML. The output from animation specification stage contains a series of n gesture units that synchronize with the input utterances:

$$G_i = (c^i, {t_{gst}}^i, {t_{grl}}^i, pr^i) \tag{4}$$

Each gesture unit $G_i$includes a gesture category c^i, timings for desired stroke strike time ${t_{gst}}^i$ and desired relax time ${t_{grl}}^i$, and a priority value pr^i that indicates the importance of this gesture. For each G_i, we need to select a gesture animation A_i from the database based on its category c^i. The selected animation must adhere to the constraint:

$${t_{rl}}^i - {t_{st}}^i \leq {t_{grl}}^i - {t_{gst}}^i \tag{5}$$

which indicates that the time span of the gesture's stroke to post stroke hold duration fits within the time duration indicated by the desired stroke and relax times from Equation 4. If the animation violates the timing constraint, the gesture G_i is discarded and does not participate in the motion synthesis. Additionally, we compute the velocity v^i of hand movements for the stroke phase of A_i. We then align each A_i with the desired timing in G_i by offsetting ${t_{st}}^i$ to ${t_{gst}}^i$ and also extending the post-stroke gesture hold period by shifting ${t_{rl}}^i$ to ${t_{grl}}^i$. This planning step schedules all gesture animations on the timeline.

There could be timing conflicts between some adjacent gesture animations. Our approach chooses not to timewarp the gestures but to handle conflicts between overlapping or adjacent gestures using the following heuristics to prune or concatenate gestures:

- If ${t_{ss}}^{k+1} < {t_{grl}}^k$, the two gesture animations could not be played in sequence, lest the stroke phase of the A_{k+1} must be truncated. Since changing the stroke phase would change the meaning of the gesture, we remove the gesture animation with the lower priority of the two from the schedule.
- If ${t_{ss}}^{k+1} > {t_{grl}}^k$, the two gesture can be played in sequence together. We create a blending transition from ${t_{grl}}^k$ in A_k to ${t_{ss}}^{k+1}$ in A_{k+1} and remove the portion $t < {t_{ss}}^{k+1}$ in A_{k+1} from the timeline. This way the gesture animation A_{k+1} would start directly from the stroke start point, instead of from the idle pose.

We also compute the transition velocity $v_{k \to k+1}$ using the time interval ${t_{ss}}^{k+1} - {t_{grl}}^k$ and distance between hand positions in ${t_{grl}}^k$ and ${t_{ss}}^{k+1}$.

- If $v^{k \to k+1} - v^{k+1} > v^{th}$, where v^{th} is the gesture blend speed threshold, then there is not enough time to blend from A_k to A_{k+1} without incurring obvious velocity change. This will impact the animation quality when concatenate two gestures together. Therefore we either replace A_{k+1} with a new animation A'_{k+1} in category c^{k+1} that fits the velocity constraint, or remove the gesture animation with lower priority from the time-line.

- If $v^{k \to k+1} < v^{k+1}$, we need to reduce $v^{k \to k+1}$ to avoid velocity discontinuity artifacts. As shown in Figure 4, we do this by postponing ${t_{grl}}^k$ to ${t_{grl}}^{k\star} = {t_{grl}}^k + t_h$ so that $v^{k \to k+1} = v^{k+1}$ with the new time interval ${t_{ss}}^{k+1} - {t_{grl}}^k - t_h$. This will yields more gesture holding to match the velocity $v_{k \to k+1}$ to the velocity v^{k+1} of next gesture. We add Perlin noise [Perlin 2002] on upper body channels to remove "freezing" effect during the gesture hold.

Post-stroke holds are done to ensure synchronization with coexpressive parts of speech [McNeill 1992]. Additionally they can be used to emphasize parts of the speech. For example, the speaker may point at the listener and hold that gesture to express strong anger.

Gazing and Head Movements. Our animation component synthesizes gazes by controlling any number of joints along the character's spine, neck and eyes. The gaze controller overrides values generated from previous animation stages, since the gazing control needs to orient the body or head in a particular direction.

Head movements play an important role during conversation. The head controller produces head movements by applying a phase-shifted sine wave on head and neck joints. Complex head movements that involve several degree of freedom can be synthesized by combining multiple joint rotations along different axes.

