TESIS DOCTORAL

Título	Prosodic and Voice Quality Cross-Language Analysis of Storytelling Expressive Categories Oriented to Text-To-Speech Synthesis
Realizada por	Raúl Montaño Aparicio
en el Centro	Escola Tècnica Superior d'Enginyeria Electrònica i Informàtica La Salle
y en el Departamento	Grup de Recerca en Tecnologies Mèdia (GTM)
Dirigida por	Dr. Francesc Alías Pujol

To my parents, Esperanza & Juan.

ABSTRACT

For ages, the oral interpretation of tales and stories has been a worldwide tradition tied to entertainment, education, and perpetuation of culture. During the last decades, some works have focused on the analysis of this particular speaking style rich in subtle expressive nuances represented by specific acoustic cues. In line with this fact, there has also been a growing interest in the development of storytelling applications, such as those related to interactive storytelling. This thesis deals with one of the key aspects of audiovisual storytellers: improving the naturalness of the expressive synthetic speech by analysing the storytelling speech in detail, together with providing better non-verbal language to a speaking avatar by synchronizing that speech with its gestures. To that effect, it is necessary to understand in detail the acoustic characteristics of this particular speaking style and the interaction between speech and gestures.

Regarding the acoustic characteristics of storytelling speech, the related literature has dealt with the acoustic analysis of storytelling speech in terms of prosody, being only suggested that voice quality may play an important role for the modelling of its subtleties. In this thesis, the role of both prosody and voice quality in indirect storytelling speech is analysed across languages to identify the main expressive categories it is composed of together with the acoustic parameters that characterize them. To do so, an analysis methodology is proposed to annotate this particular speaking style at the sentence level based on storytelling discourse modes (narrative, descriptive, and dialogue), besides introducing narrative submodes. Considering this annotation methodology, the indirect speech of a story oriented to a young audience (covering the Spanish, English, French, and German versions) is analysed in terms of prosody and voice quality through statistical and discriminant analyses, after classifying the sentence-level utterances of the story in their corresponding expressive categories. The results confirm the existence of storytelling categories containing subtle expressive nuances across the considered languages beyond narrators' personal styles. In this sense, evidences are presented suggesting that such storytelling expressive categories are conveyed with subtler speech nuances than basic emotions by comparing their acoustic patterns to the ones obtained from emotional speech data. The analyses also show that both prosody and voice quality contribute almost equally to the discrimination among storytelling expressive categories, being conveyed with similar acoustic patterns across languages. It is also worth noting the strong relationship observed in the selection of the expressive category per utterance across the narrators even when, up to our knowledge, no previous indications were given to them.

In order to translate all these expressive categories to a corpus-based Text-To-Speech system, the recording of a speech corpus for each category would be required. However, building ad-hoc speech corpora for each and every specific expressive style becomes a very daunting task. In this work, we introduce an alternative based on an analysis-oriented-to-synthesis methodology designed to derive rule-based models from a small but representative set of utterances, which can be used to generate storytelling speech from neutral speech. The experiments conducted on increasing suspense as a proof of concept show the viability of the proposal in terms of naturalness and storytelling resemblance.

Finally, in what concerns the interaction between speech and gestures, an analysis is performed in terms of time and emphasis oriented to drive a 3D storytelling avatar. To that effect, strength indicators are defined for speech and gestures. After validating them through perceptual tests, an intensity rule is obtained from their correlation. Moreover, a synchrony rule is derived to determine temporal correspondences between speech and gestures. These analyses have been conducted on aggressive and neutral performances to cover a broad range of emphatic levels as a first step to evaluate the integration of a speaking avatar after the expressive Text-To-Speech system.

Ah, they'll never ever reach the moon,
at least not the one that we're after.

— Leonard Cohen

ACKNOWLEDGMENTS

Quisiera empezar agradeciendo a quienes va dedicada esta tesis, mis padres, por todo su apoyo para que hiciera lo que quisiera con mi vida académica y profesional, además de la ayuda diaria para que todo me fuera un poco más fácil. También al resto de mi familia cercana, que siempre me han animado a seguir con mis objetivos.

Seguidamente quiero agradecer a mi director de tesis, Francesc Alías, por ser muy posiblemente el mejor director de la historia. Tenerle al lado en esta aventura ha sido toda una suerte. He aprendido muchísimo con él, y considero que soy mejor profesional y persona en gran parte por sus consejos. Es también seguramente la persona más trabajadora que conozco y, por qué no decirlo, es un tío enrollado.

Muchas otras personas también me han acompañado casi dirariamente en esta aventura. Si se me permite, querría destacar primero que he tenido un compañero de batallas, Marc Freixes, una gran persona con la que he compartido escritorio, trabajos y buenos ratos. Deseo que todo te vaya bien y seas doctor el año que viene. También esta la gente que ha estado desde el principio o desde algún momento. Àngels, Lluís, Ramón, Arnela, Patri, Sevillano, Rosa, Adso, Alan, Marc Antonijoan, Àngel, Ale, Davide, Marcos, Ferran, Barco, Alejandro, Pascal (espero no dejarme a nadie)... Como digo muchas veces, gracias por ser cómo sois, será difícil encontrar a un grupo de gente tan especial como vosotros.

También han habido profesores que me han ayudado y enseñado muchas cosas, como Joan Claudi Socoró, Ignasi Iriondo o David Miralles. A todos ellos mi más sincero agradecimiento por guiarme y valorar positivamente mi trabajo.

Otra gente que me ha animado a su manera es la gente del restaurante de La Salle, como Lídia, Carmen, Lucía, Gustavo y Carlos. Gente sana y divertida, ¡y que hacen muy buenos cafés!

Gracias a mis amigos de toda la vida, que aunque muchas veces me dijeran que no sabían lo que yo hacía ni lo que haría después (bueno, realmente no eran los únicos), son un grupo increíble del que me siento muy afortunado de formar parte. Celebraremos la finalización de esta tesis con unas *birritas* en el KPC seguro.

He de agradecer también a los que me han otorgado la beca con la que he podido llevar a cabo mi doctorado, es decir, a la *Secretaria d'Universitats i Recerca* y al *Departament d'Economia i Coneixement* de la *Generalitat de Catalunya* y al Fondo Social Europeo por la beca FI (No. 2013FI_N 00790, 2014FI_B1 00201, y No. 2015FI_B2 00110), además de la financiación al Grup de recerca en Tecnologies Mèdia de La Salle - Universitat Ramon Llull (refs. 2009-SGR-293 y 2014-SGR-0590), y a la ayuda parcial por el proyecto CEN-20101019, otorgado por el Ministerio de Ciencia e Innovación de España.

Finalmente, quisiera agradecer a Laia por estar a mi lado estos últimos dos años y medio. Estar contigo ha hecho mi vida mucho más feliz y me has hecho mejor persona. Ahora que el doctorado ha acabado, se cierra una etapa de nuestras vidas y nace otra donde viviremos muchas más experiencias juntos.

PUBLICATIONS

This thesis is based on the following publications:

— R. Montaño, F. Alías, and J. Ferrer. Prosodic analysis of storytelling discourse modes and narrative situations oriented to Text-to-Speech synthesis. In *8th ISCA Workshop on Speech Synthesis*, pages 171–176, Barcelona, Spain, 2013

Contributions of the author to the work
- Annotation methodology definition
- Speech corpora annotation and segmentation
- Prosodic and statistical analyses
- Writing of the publication, including the creation of Tables and Figures.

— A. Fernández-Baena, R. Montaño, M. Antonijoan, A. Roversi, D. Miralles, and F. Alías. Gesture synthesis adapted to speech emphasis. *Speech Communication*, 57:331–350, 2014

Contributions of the author to the work
- Gestures and speech annotation and segmentation
- Classification analyses of speech and gestures
- Gestures and speech correlation analysis
- Writing of some parts of the publication, including the creation of some Tables and Figures.

— R. Montaño and F. Alías. The role of prosody and voice quality in text-dependent categories of storytelling across languages. In *Proc. Interspeech*, pages 1186–1190, Dresden, Germany, 2015

This thesis is also based on the following publications currently under review:

— R. Montaño, M. Freixes, F. Alías, and J. C. Socoró. Generating Storytelling Speech from a hybrid US-aHM Neutral TTS synthesis framework using a rule-based prosodic model. In *Proc. Interspeech*, San Francisco, USA, 2016. Submitted

Contributions of the author to the work
- Annotation, segmentation, and analysis of the set of utterances
- Statistical analysis of the perceptual test results
- Writing of the publication, including the creation of Figures.

— R. Montaño and F. Alías. The Role of Prosody and Voice Quality in Indirect Storytelling Speech: Analysis Methodology and Expressive Categories. *Speech Commun.*, 2016a. Second Round of Revisions

— R. Montaño and F. Alías. The Role of Prosody and Voice Quality in Indirect Storytelling Speech: A Cross-language Perspective. *Speech Commun.*, 2016b. First Round of Revisions

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AHM	adaptive Harmonic Model
AR	Articulation Rate
CB	Codebook
CMOS	Comparative Mean Opinion Score
DB	decibels
EV	Evidence Variable
F0	Fundamental Frequency
GCI	Glottal Closure Instant
GNE	Glottal-to-Noise Excitation Ratio
GP	Gesture Phrase
GSI	Gesture Strength Indicator
GTM	Grup de Recerca en Tecnologies Mèdia
HAMMI	Hammarberg Index
HCI	Human-Computer Interaction
HMM	Hidden Markov Models
HNM	Harmonic plus Noise Model
HNR	Harmonics-to-Noise Ratio
HSD	Honestly Significant Difference
HTK	Hidden Markov model ToolKit
INTSINT	INternational Transcription System for INTonation
IP	Intonational Phrase
IQR	Inter-Quartile Range
LAICOM-UAB	Laboratory of Instrumental Analysis-Universitat Autònoma de Barcelona
LDA	Linear Discriminant Analysis
LPC	Linear Predictive Coding
LS-URL	La Salle-Universitat Ramon Llull
MANOVA	Multivariate ANalysis of VAriance
MAUS	Munich AUtomatic Segmentation

MCC Matthews Correlation Coefficient

MDN Median

MDQ Maxima Dispersion Quotient

MDS Multi-Dimensional Scaling

MOCAP MOtion CAPture

MOMEL MOdélisation de MELodie

MS milliseconds

NAQ Normalized Amplitude Quotient

NLP Natural Language Processing

NNE Normalized Noise Energy

NSP Number of silent pauses

PAPT Pitch Accent Peak Time

PE1000 Relative Amount of Energy above 1000 Hz

POS Part Of Speech

PSI Prosody Strength Indicator

PSP Parabolic Spectral Parameter

SAMPA Speech Assessment Methods Phonetic Alphabet

SG Stress Group

SLAM Stylization and LAbelling of speech Melody

SMO Sequential Minimal Optimization

SPPAS SPeech Phonetization Alignment and Syllabification

SS Spectral Slope

ST Semitone

STSA Speech-To-Speaking Avatar

STD Standard deviation

SVM Support Vector Machine

SYLL/SEC Syllables per second

T0 Fundamental Period

TOBI Tones and Break Indexes

TRUE Testing platfoRm for mUltimedia Evaluation

TTS Text-To-Speech

US Unit Selection

VOQ Voice Quality

1 INTRODUCTION

This thesis has been developed under the doctoral program "Information and communications technologies and their management" of La Salle-Universitat Ramon Llull (LS-URL). It has been carried out within the Grup de Recerca en Tecnologies Mèdia (GTM) of LS-URL under the supervision of Dr. Francesc Alías. In this chapter, the thesis is introduced by explaining the framework (Section 1.1), the objectives (Section 1.2), and the structure of the document (Section 1.3).

1.1 FRAMEWORK OF THE THESIS

The affective communication channel in Human-Computer Interaction (HCI) has been a topic of interest for both the speech recognition and speech synthesis research communities. Analysing the expressiveness present in affective speech has been of great interest for both lines of work and it is still a field under investigation. During the last decades, expressive speech has been explored in myriad studies devoted to speech analysis. Although the main focus was placed on prototypic expressions of basic emotions in initial works (see, Schröder, 2004, and references therein), the interest of the investigations has moved to the analysis of subtler and context-specific speaking styles more recently (cf., Nicolaou et al., 2011; Grawunder and Winter, 2010; Cheang and Pell, 2008; Govind and Prasanna, 2013). However, the storytelling speaking style oriented to an audience has not obtained so much attention, even though it is a good example of context-specific speaking style rich in expressiveness (specially, when aimed at children) with a great variety of applications. For instance, storytelling speech can be applied to entertainment (e.g., interactive storytelling, Silva et al., 2004; Merabti et al., 2008; Alofs et al., 2015, or digital talking books, Suchato et al., 2010), human-robot interaction (Mutlu et al., 2006; Gelin et al., 2010; Min et al., 2013; Leite et al., 2015), or education (Weng et al., 2011; Marchetti, 2012; Leite et al., 2015).

In general, during the last decades there has been a growing interest in the development of storytelling applications. One of the most representative applications are those related to interactive storytelling (Silva et al., 2004; Cao et al., 2010; Alofs et al., 2015), which beyond the obvious entertainment purpose can be used to develop the cognitive abilities of children or people with special needs, such as autistic children (van Santen et al., 2003; Grynszpan et al., 2005).

1.1.1 *Why storytelling?*

The GTM of LS-URL has a long track record in the development of HCI applications developed in several R&D projects, such as IntegraTV-4all (FIT-350301-2004-2), Sam - The virtual weatherman (RDITSCON04-0005), SALERO (IST-FP6-027122), INREDIS (CEN-2007-2011), evMIC (TSI-020301-2009-25), THOFU (CEN-2010-1019), etc. Among them, the SAVE project (TEC2006-08043 /TCM) was focused on the development of an expressive audiovisual synthesis system based on a photo-realistic talking head to provide a natural interaction. That project built on a similar project called *Locutor Virtual* (MCyT PROFIT FIT-150500-2002-410), which developed a system with a customizable virtual speaker from 2D-images and also included speech synthesis based on diphonemes. Later, a similar philosophy was translated into the storytelling world in the *CuentaCuentos* 2.0 project (TSI-070100-2008-19). The general characteristics of this system are depicted in Figure 1. Firstly, the user can personalize a 3D avatar with his/her face through the extraction and post-processing of facial key-points. Next, a tale/story can be selected and the text can be annotated with different emotional tags indicated

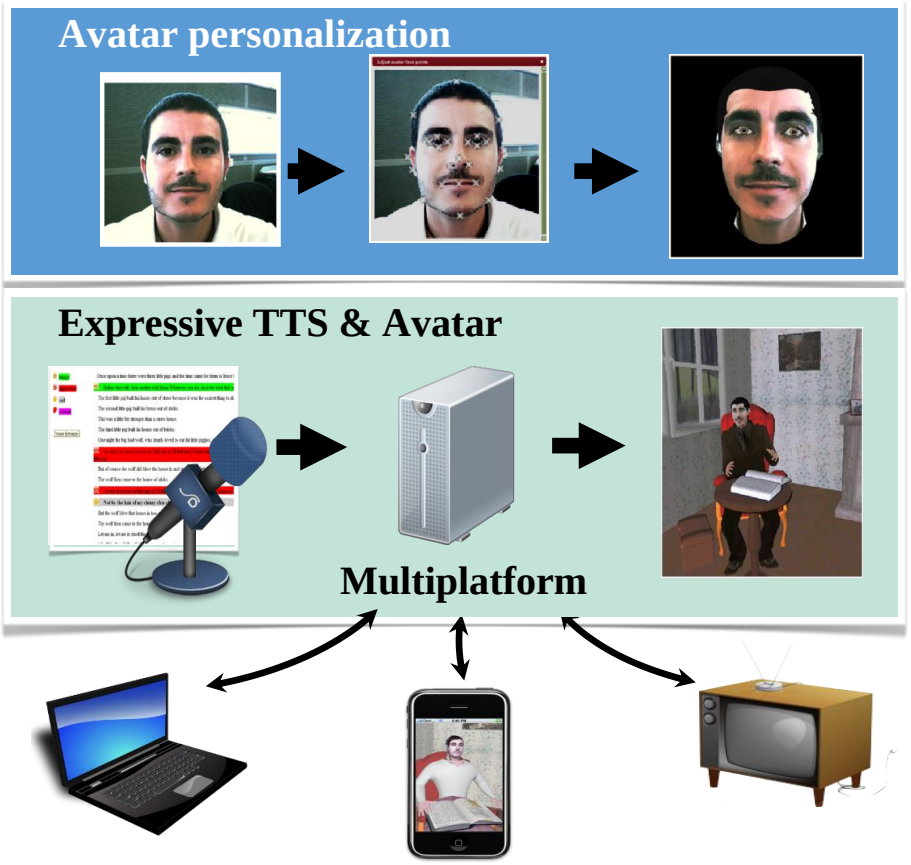

FIGURE 1: OVERVIEW OF CUENTACUENTOS 2.0 – LA SALLE-UNIVERSITAT RAMON LLULL GRUP DE RECERCA EN TECNOLOGIES MÈDIA.

by the user. Finally, the expressive Text-To-Speech (TTS) synthesis (taking into account the emotional tags) together with the animation of the personalized avatar are executed providing an audiovisual storytelling output. As it can be observed in Figure 1, it is a multi-platform system, capable of providing the output for personal computers, mobile phones, and television.

Although *CuentaCuentos* 2.0 was a successful first attempt to generate audiovisual storytelling, we could observe that there were key points that should be readdressed to achieve better results. Among them, this thesis deals with one of the key aspects of audiovisual storytellers: improving the naturalness of the expressive speech by analysing the storytelling speech in detail, together with providing better non-verbal language to the avatar by analysing the interaction between the speech with its gestures.

1.1.2 *The analysis of storytelling speech: Main research gaps*

Tales and stories, in general, should awake noble and/or repulsive emotions and empathy for the characters of the story together with entertainment (Kready, 1916; Brewer and Lichtenstein, 1982). Thus, in the oral interpretation of stories and tales, the emotional response is present in the target audience. For instance, if the main characters face problematic situations, narrators may use a negative or a suspenseful tone of voice so that the audience empathizes with such characters. In that case, the audience might feel afraid or sad for the situations presented by the narrator. In this sense, a good storyteller must convey the story with a wide range of expressive variability to engage and entertain the audience, from subtler speech nuances in the indirect discourse (Theune et al., 2006), to more exaggerated affective states when interpreting the characters' interventions (Greene et al., 2012), i.e., using a direct discourse containing acted emotional speech.

As indicated in Section 1.1, there is still room for further investigating storytelling speech. One of the main limitations of this line of research is the lack of standardized analysis guidance, which has derived in a considerable variety of analysis approaches. Such approaches range from the study of very specific expressive aspects of storytelling speech like suspense situations (Theune et al., 2006) or emotional categories (Alm and Sproat, 2005a; Buurman, 2007; Sarkar et al., 2014) to more general analyses using an annotation based on the structure of tales (Doukhan et al., 2011; Adell et al., 2005), or even a mixture of both approaches (Eyben et al., 2012). In fact, the complexity of the affective information involved in narrative texts and the consequent expressive variability of their oral interpretation has been already highlighted by Francisco et al. (2011) and Greene et al. (2012).

Some studies have opted for using basic emotions (or "story-specific" emotions, Sarkar et al., 2014) as a means to analyse storytelling speech (Alm and Sproat, 2005a; Sarkar et al., 2014). In contrast, other works have avoided the use of emotional tags as the authors consider that they do not match the expressive characteristics of (at least) the indirect discourse components (Eyben et al., 2012; Theune et al., 2006; Doukhan et al., 2011; Adell et al., 2005). Nonetheless, some works have considered basic emotions for characters' interventions (Buurman, 2007; Burkhardt, 2011). Definitely, in the literature there are many different approaches to be able to draw clear conclusions about the definition of a general annotation methodology of this particular expressive speaking style.

Another research gap regarding the analysis of storytelling speech is the lack of studies considering both prosody and Voice Quality (VoQ)[1] and evaluating the statistical significance of the findings. Certainly, prosody has proven crucial (Theune et al., 2006; Doukhan et al., 2011; Adell et al., 2005). However, up to our knowledge, the principal contribution to this aim is just the suggestion that VoQ may also play an important role in this speaking style (Theune et al., 2006; Doukhan et al., 2011), although perturbation measures have already proved useful for unsupervised clustering of some parts of a story (Eyben et al., 2012). It seems plausible that, in the same way VoQ is relevant for emotions (Patel et al., 2010), VoQ may be of great interest for the characterization of subtle affective variations like those present in storytelling speech.

1.1.3 *Synthesizing expressive speech: What about storytelling speech?*

During the last decades, expressive speech synthesis has mainly been addressed by corpus-based approaches, following Unit Selection (US) (Black, 2003; Iriondo et al., 2007; Alías et al., 2008; Tsiakoulis et al., 2014), or Hidden Markov Models (HMM) techniques (Yamagishi et al., 2005; Zen et al., 2009; Latorre et al., 2012). The former generally yields good naturalness if sufficiently expressive speech data are available. The latter typically allows smaller corpora (hence, entailing a less laborious and costly building process), but suffers from a decrease in naturalness due to the inherent over-smoothing of this statistical approach (Barra-Chicote et al., 2010). As described in Section 1.1.2, storytelling speech entails many expressive nuances that would require a proper acquisition of the data from the expressive corpora.

In what concerns synthesis of storytelling speech, some studies have directly used audiobooks containing stories to generate expressive speech (Jauk et al., 2015; Charfuelan and Steiner, 2013; Prahallad and Black, 2011). However, even though audiobooks can be used to generate expressive speech with good quality in average, there are several subtle expressive nuances within the storytelling speaking style that need further analysis and model to fully accomplish the requirements of storytelling applications. For instance, descriptive utterances of storytelling speech contain a specific prosodic emphasis on stressed vowels of certain adjectives and adverbs, and along the story there may be several types of suspenseful speech (Theune et al., 2006). Ideally, this kind of subtle expressiveness should be considered in certain parts of the story to maximize the user experience.

[1] More information regarding prosody and VoQ in Chapter 2

In order to bridge the daunting task of building ad-hoc corpus for each and every expressive speaking style, some works have tackled the generation of synthetic expressive speech following quite diverse approaches. Some authors have used basic fixed acoustic rules to transform neutral to expressive synthetic speech (Theune et al., 2006; Zovato et al., 2004). Differently, adaptation techniques have been considered in HMM-based synthesizers to interpolate between statistical models trained on different expressive databases (Yamagishi et al., 2007). Hybrid approaches have also been introduced with the same aim. An US-based conversion system using Harmonic plus Noise Model (HNM) was introduced to generate emotions from neutral speech by Erro et al. (2010). Later, an emotion transplantation approach that used the adaptation functions as pseudo-rules for modifying the HMM-based models was presented by Lorenzo-Trueba et al. (2015). Although both approaches are based on rather small corpora, they still need non-negligible speech data for each expressive style (e.g., 6–30 min. Lorenzo-Trueba et al., 2015 and around 10 min. in, Erro et al., 2010 per style), besides presenting other limitations such as the need of parallel corpora (Erro et al., 2010).

Finally, it is worth noting that the original TTS system of LS-URL is based on the US approach. Previous works of GTM researchers in the US-TTS framework have been focused in optimizing Natural Language Processing (NLP) (Trilla et al., 2010; Trilla and Alías, 2013), and US modules (Formiga et al., 2010; Alías et al., 2011), incorporating new expressive speech styles (Iriondo et al., 2007), while defining a multi-domain approach that enables the managing of many different speech domains within the same synthesis framework (Alías et al., 2008). Furthermore, some versions of the system have included HMM (Gonzalvo et al., 2007; 2010), and HNM techniques (Calzada and Socoró, 2011; Calzada and Socoró, 2012).

This thesis aims at incorporating new expressive speech styles (in this case, storytelling expressive categories) within the TTS system as rule-based prosodic models to be applied to the neutral database.

1.1.4 *Analysis of the gestures-speech relationship oriented to synthesizing animations*

As aforementioned, one of the objectives is to provide better non-verbal language to a storyteller avatar by synchronizing its speech with its gestures. Thus, a key aspect is the synchronization of both phenomena in order to maintain the naturalness, which makes necessary to study their interaction.

Little attention was given to study the insights of gesture and speech synchronization until the end of the twentieth century. Kendon (1980) defined the following synchronization rule between gesture and speech: The extension phase of a beat gesture would coincide with, or slightly precede, the onset of a stressed syllable. McNeill (1992) suggested a similar rule: The stroke of a gesture phrase is always completed either before or at the same time as the tonic syllable of the concurrent tone unit. Nobe (1996) observed that this statement not only holds for tonic syllables, but also for peaks of intensity. Subsequent studies proposed other anchor points of synchrony between gestures and speech. For instance, Valbonesi (2002) defined the speech focal points, which are computed using five prosodic features of speech signal; and gesture focal points, which are computed as local maxima or minima of hand traces. They observed that speech focal points occurred either near a gesture focal point or within the duration of a stroke.

