GESTURE–PROSODY CORRELATIONS ANALYSIS
ACTA DE L'EXAMEN
DEL TREBALL FINAL DE MÀSTER

Reunit el Tribunal qualificador en el dia de la data, l'alumne

D. Marc Antonijoan Tresens

va exposar el seu Treball Final de Màster, el qual va tractar sobre el tema següent:

GESTURE–PROSODY CORRELATIONS ANALYSIS

Acabada l'exposició i contestades per part de l'alumne les objeccions formulades pels Srs. membres del tribunal, aquest valorà l'esmentat Treball amb la qualificació de

Barcelona,

VOCAL DEL TRIBUNAL VOCAL DEL TRIBUNAL

PRESIDENT DEL TRIBUNAL
Abstract

Current 3D character animation systems are very powerful. They are able to generate many new animations from a limited input set of pre-recorded animation clips. When synthesizing speaking avatars animations, a speech file is played along with the animation. If not care is taken in the alignment between gestures and speech, results may present noticeable misalignments. These may decrease the quality of the videos. The current work presents a set of rules that synchronize gestures and speech. These rules can be used in an unsupervised animation system. The impact on naturalness perception of the synchronization rules is evaluated with a subjective test. Results show that synchronization rules improve the quality of the animations, but more rules are needed to achieve a quality similar to non-processed videos.

Els sistemes d’animació de personatges 3D són molt potents. Aquests permeten generar un gran nombre d’animacions des d’una entrada limitada de clips d’animacions pre-gravades. Quan es sintetitzen animacions d’avatars parland, es reproduex un fitxer amb la veu juntament amb les animacions. Si no es tenen en compte els alineaments entre els gestos i la veu, els resultats poden presentar desalineaments perceptibles. Aquests desalineaments poden minvar la qualitat dels vídeos. El treball aquí exposat presenta un conjunt de regles de sincronització entre els gestos i la veu. Aquestes regles es poden emprar en un sistema d’animació no supervisat. L’impacte en la naturalitat percebuda de les regles de sincronització s’ha avaluat mitjançant un test subjectiu. Els resultats mostren que les normes de sincronització milloren la qualitat de les animacions, però es necessita un nombre més gran de regles per tal d’aconseguir una qualitat similar a la que tenen els vídeos no processats.

Los sistemas de animación de personajes 3D son muy potentes. Estos permiten generar un gran número de animaciones a partir de una entrada limitada de clips de animaciones pre-grabadas. Cuando se sintetizan animaciones de avatares hablando, se reproduce un fichero con la voz juntamente con las animaciones. Si no se tienen en cuenta los alineamientos entre los gestos y la voz, los resultados pueden presentar desalineaciones perceptibles. Estas desalineaciones pueden afectar negativamente en la calidad de los videos. El trabajo aquí expuesto presenta un conjunto de reglas de sincronización entre la voz i los gestos. Estas reglas se pueden utilizar en un sistema de animación no supervisado. El impacto en la naturalidad percibida de las reglas de sincronización se ha evaluado mediante un test subjetivo. Los resultados muestran que las normas de sincronización mejoran la calidad de las animaciones, pero se necesitan más normas para conseguir un nivel de calidad similar al de los vídeos no procesados.
Summary

In a previous work, Antonijoan et al. [1] observed that animations from a MOCAP corpus might be used together with speech audios, which were recorded in different sessions, to create videos of speaking avatars. The resulting videos were of acceptable quality most of the times. They were only perceived as unrealistic when there were clear asynchronies between the gestural and the speech channels.

The aim of this work was to test if adding some speech-to-gesture synchronization rules would produce higher quality animations. In order to achieve this goal two objectives were identified: finding the synchronization rules that better define the alignments between gestures and speech features; and to test if applying this set of rules actually increased the perception of naturalness of the animations.

In order to define the set of synchronization rules it was done a review of the literature related to the field. The review covers old studies in order to provide some background, but focuses in state-of-the art publications. Among all proposals, two synchronization rules were selected. These define alignments between gestural and prosody prominences situated in an instant of time.

In order to validate the synchronization rules, it was constructed a gesture synthesizer based in the one presented in [1]. The system takes as input a database of MOCAP data with its prominence points tagged, and speech audio files also accompanied by its prominence points. The synthesizer generated a total of 40 videos of three types: original videos, which used the same MOCAP data that was recorded during the audio capture; random videos, which used a random MOCAP sequence to be reproduced with the audios; and synthesized videos, which used a different MOCAP sequence that the one captured with the audio, and ensured that the selected sequence followed the synchronization rules. Users were asked to take a subjective test. In this test, they had to grade the videos in terms of perceived naturalness of the synchronization between gesture and speech.

Results show that synchrony rules improve the quality of the animations. However, the perceived naturalness for synthesized videos is still far from the inherent naturalness of original videos. This suggests that more restrictions have to be added in the gesture synthesizer to capture all relevant synchronies between gestures and speech.

I like to acknowledge the people without whom this work would not be possible. First of all I want to thank Dr. David Miralles, who supervised this work. He guided me in many aspects of the Thesis. I particularly appreciate his full availability when asked for help. I like to mention also Dr. Francesc Alias, whose advice helped me to approach problems with a scientific perspective. This Thesis is part of the MITMA program of La Salle–Ramon Llull University. I like to acknowledge all professors of the program, because with perspective I can identify helpful lessons learned in all courses. I want to recognize the work done by Raúl Montaño and Alejandro Del Aguila, who assisted me
with the feature identification process of speech and gesture signals. I like to mention Santi Planet, the author of TRUE platform. He not only made the tool freely available but also assisted me when needed. I like to thank Dani Arguedas and Rubén Bernal who volunteered to be recorded, the Medialab staff, Meritxell Aragonés and Eduard Ruesga, who cleaned all the MOCAP data, and the colleagues from the Video Department who recorded the MOCAP sessions in video. Before finishing writing this document I presented the work to the HCI group colleagues. Their feedback helped me to clarify the conclusions of this work. I want to acknowledge the help of all friends and colleagues who agreed to take the subjective test. Finally I like to remember my family and my fiancée Maite, for their patience and love.
# Index

1. Introduction ........................................................................................................... 7
   1.1. Problem Statement ....................................................................................... 7
   1.2. Goal identification ....................................................................................... 8
   1.3. General perspective of the project ............................................................. 9
2. Review of Related Literature ................................................................................ 11
   2.1. Definition of Gesture .................................................................................. 11
   2.2. Intonation analysis ..................................................................................... 13
       2.2.1. Pitch accents ....................................................................................... 14
       2.2.2. Prosodic phrases ................................................................................. 16
       2.2.3. ToBI .................................................................................................. 17
   2.3. Earlier Studies of Gesture and Prosody Correlations ................................... 17
       2.3.1. Historical view .................................................................................... 17
       2.3.2. Valbonesi ............................................................................................ 19
       2.3.3. Loehr .................................................................................................. 21
       2.3.4. Yasinnik, Renwick and Shattuck-Hufnagel ......................................... 24
       2.3.5. Esposito .............................................................................................. 26
       2.3.6. Jannedy and Mendoza-Denton .......................................................... 27
       2.3.7. Karpiński ........................................................................................... 28
       2.3.8. Leonard and Cummins ....................................................................... 30
       2.3.9. Ferré ................................................................................................. 32
       2.3.10. Summary of Speech to Gesture Synchronization Studies .................. 33
3. Research objective ................................................................................................. 37
4. Methodology .......................................................................................................... 39
   4.1. Gesture synthesizer ...................................................................................... 39
   4.2. Data acquisition ........................................................................................... 42
       4.2.1. Data alignment ................................................................................... 44
   4.3. Gestural alignment points identification ..................................................... 45
       4.3.1. Gestural signals ................................................................................... 45
       4.3.2. Gesture alignment points ................................................................... 48
       4.3.3. Identification methodology ................................................................. 48
   4.4. Prosody alignment points identification ...................................................... 53
       4.4.1. Prosody signals .................................................................................... 53
1. Introduction

1.1. Problem Statement

People, when speaking tend to move their body forming gestures. Several studies, such as Kendon’s [29] and McNeill’s [44], pointed out that some parameters of gestures tend to align in time with some features in speech. Moreover, a lot of work has been done analyzing alignment rules between kinematic parameters of gestures and prosodic features of speech [11], [60], [40], [61], [19], [27], and [37]. 3D animation artists take into account gesture to speech alignment rules when manually creating realistic animations of speaking 3D avatars. Animations may also be obtained using MOCAP systems. These allow capturing animations and speech concurrently, thus obtaining animations that naturally synchronize to its accompanying speech. In both scenarios, resulting animations align with a specific speech sequence but not necessarily align with other speech sequences. Reusing these animations for different speech sequences cannot be done arbitrarily.

Cassell et al. [13], Pelachaud [47] and Neff [45] have proposed animation systems that synthesize gestures that align in time with some speech features. Gestures are completely synthetic and are created with computer algorithms. Resulting animations are convincing, although not realistic. Completely synthetic animations do not have the subtleties of real animations, which, in the other hand, are inherently captured in MOCAP systems and carefully created by artists in keyframe animations.

Stone et al. [56] and Levine et al. [39] [38] have proposed animation systems that use pieces of recordings of real human gestures using MOCAP systems. MOCAP clips are used to create new clips that align in time with speech. Stone’s system relies in human supervision for ensuring speech-gesture alignment, which makes it unsuitable for automated applications. Levine proposed an unsupervised method that uses statistical trained models to perform the synthesis. These models require splitting MOCAP sequences in small pieces that coincide with the gesture phases limits defined by McNeill [44]. The synthesis system combines these pieces of MOCAP to create completely new animations that align in time with prosodic parameters of the accompanying speech. Levine’s results show that their synthesized animations are perceived as less realistic than original motion captured animations. Real captured sequences have longer and richer gesture utterances. Using gesture utterances rather than gesture phases as indivisible unit for gesture synthesis may increase perception of naturalness of the animations.

Antonijojan et al. [1] presented an unsupervised system that synthesized gesture and speech for talking 3D avatars. The system uses MOCAP clips as input for the animation synthesis. A simple animation system uses portions of the original MOCAP recordings maintaining continuity along all gesture utterances. These utterances are reproduced in
a random order one after the other whenever the avatar is speaking. No speech-gesture alignment rules are taken in account. Synthesized animations resulted to be perceived as realistic most of the times. Only occasionally it was evident that there were alignment problems between speech and its accompanying speech, for example moving the arms up and down strongly when the speech was plain without accents. However, most of the time the alignment resulted convincing. Even though these observations are to be tested in a rigorous study they point that humans may be tolerant to slight speech to gesture misalignments. It is possible that following few alignment rules is enough to create realistic animations of talking avatars. If this is the case, which are these rules?

1.2. Goal identification

The aim of the present study is to test if adding some time alignment rules between gesture and speech into an animation synthesizer will help producing higher quality animations of speaking avatars. Animation quality is defined as how much realism is perceived in the animations. A high quality animation will be perceived as being performed by a real human.

The animation synthesizer system will use gesture and speech signals captured from human actors. Using real captures help avoiding limitations of fully synthetic animation systems. Gesture data will be split into gesture utterances and re-used for different speech utterances in synthesis. The outcomes of the study may set the basis for a high quality unsupervised animation synthesizer.

The study has two goals that will be pursued sequentially in time. The first goal is to define a set of time alignment rules between speech and gesture. Many studies pointed out the presence of these time alignments [29] [44], however there is not a consensus in which are the parameters that better define these correlations. Some studies have focused in finding time alignments between prosodic parameters for speech and kinematic parameters for gesture [60] [40] [61] [19] [28] and [37]. This type of alignments has the advantage that usually can be extracted without human supervision. We will review the current literature on the subject and select the set of rules that appear to have a more dramatic effect on the perception of animation quality.

The second goal is to test the impact on quality perception of these rules. An animation system that selects gesture utterances based on time alignment rules to speech will be constructed. The system will only use parameters that can be inferred from prosodic and kinematic signals to ensure that alignment rules are fulfilled. The quality of the animations generated by this system will be tested.
1.3. **General perspective of the project**

In chapter “2. Review of Related Literature” it is described what a gesture is, and which elements conform it. Moreover, it is described which intonation analysis parameters are relevant for this work. This chapter covers a review of works in gesture to speech synchronization from the Roman era until now. However a greater focus is set to recent works.

The chapter “3. Research objective” describes which are the goals of this study and introduces how these would be pursued.

In chapter “4. Methodology” there is a definition of the gestural and speech alignment points that are used by the gesture synthesizer. Moreover, it is described a method to identify these points in the original data. In addition, it explains how the subjective test is, and how it was created.

Chapter “5. Results” presents the outcome of the subjective test in a set of tables and figures.

The chapter “6. Discussion” contains a detailed analysis of the results presented in the previous chapter. Several hypotheses are depicted from expected or unexpected outcomes.

In the chapter “7. Conclusions and Future Work” there is a list of conclusions taken from the work. In addition, it is described a set of possible new synchronization rules and other techniques, which might improve the quality of the generated animations.
2. **Review of Related Literature**

2.1. **Definition of Gesture**

A gesture is a movement or a succession of movements performed by the human body, primarily but not always with the hands and arms. Gestures differ from other body movements because these entail communicative goals. Other movements pursue different goals instead, such as transportation (e.g. walking, running, crouching), handcrafting (e.g. painting, polishing, scratching), and many more. Gestures are part of the body language.

Kendon [31] defined a taxonomy for gestures, which was called the Kendon’s Continuum by McNeill [44]. This continuum classifies gestures in terms of its communication function and dependence to its accompanying speech. The classes of the continuum are:

- **Gesticulation.** These are gestures produced concurrently with speech. Their expressivity is related to the accompanying speech. This type of gestures are the ones that are used the most. The stroke of the gesture, which is the most prominent part of the gesture, aligns temporally with its related expressive speech. Gesture strokes tend to slightly precede or co-occur with their co-expressive speech [29]. Gesticulation share invariant properties across many different languages.