Lip Syncing. Our method uses a diphone-based approach similar to [Deng et al. 2006]. The audio signal is translated into a phoneme (word sound) schedule by a commercial tool [FaceFX 2012] which generates phonemes and aligns them according to the speech signal. Offline, we hand-animate curves using static facial poses that represent the lip and facial poses needed to express a particular diphone. Thus, each diphone corresponds to a small set of facial curves, which are sequenced together, then smoothed to produce the final lip syncing result during runtime. We also apply the word stress value obtained in Section 3.1.5 to drive the open-mouth shape. Thus the mouth would open wider while speaking more loudly.

Emotion, Facial Expression and Eye Movements. Behavior Mapping stage (see 3.4) can specify changes to parts of the face, such as eyebrows, cheeks, eyelids, nose, and so forth. Such behaviors are typically done in parallel with gestures, such as lowering the eyebrows and shaking the head at the same time. In addition, eye movements, such as saccades, can be triggered from the Function Derivation Classes. Note that the eye saccades, like head movements, are additive motions, in order to cooperate with the eye gazing, which is specified during an earlier animation stage.

System Performance. Our acoustic analysis stage takes approximately 250ms per second of audio. Once processed, the analysis phases typically takes less than 1s to process an utterance. Animation synthesis from the behavioral description occurs in real time. Thus, our entire system can generate results in real time, assuming that the audio transcription, acoustic processing and analysis phases are preprocessed. Our system could be used for near-real time uses, such as instant messaging and virtual conferencing, assuming that delays of a matter of seconds between vocalization and transmission are acceptable.

4 Applications

4.1 Previsualization and Animatics

Previsualization is a technique used in film and television production that allows a director to experiment with various camera, scene, lighting and staging options. An extension to storyboarding, 3D visualization tools typically allow the incorporation of nonmoving elements, such as buildings and props, as well as simple animatics to convey motion and timing, such as characters placed in rigid poses. Our method allows the incorporation of an entire character's performance based on the audio signal. Thus a previsualization can become a rich representation of the final video results, as shown in Figure 1 and Figure 5. Our video results present a comparison with feature films. We do not claim that our method produces comparable results to trained actors, rather that our performances are reminiscent of the final performance given only audio recordings and transcriptions. It is conceivable that an entire 3D performance could be generated given a script, camera angles and audio performance. Similarly, table read throughs are common practices when evaluating a script for television or film. Our video shows an example of how our method can be used to visualize a table read through.

4.2 Video Game NPC Performance Generation

There are many non-player characters (NPCs) who appear in modern video games for whom generation of an entire animated performance would be overly costly to generate. Random gestures are commonly used in combination with audio to animate NPC movements. Our study in Section 5 also shows that our method performs significantly better than random gestures. In addition, our animation specification is done using BML, which could be easily interpreted, edited, and replayed if changes to the performance were desired. Thus, our method is well suited for the generation of such characters performances' based only on recorded audio assets and an audio transcription.

5 Study

We perform a study in order to test the influence of semantics gestures on the appropriateness of a performance. For each one of 5 audio files extracted from features films, we generated 3 videos of virtual human performance. The first video **(our method)** shows the suggested gestures generated by using the semantic analysis capacity of our system and. On the second video **(random)**, we replaced the gestures suggested by our study by gestures randomly selected from our animation gesture database. Finally, the third video **(prosody)** shows beats based on prosody. Head movements and facial expression were appropriately selected and not affected by the video variations. However, since head movements are synthesized as additive motion, the exact position and orientation of the head varied slightly according to the underlying gesture pose being simultaneously executed. Figure 6 gives an example of the three variations.

Amazon Mechanical Turk platform [Amazon 2012] was used to recruit 69 participants. They were ask to watch the three performances in a row and to rate the appropriateness of each of them, using a scale from 1 (not at all appropriate) to 10 (very appropriate).