Bolinger (1986) observed a synchrony rule between speech and gesture: pitch and body parts rise and fall together, to reflect increased or decreased tension. However, in a latter work, Loehr (2004) did not find evidence of this. However, he observed that apexes of gestures (points of maximum extension) tend to co-occur with pitch accents. Renwick et al. (2004) also used pitch accents as speech anchor points, but their corresponding gesture anchors were hits. Hits are abrupt stops or pauses in movement, which break the flow of gestures. These are only present in discrete gestures, and correspond to apexes of these type of gestures. Leonard and Cummins (2011) studied the time alignments between several anchor points, and concluded that the apex of the beat gesture shows less temporal variability with respect to speech than with any other point within the gesture.

Antonijoan (2012) performed a study of the impact on the perceived quality of animations when using a temporal synchrony rule that aligns pitch accents and apexes of beat gestures. Although an increase of perceived quality in synchronized animations versus not synchronized ones was observed, it was evident that the temporal synchrony rule alone was not sufficient to ensure a proper synchrony for all cases. The perception of animation quality dropped when gesture and speech emphasis levels did not match each other. Levine et al. (2009; 2010) proposed a gesture animation system that, besides aligning syllables and gestures temporally, employed the correlation between prosody and kinematics of motion to select appropriate gestures. Levine system uses a probabilistic model from which it is not possible to infer general synchrony rules directly.

This thesis proposes a new synchrony rule that correlates strength levels of speech signal with strength levels of gestures, used to match emphasis between the two modalities.

1.2 OBJECTIVES OF THE THESIS

In this thesis, after considering the research framework and state of the art explained in the previous Sections, indirect storytelling speech corpora is analysed in order to address the following three main objectives:

1. Define and validate a methodology suitable for annotating indirect storytelling speech in terms of expressive categories.
2. Evaluate the relevance of both prosody and VoQ to characterize the expressive categories of indirect storytelling speech across languages.
3. Define a TTS synthesis framework capable of generating storytelling speech from rule-based acoustic models, together with the first steps towards developing a storytelling speaking avatar using speech-gestures synchrony rules to drive its animation.

1.3 STRUCTURE OF THE THESIS

This thesis is structured as follows. Chapter 2 starts by introducing a general description of the voice production model and continues with an overview of the analysis of both prosody and VoQ features from speech. That Chapter includes a definition of some of the most common prosodic and VoQ parameters, and a description of the considered analysis tools. Next, Chapter 3 contains a detailed review of studies that have tackled the linguistic-structural analysis of narratives, together with other works that have analysed and/or synthesized storytelling speech.

Then, the main contributions of this thesis are presented. They have been organized in three parts, which are the following:

Part I: A sentence-level annotation methodology based on storytelling discourse modes (narrative, descriptive, and dialogue) is introduced to analyse indirect storytelling speech. In a first step, such methodology is applied to a Spanish storytelling speech corpus as a proof of concept to evaluate its viability.

Part II: A cross-language study of storytelling speech is conducted in terms of prosody and VoQ, following the annotation methodology validated for Spanish in Part I. Concretely, the same story is studied considering three more languages: British English, German, and French. This analysis is performed with the objective of evaluating to what extent the evidences obtained in the proof of concept for the Spanish version can be generalized to other languages by considering parallel corpora (i.e., the same story).

Part III: In this part, the analyses oriented to storytelling synthesis are performed. The main focus is placed on expressive TTS synthesis, which then includes a Speech-To-Speaking Avatar (STSA) module. In the first place, averaged rules derived from the previous analyses are applied as a post-processing stage of US-based neutral speech synthesis. Next, taking the most of the conclusions obtained in that preliminary experiment, an analysis-oriented-to-synthesis methodology together with an improved synthesis framework is presented with the aim of generating synthetic storytelling speech from small corpora. Finally, a study focused on the relationship between speech and gestures that derives in synchrony and emphasis rules oriented to speaking 3D avatar synthesis is presented.

Each part ends with a discussion of the obtained results together with the corresponding conclusions. The final conclusions of the thesis and the future work are collected in Part **IV**.

Finally, two Appendixes are included as last part of the thesis. Appendix A consists of a preliminary analysis of the emotional content of the direct speech of the narrator interpreting a character and the relationship of this emotions present in storytelling with respect to previous analyses conducted on basic emotions. The thesis ends with Appendix B, which shows a more detailed study of the emotional corpora analysed in Part I.

2

ANALYSIS OF PROSODY AND VOICE QUALITY

2.1 SPEECH PRODUCTION

The speech production from a physiological perspective is typically described in terms of three components (see Fig. 2): the respiratory system, the larynx, and the supra-laryngeal vocal tract (Lieberman and Blumstein, 1988). The respiratory system (also often referred as sub-glottal system) produces the air flow that is filtered in the upper stages. This flow enters in the larynx, which is responsible for converting the air flow into a series of quasi-periodic pulses: the glottal flow. Specifically, the vocal cords rapidly vibrate (opening and closing) under the control of the laryngeal muscles to accomplish this goal. Note that the space between the vocal cords is the glottis. Finally, the cavities and parts of the supra-laryngeal vocal tract (the pharynx and the oral and nasal cavities) filter the glottal source emphasizing the energy of some specific frequencies, and producing the speech signal. The sounds produced when the vocal cords vibrate are called voiced sounds, while when there is no vibration (the vocal cords remain open) an unvoiced sound is produced. Thus, unvoiced sounds (e.g., the phoneme [s]) are aperiodic and voiced sounds are quasi-periodic (e.g., the phoneme [e][2]), because the glottal source of the unvoiced sounds is noisy (see Fig. 3 for an example of voiced/unvoiced speech signals).

Speech can be regarded as a medium which allows humans to communicate with each other through spoken words, but the speech signal consists of several layers (Laver, 1991): linguistic, paralinguistic, and extralinguistic. There is a strong agreement in the related literature concerning this division, although the terms and definitions sometimes are slightly different. The linguistic layer consists of the semantic information and its phonetic representation. However, quoting Laver (2003), *"we speak not just to be understood, but to communicate"*. In this sense, the paralinguistic layer contains several information, such as speaker's current affective state (e.g., a sad state tends to be related to a lowering in speech intensity), emphasizing certain words, speaker's pragmatic intent, among others. One very representative example that differentiates this layer with respect to the linguistic information is the use of sarcasm, as the actual message is intended to be interpreted as the opposite of the literal meaning of the words. The third layer (extralinguistic aspects), conveys quasi-permanent information about physical, social, and psychological details of the speaker. For instance, the Fundamental Frequency (F0) slightly vary through age (Pegoraro Krook, 1988). Roach (1998) considered paralinguistic features as those used intentionally, while they considered non-linguistic features as those that cannot be used intentionally (e.g., age, sex, etc.). Within non-linguistic features, a distinction was made between individual variation (physiology and histology traits) and reflexes (involuntary reactions to an emotional state). Thus, in this case, the term non-linguistic features can also be used to refer to extralinguistic information. Mixdorff (2002) also used the term "non-linguistic" for health condition, emotional state, etc. Nonetheless, other authors considered different terms and definitions. For example, a different typology was suggested by Traunmüller (2000), where the information within speech is categorized into linguistic (message, dialect, etc.), expressive (emotion, attitude, etc.), organic (age, sex, pathology, etc.), and perspectival (distance, channel, etc.).

[2]The phonetic alphabet used throughout the thesis to represent phonemes is the Speech Assessment Methods Phonetic Alphabet (SAMPA) introduced by Wells (1997)

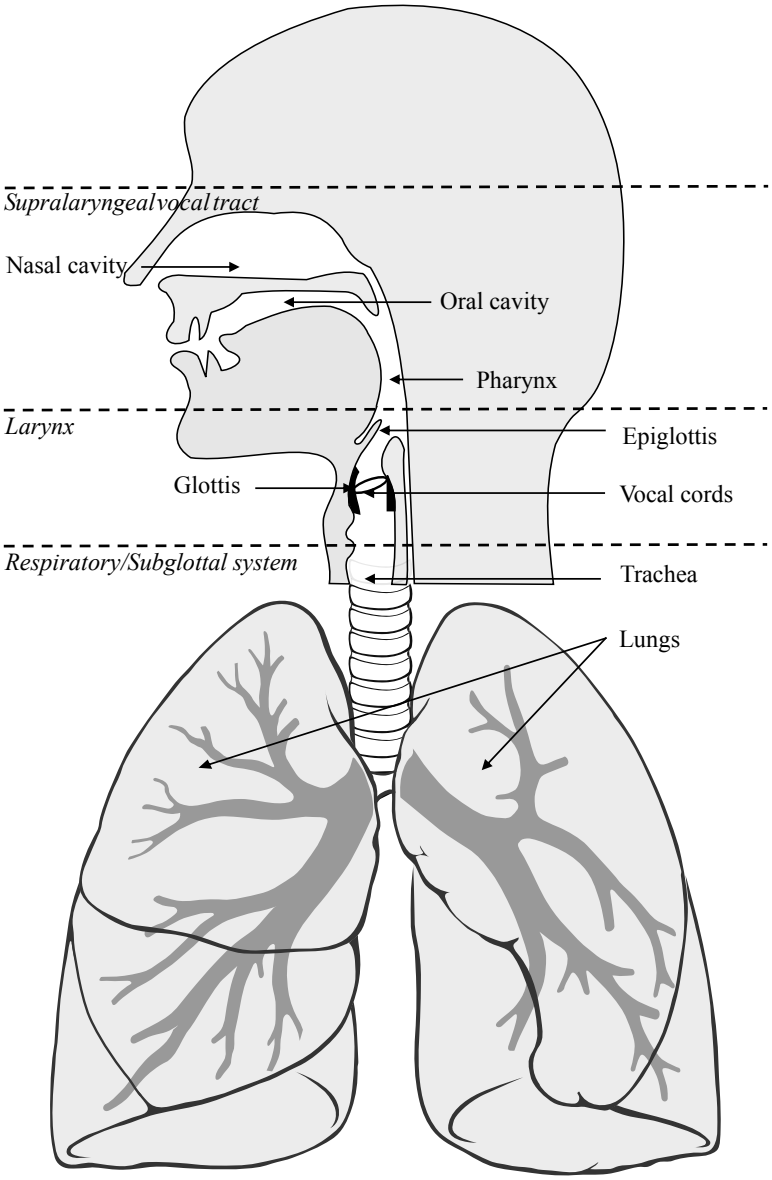

FIGURE 2: PHYSIOLOGICAL COMPONENTS OF SPEECH PRODUCTION[1].

2.2 PROSODY

Although speech can be characterized by means of many prosodic parameters, there are three main features that are consistently used for linguistic purposes: pitch, duration, and loudness (Cruttenden, 1986). These measures are defined from a perceptual point of view and they have a physical correspondence in the speech signal. Pitch, loudness, and duration are related to the F0, the amplitude, and the time, respectively. Prosody, however, may contain non-linguistic information in addition to linguistic cues. At the linguistic level, prosody is used to mark accentuation in lexically stressed syllables or differentiate among declarative, interrogative, and exclamatory utterances. Nonetheless, it can be also modified to convey emotional states such as sadness or happiness by lowering or increasing pitch, intensity, and speaking rate, respectively.

[3]Fig. 2 was made from the following original illustrations:
© Tavin / Wikimedia Commons / CC-BY-3.0
© "Lungs-simple diagram of lungs and trachea" by Patrick J. Lynch / CC-BY-2.5

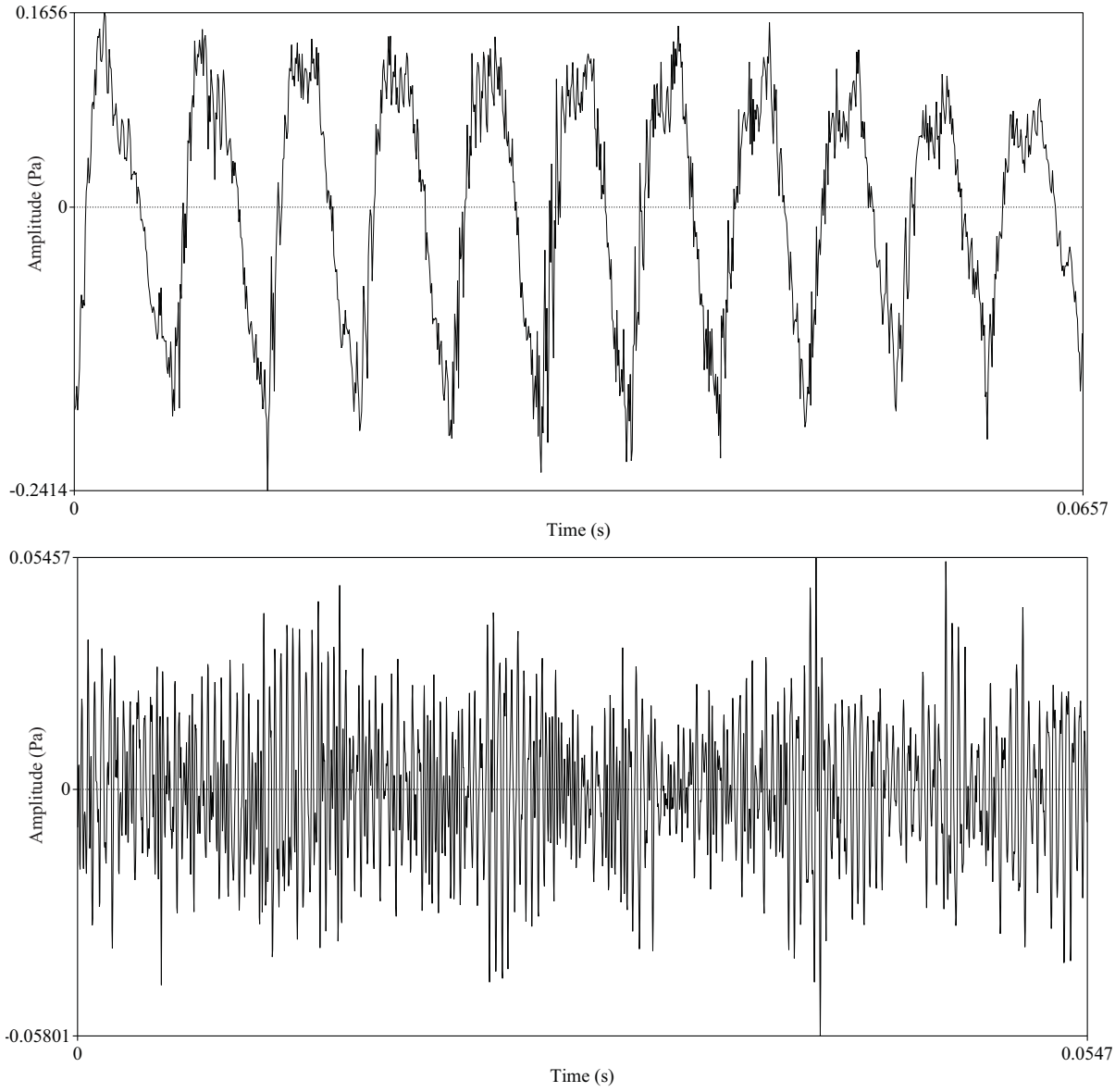

FIGURE 3: SPEECH WAVEFORMS OF THE PHONEMES [*e*] (TOP) AND [*s*] (BOTTOM).

2.2.1 *Prosodic parameters*

In this section, a description of typical prosodic and VoQ parameters is given. In addition, some intonation systems are also introduced.

2.2.1.1 *Fundamental Frequency*

F0 is the frequency at which the vocal folds vibrate. However, the final speech signal is not a perfect sine curve but a complex sinusoid, i.e., a quasi-periodic signal (e.g., see the speech signal of the vowel [*e*] depicted in Fig. 3), as a consequence of the vocal tract, which reinforces some frequencies while it dampens others. The F0 is the one that defines the fundamental frequency while the rest of emphasized frequencies by the vocal tract are denoted as formants. An important F0-related measure used to characterize the F0 register of a speaker is the F0 range. Although it can be regarded as the difference between some top and bottom values from the F0 curve used by a speaker, the definition of these top and bottom values is rather controversial (Patterson, 2000).

2.2.1.2 *Intensity*

Loudness is related to the volume of speech, and it is often referred as intensity when measuring the speech signal amplitude. Speech intensity is usually measured in decibels (DB), a logarithmic unit that gives the ratio between two values.

2.2.1.3 *Duration & Rythm*

Finally, time-related prosodic parameters are typically used to measure the tempo of speech. On the one hand, the duration of speech segments can be used (e.g., the duration of the [*e*] vowel depicted in Fig. 3 is 67.5 milliseconds (MS)) to, e.g., assess if a speaker stretches certain phonemes. Moreover, the duration of pauses can also be evaluated to obtain a representation of speech tempo. On the other hand, speaking and/or articulation rates are frequently used to measure how fast or slow a person is speaking. The former includes pauses in the measure and the latter does not. The syllable is usually considered the minimal unit to describe speech tempo, at least in syllable-based languages (Ladd and Campbell, 1991). Therefore, one of the most common measures of speaking and Articulation Rate (AR) is Syllables per second (SYLL/SEC) (Trouvain, 2004).

2.2.2 *Intonation systems*

F0 has been the most widely studied acoustic feature in studies related to speech prosody. In general, intonation is considered as the evolution of F0 within a certain speech unit (Botinis et al., 2001), although slightly different definitions have also been used (see, Garrido, 1996, and references therein). Sometimes, the terms prosody and intonation have been used interchangeably, but most usually intonation is linked to F0 alone.

Intonation may contribute to linguistic, paralinguistic, and extralinguistic functions. For example, different language communities may be classified according to their lexical function of intonation. Concretely, there are stress languages (e.g., Spanish, English, French, German, Italian, etc.), tone languages (e.g., Chinese, Vietnamese, Thai, etc.), and pitch accent languages (e.g., Swedish, Japanese, Korean, etc.).

The analysis of intonation can be conducted at a microprosodic level (e.g., segments, syllables, or stress groups) to capture local phenomena, or at a macroprosodic level (e.g., intonation groups, clauses, sentences, or paragraphs) to capture global phenomena. Thus, before the analysis, the intonational unit has to be decided. The syllable is considered in some works as a basic intonation unit (Pierrehumbert, 1980), although many others have chosen the stress group to obtain an intonation model (Thorsen, 1978; Escudero and Cardenoso, 2002; Iriondo, 2008; Lopez-Gonzalo et al., 1997).

Several systems have been proposed for the modelling (i.e., labelling and transcription) of intonation, which have been applied in many languages. These methods aim to offer a limited set of elementary F0 contours explaining the majority of situations that may occur within the F0 curve of any utterance. In the following Sections, some of the employed systems for the stylization and transcription of intonation are described.

2.2.2.1 *The Tones and Break Indexes standard*

The Tones and Break Indexes (TOBI) system is the most widely used method for intonation modelling and speech synthesis. The origin of the TOBI system is the seminal thesis of Pierrehumbert (1980), where she developed an underlying representation for English intonation. Later on, Silverman et al. (1992) developed and defined the TOBI standard for American English together with detailed guidelines (Beckman and Ayers, 1997). During the last decades, the TOBI system has been adapted to several

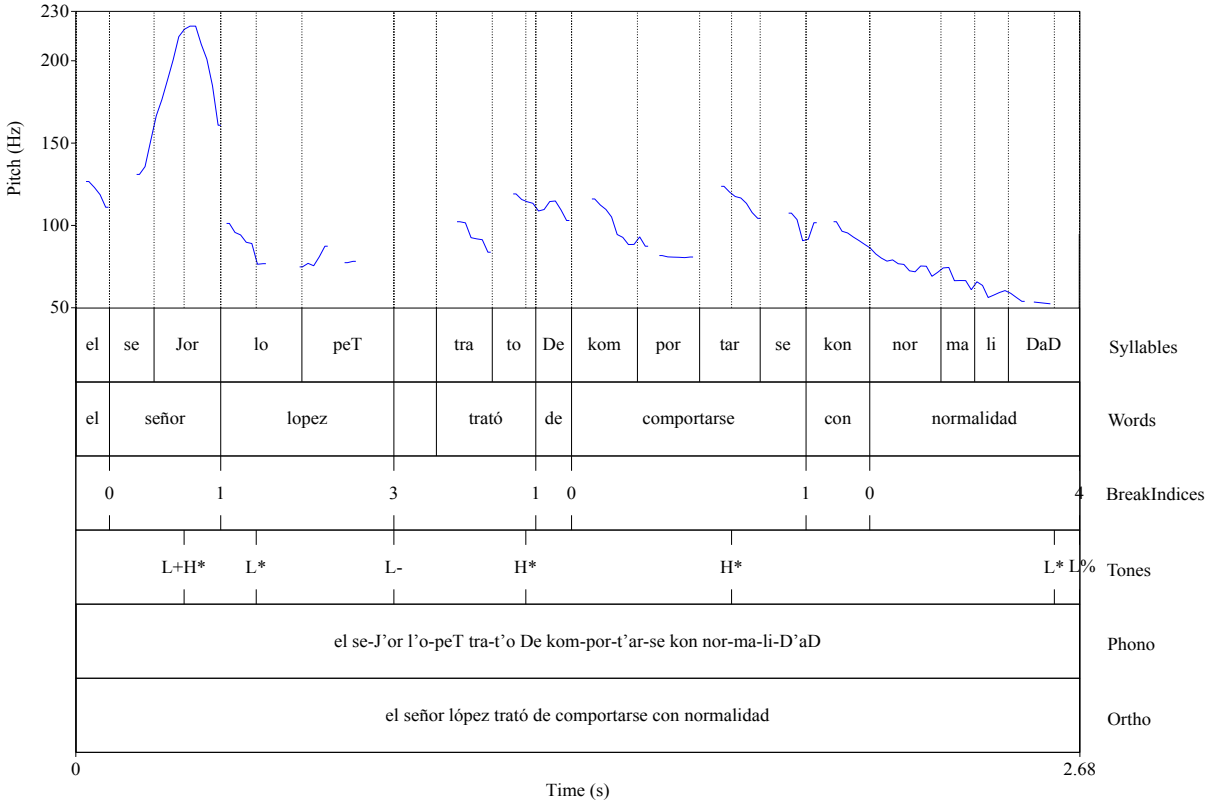

FIGURE 4: EXAMPLE OF A TONES AND BREAK INDEXES ANNOTATION.

languages, including Spanish (Beckman et al., 2002), French (Delais-Roussarie et al., 2015), German (Baumann et al., 2000), and Catalan (Prieto, 2014), among others.

In general, the ToBI labelling consists of creating four tiers of labels (Beckman and Ayers, 1997):

- **Orthographic tier**: Contains the orthographic transcription of all of the words in the utterance.
- **Break indices tier**: A value between 0 to 4 is used to mark the prosodic grouping of the words according to the subjective strength of its association with the next word. 0 represents the strongest perceived conjoining, and 4 the most disjoint grouping.
- **Tones tier**: Together with the break indices tiers, this tier represents the core of the prosodic analysis. It contains the pitch accents and boundary tones. Both pitch accents and boundary tones are transcribed as a high (H) and low (L) tones marked with an asterisk (*) in the case of pitch accents, a dash (-) in intermediate phrase boundaries, and a percent sign (%) to mark the boundary of intonational phrases. The plus sign (+) just indicates a combination of two tones. The inventory of labels and their definitions depend on the language.
- **Miscellaneous tier**: This is an optional tier and it has not been standardized. It is basically a tier for commenting non-speech events, such us laughters, coughs, or possible difficulties in the annotation.

In Fig. 4, an example of ToBI annotation for Spanish can be observed.

In spite of its popularity, the ToBI system suffers from some drawbacks that have been already pointed out by Wightman (2002). As implicitly stated above, the ToBI system is language-dependent, i.e., an ad hoc system is needed for each language. Moreover, not all researchers use the same ToBI pitch

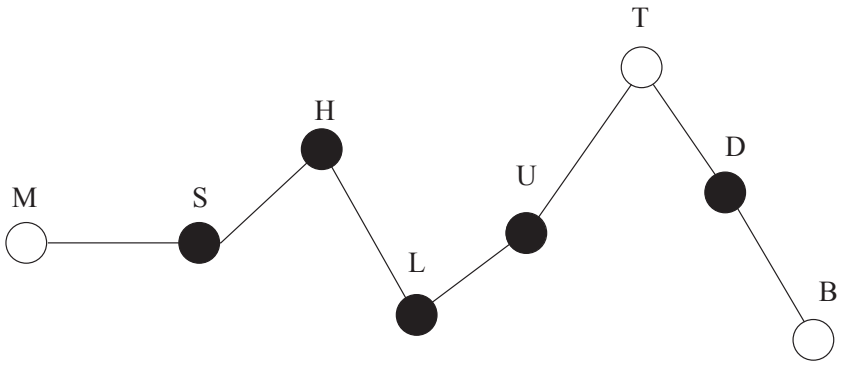

FIGURE 5: EXAMPLE OF THE INTERNATIONAL TRANSCRIPTION SYSTEM FOR INTONATION LABELING SYSTEM (EXTRACTED FROM LOUW AND BARNARD, 2004).

accents, as different opinions exist in their use and definition (Wightman, 2002). In this sense, the inter-annotator agreement can be quite low (Wightman, 2002), and the transcription requires individuals previously trained in the system. Even with trained labellers, a complete ToBI transcription may take up to 100–200 times real time (Syrdal et al., 2001). Another limitation is that ToBI is solely based on the syllabic unit, and it may be interesting to be able to explore the intonation contours of other units.