- **Speech-framed gestures.** These gestures are produced accompanying speech. However, in contrast to gesticulations, they carry expressivity that speech does not. For example, if a person says “go” and points to a door. The gesture of pointing a door carries the information about where to go, which was unspecified in the speech. Speech-framed gestures do not synchronize with speech.

- **Emblems.** These are gestures that carry meaning by themselves without the need of speech. For example, the sign of victory, raising the middle and index fingers forming a V shape, while closing all other fingers forming a clenched fist. Emblems are also called “quotable gestures” because they can be replaced by words. They can co-occur with speech or not. Emblems vary across different cultures. The same gestures may differ on meaning depending on the culture.

- **Pantomime.** These are the type of gestures used in a dumb show. A sequence of these gestures may be used to narrate a story. They are produced without speech.

- **Signs.** These gestures are part of an unspoken language. Signs may equal to words, verbs and other grammatical units. An example of signs language is the American Sign Language. The structure of such languages may differ significantly to spoken languages.
Ekman and Friesen [18] defined another gesture taxonomy, which included a new type of gesture called adaptors. These gestures fall outside the scope of gesticulation, and therefore are not of interest for this study. Three types of adaptors were presented: the so-called object adaptors, which originate with movements for manipulating objects; the alter-adaptors, which have to do with interpersonal contacts; and the self-adaptors, which have to do with satisfying bodily needs, such as rubbing the eyes.

The current study focuses on studying the impact in animation quality of time alignments between speech and gestures. Therefore, from now on the focus will be given solely to gesticulation, because these are the gestures that have a stronger correlation to speech. We will use the term gesture to refer to gesticulations as most literature does [40].

Gesticulation (gestures) were lately classified by McNeill [44] in a four classes taxonomy:

- **Iconic.** These are gestures that mean concrete entities or actions. They are closely related to what is being said. The shape of the gesture resembles the entity or the action it references, helping the audience to understand some facts about the object of description. For example, talking about a ball while forming a sphere with the hands, resembling the shape of the ball, will help identify the size of the ball.

- **Metaphoric.** These gestures refer to abstract concepts. They use the three-dimensional space to shape these abstract ideas. For example raising a hand to express that an object has a “high” price.

- **Deictic.** These are gestures used to indicate locations in the space. The most common deictic gesture is pointing with an extended finger. These gestures may be used to point to real locations or abstract locations. Using deictic gestures for locating abstract locations is a milestone in children’s development, not appearing until the age of 12 [44].

- **Beats.** These are gestures that do not carry meaning. They are movements up and down or back and forward. They tend to synchronize with speech by stopping on the prosodic peaks of speech. Beats may have discourse functionality, by pointing to important parts of the speech. Beats are the most used gestures constituting about half of all gestures.

The gestures that best synchronize with prosody of speech are beat gestures. Other gestures synchronize with their lexical affiliates in speech. In the current study we are not trying to address scenarios where it is necessary to interpret the utterances. We are focusing only in events that could be depicted from prosodic signals. This leaves non-beat gestures out of this study.

Kendon [30] proposed a segmentation scheme for gestures. He defined the terms “gesture unit”, “gesture phrase” and “gesture phase”. A gesture unit starts in rest pose, contains one or several consecutive gestures, and ends up into another rest pose. A gesture phrase is what it is generally called a gesture. A gesture phrase is composed by several phases, some of them being mandatory and some optional. These phases are:
• Preparation. Body parts move from a rest position to the initial position of the gesture. This movement can move towards the opposite direction of the main direction of the gesture. For example, in a gesture where a hand is moved left to right indicating a driver to move his or her car to the right, the preparation phase could imply moving the hand to the left, so allowing creating a longer left to right trajectory. This phase is optional.
• Stroke. It is the phase with meaning. It involves the greatest effort of all phases. A gesture phrase should contain a stroke; otherwise it will not be called a gesture phrase.
• Retraction. Body parts are moved to the rest position. This phase could not be present if the speaker concatenates a stroke phase with another stroke.

Kita et al. [34] identified another phase:

• Pre-stroke or post-stroke holds. These are temporary cessations of movement that occur immediately before or after a gesture stroke. These holds are optional.

And later on, Kipp [33] suggested another one:

• Recoil phase. Occurs after a forceful stroke when the hand lashes back from the end position of the stroke.

The stroke is the most meaningful part of a gesture, and it is the only mandatory phase. Other phases are placed before or after the stroke. A stroke may consist of multiple repeated movements, which would make it a multi-stroke [34]. The stroke phase has the property that tends to align in time with co-expressive speech.

2.2. Intonation analysis

Several studies have found strong correlations between gestures and intonation of speech. For example, Bolinger stated in [10] that “Intonation belongs more with gesture than with grammar”. It is therefore logical to focus in the analysis of this feature of speech in the present study. In this chapter it is reviewed the Autosegmental and Metrical model, one of the most popular frameworks for intonation analysis.

Pierrehumbert introduced the Autosegmental and Metrical model (AM) in her PhD thesis [48]. It was used for analyzing intonation in English. The objective of AM is the identification of contrastive elements of the intonation system. The combination of these elements produces the melodic contours found in language. In AM theory the melody or tonal modulation is treated as a separate phenomena than other phonological features. Tones are treated as autosegments, which relate to text following a set of rules that may vary from language to language.

In some languages, such as Chinese, each syllable has a tone determined lexically. However, in Spanish or English tones have a pragmatic function, conforming the
melody of utterances. Liberman presented a study in English [41], which shows that the association between tones and syllables depends on the prominence relations of syllables within words or words within sentences. For example, in the word Castillo the syllable –ti- is more prominent than the other syllables in the word. This is called the tonic syllable. However, in the sentence Castillo de arena, the tonic syllable of the second word –re- is more prominent than the tonic syllable of Castillo. This is the syllable that contains the nuclear accent.

The AM theory says that the intonation contour corresponds to the interpolation between tonal events associated to certain syllables. In Spanish, these tonal events can be associated to lexical accents (pitch accents), or to the end of certain sentences (boundary tones).

2.2.1. Pitch accents

In Spanish, almost all words have a stressed syllable. Only functional words such as prepositions, determinants and alike are rarely accentuated. Stressed syllables are characterized by having a pitch prominence, relative longer duration, and relative higher intensity that other syllables in the same word. Moreover, stressed syllables may carry a pitch prominence. In the following example, the stressed syllable –ben- in the text y bengo contains a pitch prominence, a high tone, which it is marked as H*.

![Figure 1 - The sentence y bengo with a high pitch in the syllable –ben-](image)

However, a prominence is not always associated with a high pitch. In the following example, the stressed syllable –ka- in the word cámara has low pitch, marked as L*.
As seen before, the stressed syllable is accompanied by a pitch event, which contribute to make the syllable prominent. The type of pitch event depends on the on the type of utterance, its position and pragmatic relevance of the word within the utterance.

In Spanish, the position of the stressed syllable within words carries lexical information, thus allowing differentiating the meaning of the word. Variations on the pitch define the intonation contour, which gives information about the type and meaning of the sentence within the discourse. For example, ¿compra? or ¡compra! have a different melody. These share the same lexical meaning but their function varies: interrogation or exclamation. In contrast, in tonal languages, such as Chinese, the intonation contour has a lexical function. For example, in Mandarin Chinese the syllable /ma/ may have four completely different lexical meanings depending on the tonal melody.

Not all stressed syllables in Spanish have a pitch accent. This mechanism has a pragmatic function, which allows speakers to highlight some words over the others. This unstressing technique is also used in English.

The types of pitch accents are defined as:

- $H^*$ Pitch peak on the stressed syllable
- $L^*$ Pitch valley on the stressed syllable
- $L+H^*$ Pitch peak on the stressed syllable preceded by a valley
- $L^*+H$ Pitch valley on the stressed syllable followed by a peak
- $H+L^*$ Pitch valley on the stressed syllable preceded by a peak
- $H^*+L$ Pitch peak on the stressed syllable followed by a valley

Note that another letter with an asterisk always accompanies H and L. These have not meaning by themselves.

Not all languages use all these types of pitch accents. For example, according to Beckman et al. [6], in Spanish only $L^*+H$, $L+H^*$, $H+L^*$, $H^*$ are present. Moreover, $L^*$ may be used in interrogative sentences.
To annotate pitch accents, it is necessary to know certain aspects of language. For example, in Spanish it is necessary to know which are the lexical stressed syllables. These are the anchor points that may carry pitch accents. Therefore, before analyzing the pitch curve it is necessary to identify the stresses. However, it is possible that either lexically unstressed syllables carry pitch accents, or that stressed syllables do not carry pitch accents. For example, in emphatic speech many unstressed syllables may carry pitch accents. The identification of the style of speech facilitates the annotation task.

2.2.2. **Prosodic phrases**

A prosodic phrase is a segment of speech that presents a single prosodic contour. These may be intonational phrases or intermediate phrases. The latter present lower levels of disjunction than intonational phrases. For example, in: *cruzamos montañas nevadas, desiertos calurosos y bosques selváticos*, the utterance will probably be pronounced with two little disjunctions at the comma and the *y*. In this case, the whole utterance may form a single intonational phrase, which is formed by three intermediate phrases.

It is argued whether or not Spanish presents these two levels of prosodic phrases. Sosa (Sosa 1999) proposed that it is not necessary to detect intermediate phrases, but intonational phrases are enough to capture the structure of prosodic phrases. However, Hualde (Hualde2002) defends the existence the function of intermediate phrases in Spanish. These may clear up ambiguous meanings. For example, in *la vieja lanza la amenaza*. The meaning varies whether there is a pause between *–vieja–* and *–lanza–*, or *–lanza–* and the second *–la–*. In the first case *–lanza–* is a verb, and in the second is a noun.

The AM model defines a framework to annotate tonal features at the end of intonational and intermediate phrases. Tags at the end of these phrases are known as boundary tones. Boundary tones are indicated with L% or H% at the end of an intonational phrase, and L- or H- at the end of an intermediate phrase. The end of an intonational phrase corresponds also to the end of an intermediate phrase. Therefore, a boundary tone at the end of an intonational phrase is marked with two symbols: X- X%. Both X symbols may be L or H. In English, the tone X- corresponds to a place between the last pitch accent and the end of the phrase, and the X% to the tone at the end of the phrase. Depending on the language, boundary tones may be different. In English, the end of an intonational phrase may be annotated with one of these four interpretations (based on Beckham and Hirshberg [4]):

- L- L%  Decrease of pitch
- L- H%  Incomplete decrease with an increase of pitch at the end
- H- L%  Increase of pitch followed by a lower pitch that remains flat.
- H- H%  Increase of pitch
2.2.3. **ToBI**

One of the objectives of the scientific community using the AM model is to define a standardized method for transcribing prosodic events. This will allow creating corpuses that may be reused in many studies. Since each language has differences on prosody function and behavior, transcription methods must differ from language to language.

Silverman et al. [55] proposed the transcription ToBI framework. Its goal is to define a standard transcription framework for English. ToBI stands for *Tone and Break Indices*, which underscores the importance of the indices that separate words. Later, other similar attempts were done in several languages: ToBit for Italian, K-ToBi for Korean, Sp-ToBI for Spanish among others. However, all these attempts are not free from controversy, and are not necessarily adopted by the majority of experts.

2.3. **Earlier Studies of Gesture and Prosody Correlations**

2.3.1. **Historical view**

Gesture studies took a remarkable importance during the Roman era. Cicero discussed the importance of gesture and facial expression in a discourses and oratory. However, the most complete study about gestures was the eleventh Book of *Institutio oratoria* ('Education of the Orator") by Marcus Fabius Quintilianus, written in the first century AD. Quintilian was a rhetorician from the Iberian Peninsula. Quintilianus wrote about gesture and speech synchronization. He drew a distinction between gestures that “naturally proceed from us simultaneously with our words” and those by which one indicates things by means of mimicry. Moreover, he agreed with earlier instructions that the movements of the hand should begin and end with the thought that is expressed, otherwise the gesture will anticipate or lag behind the voice, both of which produce an unpleasing effect” [50].

Gesture took the attention of philosophers of the seventeen and eighteen century. However, their studies were not focused on the contemporaneous relationship between gesture and speech. It was gesture as an autonomous medium of expression that was of interest [32]. On the latter half of the nineteenth century there was a remarkable interest in gestures and its relationship to speech. Edward Tylor, Garrick Mallery and Wilhelm Wundt drew theories about the origin of language taking in account both speech and gesture.
At the end of the nineteenth century interest in gesture and its relation to speech receded markedly [32]. In contrast, Bolinger said that language is an arbitrary, although convenient, abstraction; and we could not insist upon what this abstraction excludes before the full implications of phenomena that accompany language, such as gesture, are fully understood [9]. Bolinger observed a synchrony rule between speech and gesture: pitch and body parts rise and fall together, to reflect increased or decreased tension [11]. However, Loehr found no evidence on this [40]. Other authors suggested that language and gesture are intimately related are Zellig Harris [23], Edward Trager [58] and Ray Birdwhistell [7].

Condon studied the correlations of body movements and speech. He observed that body movements conform a hierarchy. Small and fast movements are superimposed on greater and slower movement. For example movements of eyebrows or hands are superimposed on movements of the head or arms. He observed that the smallest and fastest movements synchronized with small speech units such as phones and syllables, and slower movements synchronized with larger speech units, such as phrases. More surprisingly, Condon showed that listeners also moved synchronized with the speaker’s speech (Interactional Synchrony) [14].

Kendon continued Condon’s studies on synchrony of speech and gesture. He suggested a hierarchy of synchronies [30]:

- Strokes align to stressed syllables.
- Gestural phrases (a stroke plus optional preparation and retraction phases) align to tone units (completed intonation tune).
- Gestural units (a set of gestural phrases that start and end in a rest pose) align to sentences.