Figure 7 shows the results of this study. Overall, the resulting grades of the left graph demonstrate that our method performs considerably better. When using a normalized scale to obtain a relative notation (purple bars), this is even more apparent. The normalization was conducted by normalizing the scale of each subject's an-

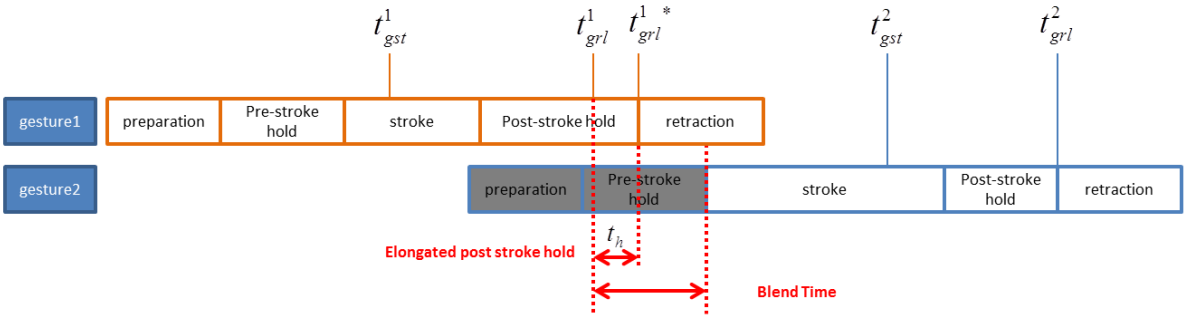

Figure 4: *We ensure that the blending time between gestures does not exceed the speed of the stroke of the latter gesture. Failure to do so might cause the viewer to perceive the transition between gestures to be of greater significance than the latter gesture's execution. We postphone the post-stroke hold phase of Gesture 1 and then blend over a shorter timespan to ensure consistent velocity of movement between the transition and the upcoming gesture's stroke phase.*

Figure 5: *Previsualization results generated from our system. Our method produce an entire character's performance from audio signal.*

swers with the maximum appropriateness selected by the respective subject. A one-way ANOVA was conducted to compare the effects between the proposed method, prosody, and random one. Across the groups a significant effect could be observed with [normalized: $F(2, 180) = 55.66$, $p < .001$; unnormalized: $F(2, 180) = 21.41$, $p < .001$]. Post-hoc comparisons using the Tukey HSD test indicated that the mean score for proposed method ($\mu = 0.89$, $\sigma = 0.24$) was significantly rated better than the other two conditions (random: $\mu = 0.28$, $\sigma = 0.36$; prosody: $\mu = 0.36$, $\sigma = 0.42$). However, there was no significant difference between the random and prosody variations.

The right part of Figure 7 shows the participants preferred video (i.e. the one that got the higher grade). The video generated using the proposed method is selected as the first choice 71% of the time. It is interesting to note than less than 2% of the subjects rated our video as the worst while prosody and random were rated worst respectively 48% and 51% of the time.

Surprisingly, the prosody videos do not receive significantly better results that the random method. One possible explanation could be related to the human tendency to attribute meaning to gestures, even when inconsistent with the content of the utterance. The prosody-based method fires beats on stressed words; except from emphasizing stressed words, beats do not convey any additional meaning. We can argue that the gestures used in the random method, even if inconsistent, could be interpreted as another communicative intent caused by another internal process, independent from speech, such as cognitive processing or emotional arousal. Therefore, this performance could be judged more informative than the one obtained using the prosody-based method.

6 Discussion

Our method utilizes both the lexical content of an utterance as well as the audio signal and its acoustic information in order to generate a virtual character performance. For example, the method can generate different performances based on the timing of the words from the audio signal. The proximity of the words spoken can cause behaviors that would ordinarily be triggered by each word to be ignored, coarticulated, or blended together to synthesize a novel motion. The performance may also differ according to the emotional state of the character, which can be automatically detected from the prosodic information. Each emotional state triggers a different set of behaviors through rules for that state. In addition, the stress detected on each word through the audio signal changes the performance by triggering an additional stressed word behavior, such as a beat gesture, which in turn could cause nearby behaviors to be altered or dropped from the performance. Thus, many different performances can be generated while holding the words in the utterance constant. The human subject studies demonstrate that drawing inferences about the functional content of the utterance in addition

Figure 6: *Study comparing the same utterance using (left) our method, (middle) prosody-based beats, and (right) random gestures.*