2.2.2.2 *INternational Transcription System for INTonation*

The INternational Transcription System for INTonation (INTSINT) developed at the Institut Phonétique d'Aix-en-Provence describes the intonation using a set of abstract tonal symbols and it is applicable to any language (Hirst and Di Cristo, 1998). This system can be applied after a MOdélisation de MELodie (MOMEL) stylization (Hirst and Espesser, 1993), and the process can be reverted because from the INTSINT labels it is possible to recover the intonation contour.

Within the INTSINT system, several abstract tonal symbols are considered. In Fig. 5, an example of the symbols is depicted, which are defined as follows:

- **Absolute tones**: These symbols refer to the overall pitch range of the speaker, which are denoted as T (Top), M (Mid), and B (Bottom).
- **Relative tones**: These symbols refer to the relation with respect to the previous point, which are denoted as S (Same), H (Higher), L (Lower), U (Up-stepped), and D (Down-stepped).

The MOMEL and INTSINT algorithms have been implemented as a plug-in within the Praat tool (Hirst, 2007).

2.2.2.3 *Stylization and LAbelling of speech Melody*

Stylization and LAbelling of speech Melody (SLAM) was recently created by Obin et al. (2014) with the objective of offering a transcription system generalizable to any linguistic or syntactic unit. The labelling of the melodic curve is performed automatically by means of a freely available (and open-source) python program[4]. SLAM has some advantages with respect to other existing methods:

- The alphabet of melodic contours is fully **data-driven**.
- A **time-frequency representation** is used to capture with more detail complex melodic contours.
- It can be applied to a large **variety of prosodic or syntactic units**.
- The method can be used to analyse **other stressed languages**.

[4] https://github.com/jbeliao/SLAM/

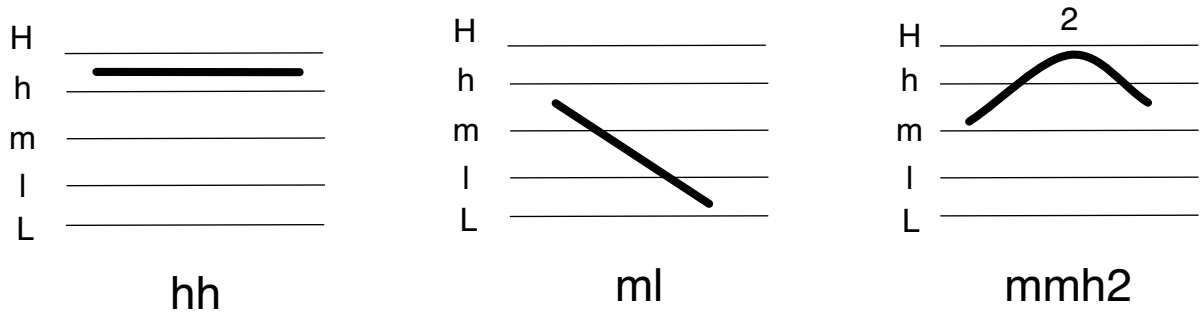

FIGURE 6: EXAMPLE OF THE STYLIZATION AND LABELLING OF SPEECH MELODY LABELLING SYSTEM (EXTRACTED FROM OBIN ET AL., 2014).

- The representation is normalized with respect to the **average range of the speaker** under analysis.

Each F0 contour is represented with a symbolic label with the following format:

"Initial frequency value" "Final frequency value" ["Main saliency" "Main saliency position"]

Frequency values are represented with respect to 5 pitch levels covering the whole F0 range of the speaker. The characters used for the symbolic representation of frequency values and the definition of levels can be observed in Table 1.

TABLE 1: PITCH LEVELS USED FOR THE SYMBOLIC REPRESENTATION (OBIN ET AL., 2014).

Pitch levels	Description	Range (Semitones (STs))
H	extreme-high	>+6
h	high	+2/+6
m	medium	−2/+2
l	low	−2/−6
L	extreme-low	<-6

The optional symbols "Main saliency" and "Main saliency position" (represented above between square brackets) correspond to the most salient F0 peak (if one exists), and the position of the saliency within the unit, respectively. For the saliency position, the unit is divided into 3 equal parts and a value of 1 is given is the salience is in the first part of the unit, 2 if it is in the middle, and 3 if it is located in the last part. An example of SLAM labelling is depicted in Fig. 6.

2.3 VOICE QUALITY

Laver (1980) defined VoQ as the auditory colouring of an individual speaker's voice. More concretely, as the suprasegmental attributes of both the laryngeal (or phonatory) and the supralaryngeal (or vocal tract) components of the articulatory settings adopted during speech production. Abercrombie (1967) regarded VoQ as a quasi-permanent characteristic that defines a person's voice. In this sense, VoQ can be used to detect pathological voices since it represents the state of a person's voice (Arias-Londoño et al., 2011).

Although regarding VoQ as a psychoacoustic phenomenon is quite common, its definition is problematic and there is no universal consensus (cf., Kreiman et al., 2004). Some studies like the one performed

by Laver (1980) exhaustively describe VoQ in terms of the global physiological configuration, but lack insight in how listeners may use different features to evaluate VoQ. The latter is very important as the reliability of listeners judgements on VoQ analyses can be questioned because listeners rarely agree in their ratings (Kreiman and Gerratt, 1998). This variability could be explained by the type of analysis, the mood of the listener, or the fact that different listeners may use different cues for evaluating VoQ.

One of the most common approaches oriented to the analysis of VoQ is to perform analyses by using a series of terms (breathy voice, hoarseness, etc.) to categorize different types of VoQ. Often, redundancies and ambiguities are found in this lists of terms and, as mentioned above, it may be difficult for listeners to agree in their interpretation. Nonetheless, in the absence of a universal definition and methodology, this approach may be considered an acceptable solution rather than an inaccurate one (Alm, 2011).

Phonatory settings terms from the seminal work of Laver (1980) have been used in many studies to analyse different types of VoQ (e.g., Kane and Gobl, 2011; Childers and Lee, 1991; Gobl and Ní Chasaide, 2003). Laver (1980) defined, described, and studied different VoQ characteristics. In the following Section, the phonatory settings described by Laver (1980) are summarized from an acoustical point of view rather than a physiological perspective because the former is more relevant to the content of the thesis.

2.3.1 *Laver's phonatory settings*

Laver (1980) described several phonatory (or laryngeal) settings which can potentially be controlled by any speaker with a normal vocal apparatus. This settings are not restricted to linguistic purposes such as language-specific phonetic characteristics, as they can also be used for conveying an affective state (paralinguistic level).

In the first place, a neutral reference from where all the phonatory settings can be produced was described: the modal voice. Modal voice and the rest of settings may occur alone (simple phonation types) or combined with another setting (compound phonation types). The following settings described by Laver (1980) are considered simple phonation types:

- **Modal voice**: Laver described this neutral state as *"the vibration of the vocal folds is periodic, efficient, and without audible friction"*. However, some settings may be achieved by other mechanisms instead of by vocal cords alone.
- **Falsetto**: It presents a F0 significantly higher than in modal voice and the pitch-control mechanism is also different. For instance, Hollien and Michel (1968) found a pitch for falsetto in male voice of 275–634 Hz vs. 94–287 Hz for a modal voice. This high F0 produces a wider separation between harmonics, entailing less components in any frequency range than in a voice with lower pitch. Moreover, the spectral slope of the speech spectrum is much steeper with respect to modal voice. Finally, the closing of the glottal flow is much more steeper in falsetto than in modal voice.
- **Whisper**: It is acoustically characterized by higher concentration of energy in formants bands and much more noise in all frequencies (specially, in the higher band).
- **Creak**: It can also be called "vocal/glottal fry". The F0 is very low, it can be around 24–52 Hz for males (Hollien and Michel, 1968) and the vocal pulses are more complex (voice pulses are clearly spaced in the temporal domain).

The following settings are considered compound phonation types, as they are combined with modal voice:

- **Harsh voice**: Its F0 has a similar range to modal voice but higher perturbations in the F0. Harshness is also associated with irregularity of the glottal flow signal, spectral noise, and amplitude irregularities.
- **Breathy voice**: When there is breathiness in the voice, the vocal cords vibration is less efficient than in modal voice as they do not close completely, producing an audible friction. Pitch and intensity are quite limited. Therefore, it is quite similar to "Whisper" in all these terms, but they are quite different from a physiological description. Nonetheless, the friction present in whispery voice is more prominent than in breathy voice.

In Table 2, all the combinations of different settings can be observed. The reader is referred to Laver (1980) for more information about all possible combinations

TABLE 2: PHONATORY SETTINGS PRESENTED BY LAVER (1980).

Phonatory Settings	
Simple phonation types	Modal voice
	Falsetto
	Whisper
	Creak
Compound phonation types	Whispery voice
	Whispery falsetto
	Whispery Creak
	Whispery creaky voice
	Whispery creaky falsetto
	Creaky voice
	Creaky falsetto
	Breathy voice
	Harsh voice
	Harsh falsetto
	Harsh whispery voice

It is also worth to note that, although Laver's description of VoQ is based on physiological properties rather than a perceptual point of view, he pointed out that listeners must be trained in the system in order to maintain a reliable annotation. That is, keep consistency with other listeners and with their own judgements of the same material on different occasions.

2.3.2 *Acoustic Voice Quality parameters*

Next, some of the most common VoQ parameters found in related studies are described. They have been classified into perturbation, spectral, and glottal parameters.

2.3.2.1 *Perturbation parameters*

To a greater or lesser degree, speech usually contains disturbances or perturbations of both frequency and amplitude. The speech signal consists of an harmonic part and a noise part, and the latter may be

understood as a form of perturbation since a great increase is associated to pathological voices (Shama et al., 2007). Below, some of the most used perturbation parameters describing these phenomena are defined:

- **Jitter**: It can be defined as cycle-to-cycle variations of the fundamental period. These variations are originated in the vocal cords, where fluctuations in the opening and closing times introduce a noise that manifests as a frequency modulation in the speech signal. Jitter can be measured with different parameters, which vary depending on several considerations such as the number of periods or the normalization.

  Jitter has been found to be correlated with breathiness (Eskenazi et al., 1990; Wolfe and Martin, 1997), hoarseness (Eskenazi et al., 1990), and roughness (Dejonckere et al., 1995).

- **Shimmer**: Similarly to jitter, shimmer can be defined as cycle-to-cycle variations of the speech waveform amplitude. In this case, the noise manifests as an amplitude modulation.

  Shimmer has been found to be correlated with breathiness (Dejonckere et al., 1995) and hoarseness (Wolfe and Martin, 1997). However, it is to note that it has also been stated that associations between jitter, shimmer, and perceived VoQ may not be sufficiently strong (Kreiman and Gerratt, 2005).

- **Harmonics-to-Noise Ratio (HNR)**: It measures the relation between the energy of the harmonic part and the energy of the rest of the signal (typically in dB).

  A low HNR has been associated to hoarseness (Yumoto et al., 1982), or, in a broader sense, as an indicator of laryngeal pathology (Shama et al., 2007).

  Finally, it is worth to note that there are other parameters that measure additive noise in the speech signal, such as the Normalized Noise Energy (NNE) and the Glottal-to-Noise Excitation Ratio (GNE). The former is typically defined as the relation between the noise energy and the total energy of the speech signal (Kasuya et al., 1986). The latter is computed taking into account the correlation coefficient of the Hilbert transform envelope among different frequency bands and it is related to the breathiness in the voiced speech (Michaelis et al., 1997).

2.3.2.2 *Spectral parameters*

Measures of spectral energy distribution may reflect changes in respiration, phonation, and articulation, related to specific paralinguistic variations (Banse and Scherer, 1996). Typically, these parameters reflect a relative difference of energy between different frequency bands (e.g., the Relative Amount of Energy above 1000 Hz (pe1000), Scherer, 1989) or harmonics. (e.g., H1H2, Jackson et al., 1985).

2.3.2.3 *Glottal flow parameters*

As explained in Section 2.1, when the airflow passes through the glottis, the vibration of the vocal cords transforms this signal into a quasi-periodic signal, which is typically called the glottal flow[5] (see Fig. 7). The glottal flow can be estimated by means of different techniques that can be classified into two classes depending on the way they perform the source-filter separation (cf., Drugman et al., 2012): inverse filtering and mixed-phase methods. The former is the most common technique and it consists of modelling the vocal tract and apply the model to the speech signal to obtain the glottal flow via inverse filtering. The different methods within this category mainly differ in the way the vocal tract is estimated. Mixed-phase models are based on the fact that speech is composed of minimum-phase (i.e., causal)

[5] Although the term "glottal flow" is used in this thesis, it is worth remarking that in similar works the terms "voice source", "glottal source","glottal-airflow", among others, are also used.

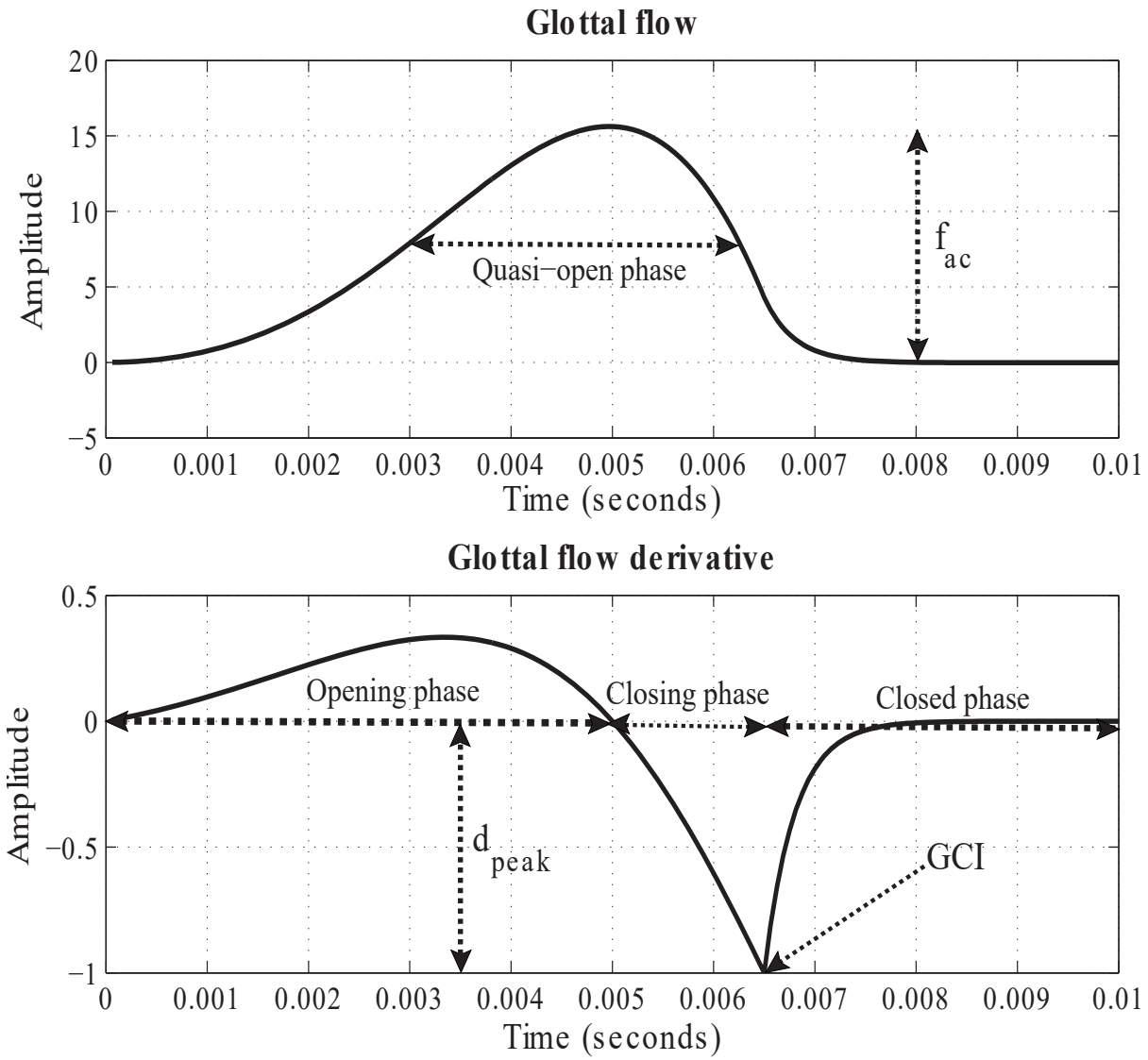

FIGURE 7: GLOTTAL FLOW (TOP) AND GLOTTAL FLOW DERIVATIVE (BOTTOM). F_{AC}: MAXIMUM OF THE GLOTTAL FLOW; D_{PEAK}: MINIMUM OF THE GLOTTAL FLOW DERIVATIVE; GCI: GLOTTAL CLOSURE INSTANT. FIGURE ADAPTED FROM KANE AND GOBL (2013).

and maximum-phase (i.e., anti-causal) components (Bozkurt and Dutoit, 2003). On the one hand, the vocal tract impulse response and the glottal return phase of the glottal component can be considered minimum-phase signals. On the other hand, the glottal open phase of the glottal flow is a maximum-phase signal (Doval et al., 2003). Therefore, the idea behind this approach is to separate minimum from maximum-phase components of speech.

In the next paragraphs, some of the commonly used glottal flow parameters are described. The reader is the referred to the work of Airas and Alku (2007) for a deeper overview of glottal parameters.

- **Normalized Amplitude Quotient (NAQ)**: It describes the glottal closing phase using amplitude-domain measurements (Alku et al., 2002). Specifically, it is defined by the ratio between the maximum of the glottal flow (f_{ac}) and the minimum of its derivative (d_{peak}), and it is normalized with respect to the glottal Fundamental Period (T0) (see Eq. 1 and Fig. 7). The parameter is robust

against distortion and it is capable of differentiating among different phonation types and emotions (Alku et al., 2002)..

$$NAQ = \frac{f_{ac}}{d_{peak} \cdot T_0} \tag{1}$$

- **Maxima Dispersion Quotient (MDQ)**: This parameter measures how impulse-like the glottal excitation is via wavelet analysis of the linear prediction residual (Kane and Gobl, 2013). Then, the dispersion of peaks across the different frequency bands are measured in relation to the Glottal Closure Instants (GCIs) (see Fig. 7) and they are averaged to be finally normalised to the local glottal period.

$$MDQ(p) = \frac{\frac{1}{K} \cdot \sum_{i=0}^{K} d_i}{T_0(p)} \tag{2}$$

where d_i is the distance from the maxima locations in the vicinity of the GCI, K is the number of scales, and p is the GCI index. MDQ is a good parameter to differentiate speech samples in the tense-lax dimension of VoQ; concretely, breathy, modal, and tense voice (Kane and Gobl, 2013).

- **Parabolic Spectral Parameter (PSP)**: It is a frequency-based parameter that describes the spectral decay of the glottal flow (a in Eq. 3) with respect to the maximal spectral decay (a_{max} in Eq. 3).

$$PSP = \frac{a}{a_{max}} \tag{3}$$

2.3.3 *Voice Quality in expressive speech*

The relevance of prosody and VoQ when it comes to perceptually identify affective states from speech has been recently investigated by Grichkovtsova et al. (2012). In that work, it is stated that in general each affective state entails particular patterns of both prosody and VoQ. Moreover, it has been observed that VoQ is key to identify subtle affective variations in oral communications (Gobl and Ní Chasaide, 2003). Several works have analysed VoQ patterns of emotions and attitudes (Monzo et al., 2007; Drioli et al., 2003; Gobl and Ní Chasaide, 2003; Sundberg et al., 2011; Wang et al., 2014).

Perceptual tests evaluating synthetic speech using different VoQs and comparing affective pairs (relaxed/stressed, friendly/hostile, sad/happy, bored/interested, etc.) were conducted by Gobl and Ní Chasaide (2003). The most readily perceived affective categories were states, moods and attitudes. These results show that the modification of VoQ is specially important for conveying milder affective states with subtle speech nuances like the ones that may be present in the storytelling speaking style. Contrarily, when conveying full-blown emotions, those results showed that prosodic patterns such as large pitch excursions are more crucial. A similar conclusion was provided by Monzo et al. (2007), who used several VoQ parameters were to discriminate among five expressive speaking styles (neutral, aggressive, happy, sad and sensual) by means of automatic classification. Recent studies have also highlighted the importance of VoQ to discriminate among emotions (Sundberg et al., 2011; Wang et al., 2014).

In consideration of the previously conducted studies, the acoustic analysis of storytelling speech conducted in this thesis takes into account both VoQ features and prosodic parameters, as a step forward from previous works where the acoustic analysis of the storytelling speaking style was only based on prosody (e.g., see Doukhan et al., 2011; Adell et al., 2005; Theune et al., 2006; Alm and Sproat, 2005a).

2.4 ANALYSIS TOOLS

In this Section, the main analysis tools that have been used during the development of this thesis are described in a nutshell.

2.4.1 *Speech segmentation software*

Typically, before analysing some speech characteristics, a segmentation of the speech corpora needs to be performed in order to extract the desired information. For example, in this thesis, a segmentation of utterances at the sentence-level is initially performed and then, such utterances are segmented at the phoneme-level as vowels is the minimal unit of analysis considered. The following speech segmentation softwares have been used during this step.

2.4.1.1 *Easyalign*

Easyalign is a Hidden Markov model ToolKit (HTK)-based automatic phonetic alignment tool for continuous speech developed as a plug-in for Praat. Speech is aligned from an orthographic transcription and a multi-level annotation composed of phonetic, syllabic, lexical, and utterance tiers is generated. The process consists of three steps:

1. **Macro-segmentation at utterance level** $\longrightarrow$ results in the "ortho" tier.
   - French, English, Spanish, Brazilian Portuguese, and Taiwan Min are supported in this step.
2. **Grapheme-to-phoneme conversion** $\longrightarrow$ results in the "phono" tier.
   - French, English, Spanish, and Brazilian Portuguese are supported in this step.
3. **Phone segmentation** $\longrightarrow$ results in the "words", "syll", "phones" tiers.
   - French, Spanish, Brazilian Portuguese, and Taiwan Min are supported in this step.

2.4.1.2 *SPeech Phonetization Alignment and Syllabification*

SPeech Phonetization Alignment and Syllabification (SPPAS) is a free audio annotation tool that produces automatically speech segmentation annotations from speech and its transcription and it is based on the Julius Speech Recognition Engine (Bigi and Hirst, 2012). Other functions are also available for managing corpora of annotated files. SPPAS uses XML-based files, but it is compatible with Praat, Elan, Transcriber, and many others. The whole process is a succession of four automatic steps: Utterance segmentation, word segmentation, syllable segmentation, and phoneme segmentation.

SPPAS is currently implemented for French, English, Italian, and Chinese.

2.4.1.3 *The Munich AUtomatic Segmentation web service*

The Munich AUtomatic Segmentation web service is an interface to provide easy access to to an application for automatic segmentation and labelling Kisler et al. (2012). The speech signal and an orthographic or phonological transcription is uploaded to perform the alignment and the result can be downloaded as a Praat's TextGrid. At the time of the development of this thesis, the Munich AUtomatic Segmentation web service supported the following languages: English, German, Hungarian, and Italian.

2.4.2 *Speech analysis software*

Once the speech corpora is prepared to be analysed, several acoustic parameters can be extracted from it by means of speech analysis software. In this thesis, the following softwares have been used.

2.4.2.1 *Praat: doing phonetics by computer*

Praat is a free computer program that can analyse, synthesize, and manipulate speech, create high-quality pictures, perform some statistical analyses, etc. (Boersma and Weenink, 2014). Praat also has its own scripting language that allows the creation of scripts to automatize tasks. It is a widely used software among linguists from a great variety of fields.

2.4.2.2 *COVAREP: A Cooperative Voice Analysis Repository for Speech Technologies*

COVAREP is a collaborative and freely available repository of advanced speech processing algorithms Degottex et al. (2014). In this thesis, the general COVAREP feature extraction script is used to extract several glottal flow parameters.

2.4.3 *Statistical analysis software*

Often, the extracted acoustic information consists of a large database. When dealing with large amounts of data, it is necessary to import such data into a statistical analysis tool in order to draw useful and robust conclusions that will (or will not) support the initial hypotheses. The following statistical analysis tools have been used during the development of the thesis.

2.4.3.1 *IBM SPSS*

IBM SPSS is a widely used program for statistical analysis (IBM Corp., 2013). The statistics included in the base software are the following:

- **Descriptive statistics**: cross tabulation, frequencies, descriptives, etc.
- **Bivariate statistics**: t-test, ANOVA, correlation (bivariate, partial, distances), non-parametric tests, etc.
- **Prediction for numerical outcomes**: e.g., linear regression.
- **Prediction for identifying groups**: factor analysis, cluster analysis (two-step, K-means, hierarchical), discriminant analysis, etc.

2.4.3.2 *StatSoft Statistica*

Statistica is a statistical analysis tool developed by StatSoft and acquired by Dell in 2014. Statistica provides a great variety of useful tasks within a user-friendly interface, e.g., data analysis and management, statistics, machine learning, data visualization, etc.

2.4.3.3 *The R Project for Statistical Computing*

R is a language and environment for statistical computing and graphics under the GNU General Public License. R provides a wide variety of statistical (linear and non-linear modelling, statistical tests, classification, clustering, etc.) and graphical techniques, and it is expanding its capabilities each day by means of new packages.