Kendon noted that the extension phase of a beat gesture would coincide with, or slightly precede, the onset of a stressed syllable. These findings about speech and gesture synchrony suggested Kendon that speech and gesture are not only somehow connected, but they are two surfaces forms of the same underlying structure.

McNeill supported Kendon’s theory [43]. He stated that speech and gesture are part of the same psychological structure and share the same origin, which he called “growth point”. He supported this statement with several observations; such as, gestures occur mainly during speech; gestures and speech share pragmatic functions; gesture and speech are synchronized in time; speech and gesture are both affected by aphasia; and, children’s development of speech language parallels to gesture’s development.

McNeill proposed three synchrony rules between speech and gestures:

- Semantic synchrony. Concurrent speech and gesture must share the same semantic information.
- Pragmatic synchrony. Concurrent speech and gesture must share the same pragmatic function.
- The stroke of the gesture phrase is always completed either before or at the same time as the tonic syllable of the concurrent tone unit [44].
The latter synchrony rule closely resembles the rule proposed by Kendon. Nobe stated that this statement also holds for peaks of intensity besides of tonic syllables [46].

Moreover, McNeill noted that the preparation phase precedes the tone unit with which it is associated. In the other hand, Karpiński noted that McNeill’s synchrony rules seem to apply mostly to narration, but it might be possible that interactivity of dialogue introduces other factors that may influence the timing of speech and gesture [28].

McNeill’s colleague Tuite, observed that there is a regular rhythmic kinetic pulse underlying both the production of speech and gestures. For speech, a pulse is manifested by a peak in intonation, or by pitch accent. For gesture the pulse corresponds to the extension phase of the gesture [59]. However, Loehr found a richer relationship, in which the three main “instruments” (hands, head, voice) interplayed much like a jazz piece, with tempos that sometimes synchronized, sometimes differed, and which included full notes, half notes, and syncopation [40].

The historical view described so far shows that there is a large number of authors that support the idea of an existing synchrony between gestures and speech. However, these authors describe vaguely how this synchrony behaves. Recently there have appeared studies that analyze what are the anchor points between both modus, and what are the rules that govern this synchrony. The following chapters draw an overview of recent studies focused in giving answers to these questions.

2.3.2. Valbonesi

Valbonesi et al. investigated the temporal relationship between prosodic focal accents and significant cues in hand traces [60]. The analysis consists of determining the temporal correlations between two types of focal points: ones obtained from speech data and others obtained from combination of information of hand motion and gesture strokes. Speech focal points are computed by using five cues within a phrase: F0 Peak, minimum and second minimum of the frequency gradient, the maximum and second maximum value of the amplitude within each segment. Two types of gesture focal points are identified: the first kind comes from extracting local maxima and minima of the hand traces motion, and the second type by hand annotating gesture’s stroke limits. They performed studying the correlations of gesture and speech focal points. Segments of speech with no hand movement were discarded, so it would be impossible to find correlations to gestures.
Table 1 - Time correlations between speech focal points and two types of gesture focal points: hand movement (Condition 1) and gesture strokes (Condition 2)

<table>
<thead>
<tr>
<th></th>
<th>% Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition</td>
<td>68.18</td>
</tr>
<tr>
<td>Condition</td>
<td>68.18</td>
</tr>
<tr>
<td>Condition</td>
<td>95.45</td>
</tr>
<tr>
<td>None</td>
<td>4.55</td>
</tr>
</tbody>
</table>

Results are shown in Table 1. Condition 1 refers to the case where speech focal points fall within windows centered in hand movement focal points. The window size was selected as a compromise of being as small as possible and giving a high detection rate. Condition 2 refers to the case where focal points occur within the duration of a gesture’s stroke. From these results, they concluded that leaving aside a small percentage, all focal positions satisfy at least one of the two proposed correlation conditions.

This study supports the idea that exists a significant correlation between gesture and speech prominences. They introduced the idea of the focal points as correlation anchor points for gesture and speech. Speech focal points and movement focal points have the interesting feature that can be identified by a computer program in a high rate. However, gesture stroke limits, which are another type of focal points, are identified by hand. The study uses a window to search for speech focal points that are close to movement focal points. It is not mentioned which is the size of this window. It is also unclear what percentage of the time line of the analyzed data is occupied by these windows. If the windows occupy a large percentage of the time, it would not be surprising to obtain a the moderately high percentage (68.18%) of speech focal points falling within these windows, even if no correlation exists. Moreover, these windows are centered in the gesture focal points, looking equally in future and past time for a speech focal point. Other studies, such as McNeill’s work [44], have observed that gesture prominences occur either during or before speech prominences, but rarely after. Taking this in account, it seems that moving the window earlier in time may result in better results. Speech focal points are searched within the gesture stroke phase. Again, that does not take in account that gesture prominences often precede speech prominences. Another criticism is that the experimental data used in the study does not seem to be large enough to be of significance. Only 22 speech focal points where identified in all data.

McClave used a different type of speech prominences, the nuclei of tone groups and the stressed syllable [42]. He states that beat gestures have a tendency for the downbeat (the stroke) to co-occur with these speech prominences in multisyllabic words. McClave also showed that in the case of a complex gesture where several movements appear in succession, the gestures are “compressed” and fronted to all finish before a stressed syllable.
2.3.3. **Loehr**

Loehr, in his PhD dissertation [40], studied the relation between gesture and intonation of speech. His analysis of intonation was based in Pierrehumbert work [48]. Loehr validated several hypotheses related to gesture to speech correlation.

In one of the studies, tried to validate if the “Bolinger’s Parallel Hypothesis” [11] holds. Bolinger suggested that pitch and body parts move in parallel, that is, they move up and down together. Loehr annotated vertical movement directions for predominant vertical direction intervals, and intonational prominences using the ToBI annotation scheme [55]. Table 2 shows co-occurrences of vertical movement and selected tones within 275 millisecond windows.

<table>
<thead>
<tr>
<th></th>
<th>H*</th>
<th>L-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hand(s) up</td>
<td>47</td>
<td>39</td>
</tr>
<tr>
<td>Hand(s) down</td>
<td>62</td>
<td>49</td>
</tr>
<tr>
<td>Head up</td>
<td>54</td>
<td>59</td>
</tr>
<tr>
<td>Head down</td>
<td>64</td>
<td>59</td>
</tr>
<tr>
<td>No vertical move</td>
<td>24</td>
<td>20</td>
</tr>
<tr>
<td>Movement totals</td>
<td>251</td>
<td>226</td>
</tr>
</tbody>
</table>

**Table 2 - Co-occurrences within 275 milliseconds of vertical movement and H* and L- tone types**

Table 2 shows that hands are more likely to move down than up when a high pitch accent (H*) is present. However, there is not a significant correlation between these two values. Moreover, in the presence of low pitch tone (L-), there are slightly more hands down movements than up movements. Again, this correlation lacks of statistical significance. Loehr noted that results on his study do not support Bolinger’s Parallel Hypothesis.

In another study, Loehr checked for time alignments between different types of gestural units and intonational units. Gestural units were apices of strokes, gestural phases (especially strokes), gestural phrases and gestural units. Intonational units were pitch accents, intermediate phrases, and intonational phrases. From all possible combinations, Loehr found alignments only between two groups: apices aligned with pitch accents, and gestural phrases with intermediate phrases. Apices are defined by Loehr as the “peak of the peak” or as the “kinetic goal of the stroke”. Intermediate phrases, defined by Beckman and Pierrehumbert [5], are intonational constituents, whose boundary corresponds to a level 3 Break Index in the ToBI system. This is defined as an intonation contour with one or more pitch accents and a phrase accent, but no (final) boundary tone. Loehr analyzed if apices co-occur within a 275 msec. window with pitch accents.
Results are shown in Table 3. Based in a chi-squared test (p<0.001), it can be inferred that there exists a significant tendency of pitch accents and apices to occur together. Moreover, pitch accents tend to occur very close in time to apices. In this study, pitch accents occur 17 msec. later than apices in average. This result is similar to McNeill’s observation [44] that stroke of the gesture phrase is always completed either before or at the same time as the tonic syllable. Anyhow, this rule is challenged in the following observation.

In Figure 3 it is shown the histogram of distances between apices and pitch accents. The histogram shows an unsymmetrical distribution, being more common pitch accents occurring after apices than occurring before. However, many pitch accents are still located before apices. Apices are parts of the strokes, therefore following McNeill’s rule, they must end before or at the same time than pitch accents. However, apices usually occur at the end of the gesture stroke, therefore, still most of the stroke may occur before the pitch accent even if the apex ends slightly later.

Loehr studied if this correlation between apices and pitch accents varies depending on the movement types, such as beats, iconics, metaphorics, deictics, emblems or adaptors, and found that all of them responded to a similar synchrony pattern.
Figure 3 – Histogram of distances of pitch accents from gestural apices. Extracted from Loehr’s PhD thesis [40].

Given gestural apex at zero.

Bar above zero counts pitch accents occurring within 1/2 frame (+/- 16.5 msec) of gestural apex.

Each bar has a width of 1 frame (33 msec).

Bars to the left of zero (negative msec) count pitch accents occurring before gestural apex.

Bars to the right of zero (positive msec) count pitch accents occurring after gestural apex.

Sum = 200 pitch accents

Mean = +17 msec

St. dev. (σ) of tone distribution = 341 msec

Average inter-pitch-accent interval (IPI) = 794 msec

Concentration of pitch accents around apices is therefore not due to frequent pitch accents in the first place, (in which case there would always be a pitch accent near an apex).

Average inter-apex interval (IAI) = 769 msec

Concentration of pitch accents around apices is therefore not due to frequent apices in the first place, (in which case there would always be an apex near a pitch accent).

Distance of nearest pitch accent from given gestural apex (msec); each tick-mark equals one frame (33 msec)
Moreover, Loehr studied the correlation between gestural phrases and intermediate phrases. He observed that the boundaries of gestural phrases aligned with the boundaries of intermediate phrases. Two out every three times, the limits of a gestural phrase aligned with the limits of a single intermediate phrase. Usually, gestural phrases start and end slightly before (100 msec. in average) than starting and ending points of the intermediate phrase. However, these alignments are looser than apex to pitch accent alignments. Often, an intermediate phrase contains several gestural phrases.

Loehr also investigated the correlation between gesture and intonation in terms of rhythm. He noted that humans may express rhythm in several models, such head nods, gestural strokes, gestural beats, pitch accents, stressed syllables, eye blinks, or a variety of other phenomena. He referred to all these elements as *pikes*. He observed that *pikes* of different modalities sometimes align, and usually don’t. They may follow a constant rhythm temporally, or syncopate from the rhythm. Other observations are that usually, intervals between hand pikes are shorter than pikes in speech, and that eyeblinks overtures usually align with pikes of other modalities. He concluded that there exists an inter-relation between body movement and intonation rhythm. But this relationship is not uniform, only manifesting at certain points.

For the present study, the most relevant contribution of Loehr’s work is the finding of the existence of a high correlation between pitch accent occurrences and apices of strokes. Other discoveries such as intermediate phrase to gestural phrase correlations, or rhythm correlations present looser correlations. These last discoveries respond to complex phenomena, which are hard to model with synchrony rules.

### 2.3.4. Yasinnik, Renwick and Shattuck-Hufnagel

Yasinnik et al. studied the temporal alignment of gestures with the prosodic structure of speech [61]. They used prosodic and gestural features to carry out the study. The used the ToBI system to label the spoken utterances. From all the labels ToBI define, they selected these two prosodic features: pitch accent locations and intonational phrase boundaries. As for gestures, they selected *hits* as prominent feature. Hits are discrete features; therefore, their alignment with pitch accents can be reliably quantified. Gestures can be classified in two categories: discrete and continuous. Discrete gestures contain hits, while continuous gestures do not. A *hit* is an abrupt stop or pause in movement, which breaks the flow of the gesture during which it occurs. Hits appear as bouncing, jerky movements, changes in the direction of movement, or as complete stops in movement. Yasinnik made an analogy of hits, as resembling to the movement of an orchestra conductor marking each beat with a sharp movement of the baton. Finally, also gesture onset and offset times were annotated as gesture features.

Yasinnik, Renwik and Shattuk-Hufnagel performed two types of studies. In the first type, they analyzed the co-occurrence of hits within syllables containing pitch accents. They observed different results for monosyllabic and polysyllabic words. While, for
polysyllabic words, hits occurred in 90% of the cases within the syllable containing a pitch accent, that only happened 65% of the times for monosyllabic words. They interpret that these results suggest that the alignment of the gestures is not with the pitch-accented syllable per se, but possibly with another prosodic constituent that groups a strong syllable with the following weak syllables. In the second type of studies, they observed that in intonational phrases containing several gestures. The pauses between gestures where longer at the end of the intonational phrase than the pauses occurring within the phrase. They stated that these observations support the hypothesis that gesture-span ensembles correspond to a higher-level type of speech constituents.

Results of these studies show that pitch accents and hits correlate better than the focal points proposed by Valbonesi. While hits and pitch accents correlated an average of 78% of the times (90% for polysyllabic words, and 65% for monosyllabic words), any of the two Valbonesis’s focal points correlated 68.18% of the times. Despite this improvement on the correlation, they suggested that there are prosodic constituents of higher order than pitch accents that align better to hits. Following this hypothesis, they observed that pauses between gestures are longer between intonational phrases than in any other place. They stated that more research in gesture to higher order prosodic constituents is necessary.

Shattuck-Hufnagel et al. [54] defined a method for the study of gesture to prosody temporal relationships. They observed that various studies of these time alignments use different prominent features in both modalities, making to compare results of different studies difficult. They proposed ToBI as the framework to annotate prosody constituents, such as intonational phrases or pitch accents. They propose to use hits as prominences in gestures. Hits are annotated by analyzing video data. A hit corresponds to a single frame in the video, which last 33 msec. for a 30 frames per second video. They proposed a method to detect hits: a hit exists only when these three situations co-occur: firstly, the image in the video focuses abruptly or at least becomes strikingly less blurry, that is due the sudden cessation of movement; secondly, there is a change on the direction of the movement; and finally, there is a beginning of a change in the shape of the hands. Speech and gestures should be annotated separately; otherwise these channels may interfere in the process of identifying features in the other channel. In this study it is also addressed the issue of determining the limits of a syllable when trying to identify if a hit occurs within its duration. Syllables in English do not have clear boundaries. They propose to use different symbols to annotate these unclear cases, thus allowing performing the analysis of these results separately.