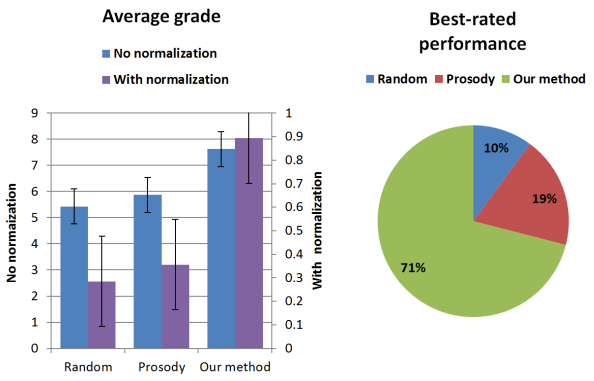

Figure 7: *Comparison of appropriateness between random gestures, prosody-based beats and our method's semantic gestures.*

to prosodic analysis leads to improved performances over gesturing using just prosodic analysis or arbitrary gesturing.

Based on our experience with this work, we feel it is absolutely critical to stress a comprehensive approach to behavior generation that spans head movements, posture shifts, gaze, facial expressions and gestures. The addition of all these behavioral pieces creates a powerful gestalt that brings the character to life. Further each of these pieces is critical. Rich head movement, beyond simple nods and shakes, is often not a focus in research on automated techniques however we have found it to be critical in conveying a sense that the character possesses internal mental states. In some respects this should not be surprising. A close observation of human speakers reveals that often the head is in constant motion. And obviously the details of a how a behavior is realized is also critical. Gesturing, for example, that cannot realize co-articulation and holds will not only look un-natural, it will also inhibit the ability of the character to build up a natural emotion in their performance.

As noted in the introduction, in human face-to-face interaction nonverbal behaviors express meaning through their form and dynamics. By inferring and exploiting that communicative function in the creation of the character's performance, the performances our method creates are more expressive and life-like than behaviors created by prosody. Moreover, techniques that rely on prosody alone run a greater risk of false implicatures, conveying meaning through the nonverbal behavior that is inconsistent with the communicative goals of the utterance.

Moving forward, one of the key issues will be to expand the functional analysis being performed by the system, especially the underlying lexical knowledge it uses, as well as exploring the feasibility of making finer grained distinctions in the prosodic analysis. Also as we noted previously, nonverbal behaviors can replace or contradict the speech message [Ekman and Friesen 1969] and, of course, they are also determined by culture, gender, personality, attitudes as well as the context in which the communication takes place [Burgoon et al. 2009]. The work presented here, is leveraging the utterance in deriving the communicative function implicit or explicit in the utterance, but even richer use of nonverbal behaviors would benefit from tagging the utterance with such information. Fortunately, the staged approach we have taken of functional derivation followed by behavior generation phases supports such tagging of additional requirements on the input.

References

ALBRECHT, I., HABER, J., AND PETER SEIDEL, H. 2002. Automatic generation of non-verbal facial expressions from speech. In *In Proc. Computer Graphics International 2002*, 283–293.

AMAZON, 2012. Amazon mechanical turk.

ARGYLE, M., AND COOK, M. 1976. *Gaze and mutual gaze.* Cambridge University Press, Cambridge.

AYLETT, M. P., AND PIDCOCK, C. J. 2007. The cerevoice characterful speech synthesiser sdk. In *Proceedings of the 7th international conference on Intelligent Virtual Agents*, Springer-Verlag, IVA '07, 413–414.

AYLETT, M. P., POTARD, B., AND PIDCOCK, C. J. 2013. Expressive speech synthesis: Synthesising ambiguity. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

BAVELAS, J. B. 1994. Gestures as part of speech: Methodological implications. *Research on Language and Social Interaction 27*, 3, 201–221.

BERGMANN, K., AND KOPP, S. 2009. Increasing the expressiveness of virtual agents: autonomous generation of speech and gesture for spatial description tasks. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 361–368.

BRAND, M. 1999. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, SIGGRAPH '99, 21–28.

BURGOON, J. K., GUERRERO, L. K., AND FLOYD, K. 2009. *Nonverbal Communication*. Allyn & Bacon.