3

STORYTELLING

In this Chapter, works related to the analysis of structural properties of stories (Section 3.1) and those focused on the analysis/synthesis of storytelling speech (Section 3.2) are described.

3.1 STRUCTURAL PROPERTIES OF NARRATIVES – A HISTORICAL PERSPECTIVE

According to Barthes (1977), narrative is international, transhistorical, and transcultural. It can take many forms [6] and there are millions of different stories, so a universal narrative structure analysis approach can be considered unattainable. In fact, the analysis of narratives have been addressed from many areas of research, such as literary sciences, theology, history, psychology, anthropology and linguistics (Oliveira, 2000). Therefore, as discussed by Barthes (1977), the analysis is often carried out taking a general theory into account and, then, deduce from there which parts are contained in such theory. Nonetheless, there has been a universal agreement in considering narrative as a succession of events that are organized in a non-random fashion (Labov, 1972; Rimmon-Kenan, 1983; Cohan and Shires, 1988; Toolan, 1988; Greimas, 1966). However, the definition of such organization is quite author-dependent, ranging from subtle to great differences among authors in terms of the labels used and their definitions.

Along the 20th many authors have analysed narrative from a literary point of view (i.e., considering the text). Propp (1928) analysed 115 Russian folklore fairy tales and identified a common structure in them. Specifically, he identified 31 functions (delivery, departure, struggle, victory, wedding, etc.) that happen after an initial situation. Although not all tales or stories may contain all these functions, it is common to find at least some of them. Propp (1928) also classified the characters of the fairy tales depending on their role in the story: villain, donor, helper, sought-for-person and her father, dispatcher, hero, and false hero. Propp has been recognized for his approach but he has also been criticized for lacking a deeper analysis of subtleties and context (Taylor, 1964). Nevertheless, his work inspired a great number of studies. Bremond (1966) also considered functions applied to actions and events as defined by Propp (1928), which grouped in sequences form the story. However, he aimed to a more universal typology of stories rather than the one presented by Propp (1928). According to Bremond (1966), the narrator may use these functions in a more flexible way, e.g., not all stories end with a marriage. When three functions are grouped together they form the *basic sequence* and, as every process, this basic sequence consists of opening, mid, and closing phases. Furthermore, in the same way, the basic sequences are combined to form *complex sequences*. Similarly to Propp, Greimas (1966) classified the characters (*actants*) according to their actions in three interrelated pairs: sender/receiver, helper/opponent, and subject/object. Therefore, there is a relationship in both approaches in the sense that characters can be classified according to their actions, as these kind of actions are present in many tales. Nonetheless, it seems obvious that some characters may evolve along some tales or stories, playing different roles.

Todorov (1966) divided the processes of the narrative discourse in three groups: time, aspects, and modes. Firstly, the *time* of the discourse is linear but the story in itself is multidimensional, i.e., many events may take place simultaneously in the story but the discourse must be linear. In addition, some authors turn to temporal dispersion for aesthetic purposes. Secondly, the term *aspects* refers to the relationship of the narrator with the characters, i.e., how much the narrator knows or which information delivers about the characters. Finally, *modes* allude to how the narrator explains the story, e.g., alternating between indirect and direct discourse. Moreover, a story may contain passages where the narrator invites the reader to share the action by using first-person plural pronouns. Authors may also use direct

[6] In this thesis the main interest is placed on fictional tales and stories

discourse within the story, i.e., character interventions, conveying situations that are not directly experienced by the fictional narrator (Chafe, 1994).

Labov (1972) differentiated six functional parts in a fully formed narrative (although all of these elements may not appear in most narratives) in the following order:

- **Abstract**: It is an optional element used for providing a statement of a general proposal exemplified later on in the story.
- **Orientation**: It gives the background necessary to understand the story, i.e., information about the characters, time, situation and place where the action occurs.
- **Complication**: The complication is the main body of the narrative and contains a series of events that describe the actions that have taken place (Labov and Waletzky, 1997). Generally, the action builds up to its climax, to be later resolved with the result.
- **Evaluation**: It is also an optional element where the storyteller controls the importance of some narratives as opposed to others. Thus, controlling how the audience may receive the story.
- **Result**: The result is basically the resolution of the conflict contained in the narrative.
- **Coda**: Another optional element used by the narrator to wrap up the story, returning the audience to the present moment, e.g., *"that's all folks"*.

Only the "complicating action" is necessary for a minimal narrative, as the minimal definition of narrative consists of a pair of temporally ordered events.

Barthes (1977) proposed to divide narratives into three levels that share their definition with some of the aforementioned works. Specifically, the *functions* level (similarly to Propp, 1928), the *actions* level (related to the characters or *actants* introduced by Greimas, 1966), and the *narration* level, which can be related to the narrative discourse structure presented by Todorov (1966). However, the *functions* are defined with much more detail than in the work of Propp (1928), and they are differentiated in two groups: distributional (actions and events) and integrational (information related to atmosphere, characters, etc.). Furthermore, these two groups are broken down into cardinal, catalysers, indices, and informants (see Fig. 8). Cardinal functions are defined as actions that open, maintain, or close an uncertainty, and these are complemented by the calaysers, i.e., small actions that enrich those functions. Finally, the indices and informants establish the mood of the story but they do not move the story forward by themselves. The former refer to character and atmosphere characteristics and the latter to very specific information, e.g., the age of a character.

Calsamiglia and Tusón (1999) performed a deep analysis of what can be considered by some authors as discourse organization modes (Adam, 1992; Charaudeau, 1992). These modes represent textual functions like narrating, describing, etc., and they can be found mixed in many texts. The fiction literature essentially contains the narrative, the descriptive, and the dialogue modes (Calsamiglia and Tusón, 1999).

The narrative mode is used to inform, educate, persuade, etc., but it has also been a source of entertainment for children and adults and an essential instrument for cultural transmission. Probably, for these reasons, narratives have been widely analysed from a literary point of view. Adam (1992) defined an internal structure of the narrative sequence composed of six constituents:

- **Temporariness**: Events happen one after another, making the narrative to move forward.
- **Thematic unit**: It is related to at least one subject-actor.
- **Transformation**: There are states that change along the narrative (e.g., from fear to relax).
- **Action unit**: From an initial situation, through the transformation process, to a final situation.
- **Causality**: An intrigue is created via causal relationships between events.

In Fig. 9 the narrative diagram that results from these six constituents is depicted. As it can be observed it is quite related to the aforementioned functional parts of narratives defined by Labov (1972).

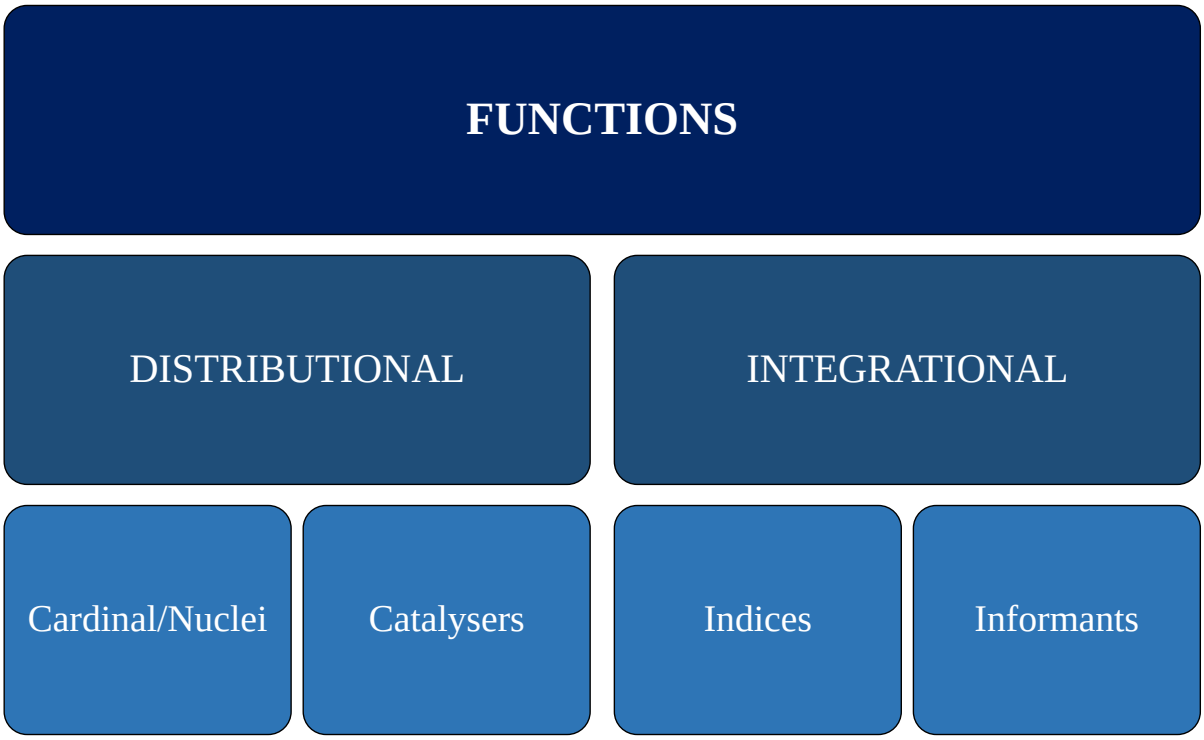

FIGURE 8: NARRATIVE FUNCTIONS AS DEFINED BY BARTHES (1977).

FIGURE 9: CONSTITUENTS OF THE NARRATIVE STRUCTURE AS DEFINED BY ADAM (1992).

The descriptive mode is used to linguistically represent a real or imaginary world: people, objects, environments, etc. The description may entail different purposes such as to inform or mock. Its structure starts with the subject, which can be established before or after the characteristics, then, the qualities and the relationship with the external world are defined (see Fig. 10). The most characteristic lexical elements of the descriptive mode are substantives and adjectives. According to Bal (1977), description occupies a marginal role in narratives, it is subordinated to the narration of action, and can be used as an ornament of the general narration.

The dialogue mode is present in many novels and tales as it is a factor that can engage the audience in the story to a large extent. Through the dialogue mode the audience can experience in a closer way what the characters are experiencing. Moreover, in oral communication the narrator may make use of different voices and styles to enhance realism and entertainment when performing different characters. Dialogues consist of an opening phase, the interaction, and a closing phase (see Fig. 11), and each turn tends to be enclosed between punctuation marks, such as quotation marks or dashes.

More recently, Alm and Sproat (2005b) analysed the sequential structure information of emotion in the text of fairy tales (from the point of view of the *reader*). The authors considered that the genre is schematic and therefore predictable. The results showed that the beginning of tales consists of a

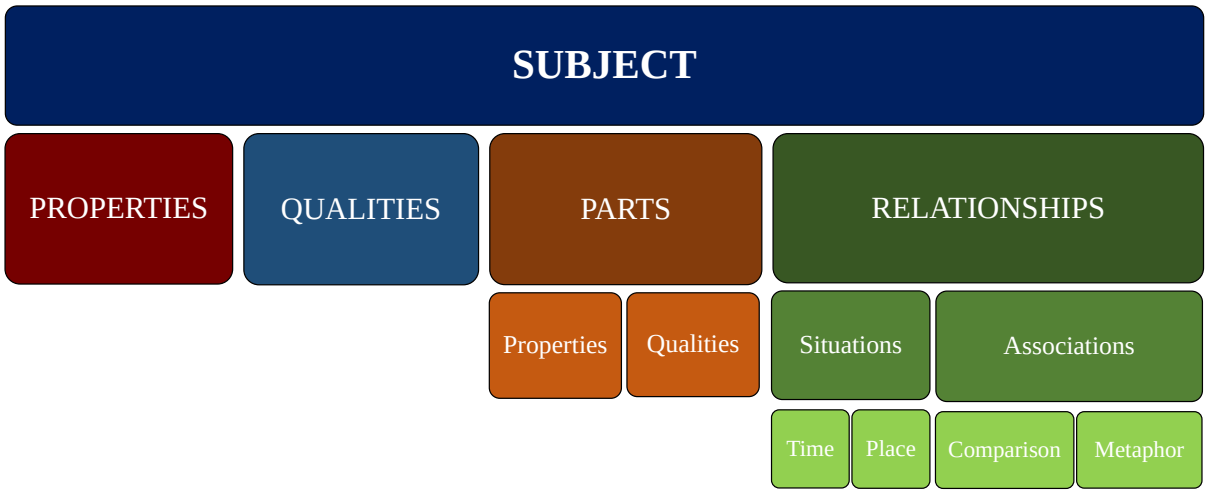

FIGURE 10: STRUCTURE OF THE DESCRIPTIVE SEQUENCE AS DEFINED BY ADAM (1992).

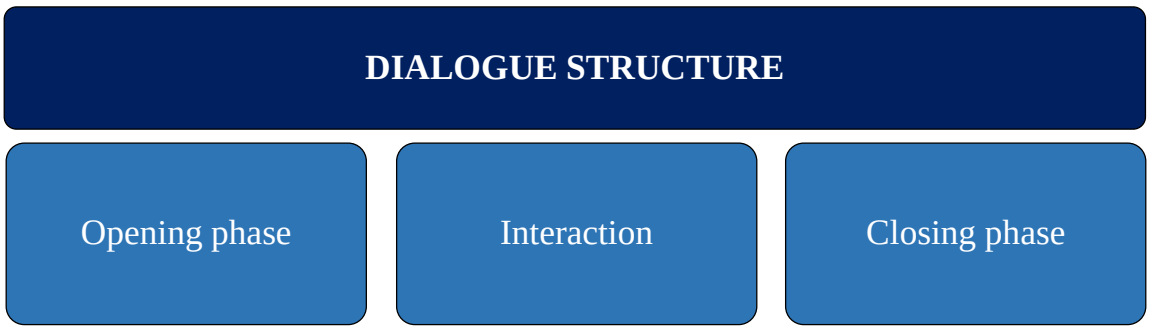

FIGURE 11: STRUCTURE OF A DIALOGUE AS DEFINED BY ADAM (1992).

neutral situation while they tend to have a happy ending in general. Moreover, it was observed that emotional sentences are surrounded by neutral sentences in most of the cases. Finally, some emotions like sadness were manifested in a greater range of consecutive sentences while others like surprise were more instantaneous. In summary, they observed that the first part of the fairy tales under analysis tends to be more neutral, then, when a series of events takes place, the emotional content is greater. It is also worth to note that it was observed that Propp's template does not always apply, e.g., some stories end with a negative emotional content.

In this thesis, the discourse modes approach described by Adam (1992) is considered as basis for the definition of the storytelling annotation methodology. Concretely, the focus is placed on the indirect discourse modes of storytelling, i.e., the narrative and descriptive modes (see Fig. 12).

3.2 ANALYSIS AND SYNTHESIS OF ORAL STORYTELLING

The literature related to the acoustic analysis of storytelling speech is very diverse, as different approaches have been considered depending on: the type of corpus under analysis, the considered annotation methodology, and the automation of the process.

3.2.1 *Types of storytelling speech data*

Firstly, concerning the storytelling speech corpus, several types of speech data have been analysed in the literature. On the one hand, there is a distinction between everyday storytelling and fictional

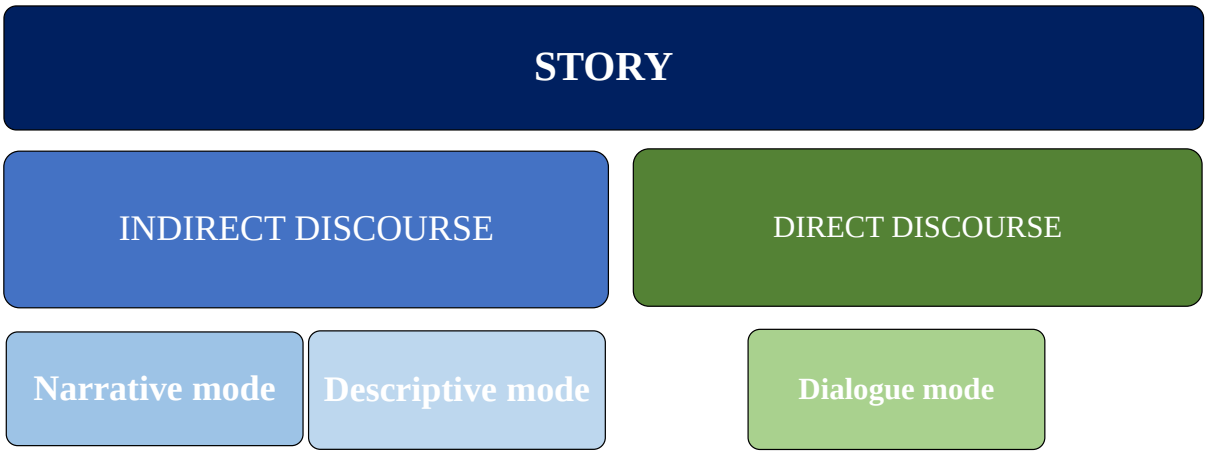

FIGURE 12: DISCOURSE MODES IN STORYTELLING.

storytelling/narratives. The former (sometimes called conversational storytelling), can be defined as those everyday situations when a person tells a story to other(s), which may be a personal or external experience. These situations may entail interruptions and comments from the listener(s) (Norrick, 2000). Differently, fictional storytelling refers to, e.g., a professional storyteller telling a fictional tale or story to an audience. On the other hand, the way the speech corpus was generated is also an important factor tied to the type of analysis. For instance, Jokisch et al. (2005) used "read-aloud" tales in a prosodic analysis together with other styles like "read-aloud" news. The "Tale" style was found to be slower and more expressive than its "News" counterpart. While "read-aloud" speech is useful for research that needs controlled data, it inevitably fails at capturing important expressive aspects present in everyday communication (Douglas-Cowie et al., 2003). Another way to generate representative speech corpora for a specific task is to use actors, as it can be assumed that actors are able to produce natural and realistic expressions (Banse and Scherer, 1996). In fact, it has been argued that some real-life expressions may also be considered acted because, depending on the social context, there may be some kind of self-control or sociocultural censure (cf., Banse and Scherer, 1996). Some spontaneous speech corpora examples related to oral narratives can be found in the works of Oliveira (2000) and Redford (2013). Oliveira (2000) analysed the prosodic characteristics of spontaneous non-elicited narratives. Specifically, the subjects selected to record the speech corpus were asked to talk freely about a certain topic, while the researcher only acted as an interviewer. Although the presence of the interviewer in principle could reduce the spontaneity, most of the participants were friends of the researcher, which contributed to the high degree of spontaneity and even some of them included narratives in their talk. Their spontaneous narratives were annotated using the *Labovian* model (see Section 3.1) to then evaluate if a prosodic structure (considering pause, speech rate, and pitch range) exists and if it is correlated with boundaries and narrative sections of the *Labovian* model. Several conclusions were drawn such as higher pitch reset and longer pause duration in narrative sections boundaries, and only pitch range yielded a significant result in the comparison between narrative sections. Redford (2013) opted for analysing spontaneous spoken narratives obtained after a storytelling/retelling task, in order to compare the pausing used by children and adults. Each subject had to look through a book of pictures and then develop the story they wanted to tell to the others. The results showed that even though the language of children is simpler than the language of adults, pausing patterns of both groups are quite similar. In any case, recording completely spontaneous speech is a very challenging task, since the quality of speech obtained outside a laboratory is lower (which may distort the acoustic analysis), and it is very difficult to remove any type of factor that may bias the spontaneity (e.g., the speaker is aware of the recording device).

From these different uses of the *storytelling* term, the present thesis focuses on storytelling as a fictional narrative interpreted by a professional actor or storyteller. Several studies have also considered

this type of speaking style for analysis and synthesis purposes, although the annotation approaches range from very specific expressive aspects of storytelling to more general analyses of the structure of narratives. In the following Section, these variety of methodologies is presented.

3.2.2 *Analysis and synthesis of storytelling speech: Different methodologies*

Adell et al. (2005) conducted a prosodic analysis of a tale after dividing the text into storytelling discourse modes (see Section 3.1) to perform automatic classification of sentences. The authors concluded that prosody is influenced by discourse modes, specially through variations of some F0 parameters (e.g., F0 mean, first derivative of F0, or its minimum value). Nonetheless, the automatic sentence classification performed on the tale achieved a 64% of correctly classified sentences, which is a moderate value leaving room for improvement. Several years later, some tales were prosodically analysed and modelled by Doukhan et al. (2011) according to a scheme derived from the *Proppian* structure of tales (see Section 3.1). Concretely, they analysed the title, the exposition, the triggering event, the scenes, the ending and the epilogue, and they also included different types of characters according to age, gender, size, etc. All the tales under analysis (a total of 12 that entailed one hour speech, performed by a professional speaker) had at least a title and a scene, but not all them contained the rest of sections considered in the scheme used. Firstly, they performed a general comparison of global storytelling speech (considering characters and narrator turns all together) and the indirect speech of the narrator with respect to the results obtained by Roekhaut et al. (2010) in their prosodic analysis of radio news, political address and conversational speech. The comparison revealed rhythmic similarities between both storytelling corpora and the political style, besides a clearly wider pitch range in storytelling (around +6 STs). Secondly, the prosodic characteristics of each narrative structure were analysed. The titles showed exaggerated prosodic values (highest mean intensity, highest pitch range, slowest speaking rate, etc.), while the beginning of tales (expositions and triggering events) shared prosodic similarities (although the mean intensity of triggering events was significantly lower). Furthermore, the scenes showed average values in all prosodic parameters, and tales tended to end (refrains and epilogues) with a quiet and flat intonation. Finally, it was observed that the storyteller tended to increase his pitch and intensity when interpreting a character (direct speech) than in the narration parts (indirect speech). The authors left for future work the refinement of the model, which was foreseen through a prosodic analysis at the sentence level together with the expansion of the set of acoustic features with VoQ parameters. This claim together with the results obtained by Adell et al. (2005) are indicators that considering an annotation scheme based on the global narrative structure of tales could be not sufficient to capture the diversity of conveyed expressive nuances in storytelling speech, which in turn, is key to create a good storytelling acoustic model for TTS systems.

Other authors opted for drifting away from the use of narrative structure theories to annotate and analyse storytelling speech oriented to TTS applications. Concerning studies using manual annotations, Theune et al. (2006) modelled what they referred to as global storytelling style and suspense situations (increasing and sudden suspense) present in some sentences of a particular tale. The global storytelling speaking style was derived after observing that in storytelling there is much more variation of pitch and intensity than in newsreaders speech[7]. More concretely, the authors observed that storytellers speak slower than the newsreaders, take longer pauses, and sometimes add an emphasis within certain adjectives and adverbs by increasing their pitch and duration. Although the resynthesized speech (applying the prosodic rules to neutral synthetic speech) of global storytelling and suspense utterances obtained good results from a subjective point of view, the reported prosodic rules (derived from a trial-and-error approach) are difficult to generalize because of the analysis was informal and because of the very small

[7] Theune et al. (2006) associated newsreader speech to the default TTS system speech, but this assumption was not checked.

amount of speech data considered in the analysis of suspense. A similar procedure was recently conducted to convert neutral synthetic speech to storytelling speech for three Indian languages using a TTS system and a set of prosodic rules (Sarkar et al., 2014). To that effect, the authors tackled the annotation of storytelling speech using "story-specific emotions" labels. Concretely, they observed five different "story-specific emotions" in their corpora: anger, fear, happy, sad, and neutral. Twenty phrases from each category were perceptually analysed to derive a set of prosodic rules by a trial-and-error methodology. The resynthesized utterances obtained slightly better results than their neutral synthesized counterparts. A later work of some of the authors has introduced a pause prediction model for storytelling speech in Hindi language (Sarkar and Sreenivasa Rao, 2015). Differently than Theune et al. (2006) and Sarkar et al. (2014) but with the same aim, other researchers have opted for automatic approaches. Eyben et al. (2012) showed an alternative to using hand-crafted definitions of expressive classes by means of an unsupervised clustering approach taking into account acoustic features that automatically classified different parts of an audiobook: 'narration' (a sentence without quotes), 'carrier' (a part not in quotes of a sentence containing quotes, e.g., "he said") and 'direct speech' (character intervention between quotation marks). After evaluating the results, it was observed that the feature set containing prosodic and perturbation features (local jitter and shimmer, and HNR) showed the highest automatic classification performance. While that study shed light on the importance of VoQ in storytelling, the three considered classes could fall short in describing the diversity of expressive nuances in storytelling speech. For example, such 'narration' category may contain utterances with a suspenseful tone (Theune et al., 2006), whereas 'direct speech' involves many characters interventions that sometimes entail emotional speech (Buurman, 2007; Burkhardt, 2011).