This work presented by Shattuck-Hufnagel et al. [54] proposes an interesting framework to perform speech to gesture correlation analysis. Following this method would facilitate comparing results across different studies. Moreover, they proposed a detailed methodology to obtain the annotations.
2.3.5. Esposito

Esposito et al. performed a study [19] similar to the work by Yasinnik [61] but applied to Italian speakers. Besides in language these works differ in that hits are annotated differently. While in Yasinnik’s work a hit was annotated as a single frame (33 msec.), in Esposito’s work all the extent of the gesture was annotated. The onset of the hit was considered to be the time when the bouncing, and/or jerky movement, and/or the changes in the direction of movement begin. The hit’s end point corresponds to the frame where the movement abruptly stops. The co-occurrence between gesture and speech was redefined to considering that a hit and a prosodic feature overlapped whenever part of the hit extent overlapped with the prosodic feature duration.

For speech they used the ToBI’s annotation scheme to mark intonational features such as pitch accents, intonation boundary tones, and intermediate boundary tones. Moreover other labels were added to mark speech events such words; empty and filled pauses, which indicated hesitation phenomena; and word repetitions and fresh starts, which indicated repairs.

Two video recordings were annotated with hits and speech features. The first video contained 140 hits and, and the second video 73 hits. A summary of the number of speech features (speech and prosodic entities) is shown in Table 4.

<table>
<thead>
<tr>
<th></th>
<th># Pitch Accents</th>
<th># Intonation Boundary Tones</th>
<th># Empty Pauses</th>
<th># Filled Pauses</th>
<th># Repairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker 1</td>
<td>258</td>
<td>93</td>
<td>128</td>
<td>83</td>
<td>51</td>
</tr>
<tr>
<td>Speaker 2</td>
<td>277</td>
<td>81</td>
<td>115</td>
<td>51</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 4 – Summary of speech features

<table>
<thead>
<tr>
<th></th>
<th># Pitch Accents with hits</th>
<th># Intonation Boundary Tones with hits</th>
<th># Empty Pauses with hits</th>
<th># Filled Pauses with hits</th>
<th># Repairs with hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker 1</td>
<td>109</td>
<td>36</td>
<td>30</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td>Speaker 2</td>
<td>61</td>
<td>25</td>
<td>32</td>
<td>14</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5 – Occurrences of hits over speech and prosodic entities

Co-occurrences between hits and speech features are shown in Table 5. From these results Esposito et al. concluded that there is a tendency toward temporal
synchronization between gestural hits and prosodic events in Italian, because the majority of these gesture occurs in synchrony with prosodic events, such as pitch accents: 78% for Speaker 1 (109/140), and 84% for Speaker 2 (61/73).

<table>
<thead>
<tr>
<th></th>
<th># Boundary tones within pitch accented words with hits</th>
<th># Boundary tones outside of pitch accented words with hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker 1</td>
<td>59</td>
<td>1</td>
</tr>
<tr>
<td>Speaker 2</td>
<td>45</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6 - Occurrences of hits over boundary tones, which belong or not to pitch accented words

Moreover, they observed a positive relationship between hits and boundary tones. Table 6 shows that this relationship only holds when boundary tones belong to words with a pitch accent.

Esposito’s work reinforces the idea that production of gestural hits and pitch accents show a strong time alignment. As suggested by Yasinnik [61] this work also supports the idea that there are other speech constituents rather than pitch accents that play a role in this alignment. However, no details about these other possible rules of synchronization are provided.

2.3.6. Jannedy and Mendoza-Denton

Jannedy and Mendoza-Denton [25] studied the co-occurrence of pitch accents and gestural apices. In an experiment, they analyzed a video recording of a real situation. Pitch accents were annotated using the ToBI framework. As for gestures, gestural apices were transcribed. They used Loehr’s definition for apices [40], which states that an apex is the “kinetic goal of the stroke”. In task dynamics terms this is defined as the target of the gesture. Both pitch accents and gestural apices were annotated as points in time.

<table>
<thead>
<tr>
<th># Apices</th>
<th>Apices if pitch accent</th>
<th>Apices but not pitch accent</th>
<th># Pitch accents</th>
<th>Pitch accents if apex</th>
<th>Pitch accents but not apex</th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>68 (95.7%)</td>
<td>3 (4.2%)</td>
<td>98</td>
<td>68 (69.4%)</td>
<td>30 (30.6%)</td>
</tr>
</tbody>
</table>

Table 7 - Co-occurrence of Gestural Apices and Pitch Accents
Results of the study are shown in Table 7. These results reveal that pitch accents are more common than gestural apices. It can be noted that whenever there is a gestural apex, there is also a pitch accent. However, pitch accents occur without accompanying gestural apices. They stated that this data shows that there is a robust co-occurrence between apices and pitch accents in spontaneous speech. That was previously observed by Loehr [40] but with laboratory data.

It is not clear how it was decided that a gestural apex and a pitch accent co-occur. Both features correspond to points in time; therefore a time window is necessary to define co-occurrence. However, the study supports again that there is a time alignment between gesture and speech prominences. It is remarkable the observation that apices almost always align with pitch accents but the reverse do not always hold.

2.3.7. Karpiński

Karpiński et al. performed an analysis of the temporal coordination of prosodic and gestural units in Polish task-oriented dialogues [28]. For that matter, they selected two prosodic units: intonational phrases and prosodic prominences. Intonational phrases (IPs) were annotated according to guidelines by [26], and tagged as major IPs and minor IPs. Moreover, phrases that were coherent and grammatically acceptable were tagged as well-formed. Well-formed phrases have no interruptions, no hesitations and are not exaggerated. Prosodic prominences were annotated according to guidelines of the RaP system [16]. In the RaP system, prosodic prominences are called “beats” and two types were defined: “major beats”, which are syllables especially strong and salient perceptually in their contexts, and “minor beats”, which are only moderate strength perceptually. Gestures were annotated including gestural phases and gestural phrases boundaries. These gesture annotations follow Kendon’ segmentation scheme.

In a first study, it was analyzed the synchrony between gestural phrases and major intonational phrases. The number of overlaps between gestural and intonational phrases was counted. A gestural phrase may overlap with the beginning of a major IP, contains a complete major IP, or overlaps with the ending of a major IP. A gesture may overlap with different major IPs at the same time, which explains why the number of g-phrases in these three situations adds up more than the total number of g-phrases. Results, in Table 8, do not show a clear tendency of gestural phrases for aligning with the beginning or the ending of intonational phrases, if anything it can be observed a general “centering tendency” for the alignment. The authors stated that this centering tendency was especially true for g-phrases and major intonational phrases semantically related.
In a second study, they tested if the phonological synchrony rule proposed by McNeill [44] holds in their data. This rule states that the stroke of the gesture phrase is always completed either before or at the same time as the tonic syllable of the co-occurrent tone unit. This rule is also called PSR. Beats were selected as the tone salient unit, which are also called prosodic prominences or PPs. It was observed that 96% of strokes overlapped with any PP, however only 75% overlapped with strong PPs. Table 9 shows where these overlaps occur. Initial overlaps refer to the situation where only the initial boundary of the stroke lies within its respective strong PP; inclusion refers to cases where strong PP stays within the limits of a stroke; and final overlaps refer to the case that only the final boundary of the stroke lies within the PP. The only of these cases that supports the PSR are the final overlaps, the other cases break the rule. Results show that the rule is broken many times. Karpiński et al. identified certain situations that predispose strokes lasting after the prosodic prominence ends:

- A long stroke with a continuous gestural excursion
- A long stroke with repetitions. The first realization always meets the PSR but the following ones not necessarily.
- Strokes realized not fully consciously, being a copy of a previous ones.
- The recoil phase may exceed the boundary of the strong PP syllable.
- When phrases have initial strong prominences and a pause immediately after. If the stroke starts within the first syllable, it tends to continue after its final boundary.
- With not well-formed intonational phrases.

<table>
<thead>
<tr>
<th># G-Phrases overlapping with the beginning of a major IP</th>
<th># G-Phrases overlapping with the beginning and end of a major IP</th>
<th># G-Phrases overlapping with the end of a major IP</th>
</tr>
</thead>
<tbody>
<tr>
<td>223</td>
<td>104</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td></td>
<td>106</td>
</tr>
</tbody>
</table>

Table 8 - Number of gestural phrases (G-Phrases) overlapping any major Intonational Phrase (major IP).

<table>
<thead>
<tr>
<th>Initial overlaps</th>
<th>Incision</th>
<th>Final overlaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>26%</td>
<td>46%</td>
<td>38%</td>
</tr>
</tbody>
</table>

Table 9 - Overlaps between the stroke and the strong prosodic prominence

Karpiński’s work presents an analysis of the synchrony between two pairs of gestural and intonational features: gestural phrases and intonational phrases, and stroke phases and syllables containing strong prosodic prominences (major beats). For the first pair it was observed that there is a tendency between semantically related IPs and G-phrases to overlap in time. However, little more can be said about this synchronization. For the second pair, it was observed that there was a clear overlapping between strokes and
prosodic prominences (96% in their data). However, they focused the study in strong prosodic prominences, which gave less strong results (75%). They observed that in some cases strokes end after the end of the syllable containing the prosodic prominence, which breaks one of McNeills synchrony rules.

2.3.8. Leonard and Cummins

Leonard and Cummins examined the temporal relation between beat gestures and accompanying speech [37]. In a first experiment they analyzed if subjects tolerate altered timing between gesture and speech. In the second study they evaluated how different anchor points in speech correlate to other anchor points in gesture.

For the first experiment it was recorded a set of videos of a subject speaking. The face of the speaker was hidden to avoid face to speech asynchronies cues. A group of people evaluated the naturalness of the videos, which had the speech channel displaced in time relative to its original position. This experiment shows that there is an asymmetric perception of naturalness depending on which is the displacement direction. If gestures precede the speech from 0 to 0.4 seconds, the perception of naturalness is relatively equal to non-displaced data. However, displacements of as little as 0.2 seconds of anticipation of speech over gesture result in large naturalness droppings. In Figure 4 it is shown a distribution of how gestures should be temporally aligned in respect to speech. Parameters that define this distribution are not provided because the study was not precise enough to calculate them.

![Figure 4](image)

**Figure 4 – Probability distribution of the timing of a gesture with respect to a speech anchor presented by Leonard and Cummins.**

In the second experiment it was analyzed the temporal relations between several gesture and speech anchor points. Several subjects were asked to read a short text and perform solely three beat gestures on three prominent syllables chosen beforehand. The motion and the speech of subjects were recorded to be able to analyze hand velocities and prosodic parameters of speech. Five kinetic landmarks for movement were obtained:
movement onset, peak velocity of extension phase, point of maximum extension, peak velocity of retraction phase, and termination of gesture. From the speech waveform, these three landmarks were calculated: vowel onset of the stressed syllable in each word, the estimated P-center [52], and the pitch peak within the stressed syllable. P-centers were computed following the method proposed by Cummins and Port [15]. These are placed half way through a local rise in the intensity envelope of the filtered waveform, where band pass filtering is first used to eliminate energy below 500 Hz and above 1500 Hz.

Results of the study are shown in Figure 5. These suggest that regardless of the speech anchor considered, the intervals between the gestural and speech landmarks display a fixed pattern of variability, with the point of maximum extension showing the most consistency in its relative timing with respect to each potential speech anchor. The variance of relative timing between the apex of the gesture and any of the three speech anchor points is approximately of 0.004 seconds. Moreover, it was observed that the closest speech landmark to the point of maximum extension is the peak of pitch accent on the stressed syllable. Another interesting observation is that, considering that the limit of the box plots show the data range, all apexes occurred between the instant of the pitch peak and 0.2 seconds later.

![Figure 5 - Distribution of the intervals between each of five kinematic landmarks in a beat gesture and three possible speech anchor points.](image)

Leonard and Cummings noted that their studies are highly constrained. In all recordings, subjects were asked to perform beats synchronized with pre-defined prominent vowels. In real scenarios, gesture might occur in other places and shows more complex synchronies with respect to speech landmarks.

From these studies it can be depicted that pitch peaks and apexes of gesture might be appropriate candidates to be used as speech and gesture anchor points for synchronization of gestural and speech channels. From the second study it can be depicted that apexes occur between pitch peaks and 0.2 seconds later. Moreover, the first
study suggests that anticipating up to 0.4 seconds the gesture landmark would not decrease the perception of naturalness in the combined animation. Therefore it can be said that gesture apexes may be located from 0.4 seconds before pitch accents to 0.2 seconds after pitch accents.

2.3.9. Ferré

Ferré studied the timing relationships between iconic gestures and their lexical affiliates in spontaneous French [20]. A lexical affiliate is the word or words deemed to correspond most closely to a gesture in meaning. They annotated intonational phrases and word limits for speech, and stroke and gesture phrase limits for gesture. Speech annotations were made in Praat and then imported to Anvil, which contained the annotation of gestural phenomena. They analyzed the timing relations between strokes and its word affiliate, and gestural phrases and intonational phrases. The study focuses only on iconic gestures. Results show that there is a marked preference for anticipation in the production of strokes in contrast to their word affiliates. However, even though most strokes ended after word affiliate, the difference of both signals ending is not significant. At phrase level, gestural phrases versus intonational phrases, the anticipation tendency remain the same, with a clear preference to start gestural phrases before starting intonational phrases. However, as opposed to word affiliates, intonational phrases present a clear tendency to end before to gestural phrases. Table 10 shows the percentages of gesture features stating or ending before speech.