BUSSO, C., DENG, Z., GRIMM, M., NEUMANN, U., AND NARAYANAN, S. 2007. Rigid head motion in expressive speech animation: Analysis and synthesis. *Audio, Speech, and Language Processing, IEEE Transactions on 15*, 3, 1075–1086.

CALBRIS, G. 2011. *Elements of Meaning in Gesture*. John Benjamins Publishing, Nov.

CAO, Y., FALOUTSOS, P., AND PIGHIN, F. 2003. Unsupervised learning for speech motion editing. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, SCA '03, 225–231.

CASSELL, J., PELACHAUD, C., BADLER, N., STEEDMAN, M., ACHORN, B., DOUVILLE, B., PREVOST, S., AND STONE, M. 1994. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. 413–420.

CASSELL, J., VILHJLMSSON, H. H., AND BICKMORE, T. 2001. BEAT: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 477–486.

CHARNIAK, E. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, Association for Computational Linguistics, Stroudsburg, PA, USA, NAACL 2000, 132–139.

CHUANG, E., AND BREGLER, C. 2005. Mood swings: expressive speech animation. *ACM Transactions on Graphics (TOG) 24*, 2, 331–347.

DAVIS, S., AND MERMELSTEIN, P. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing 28*, 4, 357–366.

DECARLO, D., STONE, M., REVILLA, C., AND VENDITTI, J. J. 2004. Specifying and animating facial signals for discourse in embodied conversational agents. *Computer Animation and Virtual Worlds 15*, 1, 27–38.

DENG, Z., SOCIETY, I. C., YONG KIM, T., BULUT, M., NARAYANAN, S., AND MEMBER, S. 2006. Expressive facial animation synthesis by learning speech co-articulation and expression. *Space, IEEE Transaction on Visualization and Computer Graphics 12*, 2006.

DRUGMAN, T., AND ABEER, A. 2011. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Proceedings of Interspeech 2011*, ISCA, 1973–1976.

EKMAN, P., AND FRIESEN, W. V. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica 1*, 49–98.

FACEFX, 2012. Facefx speech targets, http://www.facefx.com/documentation/2010/w76.

HELOIR, A., AND KIPP, M. 2009. Embr–a realtime animation engine for interactive embodied agents. In *Intelligent Virtual Agents*, Springer, 393–404.

JU, E., AND LEE, J. 2008. Expressive facial gestures from motion capture data. *Comput. Graph. Forum 27*, 2, 381–388.

KANE, J., SCHERER, S., AYLETT, M., MORENCY, L.-P., AND GOBL, C. 2013. Speaker and language independent voice quality classification applied to unlabelled corpora of expressive speech. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

KENDON, A. 2000. Language and gesture: Unity or duality. In *Language and gesture*, D. McNeill, Ed., no. 2 in Language, culture & cognition. Cambridge University Press, 4763.

KENDON, A. 2002. Some uses of the head shake. *Gesture 2*, 2, 147–183.

KIPP, M., NEFF, M., AND ALBRECHT, I. 2007. An annotation scheme for conversational gestures: how to economically capture timing and form. *Language Resources and Evaluation 41*, 3-4, 325–339.

KIPP, M., NEFF, M., KIPP, K. H., AND ALBRECHT, I. 2007. Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. In *Proceedings of the 7th international conference on Intelligent Virtual Agents*, Springer-Verlag, Berlin, Heidelberg, IVA '07, 15–28.

KLEIN, D., AND MANNING, C. D. 2003. Accurate unlexicalized parsing. In *IN PROCEEDINGS OF THE 41ST ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, 423–430.

KOPP, S., JUNG, B., LESSMANN, N., AND WACHSMUTH, I. 2003. Max - a multimodal assistant in virtual reality construction. *KI - Knstliche Intelligenz 4*, 3, 11–17.

KOPP, S., KRENN, B., MARSELLA, S. C., MARSHALL, A., PELACHAUD, C., PIRKER, H., THRISSON, K. R., AND VILHJLMSSON, H. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *Proceedings of the Intelligent Virtual Humans Conference*.

LAFFERTY, J., SLEATOR, D., AND TEMPERLEY, D. 1992. Grammatical trigrams: A probabilistic model of link grammar. In *In Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, 89–97.