On the other hand, some works have linked basic emotions to storytelling. Alm and Sproat (2005a) prosodically analysed two speech corpora of children's stories (around 10 minutes of speech by a professional speaker) according to the "Big Six" emotions approach (anger, disgust, fear, happy, sad, and surprise, Cornelius, 1996), besides adding a neutral category as reference. The labelling of the 128 sentences was performed by means of perceptual tests presented to several participants, and resulted in moderate-low agreement (46 representative utterances were retained from the original corpus). Their prosodic analysis showed a certain degree of correspondence with previously reported emotional acoustic profiles in the literature. Nevertheless, some contradictory results were also highlighted, e.g. the authors observed a pitch decrease for anger. This result could be an indicator that modelling indirect storytelling speech only by means of emotions may not the best option. Moreover, the aforementioned categorization into "story-specific emotions" used by Sarkar et al. (2014), already denotes that some authors are not entirely sure of completely relating the indirect storytelling speaking style to basic emotions. Contrarily, it is worth noting that emotional acoustic models borrowed from the literature have been satisfactorily used for synthesizing tale characters' interventions. Firstly, Buurman (2007) analysed how storytellers modify their voice when interpreting the characters of the story, thus, conveying the emotions that such characters are experiencing. Concretely, four characters (a total of 30 sentences) interpreted by a professional storyteller were acoustically analysed with the objective of creating a prosodic model to be implemented in a TTS system. However, due to the fact that the analysed corpus was too small, the obtained results were not robust enough to derive a reliable acoustic model. Therefore, the author opted for the implementation of emotional acoustic models from the literature (Schröder, 2004). The results of the emotional recognition of the synthetic speech were relatively low, but comparable to the recognition rate of the original fragments. Buurman (2007) suggested that for future work it would be interesting to include some acoustic resorts unrelated to speech such as coughing, laughing, or sobbing inside speech fragments and a deeper analysis considering VoQ. Later on, Burkhardt (2011) presented an affective spoken storyteller system capable of providing emotional expressiveness to characters of stories and tales. The system was a speech synthesis software (part of the *Emofilt* open source software, Burkhardt, 2005), which allowed text tagging based on different emotional speaking styles, and thanks to different acoustic rules, different emotional outputs could be generated. The software used emotional

rules from the related literature, while rules for other speaking styles were adjusted manually. The system was evaluated with a perceptual experiment in which 15 children took part. Some of them criticized the lack of understandability resulting from the emotional way of speaking. However, they liked the resulting variation of expressiveness.

Other works related to storytelling have been more focused on the synthesis part (omitting the analysis) for generating expressive voices in storytelling contexts (e.g., interactive storytelling or reading audiobooks). Silva et al. (2001) reported several problems associated with the generation of this speaking style. The authors claimed that the lack of flexibility of the TTS engine embedded in their embodied digital storyteller was the main reason for not achieving the desired expressiveness. A later work of some of these authors centred on interactive storytelling also remarked that the synthetic quality of the employed TTS system was not good enough to achieve their goal (Silva et al., 2004). Although not highlighted by the authors, the fact that both works used a prosodic model based on emotional acoustic profiles could had been another reason for achieving those undesired results.

Cabral et al. (2006) conducted a perceptual experiment to study the impact of synthetic storytelling speech in users' perception compared to real storytelling speech. For that purpose, they recorded a real storyteller and transplanted his gestural and prosodic information to a synthetic character. Then, four different configurations were implemented: real character with real voice, real character with synthetic voice, synthetic character with real voice and synthetic character with synthetic voice. The speech synthesis framework was based on a Linear Predictive Coding (LPC) diphone synthesizer and used a pitch-synchronous time-scaling method to avoid distortion caused by the pitch-scale transformations. The results showed that the real voice was much preferred. The synthetic voice achieved good intelligibility, and the emotional content was partially detected, but the overall feeling was that it was not close to the real voice.

Suchato et al. (2010) described an application to automatically generate digital talking books that, e.g., can be very helpful for blind people. The motivation behind this was to reduce the need of human effort, as many recording time is needed when reading aloud entire books. Moreover, they highlighted that enhancing the quality and naturalness of the synthetic voice of nowadays TTS systems is still challenging, especially for storytelling purposes. In their TTS synthesis sytem, F0, syllable durations, and VoQ parameters (degree of breathiness and creakiness) could be manipulated. Recorded speech signals from two speakers (male and female) were used to train two sets of acoustic models for the speech synthesizer based on HMM. The user could tag the text choosing between these two standard voices in combination with one of the predefined 12 expressive styles: Neutral, Monotonic, Lowly-pitched, Highly-pitched, Rising-pitched, Falling-pitched, Crying, Whispering, Robot-like, Randomly-pitched, and Melody-aligning (singing). The results showed that the HMM-based TTS system developed by the authors was always preferred when compared to a concatenative TTS system. It is also worth to say that the configuration with manual voice modifications is the one that obtained best results.

Székely et al. (2011) used a unsupervised clustering technique considering glottal source parameters to group utterances from a 50 minute long audiobook into clusters associated with different voice styles. After a perceptual evaluation of the clustering results, the authors found out that the method successfully separated groups of sentences associated with different voice styles. Similarly, Jauk et al. (2015) performed unsupervised k-means clustering using prosodic, spectral and an iVector representation (Lopez-Otero et al., 2014) of acoustic features on a juvenile narrative audiobook with the aim of creating expressive voices for speech synthesis. The authors also designed a novel subjective evaluation technique to assess the performance of their resulting system, which consisted of editing a paragraph with a set of different expressive voices.

A data-driven approach was presented by Greene et al. (2012) for predicting the most appropriate voices for characters in children's stories based on salient character attributes. They remarked that current TTS systems oriented to produce storytelling speaking style fail to deliver what the user expects. As children's stories contain multiple characters of different gender, age, etc., the use of specific voices

for each character enhances the experience of storytelling and the story may be more entertaining and understandable for the audience. They looked for representative voices of reoccurring classes of characters in children's stories and tried to map these voices to AT&T TTS voices[8], although they needed to expand these voices with other publicly available sources. A perceptual test was conducted to associate certain adjectives representing gender, age, emotion, intelligence, etc. to character voices. With the final attributes, the character voices of the TTS system were annotated. Then, a vector of attributes was introduced in a Naive Bayes classifier and a ranking of the voices was returned, from most likely to least likely. The system has over a 50% chance of returning the correct voice as one of its top 5 guesses and an 85% chance of returning the correct voice in the top 10 guesses. Gender and age information by itself outperformed random selection, but the fully system offered the best results. In the future, the authors plan to design a perceptual test in order to obtain information about how appropriate people find the character voices.

As it can be observed, there is a great variety of approaches in the literature. The following parts present the contributions of this thesis to tackle the analysis and synthesis of this particular expressive style.

[8] http://wizzardsoftware.com/text-to-speech-sdk.php

Part I

THE ROLE OF PROSODY AND VOICE QUALITY IN INDIRECT STORYTELLING SPEECH: ANNOTATION METHODOLOGY AND EXPRESSIVE CATEGORIES

In the previous Sections, it has been observed that in addition to prosody, Voice Quality may also play an important role for the modelling of speech nuances present in oral storytelling. However, little work has been performed to corroborate this assumption. In this first part of the thesis, the role of both prosody and Voice Quality in indirect storytelling speech is analysed to identify the expressive categories it is composed of and the acoustic parameters that characterize them. To that effect, an analysis methodology to annotate this particular speaking style at the sentence level is proposed based on storytelling discourse modes (narrative, descriptive and dialogue), besides introducing narrative sub-modes. Once the annotation process is finished, prosodic and Voice Quality parameters are automatically extracted for each uttered sentence. Next, the acoustic parameters that are significantly characteristic of each expressive category are determined through statistical and discriminant analyses.

4

ANNOTATION METHODOLOGY AND EXPRESSIVE CATEGORIES

An annotation methodology based on storytelling discourse modes that blends perceptual and structural approaches is proposed. This approach considers text-dependent and perception-dependent categories.

4.1 ANNOTATION METHODOLOGY FOR INDIRECT STORYTELLING SPEECH

Tales and stories typically contain narrative, descriptive and dialogue modes (see Section 3.1). The narrative mode is generally used to inform the listener/reader about the actions that are taking place in the story, whereas the descriptive mode has the function of describing characters, environments, objects, etc. The dialogue mode is present when the characters have a conversation and their turns explicitly appear in the story. Thus, the narrative and descriptive modes belong to the indirect discourse, whereas the dialogue mode represents a direct discourse. If a storyteller interprets the characters, he/she may perform particular voices for the different characters of the story using a more exaggerated expressiveness to enhance realism and entertainment, often including emotional content (Buurman, 2007). Thus, it seems reasonable to annotate the storytelling direct discourse using emotional category labels as it contains exaggerated acted speech (Douglas-Cowie et al., 2003).

In tales and stories, the narrative mode generally is the predominant mode. A professional storyteller conveys the information present in this mode using diverse expressive styles, which can be changed from sentence to sentence (Theune et al., 2006; Alm and Sproat, 2005a). For this reason, the proposal is to divide the narrative mode into sentence-level sub-modes or expressive categories.

The introduced annotation methodology to analyse storytelling speech corpora is depicted in Fig. 13. It consists of a first stage of text-level annotations followed by perceptual analyses together with some verifications to improve the reliability of the annotation. As a proof of concept, the introduced annotation methodology is applied on an audiobook interpreted by a Spanish professional male storyteller. This story belongs to the fantasy and adventures genres, with children and pre-teenagers as its main target audience. The audiobook contains approximately 20 minutes of indirect storytelling speech, composed of 263 sentence-level utterances.

4.1.1 *Text-dependent categories*

There are two widely employed storytelling categories that can be annotated at the text level (Adell et al., 2005; Mamede and Chaleira, 2004): descriptive sentences, which compose the descriptive mode (see Section 3.1), and sentences that specify a character intervention.

On the one hand, an important lexical characteristic of the descriptive sentences in stories and tales is the large number of adjectives used. Moreover, verbs like “to be” and “to have” in the past and present tenses abound along this mode (e.g.,*“He was a tall, strong boy with faded bluish eyes”*). On the other hand, usually right after characters’ interventions, the narrative mode contains sentences that specify the character turn (henceforth post-character sentences). These sentences show specific lexical cues, e.g., they usually start with a declarative verb in the third person or “speaking word” (“said”, “answered”, “murmured”, “asked”, etc.) and, typically, the character who intervened in the previous direct discourse is named (Mamede and Chaleira, 2004; Zhang et al., 2003). Although post-character sentences have been automatically classified in previous works (Mamede and Chaleira, 2004), their annotation is manually addressed to ensure the reliability of subsequent analyses.

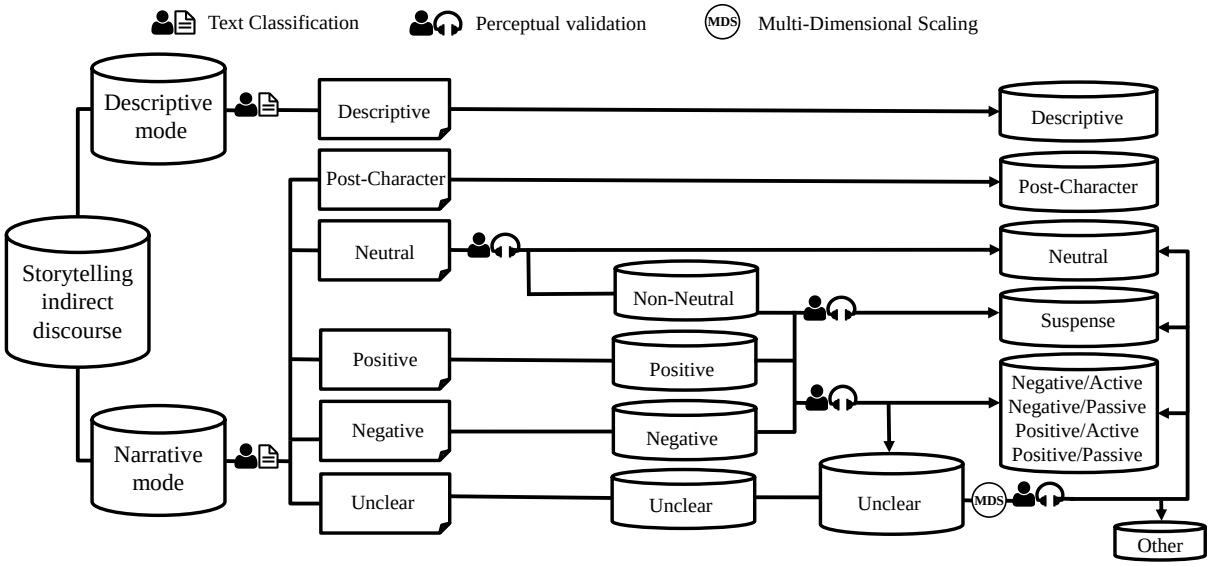

FIGURE 13: ANNOTATION METHODOLOGY FOR THE CLASSIFICATION OF UTTERANCES INTO EXPRESSIVE CATEGORIES OF THE INDIRECT STORYTELLING SPEAKING STYLE.

The text-level annotation with no audio input of the story was entrusted to two expert annotators (with Spanish as native language) after a thorough briefing of the annotation goal. Specifically, they were asked to annotate post-character, descriptive, neutral, positive and negative sentences. The motivation of the neutral, positive and negative sentences annotation is explained in Sections 4.1.2.1 and 4.1.2.3. Sentences where the annotators did not agree were initially labelled as unclear sentences, being left aside for subsequent analyses as unclear sentences (see Fig. 13). The inter-rater agreement was found to be Kappa $\kappa = 0.760$ ($p < 0.001$), due to some disagreements in several sentences within the neutral, positive, and negative sentences annotation (no disagreement in descriptive and post-character sentences was observed). From the 263 sentences, 24 were labelled as descriptive while 40 belonged to the post-character category. Regarding the neutral category, 73 sentences were extracted from the text together with 27 positive and 84 negative sentences. The number of unclear sentences after the text-level annotation resulted in 15 (i.e., 5.7% of the total).

4.1.2 *Perception-dependent categories*

The motivation of the perceptual analysis is to identify expressive categories that may not be easily identified from text-only annotation, for example, suspenseful utterances that entail very particular expressive nuances (Theune et al., 2006). It is worth noting that this task lacks clear ground truth, in the sense that there is no standard set of classes for annotating storytelling speech (Eyben et al., 2012). However, it seems reasonable to group perceptually similar sentences and investigate which acoustic parameters (in terms of prosody and VoQ) better represent their expressiveness. This subjective analysis performed after the text-level annotation can be regarded as the manual alternative of automatic methods such as unsupervised clustering of expressive categories (Eyben et al., 2012). Since the major drawback of the approach is the inherent subjectivity of the annotation (Alm, 2011), annotation agreement verifications are considered to mitigate this fact.

In the following paragraphs, the different perception-dependent categories within storytelling speech are detailed together with the results of their annotations.

4.1.2.1 *Neutral reference category*

In line with most affective speech analysis and synthesis studies (see Schröder, 2004, and references therein), a reference neutral category is also considered in this work. Within the narrative mode of tales and stories, there are merely informative sentences about actions or facts containing neutral lexical elements (e.g., *"The boy came into the living room"*). While these sentences can be identified via text only (as explained in Section 4.1.2), a posterior perceptual validation step is included to ensure the neutral expressiveness of their oral interpretation. This way, those sentences with non-neutral expressiveness uttered by the narrator (e.g., the narrator has added a suspenseful style) are not included to ensure the homogeneity of the neutral category.

For the perceptual validation of the neutral category, the two experts listened to the uttered neutral sentences of the story. As a result, the number of valid neutral exemplars was reduced from 73 to 53 due to some disagreements (and agreements on non-neutrality) between the two experts ($\kappa = 0.580, p < 0.001$). The remaining 20 sentences are labelled as non-neutral (see Fig. 13).

4.1.2.2 *Suspense category*

Suspense is a very important component of narratives, and it is related to the feeling of excitement or anxiety that the audience (listeners or readers) feels because of waiting for something to happen, i.e., the outcome is uncertain (Brewer and Lichtenstein, 1982). According to Brewer and Lichtenstein (1982), a suspense discourse organization typically consists of an initiating event and the outcome event and, the more additional discourse material placed between them the more build up of the suspense (prolonging of resolution). Therefore, suspense is evoked when the events are narrated in chronological order and the outcome is not known until the end.

In general, the feeling of suspense contributes significantly to the entertainment of the audience and, thus, it is present in many stories. This description of the discourse structure of suspense correspond to a global view within the narrative, but there is no description of suspense at, e.g., the sentence level. Furthermore, it has already been noted that determining which parts of a text are suspenseful is still an unsolved issue (Theune et al., 2006), even though some words may be useful for the identification of suspense utterances from text (e.g., "then", "suddenly", "but", etc.). As a result, utterances should be perceptually analysed in order to find out whether the narrator has conveyed a suspenseful style or not.

Similarly to the validation of the neutral utterances, the same two experts were asked to listen to the rest of the corpus with the objective of detecting suspenseful utterances. Concretely, 20 non-neutral, 27 positive and 84 negative sentence-level utterances are considered in this study. Among them, the annotators agreed (with $\kappa = 0.604$, $p < 0.001$) that 21 utterances contained a suspenseful style, being 10 of them initially labelled as non-neutral and 11 as negative, respectively.

Although two specific suspense situations with particular acoustic patterns were previously analysed by Theune et al. (2006), in this analysis suspense situations are considered in a more global sense. Concretely, utterances where it is evident that the storyteller wants to elicit uncertainty in the audience. No utterances fitting the description of the two specific suspense categories given by Theune et al. (2006) have been detected in the corpus at hand. However, the authors discussed that other forms of suspense, besides those already analysed in that work, may be distinguished in tales and stories. Furthermore, some authors have suggested that using a soft voice (lower mean intensity) may induce intimacy and suspense (Doukhan et al., 2011). This fact will be checked during the subsequent acoustic analyses.

4.1.2.3 *Affective categories*

At this point of the annotation process, a large number of utterances have already been labelled. However, there is a significant number of utterances within the narrative mode that may remain unclassified. For such utterances, a labelling scheme based on valence and activation is used. This is not an ideal but a

TABLE 3: GATHERED UTTERANCES FROM THE SPEECH CORPUS AFTER THE ANNOTATION PROCESS.

Category	# Utterances
Neutral	53
Post-Character	40
Descriptive	24
Negative/Passive	36
Negative/Active	23
Positive/Passive	13
Positive/Active	13
Suspense	21
Other	40

simple yet effective solution to deal with the annotation of affective situations where the emotional state is not fully defined (Schröder, 2004). The rationale behind this idea is that the narrator is neither self-experiencing the emotions that affect the characters nor acting them. Instead, he/she tries to entertain the audience and engage them in the story, awaking empathy for the characters of the story (Brewer and Lichtenstein, 1982). For instance, if the storyteller is narrating a passage where the characters are experiencing a joyful situation, he/she will convey the information with a positive tone rather than using the joy emotion per se. Thus, an annotation methodology based on emotions seems inappropriate for the indirect discourse of storytelling, as previously discussed in Section 3.2.

Firstly, the valence (positive or negative) of the sentences is determined at the text level (see Section 4.1.1). While discerning valence is relatively feasible via text classification, discerning activation only from text is a difficult task (see Trilla and Alías, 2013, and references therein). To that effect, a perceptual analysis is conducted by the annotators to determine if the utterances are active or passive in terms of activation. As a result, this process generates four expressive categories: negative/passive, negative/active, positive/passive, and positive/active (see Fig. 13).

From the previous annotation stage, the remaining 27 positive, 73 negative and 10 non-neutral utterances from the corpus at hand were perceptually classified by the two experts into the aforementioned four categories according to their perceived activation. 81 utterances where agreement between annotators was found (with $\kappa = 0.650$, $p < 0.001$) were retained for the subsequent analyses, being 21 negative/active, 35 negative/passive, 13 positive/active, and 12 positive/passive utterances.

4.1.2.4 *Reallocation of unclear utterances*

At this stage of the annotation process, some utterances may be still unclear in terms of expressiveness (see Fig. 13). Such utterances are considered in this final stage with the aim of reallocating the most of them (if possible). To that effect, a perceptual analysis together with a Multi-Dimensional Scaling (MDS) asked to place each unclear utterance in a common acoustic space is conducted. MDS is a good option for objectively measuring the distance (in this case, the Euclidean distance) of one item to a defined category in a common space (Kruskal and Wish, 1978). If an unclear utterance is close to a category centroid in the MDS representation and it is perceptually similar to that category, the utterance is included in that category. We discarded using automatic classification methods such as neural networks or Support Vector Machines (SVMs), since training a classifier with such low number of samples seemed inappropriate to us, although it could have been an alternative.

The remaining unlabelled utterances, i.e., 17 negative, 2 positive and 10 unclear utterances from the affective perceptual validation together with 15 unclear sentences from the text-level annotation (a total of 44 utterances) were considered in this final stage. However, only 4 from the 44 utterances were finally retained: 2 were reallocated in the negative/active category, 1 in the negative/passive category and 1 in the positive/passive category. The remaining 40 utterances were classified as 'Other' because they contained undesired expressive elements for the subsequent analyses. For instance, some utterances contained laughter or a yawn within the running speech, whereas in others the narrator slightly imitated the voice of the character he was referring to.

The final number of gathered utterances for the considered storytelling speech corpus as a proof of concept of the annotation methodology is shown in Table 3. From the original 263 utterances, 223 have been classified satisfactorily (a 84.8% from the total). This indicates that the introduced annotation methodology fulfills the objective to a large extent. Nonetheless, 15.2% of the utterances were classified as 'Other', since they contained specific expressive cues (as aforementioned) that lay out of the scope of the present thesis.

4.2 ACOUSTIC ANALYSIS FRAMEWORK

In this Section, firstly, the methodology for the acoustic analysis of storytelling expressive categories is described. Next, all the prosodic and VoQ parameters used in the present work are enumerated. Finally, the methodology for the parameters extraction from the speech corpus is explained.

4.2.1 *Acoustic analysis methodology*

The analysis methodology used in this work follows similar steps to previous studies devoted to the analysis of affective speech (Monzo et al., 2007; Pell et al., 2009b; Liu and Pell, 2014). Firstly, it is worth to note that relative acoustic differences among storytelling categories are studied. This way, speaker-dependent profiles are avoided. Moreover, the results are also normalized within each corpus using z-scores. Then, a series of statistical and discriminant analyses are conducted using SPSS in order to assess if the different storytelling categories under analysis could be acoustically differentiated. Statistical significance is assumed to be achieved at $p < 0.05$ for all the statistical analyses. As a first step, a Multivariate ANalysis of VAriance (MANOVA) is performed considering all acoustic parameters as dependent variables. Then, a series of univariate analyses are conducted to evaluate differences among categories for each parameter. To conclude the statistical analyses, Tukey's Honestly Significant Difference (HSD) post-hoc tests (i.e., pairwise comparisons between categories) are performed. However, due to the considerable number of parameters under analysis, only the results from those parameters that in the discriminant analysis strongly correlate with some significant canonical function, explaining a major portion (around 95%) of the variance among categories, are discussed. The discriminant analysis is also carried out to assess how the different storytelling categories can be discriminated based on the acoustic parameters taken into account. Wilks' lambda is reported as a measure of discriminating capability of each parameter, as the smaller this value the more important the parameter to the discriminant function (Klecka, 1980). In addition, a Linear Discriminant Analysis (LDA) is also conducted following previous works (Monzo et al., 2007; Pell et al., 2009b; Liu and Pell, 2014). The classification results are reported using the F1 measure, as it combines both precision and recall into a single metric giving a compact vision about how good the classification is (Sebastiani, 2001). All these analyses are then also performed on the emotional corpus in order to compare emotions with respect to storytelling categories.

4.2.2 *Considered prosodic and Voice Quality parameters*

The considered acoustic features describing prosodic and VoQ information for the analyses in this thesis represent a mix of voice source, waveform, and spectral parameters, which has been recommended previously (Sundberg et al., 2011). In addition, they have also proven useful in previous works devoted to affective speech analysis (Monzo et al., 2007; Kane and Gobl, 2013; Banse and Scherer, 1996; Monzo et al., 2014). Furthermore, most of them are part of the recently introduced Geneva minimalistic acoustic parameter set (Eyben et al., 2015).

Information related to prosody is extracted in terms of mean F0 ($F0_{mean}$ in Hz), F0 range (considering range as the Inter-Quartile Range (IQR) following Doukhan et al., 2011), mean intensity (int_{mean} in dB), and speech tempo related measures such as AR (in syllables per second) and Number of silent pauses (NSP) within the utterance.

Regarding perturbation measures, local jitter (in %), local shimmer (in %), and mean HNR (HNR_{mean}) in dB, are taken into account. Jitter is computed as the average absolute difference between consecutive periods, divided by the average period, in % (see Equation 4).

$$jitter(local) = \frac{\frac{1}{N-1} \cdot \sum_{i=2}^{N} |T_{0i} - T_{0i-1}|}{\frac{1}{N} \cdot \sum_{i=1}^{N} T_{0i}} \cdot 100 \ [\%] \tag{4}$$

where N is the number of signal periods, i refers to each one of them, and T_{0i} its fundamental period. Similarly, Equation 5 shows how shimmer is computed.

$$shimmer(local) = \frac{\frac{1}{N-1} \cdot \sum_{i=2}^{N} |U_i - U_{i-1}|}{\frac{1}{N} \cdot \sum_{i=1}^{N} U_i} \cdot 100 \ [\%] \tag{5}$$

where N is the number of signal periods, i refers to each one of them, and U_i is the peak-to-peak amplitude value within each i period.

For HNR, the definition of Boersma (1993) is considered:

$$HNR(dB) = 10 \cdot log_{10} \cdot \frac{r'_x(\tau_{max})}{1 - r'_x(\tau_{max})} \tag{6}$$

where $r'_x(\tau_{max})$ is the energy of the periodic component and its complementary value $(1 - r'_x(\tau_{max}))$ represents the relative power of the noise component.