<table>
<thead>
<tr>
<th></th>
<th>% of gestures starting before speech</th>
<th>% of gestures starting after speech</th>
<th>% of gestures ending before speech</th>
<th>% of gestures ending after speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gestural strokes vs. its word affiliates</td>
<td>72</td>
<td>28</td>
<td>12</td>
<td>87</td>
</tr>
<tr>
<td>Gestural phrases vs. intonational phrases</td>
<td>70</td>
<td>30</td>
<td>39</td>
<td>61</td>
</tr>
</tbody>
</table>

Table 10 Percentage of gestures starting/ending before or after speech
2.3.10. **Summary of Speech to Gesture Synchronization Studies**

The following table draws an overview of all the works presented above. This table shows the gestural and speech anchor points, and the synchrony rules presented by the authors.
<table>
<thead>
<tr>
<th>Study</th>
<th>Gestural anchor point</th>
<th>Speech anchor point</th>
<th>Synchrony rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valbonesi et al. [60]</td>
<td>Gesture focal points (local maxima and minima of hand traces motion, or gesture’s stroke)</td>
<td>Speech focal points (Computed as a combination of these audio values: F0 Peak, minimums of the frequency gradient, and maximums the amplitude)</td>
<td>Either speech focal points occur within time windows centered in motion gesture focal points, or occur within the duration of gesture stroke.</td>
</tr>
<tr>
<td>Loehr [40]</td>
<td>Apices of strokes, and gestural phrase boundaries</td>
<td>Pitch accents, and intermediate phrase boundaries</td>
<td>Pitch accents and stroke apices co-occur within 275 msec., and gestural phrases start and end slightly before than start and end points of intermediate phrases. Although the latter synchrony rule is looser than the first one.</td>
</tr>
<tr>
<td>Yasinnik et al. [61]</td>
<td>Hits (single frame), and pauses between gestures</td>
<td>Pitch accents, and intonational phrases</td>
<td>Hits occur within syllables containing a pitch accent, and pauses between gestures are longer between intonational phrases than within them.</td>
</tr>
<tr>
<td>Shattuck-Hufnagel et al [54]</td>
<td>Hits (single frame)</td>
<td>Pitch accents and intonational phrases</td>
<td>No synchrony rules presented</td>
</tr>
<tr>
<td>Esposito et al. [19]</td>
<td>Hit extent</td>
<td>Pitch accent and boundary tones within pitch accented words</td>
<td>Hit extents tend to overlap with pitch accents or boundary tones within pitch accented words.</td>
</tr>
<tr>
<td>Jannedy and Mendoza-Denton [25]</td>
<td>Apices of strokes</td>
<td>Pitch accents</td>
<td>Apices always co-occur with pitch accents</td>
</tr>
<tr>
<td>Karpiński [28]</td>
<td>Gestural phrases and strokes</td>
<td>Major intonational phrases (IPs), syllables containing strong prosodic prominences (PPs)</td>
<td>Gestural phrases tend to align to semantically related IPs. Strokes tend to overlap with syllables containing strong PPs</td>
</tr>
<tr>
<td>Leonard and Cummings [37]</td>
<td>Apices of beat gestures</td>
<td>P-center, vowel onset, and pitch peak within the stressed syllable</td>
<td>Gestural apices occur very close to pitch accents, following them up to 0.2 seconds. Gestures may anticipate up to 0.4 seconds its co-recorded speech and still be perceived as synchronized to speech</td>
</tr>
<tr>
<td>Ferré [20]</td>
<td>Strokes of iconic gestures and its gestural phrases</td>
<td>Word affiliates and intonational phrases</td>
<td>Strokes tend to anticipate word affiliates. Gestural phrases tend to anticipate and last longer that intonational phrases</td>
</tr>
</tbody>
</table>

Table 11 – Overview of the speech to gesture synchronization studies
Anchor points presented in previous works can be classified in two groups: discrete (occurring in a single instant of time), and the ones with duration. Anchor points falling in the first group are the gestural focal points presented by Valbonesi, the hits presented by Shattuck-Hufnagel, and stroke apices for gestures; and Valbonesi’s speech focal points, and pitch accents for speech. All synchrony rules that use discrete anchor points state that these occur very close to their correspondent anchor points in the other modus. For example, Loehr says that the apex of the stroke occur mostly within 275 msec. of distance of its correspondent pitch accent. The data analyzed by Leonard and Cummings seems to contradict this rule. Their data shows that the apex of the stroke occurs at the same time than the pitch accent or after it up to 200 msec. Leonard and Cummings also observed that subjects in their study tolerated well anticipations on the gestures. It was possible to anticipate up to 400 msec. the original gestures in respect to its corresponding speech without perceiving incorrect synchronizations. The latter observation may be used to modify their original synch rule, by allowing apices of gesture from 200 msec. before to 200 msec. after pitch accents. Thus resembling Loehr’s results.

Anchor points associated to a period of time are strokes and gestural phrases for gestures; and intermediate phrases, intonational phrases, syllables and pitch accents for speech. There is a larger divergence about synchrony rules between anchor points with duration. One of the few rules that is depicted in several studies is the observation that gestural phrase onsets tend to anticipate intonational phrase onsets [36][20].

Synchrony rules for discrete features are more deterministic than continuous anchor points. For this reason, in the present study it was decided to work only with discrete anchor points. Apexes of discrete gestures (hits) are used as gesture anchors, and pitch accents as speech anchor points.

From all synchrony rules presented above, it was decided to use Loehr’s and Leonard’s approach that suggested that hits occur within a time window anchored in the pitch peak point. The size of the window defined by Loehr is 550 msec. large, and from Leonard’s work can be depicted a 400 msec. window. In the present study it has been used an averaged window of 500 msec. of length. In terms of how the window is anchored to the pitch accents, McNeils suggested in [30] that “the stroke of the gesture phrase is always completed either before or at the same time as the tonic syllable of the concurrent tone unit”. This suggests that the anchor point of the window should be defined at the end of it. Therefore, hits may occur only from 500 msec. before a pitch accent until the exact time of the peak of the pitch accent. Further analysis showed that either in Loehr’s and Leonard’s data, and also in the data used in this study, hits occur in a distribution centered closely to the peak of the pitch accent. Therefore, it had been better to anchor the time window at its central point. This will be discussed in the Future Work chapter.
3. Research objective

The work presented here tries to achieve two objectives, which were presented in chapter 1.2. The first objective is to define a set of synchrony rules between gesture and speech. The second objective is to analyze the impact on the perception of naturalness of these rules, when these are used in a gesture synthesis system.

The first objective it has been solved in the previous chapter. To summarize it, the synchronization rules that are used are:

- A gesture alignment point should either anticipate or co-occur a speech alignment point. Anticipation must be less than 500 msec. No other hits may appear between one hit and its associated pitch accent.
- A speech alignment point corresponds to the peak of a pitch accent. A gesture alignment point is situated at the apex of the stroke of a hit gesture.

The second objective requires of a qualitative study to be accomplished. A set of people has evaluated the impact of the synchrony rules. The gesture synthesizer presented in [1] was taken as a basis for a new animation system.
4. Methodology

4.1. Gesture synthesizer

The animation synthesizer used in this study is similar to the one presented in [1]. This synthesizer uses a set of pre-recorded animation sequences, which were obtained using a MOCAP system. It reproduces these animations at the same time that reproduces speech audio data. The audio and the animations were recorded separately, and do not correspond to the same text. Even though the synthesis mechanism is simple, the generated animations are perceived as natural most of the time. Only once in a while, the misalignment between gestures and speech is perceptible. These results were surprisingly good, and encouraged the current work, which is a continuation of the previous one.

One of the strong points of the synthesizer used in [1], is that uses entire MOCAP sequences for utterances. Other systems break the MOCAP sequences in small pieces (for example, gesture phases), which afterwards are combined to create gesture flows. When breaking the animation sequences, the natural flow of gestures is also broken, resulting in less natural synthesized animations. In order to avoid repetitiveness on animation playback, the synthesizer uses a database with several animations. Every time that is necessary to reproduce gestures, the synthesizer randomly selects one of these animations, and starts reproducing it at a random instant of time. Therefore, the probability to reproduce repetitive animations is minimized.

The synthesizer is also capable of selecting appropriate animations regarding its emotional content. These are selected to accompany speech that carries the same type of emotion. However, as it will be explained below, in the present study the emotions in speech and gesture have been restricted to a single one. Therefore, this feature of the synthesizer is not used.

Figure 6 shows the block diagram of the computer application presented by Antonijoan et al. in [1]. This diagram illustrates that the entries of the system are: a text accompanied with emotional tags, and a MOCAP database. The system creates speech by using a Text To Speech system (TTS). This synthesizer provides information about when speech starts and ends for every utterance. This information is used by the animation synthesizer to be able to start and end gestural animations according to its accompanying speech. The animation synthesizer selects an animation from de database and reproduces it at the same time that plays back the audio obtained from the TTS. The output of the system is a video that shows an avatar that is animated and speaking. The only synchrony rule between gestures and speech used is that speech utterances start and end at the same time than gesture animations. If no speech is present, the synthesizer selects resting animations to be reproduced.
A new synthesizer system has been built based on the architecture of the previous one. The new system had to be able to generate these three animation modes:

- **Original animations.** These are composed by the speech and the gestures that were originally performed by the actor whose voice was recorded. The audio (speech) and the animation (gestures) are reproduced with the same time alignment as they were recorded. This animation mode serves as reference for being high quality animations. The synchronizations between speech and gestures should be naturally fulfilled.
- **Random animations.** Speech and gestures come from different captures. These are reproduced together without applying any synchronization rule. This mode serves as reference for generating the least quality animations that the synthesized mode should create.
- **Synthesized animations.** Speech and gestures come from different captures. These are reproduced by following synchrony rules presented in the present study. The animations created in this mode are compared to the other two. Ideally, this mode generates animations of greater quality than the random ones, but not as high as the original ones. The higher quality, the better this system will demonstrate to be.

Figure 7 shows the blocks diagram for the new animation synthesis system. The diagram illustrates which are the three possible outputs of the system: original animations, random animations, and synthesized animations. Original animations use the original gesture animations recorded with a MOCAP system. These animations and their associated audios are played back with the same time alignment as they were recorded. The random animations are generated using gestures stored in a database of pre-recorded MOCAP data. The same MOCAP database is used for generating synthesized
animations. However, these also use the anchor points in speech (prosody alignment points). Gesture alignment points are computed in pre-process and stored in the MOCAP database. The synchronized gesture synthesizer ensures that the synchrony rules between gestures and speech are fulfilled.

The synthesizer system is implemented in C++. It uses Irrlicht [21], an open source game engine. This engine makes it easy to load and visualize 3D models and to playback the animations selected by the animator synthesizer. The avatar is represented by a 3D rigged dataset.
4.2. **Data acquisition**

In order to perform the proposed experiments it is necessary to capture gesture and speech data. The gesture data is used to fill the database of gestures that uses the animation synthesizer system. The speech data is used to be reproduced along with the animations selected for the synthesizer system.

A corpus of MOCAP and video files was recorded. The corpus consists of 30 recordings that last slightly more than one minute each. Two actors were recorded: Dani and Ruben. For each recording, actors were asked to perform an improvised monologue. Each monologue had to be done with a single emotional state. Emotions are: aggressive, happiness, sensual, sadness and neutral (no emotion). For each emotion were recorded 3 clips per actor.

The name of an emotion, such as “aggressive”, it is not enough to precisely define it. Differences of interpretations may derive in differences on the performances of the two actors. To avoid that, topics of the monologues were discussed in advance to agree about the characteristics of the current emotion. Moreover, actors listened to a set of audio recordings before acting. These recordings contain the voice of an actor performing the exact emotions.

The MOCAP recording session took place on January 2011 at the MediaLab of La Salle. 24 MOCAP cameras were used to record at 120 samples per second. The whole session was additionally recorded with a video camera and a clip-on wireless microphone. The set-up is illustrated in Figure 8.

![Figure 8 - Set-up of the recording session. An optical MOCAP system with 24 infra-red cameras (left), a clip-on wireless microphone, and a videocamera. Actors were wearing black MOCAP suits with white markers.](image-url)
A process of “clean up” was done over the MOCAP files. This process detects and corrects possible errors on the captured data. It was performed by the MediaLab staff, and its results stored in C3D files [62]. C3D data files store 3D coordinates of real measurements of obtained by the MOCAP system. 3D coordinates correspond to positions of the markers relative to a global coordinate system. This format has been widely used in the field biomechanics photogrammetry. On the other hand, motion captured data was imported into Autodesk 3DS Max and exported to .X files [12]. These files are part of the Microsoft DirectX specification. These allow storing 3D articulated models and their associated animations. Many rendering engines or game engines are capable of loading .X files. This is the case for the Irrlicht game engine.

The video channel was captured with a fixed-position video camera, and the speech with a clip-on microphone attached to the flap of the actor. Both channels were saved in a single video file. This video has several uses, such as: extracting the prosody signals; temporally aligning MOCAP data and audio data; and analyzing the shape and timing of gestures. A frame of the captured video is shown in Figure 9.

Figure 9 - Image obtained during the capture session. Three channels of data were captured at the same time: an audio channel from a clip-on microphone, a MOCAP channel, and a video channel (a frame of which it is shown in the figure).

The current study is interested only in discrete gestures. These are gestures that their motion flow is interrupted by abrupt changes on the direction or velocity. This type of gestures is hardly present in soft movements. Aggressive captures are the ones that show a greater number of discrete gestures. Moreover, their gestures are usually
prominent and, therefore, more easily detectable. For these reasons, this study only used aggressive recordings. Other data files are kept for future studies. The complete corpus is especially useful for gesture studies involving emotion analysis or emotion synthesis.