LAVER, J. 1980. *The Phonetic Description of Voice Quality*. Cambridge University Press.

LEE, J., AND MARSELLA, S. C. 2006. Nonverbal behavior generator for embodied conversational agents. In *6th International Conference on Intelligent Virtual Agents*.

LEE, J., AND MARSELLA, S. 2010. Predicting speaker head nods and the effects of affective information. vol. 12, 552 –562.

LEVINE, S., THEOBALT, C., AND KOLTUN, V. 2009. Real-time prosody-driven synthesis of body language. *ACM Trans. Graph. 28*, 5 (Dec.), 172:1–172:10.

LEVINE, S., KRÄHENBÜHL, P., THRUN, S., AND KOLTUN, V. 2010. Gesture controllers. In *ACM SIGGRAPH 2010 papers*, ACM, New York, NY, USA, SIGGRAPH '10, 124:1–124:11.

LI, Y., AND SHUM, H.-Y. 2006. Learning dynamic audio-visual mapping with input-output hidden markov models. *Multimedia, IEEE Transactions on 8*, 3 (june), 542 – 549.

MCCLAVE, E. Z. 2000. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics 32*, 7 (June), 855–878.

MCNEILL, D. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

MILLER, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM 38*, 11, 39–41.

MISHRA, T., SRIDHAR, V. K., AND CONKIE, A. 2012. Word prominence detection using robust yet simple prosodic features. In *Proceedings of Interspeech 2012*, ISCA.

NEFF, M., KIPP, M., ALBRECHT, I., AND SEIDEL, H.-P. 2008. Gesture modeling and animation based on a probabilistic recreation of speaker style. *ACM Transactions on Graphics 27*, 1, 5.

NIEWIADOMSKI, R., BEVACQUA, E., MANCINI, M., AND PELACHAUD, C. 2009. Greta: an interactive expressive eca system. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '09, 1399–1400.

PERLIN, K. 2002. Improving noise. *ACM Trans. Graph. 21*, 3 (July), 681–682.

SCHERER, S., KANE, J., GOBL, C., AND SCHWENKER, F. 2013. Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. *Computer Speech and Language 27*, 1, 263–287.

SHAPIRO, A. 2011. Building a character animation system. In *Proceedings of the 4th international conference on Motion in Games*, Springer-Verlag, Berlin, Heidelberg, MIG'11, 98–109.

STONE, M., DECARLO, D., OH, I., RODRIGUEZ, C., STERE, A., LEES, A., AND BREGLER, C. 2004. Speaking with hands: creating animated conversational characters from recordings of human performance. *ACM Trans. Graph. 23*, 3 (Aug.), 506–513.

THIEBAUX, M., MARSELLA, S., MARSHALL, A. N., AND KALLMANN, M. 2008. Smartbody: behavior realization for

embodied conversational agents. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems - Volume 1*, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '08, 151–158.

VAN WELBERGEN, H., REIDSMA, D., RUTTKAY, Z., AND ZWIERS, J. 2009. Elckerlyc. *Journal on Multimodal User Interfaces 3*, 271–284.

WANG, Z., LEE, J., AND MARSELLA, S. 2013. Multi-party, multi-role comprehensive listening behavior. *Autonomous Agents and Multi-Agent Systems*, 1–17.