In what concerns spectral parameters, the following are also explored:

- **PE1000**: measures the amount of relative energy in frequencies above 1000 Hz with respect to those below 1000 Hz in dB (Scherer, 1989).
- **Hammarberg Index (HAMMI)**: difference between the maximum energy in the band frequencies [0, 2000] Hz and [2000, 5000] Hz expressed in dB (Hammarberg et al., 1980).
- **H1H2**: difference of amplitude between the first two harmonics in dB (Jackson et al., 1985).
- **Spectral Slope (SS)**: spectral slope computed as the energy difference between the [0, 500] Hz and [500, 4000] Hz bands in dB with the energy band difference function of Praat.

Finally, concerning glottal flow parameters, the NAQ, the PSP, and the MDQ (see Section 2.3.2.3) are extracted.

4.2.3 *Speech segment selection and parameters extraction*

The acoustic analysis is conducted at the utterance level. However, the acoustic measures are extracted from vowels and then averaged to ensure their reliability (specially, for VoQ parameters computation, Choi et al., 2012), as they represent stable speech segments (Patel et al., 2010; Monzo et al., 2007; Drioli et al., 2003).

All acoustic parameters have been extracted using a Praat script specifically developed for this task, except for the glottal flow parameters (NAQ, PSP, and MDQ), which have been extracted using COVAREP (version 1.3.1) algorithms. The segmentation of the storytelling speech corpora into words, syllables and phonemes has been carried out with the EasyAlign tool. This automatic segmentation was manually corrected (if necessary) afterwards in order to dispose of reliable data (nearly 16,000 phonemes were revised). Finally, in order to balance the acoustic data across the expressive categories, only 30 out of the 53 neutral utterances collected after the annotation process (see Table 3) were randomly selected for the subsequent analyses.

4.2.4 *Checking Statistical tests assumptions*

Statistical tests assume certain assumptions regarding the data under analysis, e.g., whether variable distributions are normally distributed. Usually, when dealing with real data it is very difficult to fulfill all the assumptions, so several tests and methods exist to minimize deviations from certain assumptions.

The one-way MANOVA is a statistical test used to assess whether there are any differences between independent groups on more than one continuous dependent variable. Following the MANOVA section of the SPSS tutorial website[9], nine assumptions are considered to be satisfied. Below, these assumptions are described together with a comment regarding their impact on the data under analysis. In the end of this Section, the strategy to cope with deviations from the assumptions and the justification is discussed.

- **Assumption 1**: *All the dependent variables should be measured at the interval or ratio level (i.e., they are continuous).* — This assumption is fulfilled in the data under analysis, because the data is numerical and continuous.

- **Assumption 2**: *The independent variable should consist of two or more categorical, independent groups.* — The independent variable of the analysis is "Expressive Category", which consists of 8 groups that are categorical and independent.

- **Assumption 3**: *Independence of observations must exist, i.e., there is no relationship between the observations in each group or between the groups themselves.* — The data is measured at the sentence-level, thus, independence is assumed as the narrator may opt for changing the expressiveness (which reflects in the acoustic parameters) of each utterance, independently of the previous or the next utterance.

- **Assumption 4**: *Have and adequate sample size; the larger the sample size, the better.* — Although the sample size of the analysis of the storytelling speech corpora at hand is not large as it would be extremely time consuming to obtain, it is sufficient to perform the analysis.

- **Assumption 5**: *There are no univariate or multivariate outliers.* — Univariate outliers are extreme values within a distribution (i.e., a dependent variable) due to measurement variability or some error. In contrast, multivariate outliers are cases which have an unusual combination of scores on the dependent variables. A visual inspection of each dependent variable has been conducted in

[9] https://statistics.laerd.com/spss-tutorials/one-way-manova-using-spss-statistics.php

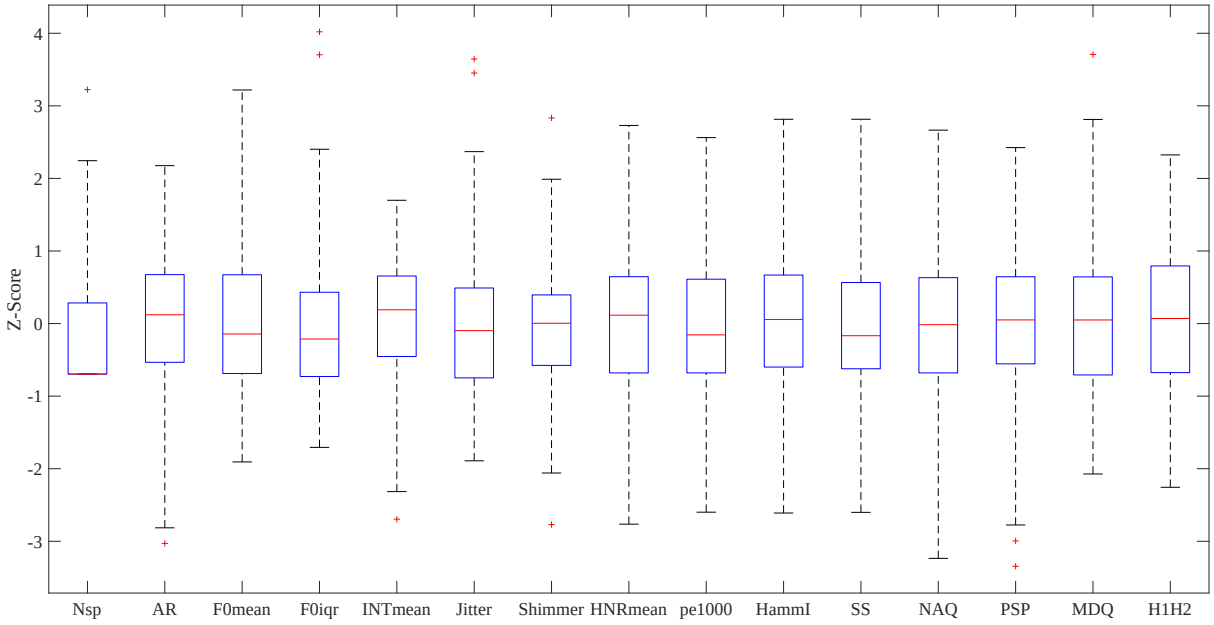

FIGURE 14: BOXPLOTS OF ALL THE PARAMETERS DISTRIBUTIONS (DEPENDENT VARIABLES) UNDER ANALYSIS.

order to spot potential outliers. As it can be observed in Fig. 14, there are few outliers[10] in the data (red crosses). Nonetheless, they are not produced by measurement errors, thus, they represent real phenomena. Taking this into account and the fact that MANOVA is robust to outliers if there are not many of them, the assumption can be considered satisfied to avoid reducing the data size. Furthermore, after computing the Mahalanobis distance (a method to spot multivariate outliers), only 3 from the total 200 cases (all 3 are Post-Character utterances) are found to be multivariate outliers ($p < 0.001$).

- **Assumption 6**: *There is multivariate normality.* — Assert if there is multivariate normality is very difficult, but confirming normality in each dependent variable separately it is considered as a good indicator of multivariate normality (Huberty and Petoskey, 2000). At first glance, all the distributions depicted in Fig. 14 look approximately normal with the exception of the NSP distribution. However, after applying a Lilliefors test to statistically evaluate normality, 7 parameters distributions do not pass the test: Nsp, $F0_{IQR}$, int_{mean}, jitter, shimmer, PE1000, and SS. However, ANOVA tests are pretty robust to violations of normality (Glass et al., 1972; Lix et al., 1996), even with unequal number of cases per group. For instance, Seo et al. (1995) showed robustness to deviations from normality with overall sample size of 40 (ten per group).

- **Assumption 7**: *There is a linear relationship between each pair of dependent variables for each group of the independent variable.* — If the variables are not linearly related, the power of the test is reduced as the linear combinations of dependent variables do not maximize the difference between independent variable's groups (Huberty and Petoskey, 2000). After observing a scatterplot matrix between the dependent variables, the linearity is only slightly violated in few cases.

- **Assumption 8**: *There is a homogeneity of variance-covariance matrices.* — This assumption can be tested using the Box's M test of equality of covariance. Although this test is very sensitive to normality and sample size and has been criticised (Huberty and Petoskey, 2000), a violation of the assumption should be considered in the present analysis since the significance value results

[10] In this case, outliers are defined as values outside the $Q3 \pm 1.5 * (Q3 - Q1)$ range, where Q3 and Q1 are the third and first quartiles, respectively.

in $p < 0.001$. However, on a preliminary visual inspection of sample variances across groups to evaluate robustness (dependent variables with largest to smallest variance ratio under 10:1 could be considered acceptable, Huberty and Petoskey, 2000), it can be observed that the main problem is present in the Nsp, $F0_{IQR}$, and int_{mean} parameters. Nevertheless, discarding them is avoided as they are of interest in the analysis to explore their role in storytelling. In the last paragraph of this Section, the solution to overcome this issue is detailed.

- **Assumption 9**: *There is no multicollinearity*. — Ideally, to perform a MANOVA, the dependent variables should correlate between each other in a moderate range (e.g., 0.2–0.6). If the correlation is too low it is not a big issue, as that would only indicate that it may be better to run separate one-way ANOVAs, which are conducted in subsequent analyses. Contrarily, a very high correlation (e.g., > 0.9) between dependent variables indicates the presence of collinearity, in the case of one occurrence, or multicollinearity if multiple occurrences of high correlations are present, which is very problematic in a MANOVA. In Table 4, it can be observed that the maximum Pearson correlation coefficient obtained is 0.684 (between SS and PE1000), so it can be concluded that there is no multicollinearity.

TABLE 4: CORRELATION MATRIX OF DEPENDENT VARIABLES.

	Nsp	AR	$F0_{mean}$	$F0_{IQR}$	int_{mean}	Jitter	Shimmer	HNR_{mean}	pe1000	HammI	SS	NAQ	PSP	MDQ	H1H2
Nsp	1	0.055	0.263	0.136	0.209	-0.034	-0.073	0.225	-0.194	0.064	-0.182	-0.033	0.181	0.178	-0.114
AR	0.055	1	0.238	0.09	0.415	-0.144	-0.016	0.202	0.01	0.041	-0.05	-0.04	0.092	0.017	-0.178
$F0_{mean}$	0.263	0.238	1	0.453	0.631	-0.286	-0.091	0.542	-0.104	0.003	-0.127	0.034	0.145	0.274	-0.255
$F0_{IQR}$	0.136	0.09	0.453	1	0.273	0.202	0.075	0.232	-0.01	-0.072	-0.121	-0.069	0.158	0.09	0.04
int_{mean}	0.209	0.415	0.631	0.273	1	-0.319	-0.215	0.392	-0.178	0.062	-0.104	-0.065	0.071	-0.016	-0.242
Jitter	-0.034	-0.144	-0.286	0.202	-0.319	1	0.192	-0.22	0.075	0.104	-0.097	-0.178	-0.184	-0.002	0.216
Shimmer	-0.073	-0.016	-0.091	0.075	-0.215	0.192	1	-0.254	0.201	-0.071	0.082	-0.119	-0.038	-0.115	-0.124
HNR_{mean}	0.225	0.202	0.542	0.232	0.392	-0.22	-0.254	1	-0.548	0.529	-0.671	0.435	0.247	0.427	0.118
pe1000	-0.194	0.01	-0.104	-0.01	-0.178	0.075	0.201	-0.548	1	-0.38	0.684	-0.277	-0.192	-0.302	-0.232
HammI	0.064	0.041	0.003	-0.072	0.062	0.104	-0.071	0.529	-0.38	1	-0.763	0.357	0.076	0.39	0.392
SS	-0.182	-0.05	-0.127	-0.121	-0.104	-0.097	0.082	-0.671	0.684	-0.763	1	-0.327	-0.213	-0.499	-0.348
NAQ	-0.033	-0.04	0.034	-0.069	-0.065	-0.178	-0.119	0.435	-0.277	0.357	-0.327	1	0.31	0.422	0.333
PSP	0.181	0.092	0.145	0.158	0.071	-0.184	-0.038	0.247	-0.192	0.076	-0.213	0.31	1	0.067	0.117
MDQ	0.178	0.017	0.274	0.09	-0.016	-0.002	-0.115	0.427	-0.302	0.39	-0.499	0.422	0.067	1	0.289
H1H2	-0.114	-0.178	-0.255	0.04	-0.242	0.216	-0.124	0.118	-0.232	0.392	-0.348	0.333	0.117	0.289	1

In conclusion, most of the assumptions are satisfied, but some of them are violated to different extents. Therefore, the MANOVA results are reported using the *Pillai's Trace* statistic (Pillai, 1955), which is recommended when there are violations of assumptions like the ones observed, as it is robust to such violations (Tabachnick and Fidell, 1983; Huberty and Petoskey, 2000).

4.3 ANALYSIS OF INDIRECT STORYTELLING SPEECH

In this Section, firstly the role that prosody and VOQ play in the different expressive categories of the Spanish storytelling corpus is explored through statistical and discriminant analyses. Next, it is investigated if relating emotions to indirect storytelling speech is appropriate by comparing emotions and expressive categories in a common acoustic space derived from MDS.

4.3.1 *Acoustic analysis*

The MANOVA revealed statistically significant results ($F(105, 1288) = 4.546$, $p < 0.001$). Posterior univariate analyses show that all parameters exhibit one or more statistical significant differences among

TABLE 5: NORMALIZED AVERAGED ACOUSTIC MEASURES OF THE STORYTELLING EXPRESSIVE CATEGORIES. NEU: NEUTRAL CATEGORY OF STORYTELLING, P-C: POST-CHARACTER, DES: DESCRIPTIVE, SUS: SUSPENSE, N/P: NEGATIVE/PASSIVE, N/A: NEGATIVE/ACTIVE, P/P: POSITIVE/PASSIVE, P/A: POSITIVE/PASSIVE.

Parameter	**Category**							
	P-C	**DES**	**NEU**	**N/P**	**N/A**	**P/P**	**P/A**	**SUS**
Nsp	-0.67	0.94	0.02	-0.12	0.11	0.28	0.51	-0.23
AR	-0.34	0.00	0.29	-0.02	0.15	0.14	0.57	-0.33
$F0_{mean}$	-0.83	0.72	0.42	-0.55	1.22	-0.49	1.24	-0.72
$F0_{IQR}$	-0.84	0.04	-0.13	0.72	1.04	-0.67	0.22	-0.35
int_{mean}	-0.71	0.38	0.51	-0.36	0.72	-0.28	0.99	-0.43
Jitter	0.16	-0.24	-0.35	0.68	0.09	-0.17	-0.69	-0.27
Shimmer	0.47	-0.23	-0.11	0.19	0.21	-0.22	0.07	-0.93
HNR_{mean}	-0.99	0.67	0.50	-0.16	0.17	0.57	0.62	-0.25
pe1000	0.71	-0.21	-0.26	-0.02	0.24	-0.53	-0.34	-0.44
HammI	-0.57	0.18	0.25	0.08	-0.05	0.64	-0.34	0.26
SS	0.94	-0.39	-0.14	-0.14	-0.06	-0.67	0.09	-0.48
NAQ	-0.29	0.39	0.29	-0.04	-0.47	0.69	-0.46	0.14
PSP	-0.68	0.37	0.30	-0.07	0.21	0.43	-0.06	0.10
MDQ	-0.63	0.47	0.01	-0.05	0.30	0.41	0.00	0.16
H1H2	-0.33	0.00	-0.32	0.38	-0.08	0.34	-0.79	0.79

categories. Further inspection of the Tukey's HSD post-hoc tests was conducted on all relevant parameters. The most relevant parameters according to the criteria explained in Section 4.2.1 were $F0_{mean}$, int_{mean}, SS, H1H2, $F0_{IQR}$, jitter and HNR_{mean} (see Section 4.3.2 for more details). The corresponding normalized averaged results together with their standard deviations can be observed in Table 5.

Concerning $F0_{mean}$, post-character, negative/passive, positive/passive and suspense utterances were expressed with significantly lower $F0_{mean}$ than the rest of categories, but with no statistical significance among them. On the other hand, descriptive, negative/active and positive/active categories show significantly higher values of $F0_{mean}$ with respect to the rest of categories with the exception of descriptive utterances, which show similar $F0_{mean}$ to neutral utterances. F0 patterns are slightly different in $F0_{IQR}$, where the negative/passive category shows a high value together with its active counterpart and the positive/active category. Post-character, positive/passive and suspense categories show low and similar values, although the post-character utterances are the only ones conveyed with significantly lower $F0_{IQR}$ than the other 5 categories. Regarding int_{mean}, post-character, negative/passive and suspense categories show the lowest values, significantly differing from other categories. Positive/passive utterances are similar in intensity level to all categories with the exception of both active categories. The remaining categories are similar in terms of int_{mean}, although active categories are expressed by the narrator with the highest intensity levels. In what concerns VoQ parameters, post-character utterances show the largest value of SS, significantly larger than the rest. This means that post-character utterances were expressed with a tenser voice, i.e, a flatter spectral slope, whereas the rest where expressed with a similar tension. Nonetheless, positive/passive and suspense categories show quite low values (although not significantly different), entailing a breathier voice that could be sensed at a perceptual level. The H1H2 parameter does not show

TABLE 6: WILKS' LAMBDA VALUES FOR EACH PARAMETER OBTAINED FROM THE ANALYSIS OF STORYTELLING SPEECH, LEFT-TO-RIGHT ORDERED FROM THE LOWEST (BEST DISCRIMINATION CAPABILITY) TO THE HIGHEST (WORST DISCRIMINATION CAPABILITY).

Parameter	$F0_{mean}$	$F0_{IQR}$	HNR_{mean}	int_{mean}	SS	Nsp	H1H2	pe1000	Shimmer	Jitter	PSP	MDQ	HammI	NAQ	AR
Wilks' Lambda	0.376	0.594	0.648	0.671	0.745	0.771	0.823	0.829	0.841	0.845	0.857	0.870	0.877	0.879	0.927

many statistically significant differences, although it shows four statistically significant differences in the suspense category with respect to post-character, neutral, negative/active and positive/active categories. Jitter values are higher in negative and post-character categories and lower in the rest. However, not many significant differences are observed in the pairwise comparisons. Finally, post-character utterances show the significantly lowest HNR_{mean}. Moreover, although suspense and both negative categories show lower values than positive, neutral and descriptive categories, the contrasts between them are not significant in all cases.

4.3.2 *Discriminant analysis*

The conducted discriminant analysis showed four (out of seven) significant ($p < 0.05$) canonical discriminant functions explaining a total of 95.4% of the variance (Function 1: $Wilks'\ \Lambda = 0.071$, $\chi^2(105) = 495.039$, $p < 0.0001$; Function 2: $Wilks'\ \Lambda = 0.227$, $\chi^2(84) = 277.866$, $p < 0.0001$; Function 3: $Wilks'\ \Lambda = 0.417$, $\chi^2(65) = 163.978$, $p < 0.0001$; Function 4: $Wilks'\ \Lambda = 0.657$, $\chi^2(48) = 78.721$, $p = 0.003$). The first canonical function explains 53.9% of the variance and is correlated positively with two prosodic features: $F0_{mean}$ ($r = 0.86$) and int_{mean} ($r = 0.46$). The second function involves spectral measures accounting for 20.6% of the variance and correlates positively with SS ($r = 0.50$) and negatively with H1H2 ($r = -0.46$). The third function explains 14.2% of the variance and correlates positively with $F0_{IQR}$ ($r = 0.69$) and jitter ($r = 0.40$). The fourth function shows a low 6.7% of variance and only correlates with HNR_{mean} ($r = 0.47$). The Wilk's lambdas values obtained for each parameter can be observed in Table 6.

The results of the LDA classification are shown in Table 7. When considering all the acoustic features listed in Section 6.3.2.2, post-character, suspense and negative/passive show the highest F1 scores, followed by neutral and negative/active categories. Ultimately, descriptive and positive (both active and passive) categories present the lowest F1 values. Nonetheless, since the classification task contemplates eight categories, all these F1 scores are well above the chance threshold (12.5% of correctly classified cases). Probably, the fewer amount of samples in the positive categories (see Table 3) has affected their results (similarly to what happened to Alm et al., 2005). Another interesting observation is that the macro-averaged F1 score (F_1^M) when considering only prosodic or VoQ parameters resulted in 0.363 and 0.384, respectively (see Table 7). However, combining prosodic and VoQ features leads to the highest F_1^M when classifying the expressive categories, which corresponds to 31% and 38.6% of relative improvements when compared to considering only prosody and VoQ, respectively. This result highlights that both prosody and VoQ are important for discriminating the storytelling expressive categories under analysis. Finally, it is worth remarking that VoQ appears to be very important in the suspenseful and positive/active speech, being the LDA unable to classify this category if only prosodic features are considered. Contrarily, prosody is crucial for discriminating positive/passive utterances.

TABLE 7: LINEAR DISCRIMINANT ANALYSIS F1 SCORES FOR EACH STORYTELLING EXPRESSIVE CATEGORY. P: PROSODY; VOQ: VOICE QUALITY.

Features	**P-C**	**DES**	**NEU**	**N/P**	**N/A**	**P/P**	**P/A**	**SUS**	F_1^M
P	0.69	0.36	0.49	0.70	0.44	0.12	0.00	0.00	0.363
VoQ	0.64	0.32	0.46	0.34	0.38	0.00	0.37	0.55	0.384
P+VoQ	0.76	0.34	0.55	0.63	0.52	0.24	0.33	0.64	0.503

TABLE 8: NUMBER OF UTTERANCES PER CORPUS OF THE LS-URL LAICOM-UAB EMOTIONAL SPEECH CORPORA.

Category	**# Utterances**
NEU	2349
HAP	915
SEN	841
AGG	1048
SAD	1000
TOTAL	6153

4.3.3 *Storytelling expressive categories vs. Emotions*

In this section, the indirect storytelling speaking style is compared to a Spanish speech corpus of emotions. The emotional corpus considered in the present study was developed by researchers from LS-URL and the Laboratory of Instrumental Analysis-Universitat Autònoma de Barcelona (LAICOM-UAB). The main features and the analyses are described in the following paragraph. The reader is referred to the works of Monzo et al. (2007) and Iriondo et al. (2009) for further details on the corpora, and to the Appendix B for a more detailed analysis of the corpora considering the same analysis framework of the storytelling expressive categories.

The corpus was built with texts whose semantic content helped the professional female speaker to elicit the desired emotion. To that effect, a textual database of advertisements extracted from newspapers and magazines by LAICOM-UAB researchers was used. Five topics associated to a specific expressive speaking style (see, Montoya, 1998) were recorded at LS-URL: technologies with neutral style (NEU); Education with happiness (HAP); Cosmetics with sensuality (SEN); Automobiles with aggressiveness (AGR); Travel with sadness (SAD). Table 8 shows the number of utterances per category. It is worth noting that the neutral category was expanded with more utterances with respect to the original corpus to feed the LS-URL general purpose TTS synthesis system.

In order to compare storytelling expressive categories and emotions, a MDS considering the same acoustic parameters detailed in Section 6.3.2.2 is conducted to represent them in a common acoustic space (Kruskal and Wish, 1978). MDS is a useful technique to visualize the distribution of several categories in a common space, and has been used in previous related works (Wang et al., 2014; Jovičić et al., 2006; Truong and van Leeuwen, 2007). This common space can be observed in Fig. 15, where closer points represent similar categories, whereas points that are far apart represent the opposite. The stress obtained for the three-dimensional solution was 0.03, which represents a good fit (Kruskal, 1964). This stress value represents how well (values near 0 indicate a good fit) a particular configuration repro-

duces the distance matrix (Kruskal, 1964). The MDS solution depicted in Fig. 15 was obtained from a distance matrix that contained dissimilarities among the storytelling expressive categories computed using euclidean distances. The 3D solution was selected over the 2D solution for two main reasons: its low stress value and the more detailed visualization.

From the visual inspection of Fig. 15, it can be derived that dimension 1 correlates negatively with activation, as active emotions (aggressive and happy) and active storytelling categories (negative/active and positive/active) are placed in the negative side of the axis. In contrast, passive emotions (sad and sensual) and *more passive* storytelling categories (negative/passive, positive/passive, suspense and post-character) are located in the positive side of the axis. In addition, neutral categories and descriptive utterances are placed around zero. On the other hand, dimensions 2 and 3 do not clearly correspond to any specific dimension (e.g., valence).

Regarding similarities and differences among emotions and storytelling categories, several interesting conclusions can be drawn from Fig. 15. Firstly, it can be observed that the storytelling neutral category differs from the neutral category of the emotional corpus to some extent, specially, in the third dimension. This result shows that the neutral storytelling category for the corpus under analysis entails a slight difference (apparently, a subtle increase in expressiveness) with respect to the neutral category of the emotional corpus. Furthermore, it can be observed that happy and aggressive categories are placed relatively close to positive/active and negative/active categories in the first dimension. However, all four categories are quite differentiated in the second and third dimensions. This finding is an evidence that supports the considered hypothesis that relating the indirect discourse of storytelling to full-blown emotions is inappropriate. Moreover, the positive side of dimension 1 contains *more passive* categories from which several conclusions can also be drawn. In relation to passive affective categories, similarly to their respective active counterparts, they are closer from each other in dimension 1. In this case, the gap in dimensions 2 and 3 is greater than the one present in active categories. However, there is a relationship in the sense that positive categories are represented with lower values in dimensions 2 and 3 as opposite to negative categories. As it was expected, sadness is closer to the negative/passive category than the positive/passive. Suspense is placed near the sensual and sadness categories when observed in the plane of dimensions 1 and 2, which was also expected as suspense situations entail a soft phonation. Finally, the post-character category is quite isolated in the top of dimension 2 being clearly uncorrelated with emotional categories, although their utterances are perceptually closer to the neutral category of the emotional corpus.