4.2.1. Data alignment

Video and motion data is captured with different equipment: a MOCAP system and a video camera. The data files obtained in both equipment start at slightly different times. These data channels must be aligned in time so that it is possible to reproduce the original sequence on the “Original Animations” mode. Anchor points in both video and MOCAP data is computed to make this alignment possible. These anchor points are:

- Clapping hands at the beginning of the capture. This hand movement is discarded for reproducing animations, but it is only used for aligning video and MOCAP data. The audio waveform clearly shows the instant of time where both hands come in contact, as seen in Figure 10.
- Gestures with abrupt changes on direction. The instant of time where direction change occurs is identified easily either in the video as in the MOCAP data.

Anchor points allow computing temporal differences between MOCAP and video channels. The events used to compute anchor points sometimes last for more than one consecutive frame. This leads to misalignments of as much as few hundreds of a second. In order to increase robustness on these measurements, several anchor points are computed for each file. The final time displacement for a file is computed as an average of the time alignments computed for all anchor points.
Figure 10 - The instant of clapping hands corresponds to a peak of intensity visible on the sound waveform (shown at the bottom of the figure).

4.3. **Gestural alignment points identification**

4.3.1. **Gestural signals**

Gesture signals contain information about how are performed gestures by hands and arms. There are other parts of the body, which movements may be considered gestures. For example, head nods. However, in this study only arm and hand movements are considered.

Information about gestures is contained in two channels: MOCAP data and video images.

MOCAP data contains 3D positions of several markers. Markers are reflective balls that are attached to the body suit of the actor during capture. Depending on the purpose of the capture session, the number and positions of the markers may vary. For this study a general full body configuration has been used. Figure 11 shows the marker configuration used to capture the corpus in this work. Markers LFIN and RFIN
correspond to left hand and right hand. The whole configuration of markers is used for creating full body animation.

Marker positions are represented in a global coordinate system. Hands positions are relative to a global center of coordinates. If the actor moves from its original position, the positions of his or her markers will also be displaced. For example, if the actor walks to the front with his or her hand attached to his or her chest, the marker of the hand will show the displacement to the front. It is not possible to know if this displacement is consequence of the whole body moving forward or consequence of moving only the hand. For studying gestures, it is necessary to convert the marker positions to a local coordinate system that does not vary depending on full body displacements. To do so, the marker CLAV is used as reference point of a local coordinate system. CLAV marker corresponds to top chest position. The local hand position is computed by subtracting the top chest position to the global hand position.

![Figure 11 - Markers locations that were used in the MOCAP sessions](image)

From hand positions it is possible to obtain several kinematic parameters. For example, Figure 12 shows the velocity and the elevation of hands. From these graphs it can be depicted that some gestures have been performed, and it can be deduced some information about how their shape it is. Looking at Figure 12 it can be depicted that two gestures are performed between 2 and 4 seconds. The first stroke took place between 2.2
and 2.8 seconds and involved only the left hand. The second stroke took place between 3.2 and 3.8 seconds and involved both hands. Both strokes implicate raising hands. Left hand strokes ended with lowering hands back to its original elevation (resting position), but right hand remained high, presumably to start a new gesture stroke from this position.

![Kinematic signals from hands. The top figure shows right hand signals, and the bottom figure shows left hand signals. Velocities are displayed with dotted lines, and hand elevations with continuous lines. The abscissa axis shows the time in seconds.](image)

Kinematic parameters are computed using the Matlab software from The Mathworks Inc. MOCAP files in the C3D format are loaded using the Matlab ports of the C3Dserver from Motion Lab Systems, Inc. C3Dserver is used to extract the positions of hand and top chest markers. Custom-made routines are used to compute local coordinates and kinematic parameters.

Video images are only used to be evaluated by the human eye. It is possible to use image-processing algorithms to extract signals from the video image. However, MOCAP data is more precise and is inherently tridimensional. Data obtained from image processing of a single static camera video would provide two-dimensional information.

The gestural data has to be segmented in gestural phases. This is a complex task, which is not easily accomplished by automated algorithms. Video visualization has been crucial for segmenting gestural data and to detect gesture anchor points, as it is discussed below.
4.3.2. Gesture alignment points

A gesture phrase may be segmented into several phases. Kendon suggests [32] that a gesture have these phases: preparation, stroke, hold, retraction and recoil. The stroke is the salient gestural movement. The stroke is differentiable from the other phases in velocity and acceleration. For example, in Figure 12, the velocity of both hands remains close to zero all the time except when the strokes occur (near 2.5 and 3.5 seconds). This phase has been used in many studies to find correlations with speech [30], [44], [60], and [40].

Gesture apex is the point of maximum extension of a gesture. It has been used by several authors to find correlations with prosody, such as in Rusiewicz’s [51] and Leonard’s [37] works. In the latter work, there were studied several gesture prominences as candidates of anchor points. These are: the movement onset, peak velocity of extension phase, apex of the gesture, peak velocity of retraction phase and termination of gesture. It was observed that the apex of a gesture showed less temporal variability with respect to speech that any other point within the gesture.

Gestures may be classified as discrete or continuous. What differentiates a discrete gesture from a continuous gesture is that the first one contains a “hit” while continuous do not. Hits are abrupt stops that break the flow of gesture. They appear as bouncing, jerky movements, changes in the direction, or as complete stops in movement. Hits have already been used to identify time alignment with prosody prominences in Renwik, Shattuck-Hufnagel, Yasinnik et al. works [61] and [54].

Gesture apices and hits are similar features. In Karpiński and Jarmolowicz-Nowikow work [27] these were identified as the same phenomena. However, there are some differences between them. For example, hits only occur in discrete gestures. Continuous gestures do not have hits, but they may have apices. Other differences are that apices relate to moments of maximum expansion, while hits not necessarily occur in expansion but can occur in simultaneous changes of direction of gesture, which may be contraction movements. Therefore subtle differences are present between hits and apices definitions. In this work it is chosen to work with hits, because these are proven to correlate well with prosody accents [61], these are easy to identify and time precisely due to its discrete nature. Moreover, within the subset of discrete gestures, hits are more general than apices. All discrete gestures contain a hit, while it is not necessarily an apex.

4.3.3. Identification methodology

Hits may be identified analyzing two types of data: either analyzing kinematic signals of gestures, computed from MOCAP data; or watching the video images. Below it is discussed difficulties and advantages of both strategies.
Figure 13 shows the kinematic signals obtained from a MOCAP capture. From velocity signals (displayed in dotted lines) it can be depicted that there are some gestures at the end of the shown segment. Hits are movements that end with a drastic slow down. Looking at the example, one may say that there are three hits, which will end at 5.5, 5.8 and 6 seconds respectively. However, when looking at the video, which was recorded at the same time than the MOCAP capture, it can be noted that first and third velocity ‘mountains’ correspond to strokes, but the second one correspond to a preparation phase. In the preparation phase, both hands are raised to start the consecutive stroke, which consists on lowering arms abruptly. This example illustrates how analyzing only kinematic signals it is possible to find false-positive hits.

![Figure 13 - Velocity (dotted lines) and elevation (continuous lines) of left and right hands. The abscissa axis shows the time in seconds. Gestures occur at the end of the graph.](image)

Figure 14 shows several frames from the video that correspond to the time margin that contains the three velocity ‘mountains’ of Figure 13. The first frame corresponds to the instant of time just preceding the first mountain (more or less at 5.2 seconds); the second frame occurs when hands are opening (5.4 seconds approximately); the third frame shows the instant of time when the stroke of opening hands finishes (5.5 seconds); the fourth frame occurs when hands are raising in the preparation phase (5.6 seconds approximately); the fifth frame (the first one in the second row) shows the instant of time when there is a hold before the second stroke (5.8 seconds); the sixth frame corresponds to the second stroke movement of lowering hands (5.9 seconds); finally, the seventh frame corresponds to the instant of time when the second hit has arrived to its end (6 seconds).
As seen in the previous example, the video images provide more information about the gesture phases than kinematic signals. For this reason it was decided to use the video to detect and annotate gestural prominences.

The method used is based on Shattuck-Hufnagel et al. proposal [54]. Alignment points coincide with the end of hit gestures. These are easily identifiable because the image of the video image usually sharpens at this frame. Figure 15 displays four consecutive frames corresponding at the end of a hit gesture. The first three frames show the hands of the actor blurred, this is particularly visible on the white markers. This is because the actor is moving the hands quickly. The fourth frame shows the markers perfectly in focus. This frame corresponds to the instant of time when the hit has arrived to its end, and therefore hands slowed down. The fourth frame is marked as the gesture alignment points.
Figure 15 - Four consecutive frames showing the end of a hit stroke. The image has been zoomed to better appreciate the hands of the actor.

Gesture alignment points are tagged using the ANVIL program [33]. This program allows visualize the video images frame by frame and annotate them with temporal tags. These may be labels with duration or labels that correspond to a single instant of time. To perform the annotation task, a person watched the video searching for hit gestures. Later on, it was looked which was the ending frame of each hit. This was done using the suddenly sharp frame technique mentioned above. Figure 16 shows an example of a tagged video with gesture alignment points in ANVIL. Hits are tagged at the tier “hit” and are displayed with a small square each one. In order to import videos into the Mac OS version of ANVIL, it was necessary to transform them to the QuickTime Movie file format (MOV), and to recode then with the Cinepak video codec.

Figure 16 - Screen capture of the ANVIL annotation tool. The annotation window is placed at the bottom of the image. Gesture alignment points are defined at the “hit” tier.
The tags annotated in ANVIL are export to the ANVIL export format. The animation synthesizer is capable of parsing this file format to extract the temporal positions of hits from it. This file format follows an XML structure. Two video files were annotated: *Dani_agressive_2* and *Dani_agressive_3*. Figure 17 shows an excerpt of the annotation of *Dani_agressive_2*. Gesture alignment point times are defined into the `<track name="hit">` tag. As it can be observed in the figure, 53 hits were found at this video file. The tag `<track name="valid animation">` indicates which are the limits of valid animations to be used by the animation synthesizer. Large pauses, clapping hands used for video-MOCAP synchronization, and binding MOCAP poses are set out of the valid animation interval.

```xml
<?xml version="1.0" encoding="UTF-16"?>
<annotation>
  <head>
    <specification src="estudi_hits.xml" />
    <video src="AgressiuDani02_cinepak.mov" master="true" />
    <info key="encoding" type="String">UTF-16</info>
    <info key="coder" type="String">Marc Antonijoan</info>
  </head>
  <body>
    <track name="hit" type="primarypoint">
      <el index="0" time="7.20721" />
      <el index="1" time="7.87454" />
      <el index="2" time="9.37604" />
      ...
      <el index="52" time="71.13781" />
    </track>
    <track name="valid animation" type="primary">
      <el index="0" start="6.23957" end="71.53821" />
    </track>
  </body>
</annotation>
```

Figure 17 - Excerpt of the annotation of gesture alignment points for the *Dani_agressive_2* video file.
4.4. **Prosody alignment points identification**

4.4.1. **Prosody signals**

The prosody alignment points used in this study are the pitch accents. Pitch accents occur in stressed syllables, which are characterized by having a pitch prominence, relative longer duration, and relative higher intensity than other syllables in the same word. In order to detect stressed syllables these prosody signals are obtained: pitch contour, intensity contour and duration.

All prosodic signals are computed with the speech analysis software Praat [8], see Figure 18. Intensity and Pitch signals are easily computed with Praat by using the functions “To Intensity” and “To pitch”. These functions automatically compute the intensity and pitch contours. Both signals are obtained in a 120 samples per second rate, which matches the MOCAP datasets sampling rate.

Pitch automatic computation in Praat is not infallible. Doubling and halving errors may occur. In order to reduce errors it is possible to define a maximum and minimum frequency values, Praat will discard frequency candidates outside of this range. The plugin for Praat *Momel and Intsint* [2] has a function called “Detect f0” which automatically finds appropriate maximum and minimum frequency values.

Figure 18 - Praat editor with an audio file loaded. Intensity and pitch signals are shown in yellow and blue respectively in the lower tier.
Duration of syllables is also computed using Praat. However, this process requires more supervision than processes for obtaining intensity and pitch signals. For this purpose the plugin for Praat EasyAlign [22] was used. This plugin tries to find the duration of each syllable and phoneme automatically. Results may show some inaccuracies, therefore it is necessary to review them and correct the errors. EasyAlign takes the transcription of the whole text as an input and outputs the time limits of several phonetic events, as it is shown in Figure 19. The entire process consists of these three phases:

- **Macro-segmentation**: it finds time limits of all utterances. This phase takes as input the transcription of the text split into utterances, and it creates the ortho tier. It is recommendable to check if the utterance limits are properly set. This data will be used in future phases, so errors will accumulate. If errors are found, it is possible to adjust the limits by hand.

- **Grapheme-to-phoneme conversion**: it transcribes the utterances text into phonemes. It is necessary to define the language that was used in the audio recordings. It is recommendable to review if the transcription is correct and modify it if necessary, because this will affect the next phase.

- **Phone segmentation**: it splits the utterances in words, syllables and phonemes. It automatically detects limits for all these events. It is recommended to review the results and adjust incorrect limits.

Although EasyAlign provides data with some errors that should be corrected manually, the use of this tool significantly speeds up the annotation process.

![Figure 19 - Results of the EasyAlign alignment process. The tier ortho contains utterance limits; the tier phono contains the phonetic transcription; the tier words contains the word segmentation; the tier syll contains the syllables segmentation; and the tier phones contains phoneme limits.](image)

Praat allows exporting the syllable limits into a text file. From there it is possible to calculate their duration easily.
4.4.2. **Prosody alignment points**

There are many ways to detect prosodic prominences in the literature. For example: Pitch peak of the stressed syllable; The estimation of the P-center using a method defined by Cummins and Port [15], which is based in the Scott’s model [52]; Speech focal points used by Valbonesi et al. [60], which are a combination of pitch accents, minimum and the second minimums of the frequency gradients, and the maximum and second maximum value of the audio wave amplitude; Vowel midpoints, because vowels are more influenced by stress than consonants; Tone unit nucleus, which is the last big change of pitch within a tone unit, used by McClave [42]. However, the most used prominence is the pitch accent defined in the Autosegmental and Metrical model, used by Loehr [40], Yasinnik et al. [61], Shattuck-Hufnagel et al. [54], Esposito et al. [19], Jannedy and Mendoza-Denton [25] among others.