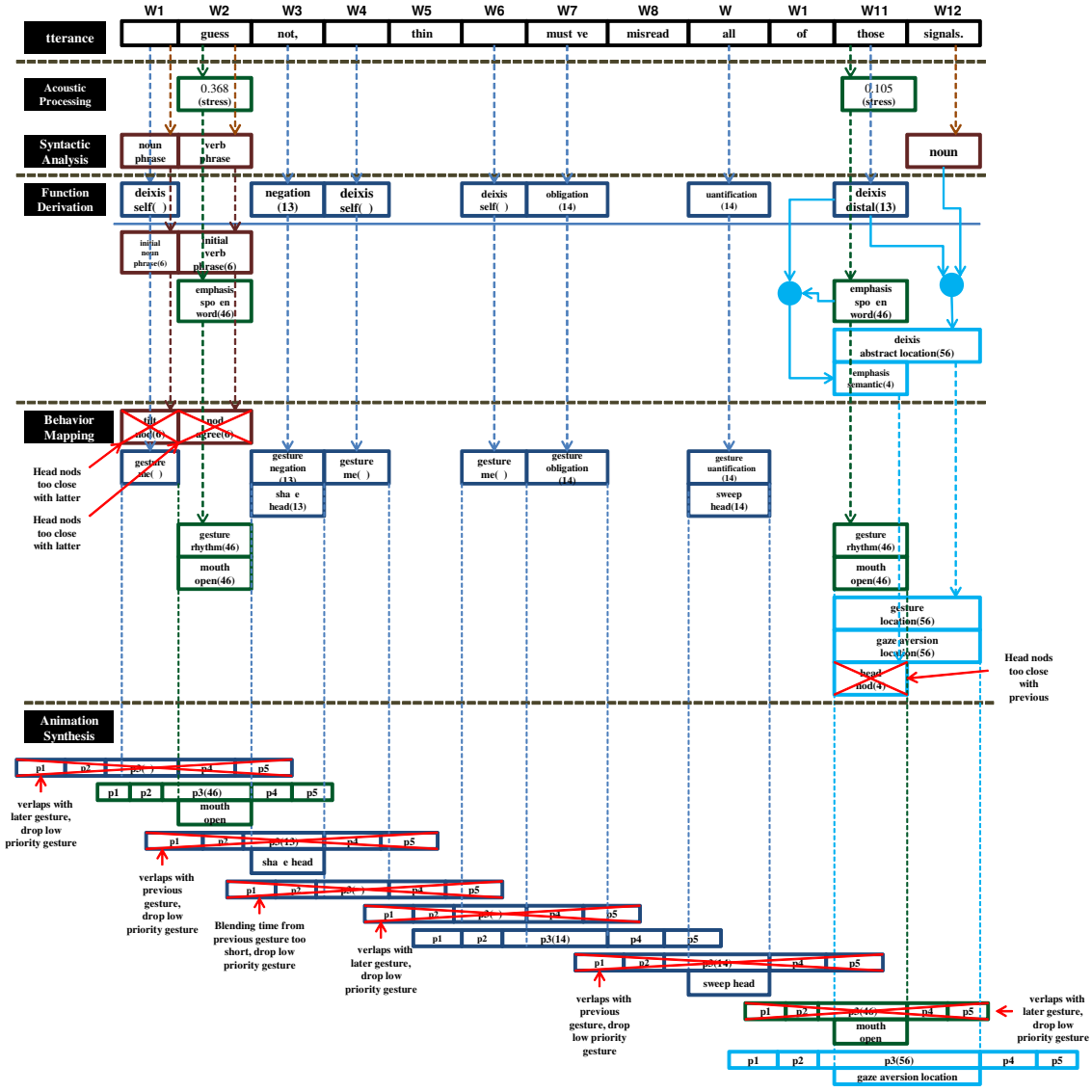

Figure 8: *Simplified processing of the utterance: "I guess not, I think I must've misread all of those signals." Acoustic processing: the audio signal is processed to generate the agitation probability of the utterance (here "high") and determine the amount of stress for each word. In this case, the threshold value of .1 causes the detection of two* stressed words *(W2, W11). Syntactic analysis: the text utterance is parsed to determine its syntactic structure.* Noun phrase *and* verb phrase *are detected (resp. W1 and W2). Function derivation: pattern-matching rules looks for comparators, quantifiers, deixis and other constructs to trigger behaviors when used in conjunction with the previous stages.* Deixis *functions are triggered when encountering expressions signifying self (e.g. pronouns and pronominal adjectives such as I, me, my) (W1, W4, W6) or* determiner suggesting distal deixis *("those" at W11). The first* noun phrase *and* verb phrase *trigger the* initial noun phrase *and* initial verb phrase *functions.* Stressed words *(W2, W11) are associated to* emphasis spoken word *functions. Complex combinations can be elaborated like at (W11, W12) when the* deixis distal *combines with the* noun*. Behavior mapping: applies the function-to-behavior mapping to generate a BML file. For example, the* deixis self *triggers a* gesture me *(W1, W4, W6). Head and eyes movements are also generated, for example at* negation *(W3),* shake head *is generated along with* gesture negation*.*