From the considered data it can be concluded that the storytelling expressive categories differ in many ways from prototypical basic emotions when represented together in a common acoustic space. The main difference appears to be the subtler use of expressiveness, as storytelling categories are clustered closer from each other than emotional categories are. As a visual representation, one could imagine a sphere placed in the 3D-MDS of Fig. 15 where emotions are located in the surface of that sphere (with the exception of the neutral category of the emotional corpus) and storytelling categories are located inside, within another sphere of lower radius. This claim is also supported by the fact that LDA classification results obtained from the emotional corpus ($F_1^M = 0.926$) are much higher than the ones obtained for the storytelling corpus at hand ($F_1^M = 0.503$), indicating that the emotional categories are characterized by more extreme and differentiated values than storytelling expressive categories.

4.4 DISCUSSION

In spite of the amount of analysed data is not negligible, it is worth to note that, at this stage, it is a proof of concept work, thus, at this point the obtained results cannot be generalized to other storytellers or languages. However, it is also worth noting that a very expressive and engaging storyteller who performs the story with no hesitations has been selected, since the objective was to analyse storytelling speech performed by a *very good* storyteller. In addition, the evaluation of the approach has been conducted in

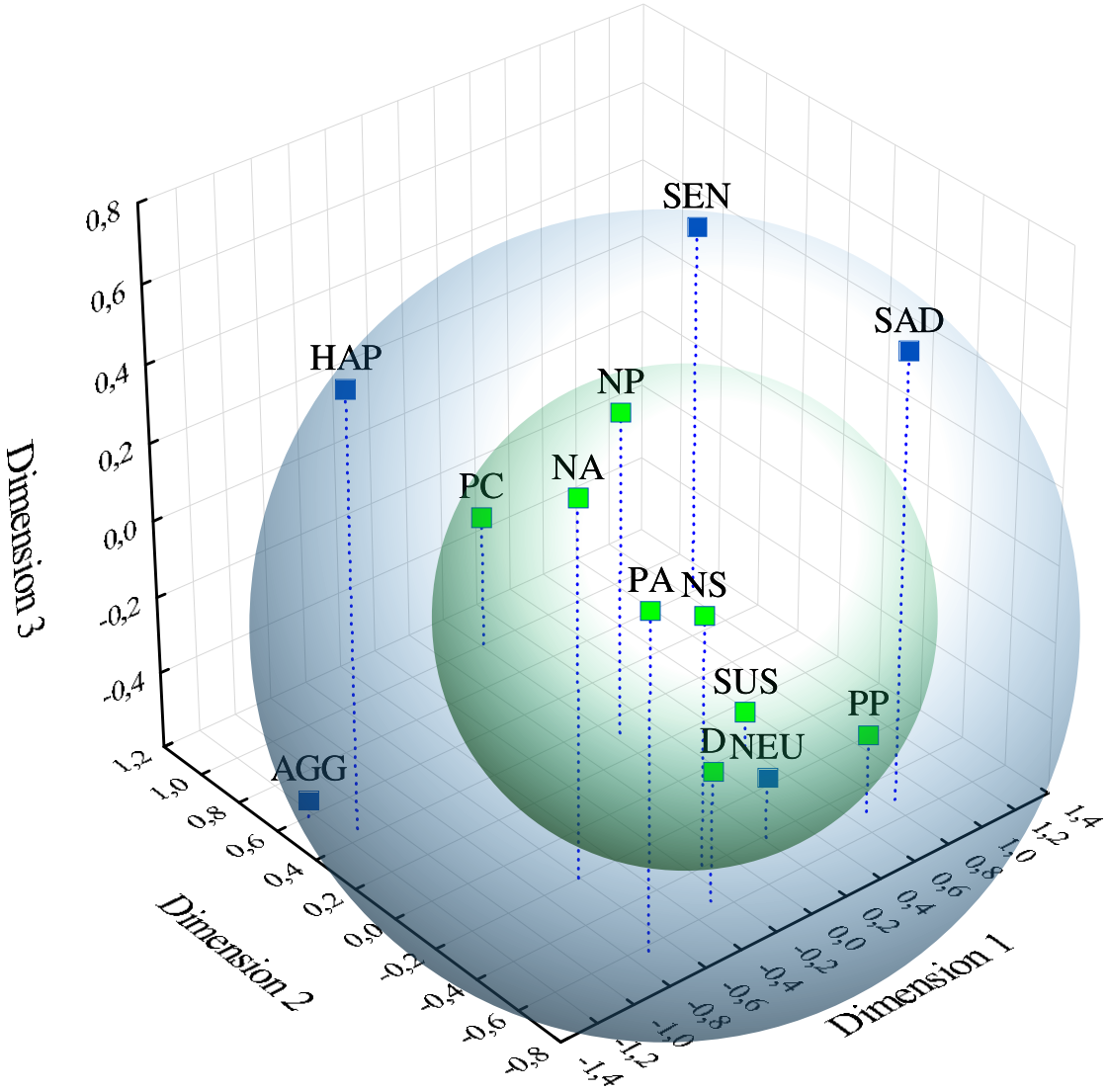

FIGURE 15: STORYTELLING AND EMOTION CATEGORIES DISTRIBUTION IN THE 3D COMMON ACOUSTIC SPACE DERIVED FROM MULTI-DIMENSIONAL SCALING USING ALL CONSIDERED ACOUSTIC FEATURES. THE SPHERES ARE ADDED FOR VISUAL PURPOSES. NS: NEUTRAL CATEGORY OF STORYTELLING, PC: POST-CHARACTER, D: DESCRIPTIVE, SUS: SUSPENSE, NP: NEGATIVE/PASSIVE, NA: NEGATIVE/ACTIVE, PP: POSITIVE/PASSIVE, PA: POSITIVE/ACTIVE, NEU: NEUTRAL CATEGORY OF THE EMOTIONAL CORPUS, HAP: HAPPY, SAD: SADNESS, AGG: AGGRESSIVE, SEN: SENSUAL.

a controlled environment (a single narrative text interpreted by a very expressive narrator together with a manually annotated corpus by experts) in order to validate the hypotheses reliably. In this sense, it is worth noting that the obtained results from the valence/activation categories are consistent with prior findings for several languages (see, Schröder, 2004, and references therein), i.e., a higher activation entails higher pitch, intensity, and tempo, and flatter spectral slope in general while valence lacks clear acoustic correlates.

In spite of the fact that some of the results could be implicitly generalizable to other languages, it is still necessary to confirm this initial results by finding explicit forms of evidence by means of subsequent cross-narrator and cross-language analyses (conducted in Part II of the thesis). At this point, the hypothesis was that at least such forms of evidence will be found among expressive narrators aimed to young

audience, independently of language or gender. It seems quite plausible that if two storytellers with similar style read the same tale or story there could be a substantial relationship in the expressive categories they choose for their interpretation, beyond their personal touches. For instance, some differences might be manifested in some utterances, but for an uttered sad sentence such as "the parents of the princess died years ago" is more than possible that both narrators would select a "negative" tone of voice.

4.5 CONCLUSIONS OF PART I

In this first part of the thesis, the subtle expressive nuances of indirect storytelling speech have been analysed. To that effect, a text-level and perceptual annotation methodology has been proposed to deal with storytelling speech annotation at the sentence level based on storytelling discourse modes besides introducing sub-modes denoted as expressive categories: neutral, descriptive, post-character, suspense, negative/passive, negative/active, positive/passive, and positive/active. This annotation methodology has been applied to a Spanish storytelling speech corpus as a proof of concept to evaluate its viability. This annotation process has allowed the classification and analysis of 84.8% of the utterances in the corpus (the remaining utterances contained very specific expressive cues out of the scope of this work). The outcomes of the statistical and discriminant analyses have proved that VoQ is as important as prosody for the discrimination among storytelling expressive categories, being VoQ specially significant in suspenseful and positive/active utterances and prosody in positive/passive utterances. In particular, $F0_{mean}$, int_{mean}, spectral slope, H1H2, $F0_{IQR}$, jitter, and HNR_{mean} have proved as relevant parameters in the discrimination among storytelling categories.

Furthermore, the appropriateness of relating basic emotions to indirect storytelling speech has been studied, as storytelling entails subtler speech nuances. To that effect, both storytelling and emotional categories have been represented in a 3D common acoustic space via MDS using the same acoustic parametrization (both prosodic and VoQ features). Results show that emotions are placed in more extreme regions of such space, while the position of the indirect speech categories supports the subtler expressiveness they contain. In line with this result, the macro-averaged F1 obtained for the classification of storytelling expressive categories is much lower ($F_1^M = 0.503$) than the one obtained from the analysis of the emotional corpus ($F_1^M = 0.926$).

In the next part of the thesis, more speech corpora is analysed via cross-language and cross-narrator analyses (the same story) following the same methodology in order to evaluate to what extent the obtained results are generalizable.

Part II

THE ROLE OF PROSODY AND VOICE QUALITY IN INDIRECT STORYTELLING SPEECH: A CROSS-LANGUAGE PERSPECTIVE

In this part of the thesis, it is studied to what extent the results obtained in the previous part (for a Spanish narrator) are generalizable to other narrators telling the same story in English, French, and German. Specifically, the same indirect speech is again analysed in terms of prosody and Voice Quality through statistical and discriminant analyses, after classifying the sentences of the story into the previously introduced storytelling expressive categories.

5

A CROSS-LANGUAGE PERSPECTIVE

The works related to the analysis of storytelling speech that have been reviewed along the thesis have only focused on a particular language, with the exception of Sarkar et al. (2014), who considered three Indian languages but without including a comparison among those languages. Thus, up to our knowledge, there are no cross-language studies specifically focused on the acoustic analysis of storytelling speech, although certain emotions and speaking styles have been investigated with a cross-language perspective. For instance, emotions have displayed many similarities in their acoustic-perceptual properties across different languages (Pell et al., 2009b; Liu and Pell, 2014), whereas certain speaking styles (e.g., polite and informal speech) have also shown several acoustic *tendencies* among speakers of different languages (Grawunder and Winter, 2010). In this sense, it is particularly interesting to study to what extent cross-language (and cross-narrator) acoustic patterns of storytelling do exist.

In the following Section, an overview of how other works have tackled the cross-language analysis of some expressive styles is given.

5.1 INTRODUCTION TO CROSS-LANGUAGE ANALYSIS

Concerning cross-language studies focused on emotions, Pell et al. (2009b) explored perceptually and acoustically six emotions (anger, disgust, fear, sadness, happiness, and pleasant surprise) together with a neutral reference in four different languages: English, German, Hindi, and Arabic. Their results highlighted that the expression of the emotions under analysis (via meaningless utterances) showed global tendencies for the considered acoustic and perceptual attributes across those languages, independent of linguistic similarity. A later work by Liu and Pell (2014) included Mandarin Chinese in the analyses, concluding that both the perceptual and acoustic characteristics were highly similar to those observed by Pell et al. (2009b). These works suggest the existence of some general pattern in the oral communication of emotions. However, other works have also evidenced language-specific patterns for different language communities. Continuous read speech from numbers and short passages of real-life topics was used by Andreeva et al. (2014) to compare two Germanic (German and English) and two Slavic (Bulgarian and Polish) languages. The results showed that speakers of Germanic languages use lower pitch maxima, narrower pitch span, and less variable pitch than Bulgarian and Polish speakers. A later work by the authors including linguistically based pitch range measures showed similar conclusions (Andreeva et al., 2015).

Some speaking styles such as infant-directed speech and polite/informal speech have also been analysed from a cross-language perspective. Infant-directed speech was prosodically analysed using fundamental frequency and speech tempo parameters in French, Italian, German, Japanese, British English, and American English by Fernald et al. (1989). The results revealed consistent prosodic patterns (higher frequency values, shorter utterances and longer pauses in infant-directed speech with respect to adult-directed speech) across languages beyond some language-specific variations (e.g., American English showed the most extreme prosodic values). Grawunder and Winter (2010) found several cross-language tendencies in polite and informal speech (voice-mail speech message to a professor and a friend, respectively) between Korean and German speakers. For instance, polite speech showed more filled pauses, a breathier phonation, and lower values of fundamental frequency, intensity and perturbation measures than informal speech. Yaeger-Dror (2002) tackled the study of prosodic prominence and contours on negatives (e.g., *not* for English, *pas* for French) in various interactive and non-interactive registers (informative, memoirs, literary readings, interviews, etc.) of Continental French. These results

were compared with those from American English negatives in similar situational contexts, concluding that polite situations entailed a lower pitch prominence than the confrontational situations in both cultures (i.e., a general characteristic of this speaking style across these languages). However, some language-specific characteristics were also observed, e.g., French informative negatives were likely to be more prominent than the US counterparts. Furthermore, vocalic hesitations were prosodically analysed from speech extracted from several national radio and TV broadcast channels in terms of duration, fundamental frequency, and formant values by Vasilescu and Adda-Decker (2007) in order to seek universal vs. language-specific characteristics in French, American English, and European Spanish. Their results showed that some cross-language characteristics like higher durations and lower fundamental frequency than regular speech are general patterns for vocal hesitations. Nevertheless, the analyses of timbre quality showed that vocalic hesitations are realized differently across languages.

5.2 STORYTELLING EXPRESSIVE CATEGORIES ANNOTATION

In order to repeat the same manual annotation methodology of Part I, it would be ideal to have native expert annotators for each language at hand, specially for the identification of the perception-dependent categories. Unfortunately, this has not been the case. In order to address this limitation, the original process has been adapted for the English, German, and French corpora by conducting several perceptual tests (see Fig. 16). These tests were taken by 18 Spanish users, considering the Spanish version of the story as reference. This approach is based on the fact that, although listeners perform best when listening to speakers of their native language, they perform well at perceptually identifying neutral and expressive categories when produced by speakers of a foreign language (Pell et al., 2009a; Scherer et al., 2001; Thompson and Balkwill, 2006; Van Bezooijen et al., 1983). Although Scherer et al. (2001) found that European speakers had more difficulties in recognizing a non-European language (Malay in that case), Pell et al. (2009a) observed that similarity among linguistic communities did not seem a relevant factor. More specifically, in the study of Pell et al. (2009a) native Spanish recognized emotions and a neutral category expressed in English, German, and Arabic between 3 and 4 time chance level, but the subjects pointed out that it was way more difficult the test is Arabic. In our case, the languages are from similar linguistic communities close to each other, so we rely on the results from both works.

Nevertheless, these cross-language tests are only conducted on the neutral and suspense categories, since applying the same approach to the annotation of affective categories was discarded. This design was taken after observing the inherent difficulty of asking to non-expert listeners to classify utterances on a valence/activation scheme through preliminary informal tests. To address this issue, a hybrid approach has been considered: an expert annotator (advanced English learner and beginner French and German learner) together with a clustering stage for the annotation of the storytelling affective categories (see Section 5.2.3). As activation is relatively feasible to detect via acoustic parameters (Schuller, 2011; Nicolaou et al., 2011), it is assumed that the clustering stage would be capable of discerning the activation of the affective utterances in a quite precise way. The rationale behind this hybrid approach is that, since neither humans nor machines can be 100% reliable in this task besides having no ground truth, the agreement may filter out unclear instances.

For the cross-language analysis of storytelling speech conducted in this part of the thesis, four audiobooks are considered, where the same story (the one considered in Part I) is interpreted by four native professional male storytellers in four European languages: Spanish, English, French and German, being 33, 42, 57 and 64 years old respectively. Up to our knowledge, no guidance regarding the way to convey the story was delivered to the storytellers, i.e., they only took the text into account. Each audiobook contains approximately 20 minutes of indirect storytelling speech, composed of 263 sentence-level utterances. As the corpora contain the same text but in a different language, the annotation of text-dependent categories (see Section 4.1.1) is the same for English, French, and German as for the Spanish narrator.

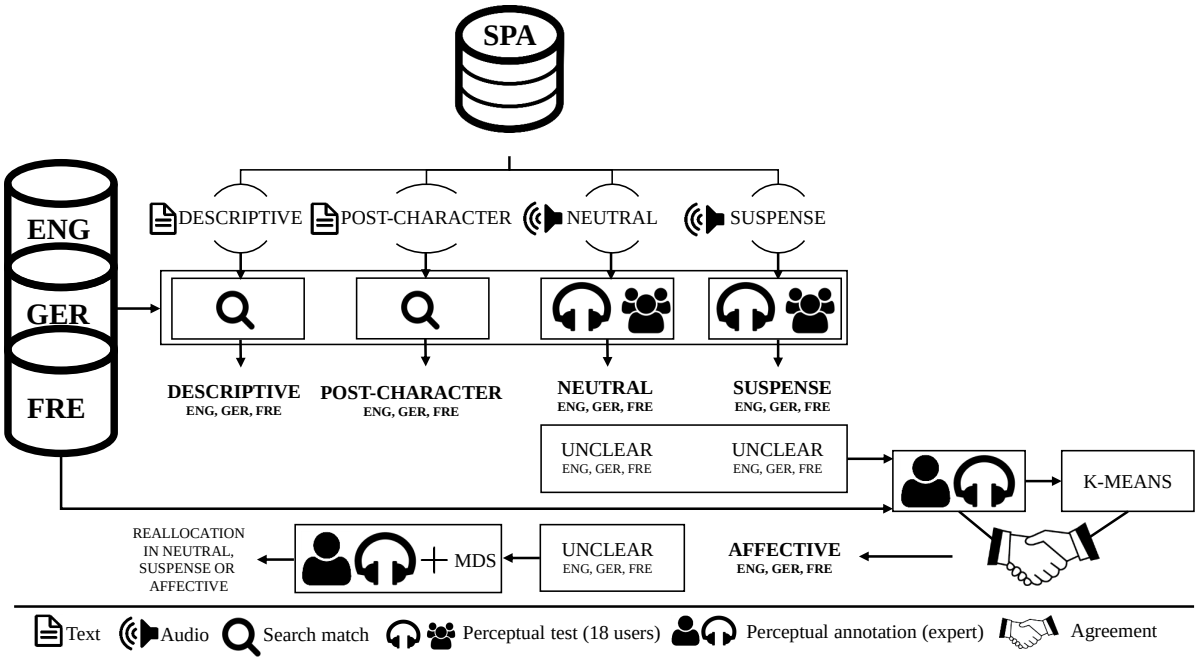

FIGURE 16: DIAGRAM SHOWING AN OVERVIEW OF THE ANNOTATION PROCESS FOLLOWED IN THE CROSS-LANGUAGE SCENARIO.

However, the annotation of the perception-dependent categories has been conducted for English, French and German languages and it is described in the following Sections.

5.2.1 *Neutral category annotation*

The neutral category present in storytelling speech can be used as a baseline for this kind of studies where relative differences of expressive categories are highlighted (Braunschweiler and Buchholz, 2011). However, the neutrality of the utterances has to be guaranteed in order to obtain representative neutral corpora. As the Spanish corpus is taken as reference to annotate the English, French, and German versions of the story, this process can be understood, at the same time, as the first analysis of cross-language similarities. To that effect, three different tests are performed confronting the neutral utterances of the Spanish narrator against the other three narrators (one test per language) using the online platform Testing platfoRm for mUltimedia Evaluation (TRUE) created by Planet et al. (2008). 18 native Spanish speakers (14 males, 4 females; mean age: 33±8.6) were recruited to take the tests. However, some of them are learners of several of the languages under analysis. Concretely, 73%, 23%, and 12% are English, French, and German learners, respectively.

The corresponding Spanish neutral audio was presented to the evaluator as reference together with the utterance to be evaluated (the same sentence uttered in the other language). The evaluators could listen to both audio signals as many times as they wanted before answering the question *"The expressiveness of the audio under evaluation compared to the expressiveness of the reference audio is:"*, by choosing among three possible answers: *"Higher"*, *"Roughly the same"*, *"Lower"*. 30 Spanish neutral utterances from the original corpus of 53 sentences (see Section 4.1.2.1) were randomly selected, in order to avoid user fatigue while maintaining balanced speech corpora for the subsequent analyses. Since more than two raters took the test and were not forced or led in any way to assign a certain number of cases to each response, the free-marginal Kappa was computed to measure the inter-rater agreement (Randolph, 2005):

$$\kappa_{free} = \frac{\left[\frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn\right)\right] - \left(\frac{1}{k}\right)}{1 - \left(\frac{1}{k}\right)} \quad (1)$$

where N is the number of cases, n is the number of raters, and k is the number of rating categories. This method, derived from the Fleiss' fixed-marginal multirater Kappa (Fleiss, 1971), avoids the prevalence and bias paradoxes of the fixed-marginal solution (Brennan and Prediger, 1981). The obtained free-marginal Kappa values were $\kappa_{free} = 0.78$, $\kappa_{free} = 0.80$, and $\kappa_{free} = 0.81$ for the English, French, and German tests, respectively. At this level of κ_{free}, the agreement is usually deemed as "substantial" (Landis and Koch, 1977). Finally, an exemplar was defined as an utterance with proportion of agreement per item greater than 0.61 (Landis and Koch, 1977), showing substantial agreement on choosing the option *"Roughly the same"*, using the following equation (Fleiss, 1971):

$$P_i = \frac{1}{n\,(n-1)}\left(\sum_{j=1}^{k} n_{ij}^2 - n\right) \quad (2)$$

where P_i is the proportion of agreement per item, n is the number of raters, k is the number of rating categories and n_{ij} is the number of raters who assigned the ith utterance under evaluation to the jth category. As a result, 29, 28, and 28 neutral utterances of English, French and German narrators, respectively, were considered for the subsequent analyses. The five utterances not included were left aside for further reallocation (see Section 5.2.4).

It is worth noting that the averaged value across tests of proportion of category assignment resulted in 0.93 for the *"Roughly the same"* category (Fleiss, 1971):

$$p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij} \quad (3)$$

where p_j is the proportion of all assignments to the jth category (in this case, *"Roughly the same"*), n_{ij} is the number of raters who assigned the ith utterance under evaluation to the jth category. Thus, this result together with the substantial values of κ_{free} are a first encounter of cross-language similarities in storytelling speech, since the four narrators used a neutral expressiveness for most of the sentences evaluated perceptually.

5.2.2 *Suspense category annotation*

Following the same approach described in Section 5.2.1, the 21 suspense utterances identified in the Spanish version of the story (see Section 4.1.2.2) are used as reference. The suspense expressiveness generated by the Spanish narrator is confronted against the rest of languages. In this case, the question posed to the 18 users was *"The feeling of suspense produced by the audio under evaluation compared to the reference audio is:"*, being asked to choose among the same pool of answers as in the neutral tests (*"Higher"*, *"Roughly the same"*, *"Lower"*). After conducting the tests, the free-marginal Kappa values were computed resulting in $\kappa_{free} = 0.77$, $\kappa_{free} = 0.76$, and $\kappa_{free} = 0.79$ for the English, French, and German tests, respectively. Thus, similarly to the neutral tests, a substantial agreement is achieved. A representative suspense exemplar was defined according to the following criteria: utterance with moderate or substantial agreement per item on *"Roughly the same"* or *"Higher"* responses. The rationale behind this approach is that utterances tagged with higher suspense level by the user (with respect to the corresponding Spanish version) can also be regarded as suspenseful sentences. As a result, 16, 18 and 16 utterances were retained for the English, French and German versions of the story, respectively. The remaining 13 utterances not considered as suspenseful were collected for the subsequent reallocation process described in Section 5.2.4. It is worth noting that, although different suspense levels may

TABLE 9: RESULTS OF THE AFFECTIVE ANNOTATION BY LANGUAGE.

Language	Case	# Utterances			
		N/P	N/A	P/P	P/A
English	Expert	35	41	10	18
	k-means	59	17	17	11
	Agreement	33	15	9	10
German	Expert	44	29	13	13
	k-means	37	36	14	12
	Agreement	27	19	8	7
French	Expert	30	26	8	6
	k-means	46	10	11	3
	Agreement	28	8	7	2

be present in storytelling (Cheong and Young, 2006; Theune et al., 2006), this investigation is left for future works.

Finally, it is also remarkable that the averaged value across tests of proportion of *"Roughly the same"* assignments resulted in 0.90, a very similar value to the one obtained from the neutral tests (i.e., 0.93). This fact is another evidence of cross-language similarities in storytelling speech at a perceptual level.

5.2.3 *Affective categories annotation*

The annotation of affective categories is divided in two stages. Firstly, the expert annotator labels those utterances that can be considered negative/passive, negative/active, positive/passive, or positive/active, leaving aside those that can not be included within any affective category (see Section 5.2.4 for further details about their annotation). Then, the clustering stage is executed which, maintaining the positive and negative labels of the expert annotator, only adds activation labels (passive and active). To that effect, k-means (computed using the SPSS software, IBM Corp., 2013) is taken into account because of three reasons (Jain et al., 1999): (1) its simplicity, (2) the number of clusters is known a priori ($k = 2$, i.e., active and passive), and (3) the data size is not very large. Finally, those utterances where agreement is obtained in terms of activation between the expert and the k-means output are retained.