Pitch accents are events that occur within the stressed syllable. However, which is exactly point where the pitch accent is located? The prosody alignment point should be situated in a single instant of time. Some of the strategies that may help finding these points are: taking the pitch or intensity maximum within the duration of the syllable, the pitch or intensity maximum within the nucleus of the stressed syllable, or selecting the middle point of the nucleus of the stressed syllable. The syllable nucleus is the most prominent part of the syllable. In Spanish, this generally corresponds to a vowel. Figure 20 illustrates an example where there have been tagged three types of pitch accent anchor points: the intensity and the pitch maximums within the stressed syllable, and the center of the syllable nucleus. The stressed syllable is indicated with an asterisk in the “syll” tier, and the syllable nucleus is also indicated with an asterisk in the “phones” tier. In this example the pitch maximum and the center of the syllable nucleus coincide in time.

![Figure 20 - Prosody alignment point candidates: Maximum of intensity within the stressed syllable (square), maximum pitch within the stressed syllable (circle), and the center of the syllable nucleus, which coincides with the maximum of pitch (circle).](image-url)
In this study it was decided to use the center of the syllable nucleus as the prosody alignment point. This decision was made only for simplicity. It is easier to detect and annotate syllable nucleus centers than other pitch accent anchor points. In a preliminary study it was observed that there were no significant differences, in terms of gesture to prosody alignment, between pitch or intensity maximums or syllable nucleus centers and hits.

4.4.3. Identification method

Before looking for pitch accents it is necessary to detect which syllables are stressed. Only stressed syllables will carry pitch accents. Almost all words contain a stressed syllable. Which is the syllable with stress within the word depends on the lexicon of the language. Therefore, detecting stressed syllables requires great knowledge of the language.

In order to annotate stressed syllables it has been used the “syll” tier, which was obtained from the plugin for Praat EasyAlign. It has been added an asterisk at the end of the phonemes of each stressed syllable. The result of this tagging is shown at the Figure 21. Functional words, such as prepositions and determiners are not tagged with stressed syllables. This process is done manually by a native Spanish speaker.

Once stressed syllables are annotated, these have been analyzed in order to decide whether they carry a pitch accent or not. Some syllables are stressed because of the lexicon but do not present any prominence in the spoken text. For example, the text in Figure 21 “estoy en el Medialab”, in the word “estoy” there is a lexical accent at the “toy” syllable. However, in reality, the actor did not stress this syllable in his discourse. The decision of which stressed syllables had a pitch accent was done manually, by listening to the recorded utterances several times.

Pitch accents are indicated with an asterisk on the syllable nucleus phoneme of the stressed syllable. This is done at the “phones” tier. Moreover, an additional tier has been added that is called “Accents”. This tier contains the precise instant where the prosodic alignment point is situated. This instant of time corresponds at the middle point of the syllable nucleus, which is the dominant vowel of the syllable. Figure 21 shows an example of utterance that has been tagged with stressed syllables, syllable nucleus phonemes of pitch accented syllables and prosodic alignment points.
Figure 21 - Utterance with stressed syllables indicated with an asterisk at the “syll” tier, pitch accents indicated with an asterisk at the “phones” tier, and prosody alignment points indicated with an asterisk at the “Accents” tier.

Not all data annotated in Praat is useful for the animation synthesis system. In fact, the only data which is necessary is the one contained in the “Accents” tier. This tier has the time position of the prosody alignment points. In order to export this data from Praat, it is used the option “Save as a short text file”. This option creates a text file with a very simple format, which is easily parsed. Figure 22 shows an example of text file with prosody alignment points.

Figure 22 - Example of a text file containing prosody alignment points. The first line indicates the number of alignment points in the file. Each floating number corresponds to the instant of time (in seconds) when an alignment point occurs.
4.5. **Subjective test**

One of the main objectives of this project is to analyze the impact on the perception of naturalness of the alignment rules. These rules define how the prosody and gesture alignment points are related in time. In order to evaluate the perception of naturalness a subjective test was performed.

The test is formed by surveys, which ask participants to evaluate the level of naturalness of some videos containing an avatar speaking. A video is considered natural if the animations of the avatar correlate with speech as if a real human performed them.

Three types of videos were generated with the animation synthesis system:

- Original: animations and speech are aligned as they were recorded, thus reproducing the original performance of the actor.
- Random: animations are randomly aligned with speech. No synchrony rules between speech and gesture are used.
- Synthetic: animations are aligned with speech using the prosody to gesture alignment rules.

The objective of the study is to evaluate if synthetic animations perform better in terms of perceived naturalness than random animations. Moreover, it was expected that synthetic animations have as much as possible the perceived naturalness of original animations.

4.5.1. **Video generation**

Videos of the test show animations of a virtual character speaking along with the audio of the speech. The avatar performs non-verbal language by means of gestures. Figure 23 shows a frame of a sample video. All videos display a male avatar engaged in a speech while performing gestures. The face of the avatar has been darkened to hide its facial expression. That is for avoiding interferences due to misalignments between facial expressions and speech. Those alignments exist between speech and lip movements, but also between speech and other phenomena. For example, Ekman et al. observed alignments between speech and eyebrow movements [17]. All these phenomena are hidden under the dark face.
In order to create the gesture animations, portions of the MOCAP data have been used. When these animations were recorded, the actors were asked to perform only “beat” gestures. This type of gestures do not carry semantic information, therefore do not align with the text due to its meaning. It was decided to use only this type of gestures because the focus of this work is in studying the alignment between gestures and prosody, and semantic alignments might mask this phenomenon.

From all animations on the captured corpus, it was decided to use only the ones of the “aggressive” type. These animations contain a great number of hits (discrete gestures), which are the type of gestures that are studied in this work. Moreover, in “aggressive” videos most of the hits are very prominent, which facilitates its identification. More precisely, “Aggressive Dani 1” was used for animations in original videos, and “Aggressive Dani 2” and “Aggressive Dani 3” for animations in random and synthetic videos. All audios with speech come from the “Aggressive Dani 1” capture.

Each video contains a single speech utterance with its corresponding animation. As mentioned above, speech utterances are taken from “Aggressive Dani 1”. This capture contains a total of 14 utterances. A list of all videos generated with these utterances is shown in Table 12.
It should be noted that videos *synth_4* and *synth_14* do not exist. This is because the gesture synthesizer system could not find any animation sequence within “Aggressive Dani 2” or “Aggressive Dani 3” captures that meets the speech to gesture alignment restrictions. Using more data from the MOCAP corpus might solve this problem. However, 12 synthesized animations were considered enough for this study.

### 4.5.2. Test description

The test is based in the P.800 recommendations of the International Telecommunications Union [24]. More precisely, the Comparison Category Rating (CCR) method was used. This method consists on giving two side-by-side stimuli every time. One of the stimuli is not processed (in this case a original video), and the other is processed (in this case a synthesized video). Users are asked to perform a test composed by several steps. Every step contains a pair of stimuli and a question. Processed and non-processed stimuli are ordered randomly. Therefore, in some steps the first stimuli is processed and the second unprocessed, and in others the first stimuli is unprocessed and the second is processed. This randomization avoids that the user learns which side of the test contains the processed stimuli, which would lead to biased results.

At every step, users are asked which of both videos contain a more natural animation in terms of synchronization between speech and gesture. They can choose one of the responses in Table 13. When a response is chosen, a quality value is given to each of the two videos in this step. Higher values represent higher naturalness of the video. Note that this is a comparative test, giving a higher value to one video involves giving a lower
value to the other. These values are also associated to the category to which each video belongs.

<table>
<thead>
<tr>
<th>Option name</th>
<th>Corresponding value for the first video</th>
<th>Corresponding value for the second video</th>
</tr>
</thead>
<tbody>
<tr>
<td>The first video is much more natural</td>
<td>+2</td>
<td>-2</td>
</tr>
<tr>
<td>The first video is more natural</td>
<td>+1</td>
<td>-1</td>
</tr>
<tr>
<td>More or less the same</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>The second video is more natural</td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td>The second video is much more natural</td>
<td>-2</td>
<td>+2</td>
</tr>
</tbody>
</table>

Table 13 - Response options for each step of the subjective test and its corresponding quality values for the evaluated videos.

In this test three different comparisons are done. The test does not only compare unprocessed videos (original) with processed videos (synthesized), but also compares synthesized videos with random videos, and original videos with random videos. Results of each comparison may lead to different conclusions. Again, in order to avoid that the user learns patterns on how the videos are organized, which may lead to biased results, the three comparisons are alternated. In some steps the user compares between an original and synthesized videos, in other cases between synthesized and random videos and in other cases between original and random videos. The user does not know which type of comparison is performing at every step.

Tests are created using the web platform TRUE (Testing platfoRm for mUltimedia Evaluation) [49]. This platform allows creating and managing subjective test with different types of stimuli, such as audios, audiovisuals, graphics and texts. This is a very convenient tool to spread the test to testers. TRUE allows creating tests conforming ITU recommendations. For the test used in this work, the ITU P.800 recommendation option was selected. Figure 24 displays the configuration page of TRUE.
There had been created two tests containing 13 comparisons (steps) each. The time to complete a test was estimated in 10 minutes. It was considered that this time was short enough to allow the contribution of participants with tight schedules. Since the participation to the tests was voluntary, larger tests would be performed by a lower number of participants. However, committed testers had the option to complete the two tests.

At the time this test was created, TRUE platform did not allow creating randomly sorted stimuli. To emulate randomization it was defined an order in which the three types of comparisons were alternated. The order for both tests is listed in Table 14.
Table 14 – Order in the pair of stimulus (videos) in the two tests.

<table>
<thead>
<tr>
<th>Order</th>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st stimulus</td>
<td>2nd stimulus</td>
</tr>
<tr>
<td>1</td>
<td>rand_1</td>
<td>orig_1</td>
</tr>
<tr>
<td>2</td>
<td>orig_2</td>
<td>synth_2</td>
</tr>
<tr>
<td>3</td>
<td>rand_3</td>
<td>synth_3</td>
</tr>
<tr>
<td>4</td>
<td>synth_5</td>
<td>orig_5</td>
</tr>
<tr>
<td>5</td>
<td>synth_6</td>
<td>rand_6</td>
</tr>
<tr>
<td>6</td>
<td>orig_7</td>
<td>rand_7</td>
</tr>
<tr>
<td>7</td>
<td>orig_8</td>
<td>synth_8</td>
</tr>
<tr>
<td>8</td>
<td>rand_9</td>
<td>synth_9</td>
</tr>
<tr>
<td>9</td>
<td>rand_10</td>
<td>orig_10</td>
</tr>
<tr>
<td>10</td>
<td>synth_11</td>
<td>orig_11</td>
</tr>
<tr>
<td>11</td>
<td>synth_12</td>
<td>rand_12</td>
</tr>
<tr>
<td>12</td>
<td>rand_13</td>
<td>orig_13</td>
</tr>
<tr>
<td>13</td>
<td>orig_14</td>
<td>rand_14</td>
</tr>
</tbody>
</table>

Figure 25 displays an example of how the stimuli were shown to testers. Each test step was shown in a single web page. This page contained two videos side-by-side. Users were able to playback videos separately as many times as desired. When a response was selected the system passed to the next page until the completion of the test. It was possible to leave the test at any time and continue it from the first unanswered step.

Figure 25 – Videos in the test were presented side-by-side. This example shows the first frame of the orig_1 (left) and synth_1 (right) videos.
5. Results

Users took an average of 30 seconds to respond every comparison of the tests. 30 people completed the first test, and 17 completed the second test. All testers are Spanish citizens and speak fluent Spanish. They are between 22 and 67 years old, are college graduates, and do not have visual impairment problems.

For each comparison it was computed the Comparative Mean Opinion Score (CMOS), which is the arithmetic mean of all quality values given by all testers in a comparative test. Since the quality values can take values between -2 and +2, the CMOS value is also within this range. Positive values mean that the first video of the comparison is evaluated as being of higher quality than the second one. In the other hand, negative values mean lower quality results for the first video. In addition, it was computed the standard deviation for all pairs of videos. Results for the comparisons of the first test are shown in Table 15, and for the second test in Table 16.

<table>
<thead>
<tr>
<th>Compared videos</th>
<th>CMOS</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>rand_1 vs. orig_1</td>
<td>0.27</td>
<td>1.21</td>
</tr>
<tr>
<td>orig_2 vs. synth_2</td>
<td>1.60</td>
<td>0.48</td>
</tr>
<tr>
<td>rand_3 vs. synth_3</td>
<td>0.17</td>
<td>1.04</td>
</tr>
<tr>
<td>synth_5 vs. orig_5</td>
<td>-0.13</td>
<td>1.09</td>
</tr>
<tr>
<td>synth_6 vs. rand_6</td>
<td>1.30</td>
<td>0.69</td>
</tr>
<tr>
<td>orig_7 vs. rand_7</td>
<td>1.33</td>
<td>0.54</td>
</tr>
<tr>
<td>orig_8 vs. synth_8</td>
<td>-0.03</td>
<td>0.95</td>
</tr>
<tr>
<td>rand_9 vs. synth_9</td>
<td>-0.23</td>
<td>0.80</td>
</tr>
<tr>
<td>rand_10 vs. orig_10</td>
<td>-1.30</td>
<td>0.64</td>
</tr>
<tr>
<td>synth_11 vs. orig_11</td>
<td>-1.20</td>
<td>0.60</td>
</tr>
<tr>
<td>synth_12 vs. rand_12</td>
<td>-0.23</td>
<td>0.84</td>
</tr>
<tr>
<td>rand_13 vs. orig_13</td>
<td>-1.17</td>
<td>0.82</td>
</tr>
<tr>
<td>orig_14 vs. rand_14</td>
<td>1.40</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Table 15 – Comparative Mean Opinion Score (CMOS) and standard deviation for video pairs of the first test.
<table>
<thead>
<tr>
<th>Video id</th>
<th>CMOS</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>synth_1 vs. rand_1</td>
<td>0.76</td>
<td>0.64</td>
</tr>
<tr>
<td>rand_2 vs. orig_2</td>
<td>-0.24</td>
<td>1.06</td>
</tr>
<tr>
<td>orig_3 vs. rand_3</td>
<td>1.6</td>
<td>0.49</td>
</tr>
<tr>
<td>rand_4 vs. orig_4</td>
<td>-0.18</td>
<td>1.04</td>
</tr>
<tr>
<td>rand_5 vs. synth_5</td>
<td>-0.53</td>
<td>1.24</td>
</tr>
<tr>
<td>orig_6 vs. synth_6</td>
<td>0.29</td>
<td>0.82</td>
</tr>
<tr>
<td>synth_7 vs. orig_7</td>
<td>-1.12</td>
<td>0.76</td>
</tr>
<tr>
<td>rand_8 vs. orig_8</td>
<td>-1.12</td>
<td>0.68</td>
</tr>
<tr>
<td>orig_9 vs. rand_9</td>
<td>1.47</td>
<td>0.70</td>
</tr>
<tr>
<td>synth_10 vs. rand_10</td>
<td>1.53</td>
<td>0.50</td>
</tr>
<tr>
<td>rand_11 vs. synth_11</td>
<td>0.35</td>
<td>0.68</td>
</tr>
<tr>
<td>synth_12 vs. orig_12</td>
<td>-1.41</td>
<td>0.60</td>
</tr>
<tr>
<td>orig_13 vs. synth_13</td>
<td>1.29</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Table 16 - Comparative Mean Opinion Score (CMOS) and standard deviation for video pairs of the second test.

Figure 26 shows CMOS and standard deviations for different types of comparisons. Every error bar in the graphs shows the results for a comparison between two videos. High values in the graph represent high quality for the first video of the comparison and low quality for the second. Table 17 displays CMOS values for all comparisons of the same type.
Figure 26 – Error bars showing CMOS and standard deviations of the comparisons between different types of videos. The left column corresponds to results of the first test, and the right column to results of the second test.
Testers were also asked to describe in which type of movement they have paid attention in order to evaluate the naturalness of speech to gesture synchronization. Not surprisingly, most testers looked at hand and arm movements. However, several of them also looked at head and torso movements, and to a lesser extent at leg movements.

<table>
<thead>
<tr>
<th></th>
<th>First test</th>
<th>Second test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random vs. Original</td>
<td>-1.30</td>
<td>-0.92</td>
</tr>
<tr>
<td>Synthesized vs. Original</td>
<td>-0.73</td>
<td>-1.03</td>
</tr>
<tr>
<td>Synthesized vs. Random</td>
<td>0.28</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table 17 – CMOS of all comparisons of the same type
6. Discussion

Several conclusions may be depicted from results in Figure 26. The comparison between random and original videos shows that original videos are perceived more natural than random ones in most cases. This observation is not striking because random videos do not follow any synchrony rule between gesture and speech. However, in three of the ten comparisons, viewers did not have a clear preference for original videos, as it would be expected. That may be explained with the hypothesis that humans have a great tolerance accepting asynchronies between gesture and speech as being natural. Leonard et al. [37] suggested that this tolerance was greater when gestures tend to precede speech stresses. In fact, the observation of this tolerance was one of the factors that motivated the current work. Since videos that do not follow any synchrony rule provide natural results in some cases, adding few rules would possibly lead to even better results, possibly resembling the naturalness of the original videos.

An example of random and original videos that present a similar level of perceived naturalness is the pair rand_1 and orig_1. Its CMOS value is shown at the first row of Figure 26 with the tag “v1”. Both videos reproduce the first utterance of the test, which has the following text: “esTOY indigNAdo, por eNEsima VEZ estoy en el MEdiaLAB”. Capital letters indicate stressed syllables that contain pitch accents. As often happens, not all pitch accents are accompanied with gestures, but all gestures are accompanying pitch accented syllables. In the original video the avatar produces three hit gestures, which align to the syllables NA, VEZ and LAB. By chance, the random video also has three hits, and these align to the same syllables. Gestures in the random video present a higher level of prominence than the ones in the original video. However, as it can be observed in the results of the test, these gestures are not perceived as being too exaggerated. A possible explanation to this is that the accented syllables are also very prominent. One example of these gestures is shown in Figure 27. The left image displays one frame that corresponds to the gesture that accompanies the syllable –NA- in the random video. The right image displays one frame of the gesture accompanying the same syllable in the original video. The gesture of the random video reaches a higher altitude, which produces a larger displacement of the hand.
When comparing synthesized and original videos, results also show that viewers had a clear preference for original videos. Either CMOS values in Table 17, or video-to-video comparisons of Figure 26 point to this fact. Moreover, as it also happened in the original to random comparison, this preference was not present in three video pairs. It should be noted that results are slightly better for synthesized videos than for random videos. There were a total number of 8 comparisons between synthesized and original videos, which gives a ratio of 3/8 synthesized videos with high degree of perceived naturalness, while random videos presented a ratio of 3/10.

One of the video pairs that showed similar levels of perceived naturalness is synth_8 and orig_8. These are tagged with a “v8” at the Figure 26. The text that was reproduced in these videos is: “riDículo, a demás alumnos miRándote, GENte que te está graBANdó” with stressed syllables in capital letters. The original video has three hit gestures, which are aligned to the syllables DÍ, RÁ and GEN. The synthesized video presents hits aligned to the same syllable, but it has a fourth hit, which aligns to the stressed syllable –BAN-. As it can be deducted from the results, adding a hit does not affect either positively or negatively to the perceived naturalness. This reinforces the idea that it is not necessary that all pitch-accented syllables have to be accompanied with a gesture. Figure 28 displays frames of the two videos when the syllable –BAN- is pronounced. As it can be observed, in the synthetic video the avatar has its arms raised due to an ongoing gesture movement. In the original video hands are close to a resting position, and these would not move right to the end.
Figure 28 - Frames corresponding to the instant of time that the syllable –BAN- from the word “grabando” in the eighth utterance is pronounced. The left image corresponds to the video synth_8 and the right image corresponds to the video orig_8.

In the other hand, the video pair synth_2 and orig_2, tagged with a “v2” in Figure 26, is an example of a clear preference for the original video. These videos reproduce the text “vesTIdo con un TRAje espanTOso riDículo que solo sirve para hacer el paYAso “. In the synthesized video, the syllables with hits are TI, DÍ and YA. However, in the original video, these are TO, DÍ and YA. This difference in the choice of syllables does not seem to be the cause that explains the great differences in perceived naturalness. By observing the videos, it can be depicted that the cause of it is the high level of prominence of gestures in the synthesized video. Gestures in the synthesized videos seem too exaggerated regarding the prominence level of stressed syllables. Figure 29 shows an example of over-exaggerated gesture. These frames correspond to the hit that accompanies the syllable –YA-. While in the original video, the avatar is performing a gesture with a short trajectory, in the synthesized video, the gesture traverses a long path from next to the right hip, passing next to the right shoulder and ending in front of the chest. This over-exaggeration of gestures is a repeating characteristic among synthesized videos with poor results.
Figure 29 - Frames corresponding to hits accompanying the syllable -YA- from the word “payaso” in the second utterance. The left image corresponds to the video synth_2 and the right image corresponds to the video orig_2.

Observing the results of the comparisons between synthesized and random videos, it can be noted that in almost all cases the synthesized videos either outperform or equal random videos quality values. This result reinforces the idea that the synchronization rules presented in this work usually improve the quality of the animations, or at least they do not make them worse.

The only case when the synthesized video has poorer quality results than the random one is marked with a “v11” in Figure 26. This is the comparison between videos synth_11 and rand_11. The text for this utterance is “porque, se supone que tú estás para haCER CLAses como profeSOR en la universiDAD “. Hits of the synthesized video align to syllables CER, SOR and DAD, while hits of the random video align to syllables CER, CLA and SOR. The hit in the syllable -CER- clearly anticipates the syllable pronunciation, however this does not present a problem to associate the gesture with the syllable. This is another example supporting the idea that there is a great tolerance of perceiving gesture anticipations as being natural. The fact that the random video is perceived as being more natural does not seem to come from the different choice of syllables accompanying gestures, but it seems to be caused by the same over-exaggeration of gesture prominences present in synth_2. Figure 30 shows frames from videos synth11 and rand_11 in the instant of time that gestures that accompany the syllable -CER- are executed. The gesture performed in synth_11 is more prominent because it uses both hands and they traverse a long trajectory, while in orig_11 the gesture involves only one hand, which follows a short path.
The pair of videos synth\_10 and rand\_10, which is tagged with a “v10” in the comparison between synthesized and random videos in Figure 26, is an example of a synthesized video clearly outperforming a random video. The text in these videos is: “la verDAD que, ES norMAL que uno se enFAde cuando le TRAen aQUI”. The video synth\_10 has four hits that correspond to the syllables DAD, MAL, FA and QUÍ. The video rand\_10 has 3 hits; two of them are associated to the syllable -DAD- and the other one to the syllable -FA-. The poorer quality of the random video is probably caused by the fact that two of its gestures correspond to a single syllable, which breaks the synchronization rules presented in this study. Moreover, there is another effect that contributes to the lack of naturalness of the video. At the end of the video, the arms of the avatar are raised, which seems to indicate that a gesture is starting (the preparation phase has been entered). Producing gestures in spaces between phrases generally leads to a decrease of naturalness of animations. Figure 31 shows one of the last frames of videos synth\_10 and rand\_10. Arms in the synthesized video are situated at a low altitude, next to a resting position. However, arms in the random video are raised, an indicator that a gesture is being performed.
To sum up, it was observed that synthesized videos are perceived slightly more natural than random videos, but are still far from naturalness levels of original videos. It seems that the effect that mostly decrease the quality of synthesized videos is the production of highly prominent gestures accompanying stressed syllables with low prominence levels. This leaded to the perception of exaggeration in the gestures. Moreover, it was proven that the synchronization between gestures and speech contributed to a higher perception of naturalness. In fact, random videos with poorer quality are the ones that present several hits for a single stressed syllable or that contain hits occurring with clearly unaccented speech, such as the end of phrases.
7. Conclusions and Future Work

This work has presented a methodology for synchronizing speech and gesture data of a speaking avatar, which may be used for creating synthetic animations. Gestures and speech come from different recordings, making possible to re-use gestural animations for a great number of different speech files. Gestures are stored in a MOCAP database.

Two synchronization rules were presented, which are based in previous studies in the subject. These rules have been tested with a subjective test that evaluated the naturalness of some animations that fulfilled the rules. Results showed that the rules improve the naturalness of the videos but the quality of the videos is still far from videos generated with speech and animation captured simultaneously.

Several observations were made that may contribute improving the naturalness of the synthesized videos. For example, synchronizing speech with other gestures than the produced by hands or arms. Sargin et al. [53] suggested that there exist correlations between head gestures, such as nods, and the prosody of speech. Other gestures that may be synchronized with speech are movements of the torso or movements of the legs.

Another observation is that when highly prominent gestures are aligned to lesser prominent speech accents, the resulting animations look exaggerated. The current work did not classify alignment points in terms of its prominence. This classification along with additional synchrony rules that ensure that prominences of similar levels are aligned together may lead to more natural animations.

One of the synchrony rules of this project states that gesture alignment points should be situated within a temporal window that is linked to a speech alignment point. Recent literature on the subject [40] agrees that this window should be centered in respect to the speech alignment point, however in this study the speech alignment point is situated at the end of the window. This was caused by a misinterpretation of McNeill’s statement [30]: “the stroke of the gesture phrase is always completed either before or at the same time as the tonic syllable of the concurrent tone unit”. McNeill compared two elements with duration, while in this study the elements correspond to a single instant of time. The statement “completed before or at the same time” is somehow more flexible for elements with duration than for single instant elements. It will be interesting to repeat the tests realized in this work but adjusting this synchrony rule.

The current study has evaluated the quality of gestural animations and ignored facial expressions. Facial expressions have a great impact on the naturalness perception of speaking avatars, and they also present temporal synchronies with speech. In order to create full body animations it will be interesting to add a facial animation system such as X-Face [3] to the animation synthesizer.
Another improvement of the system would be to use automatic or semi-automatic alignment points detection algorithms. Tools such Praat or Anvil helped in the process of detecting and annotating these alignment points, however the overall process resulted tedious and involved subjective judgments. Gesture hits may be detected by analyzing sudden changes on velocity, or trajectories of hands, and pitch accents may be detected with a system similar to Tamburini’s [57]. However, due to the complexity of both gesture and prosody channels these systems are difficult to define.

A natural way to extend this work is to add more synchrony rules that improve the quality of the produced animations. However, as more rules are defined it becomes more difficult to find adequate animations in the MOCAP database. In order to find results for the maximum number of speech utterances it will be necessary to increase the size of the database. Another way would be improving the flexibility of the animation synthesizer. The Motion Graph, presented by Kovar et al. [35], is a technique that allows transitioning smoothly between non-consecutive frames of an animation. This allows breaking MOCAP sequences and still obtaining a high quality animation. It will be interesting to create a hybrid approach that ensures that synchronization rules are fulfilled while not necessarily using consecutive MOCAP data.
8. References


[28] M. Karpinski, E. Jarmolowicz-Nowikow, and Z. Malisz, «Aspects of
gestural and prosodic structure of multimodal utterances in Polish task-oriented dialogue», *Speech and Language Technology*, pp. 113-122, 2008.