The annotation outcome for each language is shown in Table 9. A fair relationship between expert's annotations and k-means' assignments is obtained (Landis and Koch, 1977). Concretely, a 64.4% ($\kappa = 0.330$), a 61.6% ($\kappa = 0.229$), and a 64.3% ($\kappa = 0.250$) for the English, German, and French corpora, respectively.

5.2.4 *Reallocation of unclear utterances*

At this stage of the annotation process, a greater number of utterances is present in the unclear category with respect to what happened during the annotation of the Spanish version, as a logical consequence of not having native expert annotators. Precisely, in contrast to the remaining 44 utterances at this point of the annotation of the Spanish version, 87, 94, and 108 utterances remain to be classified within the English, German, and French versions of the story, respectively. For the reallocation of these unclear utterances the same methodology applied in Section 4.1.2.4 is used for each language. After the reallocation process, some utterances were discarded for the subsequent acoustic analyses for similar reasons

TABLE 10: GATHERED UTTERANCES FROM THE SPEECH CORPORA AFTER THE ANNOTATION PROCESS FOR EACH LANGUAGE. BETWEEN PARENTHESIS THERE IS THE NUMBER OF NOT CONSIDERED NEUTRAL UTTERANCES.

Category	**# Utterances**			
	Spanish	**English**	**German**	**French**
Neutral	30 (+23)	29 (+8)	28 (+30)	28 (+36)
Post-Character	40	40	40	40
Descriptive	24	24	24	24
Negative/Passive	36	43	32	33
Negative/Active	23	24	30	14
Positive/Passive	13	12	11	10
Positive/Active	13	12	10	4
Suspense	21	28	20	36
Other	40	43	38	38
Considered / Total	200 / 263	212 / 263	189 / 263	195 / 263

already observed in the Spanish version of the story, e.g., these utterances contained slight imitations of a character, yawns, laughter, etc. (those utterances are labelled as 'Other').

The final number of gathered utterances for each category per language is described in Table 10. As it can be observed, the categories are quite balanced across languages with some exceptions. For instance, the French narrator expressed few positive/active sentences, whereas he used a neutral and suspenseful style in more utterances compared to the rest. Moreover, the English narrator used a neutral expressiveness in 37 utterances while his Spanish, German, and French counterparts used it in 53, 58, and 64 utterances, respectively. In order to obtain reliable results, only the validated neutral utterances of the perceptual are considered for the cross-language analyses to have a similar amount of samples in all languages (see Table 10). The low number of samples obtained for the positive utterances may entail some difficulties in the subsequent analyses.

Finally, it is to note that the obtained percentage of satisfactorily classified utterances after applying the annotation methodology on the storytelling corpora is similar across languages, validating to some extent the adaptation of the original methodology. Concretely 84.8%, 83.7%, 83.3%, and 87.8% utterances from the Spanish, English, German, and French corpora, respectively, have been annotated within a storytelling expressive category.

5.3 CROSS-LANGUAGE ACOUSTIC ANALYSIS FRAMEWORK

In this Section, the methodology for the cross-language acoustic analysis of the storytelling corpora under study is explained together with the parameter extraction process.

5.3.1 *Cross-language acoustic analysis methodology and parameters extraction*

The cross-language analysis methodology considered in this work follows a similar approach to previous studies devoted to the analysis of affective speech across languages (Pell et al., 2009b; Liu and Pell, 2014). Acoustic results are normalized within each speaker using z-scores, and relative acoustic dif-

ferences between storytelling categories are studied in order to avoid speaker-dependent profiles. Next, in the same way as in the first part of the thesis, statistical and discriminant analyses are conducted within each language in order to assess if the different storytelling expressive categories can be acoustically differentiated for each version of the story under analysis (Pell et al., 2009b; Liu and Pell, 2014; Monzo et al., 2007). Again, due to the considerable number of parameters under evaluation, only the results from those parameters that in the discriminant analysis strongly correlate with some significant canonical function, explaining a major portion (85-95%) of the variance among categories, are discussed. The discriminant analysis is also carried out to assess how the different storytelling categories can be discriminated based on the acoustic parameters taken into account.

Furthermore, in order to have a visual representation for assessing acoustic similarities across languages using boxplots, a supervariable explaining the acoustic characteristics of each storytelling category is derived for each language. Each supervariable is computed by multiplying the raw z-scored data by the corresponding unstandardised discriminant function coefficients. Finally, univariate tests are also performed on the canonically derived supervariable following Enders (2003). However, due to the fact of dealing with multivariate derived data, a more conservative evaluation of the statistical significance tests among the storytelling categories is considered (Neufeld and Gardner, 1990), reporting the results using three levels of significance: $p < 0.001$, $p < 0.01$, and $p < 0.05$.

As one of the key goals of this part of the thesis is to study to what extent the results obtained in the first part for Spanish can be generalized to other languages and narrators, the same acoustic parameters are considered. Moreover, the parameters are again only extracted from vowels to ensure its reliability, and then, they are averaged at the utterance level. All the parameters are extracted with a Praat script (Boersma and Weenink, 2014), with the exception of the glottal flow parameters (extracted using COVAREP algorithms, Degottex et al., 2014). The segmentation of the storytelling speech corpora into words, syllables and phonemes is carried out with the EasyAlign tool (Goldman, 2011), for Spanish and French. For the English and German corpora, the SPPAS tool and the web service of the Munich AUtomatic Segmentation (MAUS) system[11] are used (Bigi and Hirst, 2012; Kisler et al., 2012), respectively. Nearly 64,000 phonemes were revised and manually corrected (if necessary) afterwards in order to dispose of reliable data.

5.4 RESULTS

In this Section, the methodology explained in Section 5.3.1 is applied for the analysis of English, German, and French versions of the story under analysis (Section 5.4.1), including the results from the Spanish version obtained in part one of the thesis. Finally, the similarity across languages is evaluated at both perceptual and acoustic levels in Section 5.4.2.

5.4.1 *Acoustic characteristics of storytelling expressive categories by language*

5.4.1.1 *English version*

Statistical analysis— The MANOVA on the English version of the story reveals statistically significant results [$Pillai's\ Trace = 1.405,\ F(105, 1365) = 3.265,\ p < 0.001$] but, differently from the Spanish narrator, the univariate analyses do not show statistically significant differences among categories on all parameters (see Table 11). The most relevant parameters according to the criteria detailed in Section 5.3.1 in the English version of the story result in $F0_{mean}$, H1H2, $F0_{IQR}$, AR, NSP, HAMMI and int_{mean}.

[11] https://clarin.phonetik.uni-muenchen.de/BASWebServices/#/services

TABLE 11: WILKS' LAMBDA VALUES OF EACH PARAMETER BY LANGUAGE. THE ASTERISK (*) INDICATES $p<0.05$ IN THE UNIVARIATE ANALYSIS.

Parameter	Wilks' Lambda			
	Spanish	**English**	**German**	**French**
Nsp	0.771*	0.732*	0.819*	0.782*
AR	0.927*	0.853*	0.831*	0.771*
$F0_{mean}$	0.376*	0.483*	0.502*	0.782*
$F0_{IQR}$	0.594*	0.793*	0.854*	0.820*
int_{mean}	0.671*	0.860*	0.728*	0.750*
Jitter	0.845*	0.938	0.787*	0.882*
Shimmer	0.841*	0.954	0.962	0.951
HNR_{mean}	0.648*	0.732*	0.779*	0.812*
pe1000	0.829*	0.951	0.835*	0.973
HammI	0.877*	0.862*	0.865*	0.763*
SS	0.745*	0.929*	0.761*	0.828*
NAQ	0.879*	0.959	0.768*	0.879*
PSP	0.857*	0.942	0.712*	0.979
MDQ	0.870*	0.742*	0.773*	0.727*
H1H2	0.823*	0.647*	0.874*	0.807*

Thus, more prosodic parameters than VoQ features. Normalized averaged values of all parameters for the English version can be observed in Table 12.

Post-character, negative/passive and positive/passive categories were expressed with similar values of $F0_{mean}$, all of them with significantly lower $F0_{mean}$ than the rest of categories. Negative/active and positive/active categories were expressed with the highest $F0_{mean}$ values (no significant differences between them), significantly higher than neutral, descriptive and suspense categories, which share similar $F0_{mean}$. Similar patterns are observed in the $F0_{IQR}$, although less statistically significant results are observed. For instance, the neutral category is similar to all categories in terms of $F0_{IQR}$ with the exception of the positive/active category, which shows the highest value (although it is similar to the one obtained in negative/active and descriptive utterances). To continue with other prosodic features, int_{mean} shows in general few statistically significant differences among categories, entailing less importance in the English version with respect to the Spanish version. In the English version, descriptive utterances were expressed with the highest int_{mean}, only differing significantly from post-character, negative/passive and suspense utterances. Suspenseful utterances were conveyed with the lowest int_{mean} but, according to the post-hoc tests, with similar int_{mean} to post-character, passive and negative/active categories. In what concerns speech tempo parameters, descriptive utterances were expressed with slow AR and large NSP. On the contrary, post-character utterances show the opposite behaviour. Nonetheless, in general, few statistically significant differences are obtained in terms of AR, only within the most extreme values and specially involving the post-character category. Affective active categories show faster AR than their passive counterparts, but the differences are not statistically significant. Finally, concerning VoQ features, H1H2 is specially relevant to differentiate the post-character situation, which shows the lowest value significantly differing from the rest of categories. In fact, it is the only negative H1H2 result among

TABLE 12: NORMALIZED AVERAGED ACOUSTIC MEASURES OF THE STORYTELLING EXPRESSIVE CATEGORIES OF THE ENGLISH VERSION. NEU: NEUTRAL CATEGORY OF STORYTELLING, P-C: POST-CHARACTER, DES: DESCRIPTIVE, SUS: SUSPENSE, N/P: NEGATIVE/PASSIVE, N/A: NEGATIVE/ACTIVE, P/P: POSITIVE/PASSIVE, P/A: POSITIVE/PASSIVE.

Parameter	**Category**							
	P-C	**DES**	**NEU**	**N/P**	**N/A**	**P/P**	**P/A**	**SUS**
Nsp	-0.74	0.92	-0.13	-0.23	0.15	0.22	0.95	0.11
AR	0.49	-0.45	0.02	-0.10	0.36	-0.21	0.32	-0.55
F0mean	-0.93	0.46	0.17	-0.47	1.27	-0.46	1.21	0.06
F0IQR	-0.35	0.38	-0.07	-0.42	0.51	-0.41	1.07	0.18
intmean	-0.36	0.51	0.43	-0.22	0.23	0.24	0.43	-0.51
Jitter	0.46	-0.32	-0.25	0.06	-0.14	-0.34	-0.10	0.09
Shimmer	-0.15	-0.16	-0.31	0.27	0.12	-0.23	0.08	0.22
HNRmean	-0.87	0.55	0.46	-0.35	0.29	0.09	0.64	0.28
pe1000	-0.02	0.35	0.29	-0.12	-0.07	0.15	-0.13	-0.34
HammI	0.55	-0.15	0.04	-0.20	-0.73	0.30	-0.07	0.15
SS	-0.18	0.24	0.15	-0.14	0.53	0.08	-0.02	-0.36
NAQ	-0.15	-0.17	0.33	-0.20	0.05	0.09	0.38	0.07
PSP	0.35	-0.23	0.10	0.05	-0.43	0.22	-0.23	-0.09
MDQ	-0.53	0.19	0.04	-0.18	0.57	-0.14	0.58	0.15
H1H2	-1.06	0.37	-0.02	-0.07	0.83	-0.25	0.54	0.48

narrators in the non-normalized form, i.e., entailing a larger amplitude of the second harmonic with respect to the first harmonic, suggesting a creakier voice phonation in these utterances (Pépiot, 2014). The rest of categories show few significant contrasts among them in terms of H1H2. The largest value is obtained in the negative/active utterances but it is only significantly greater than the ones obtained from neutral, post-character and passive categories. HAMMI shows not enough statistically significant results in order to derive clear conclusions.

Discriminant analysis— The discriminant analysis conducted on the English version results in three (out of seven) significant canonical discriminant functions [Function 1: $Wilks'\ \Lambda = 0.159,\ \chi^2(105) = 364.472,\ p < 0.0001$; Function 2: $Wilks'\ \Lambda = 0.395,\ \chi^2(84) = 184.132,\ p < 0.0001$; Function 3: $Wilks'\ \Lambda = 0.609,\ \chi^2(65) = 98.501,\ p = 0.005$]. The first canonical function explains 57.8% of the variance and correlates positively with $F0_{mean}$ ($r = 0.79$), H1H2 ($r = 0.59$) and $F0_{IQR}$ ($r = 0.36$). The second function accounts for 21.1% of the variance and correlates positively with AR ($r = 0.47$). The third function explains 10.1% of the variance and correlates positively with NSP ($r = 0.49$), HAMMI ($r = 0.38$) and int_{mean} ($r = 0.37$).

The macro-averaged F1 score (F_1^M) obtained from the LDA classification of the English version of the story is the lowest among all languages (see Table 13), suffering from the lack of correctly classified instances of the positive/passive category together with a general low classification performance, except for the post-character category. Probably, the result from the positive/passive category is due to the low number of samples (see Table 10). The addition of VOQ to the prosodic parameters improves the F_1^M slightly, from 0.330 to 0.335 (relative improvement of 1.4%). In fact, using only VOQ parameters results

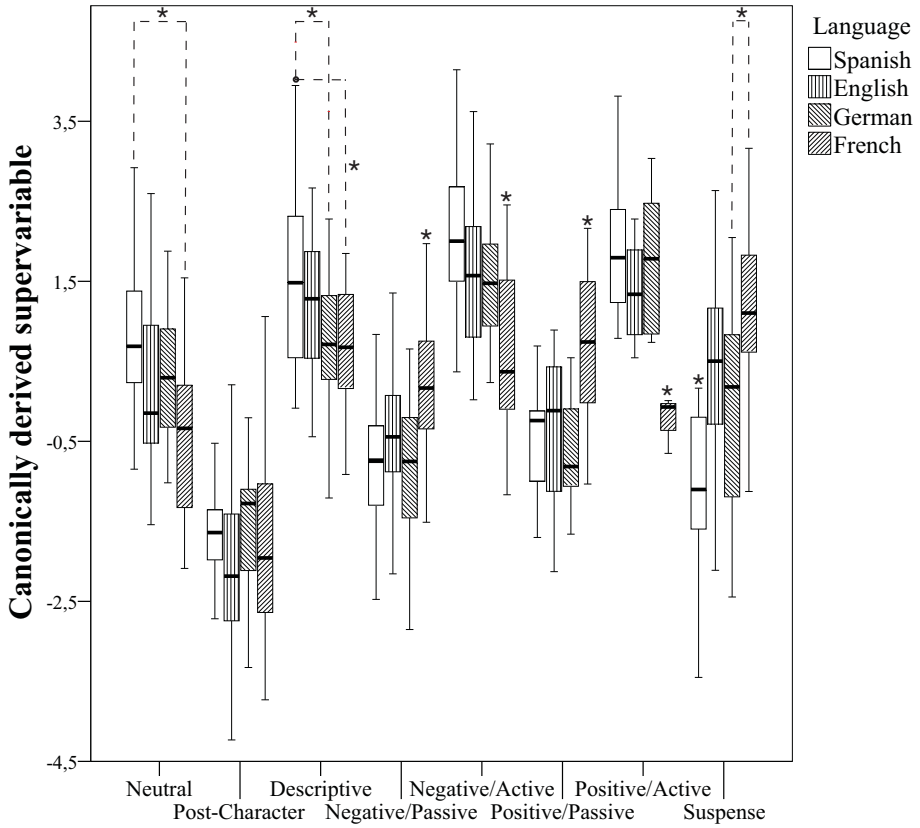

FIGURE 17: CANONICALLY DERIVED SUPERVARIABLE FOR EACH LANGUAGE. DISTRIBUTIONS WITH ONE ASTERISK ON TOP ARE STATISTICALLY DIFFERENT FROM THE REST OF DISTRIBUTIONS ($p < 0.05$), AND THOSE LINKED BY A DASHED LINE AND AN ASTERISK ALSO DIFFER SIGNIFICANTLY. NO STATISTICALLY SIGNIFICANT DIFFERENCE OTHERWISE.

in a worse classification performance than using only prosody. Hence, this result together with the fact the English version of the story only shows two relevant VoQ parameters in contrast to five relevant prosodic parameters highlight that the English narrator introduced little VoQ variability between the storytelling expressive categories in his performance (a conclusion confirmed after informal perceptual validation). Furthermore, the HAMMI parameters (one of the two relevant parameters) has not shown clear patterns in the statistical analysis, which reinforces this fact.

The canonically derived supervariable (see Fig. 17 for a boxplot representation of this supervariable) using the raw discriminant function coefficients significantly differentiates among categories [$F(7, 42.53) = 44.321,\ p < 0.001,\ \eta^2 = 0.606$]. Note that the value of η^2 implies that 60.6% of the variance in the canonically derived supervariable is accounted for by the different categories. Means and standard deviations of the supervariable can be observed in Table 16 and the post-hoc tests results performed to evaluate statistically significant contrasts between storytelling expressive categories are shown in Table 17. Overall, descriptive and both active categories show high values while the rest of categories show lower values, specially the post-character category. However, the contrasts between them entail several non-significant results (two more than the Spanish narrator).

5.4.1.2 *German version*

Statistical analysis— The MANOVA on the German version shows a statistically significant result [$Pillai's\ Trace = 1.429,\ F(105, 1253) = 3.060,\ p < 0.001$], with only shimmer as acoustic parameter

TABLE 13: LINEAR DISCRIMINANT ANALYSIS F1 SCORES PER STORYTELLING CATEGORY AND LANGUAGE. P: PROSODY.

Language	Parameters	P-C	DES	NEU	N/P	N/A	P/P	P/A	SUS	F_1^M
Spanish	P	0.69	0.36	0.49	0.70	0.44	0.12	0.00	0.00	0.363
	VoQ	0.64	0.32	0.46	0.34	0.38	0.00	0.37	0.55	0.384
	P+VoQ	0.76	0.34	0.55	0.63	0.52	0.24	0.33	0.64	0.503
English	P	0.67	0.44	0.13	0.41	0.44	0.00	0.21	0.26	0.330
	VoQ	0.67	0.24	0.23	0.36	0.32	0.00	0.00	0.23	0.260
	P+VoQ	0.66	0.42	0.24	0.38	0.42	0.00	0.33	0.21	0.335
German	P	0.60	0.44	0.33	0.17	0.58	0.00	0.00	0.16	0.294
	VoQ	0.57	0.42	0.18	0.38	0.44	0.00	0.00	0.17	0.272
	P+VoQ	0.56	0.49	0.36	0.25	0.54	0.00	0.35	0.25	0.356
French	P	0.51	0.32	0.00	0.47	0.36	0.00	0.00	0.38	0.272
	VoQ	0.57	0.30	0.00	0.07	0.07	0.00	0.00	0.31	0.177
	P+VoQ	0.67	0.43	0.27	0.39	0.48	0.35	0.00	0.31	0.364

with no statistically significant differences among categories in the subsequent univariate analyses (see Table 11). Relevant parameters consist of $F0_{mean}$, PSP, int_{mean}, NAQ, SS, PE1000, HAMMI, HNR_{mean}, NSP and AR. Normalized averaged values of all parameters for the German version can be observed in Table 14.

The German narrator expressed post-character utterances with the lowest normalized $F0_{mean}$ values, only comparable to those of positive/passive utterances. Contrarily, both active categories are conveyed with high $F0_{mean}$ (all contrasts among categories being significant except between each other). In between these categories, neutral, descriptive, suspense, and passive utterances show similar $F0_{mean}$ values. Regarding int_{mean}, negative/active utterances show the highest value with five significant contrasts (against post-character, neutral, suspense and both passive categories). The int_{mean} shows low values in post-character, suspense and both passive categories, intermediate values in neutral and descriptive categories, and high intensity in both active categories, although these differences are not always significant. Descriptive and suspense categories show the slowest AR (similar to both active categories) together with high values of NSP (in line with positive/passive and both active categories). In contrast, the post-character utterances were expressed with the fastest speech tempo (high AR and low NSP), although they only differ significantly from suspense and descriptive utterances. In what concerns VoQ measures, HNR_{mean} shows clear differentiation of the post-character category with the lowest value (only similar to the negative/passive category). NAQ and PSP are quite correlated in this case, showing similar patterns. On the one hand, post-character utterances entail a tenser voice phonation according to the lowest results in these parameters, but the results are similar to both passive categories. On the other hand, active categories were conveyed with the breathiest phonation according to NAQ and PSP, although with few statistically significant differences among categories. Finally, relevant spectral measures show very few significant contrasts among categories, being SS the one with more statistically significant differences.

Discriminant analysis— Three significant canonical discriminant functions are also obtained after the discriminant analysis on the German version of the story [Function 1: *Wilks'* $\Lambda = 0.164$, $\chi^2(105) = 329.816$, $p < 0.0001$; Function 2: *Wilks'* $\Lambda = 0.378$, $\chi^2(84) = 177.711$, $p < 0.0001$; Function 3: *Wilks'* $\Lambda = 0.539$, $\chi^2(65) = 112.868$, $p < 0.0001$]. The first canonical function explains 54.0% of the variance and correlates positively with $F0_{mean}$ ($r = 0.86$), PSP ($r = 0.54$), int_{mean} ($r = 0.51$), and NAQ ($r = 0.44$). The second function accounts for 17.7% of the variance and correlates positively with SS ($r = 0.77$), PE1000 ($r = 0.58$), and negatively with HAMMI ($r = -0.53$) and HNR_{mean}

TABLE 14: NORMALIZED AVERAGED ACOUSTIC MEASURES OF THE STORYTELLING EXPRESSIVE CATEGORIES OF THE GERMAN VERSION. NEU: NEUTRAL CATEGORY OF STORYTELLING, P-C: POST-CHARACTER, DES: DESCRIPTIVE, SUS: SUSPENSE, N/P: NEGATIVE/PASSIVE, N/A: NEGATIVE/ACTIVE, P/P: POSITIVE/PASSIVE, P/A: POSITIVE/PASSIVE.

Parameter	Category							
	P-C	**DES**	**NEU**	**N/P**	**N/A**	**P/P**	**P/A**	**SUS**
Nsp	-0.53	0.66	-0.33	-0.14	0.05	0.67	0.11	0.46
AR	0.44	-0.66	0.20	0.28	-0.22	0.22	0.10	-0.66
F0mean	-0.96	0.15	0.33	-0.44	1.06	-0.40	1.20	0.01
F0IQR	-0.38	-0.10	-0.16	0.04	0.81	-0.36	0.17	-0.04
intmean	-0.46	0.38	0.10	-0.48	0.88	-0.40	0.75	-0.38
Jitter	0.54	-0.60	-0.36	0.61	-0.19	-0.45	-0.46	-0.06
Shimmer	0.21	-0.30	0.01	0.25	-0.20	-0.05	-0.10	-0.09
HNRmean	-0.78	0.39	0.42	-0.26	0.13	0.72	0.29	0.18
pe1000	0.48	0.35	-0.46	-0.16	0.32	-0.70	-0.26	-0.46
HammI	-0.23	-0.18	0.27	0.21	-0.51	0.82	-0.29	0.43
SS	0.42	0.32	-0.37	-0.36	0.61	-0.85	0.22	-0.68
NAQ	-0.73	0.03	0.43	-0.33	0.62	-0.05	0.54	0.17
PSP	-0.74	0.09	0.36	-0.39	0.82	-0.18	0.68	0.02
MDQ	-0.72	0.26	0.34	-0.19	0.63	-0.19	0.62	-0.19
H1H2	0.05	-0.53	0.09	0.52	-0.25	0.42	-0.69	0.08

($r = -0.53$). The third function explains 13.3% of the variance and correlates with speech tempo measures, i.e., NSP ($r = 0.48$) and AR ($r = -0.43$).

The F_1^M from the LDA classification of German utterances results in 0.356, with best results in the post-character and negative/active categories (see Table 13). As in the English version, the method also fails to classify positive/passive utterances. The F_1^M of the classification using prosodic features is 0.294, increasing up to a 20.9% to the final F_1^M value when considering all parameters. Similarly, if only VOQ features are considered the F_1^M is 0.272, improving up to a value of F_1^M of 30.4% when including prosodic features. Thus, both set of parameters show equivalent importance in the discrimination between storytelling expressive categories.

The canonically derived variable also shows a statistically significant result on the German corpus $[F(7, 33.70) = 46.079,\ p < 0.001,\ \eta^2 = 0.641]$ and it is depicted in Fig. 17. Means and standard deviations are shown on Table 16, and post-hoc tests results can be observed in Table 17. Similarly to the other versions, in general, active and descriptive categories entail higher values than the rest of expressive categories. The German version of the story shows a similar amount of statistically significant contrasts to the Spanish and English counterparts (see Table 17).

5.4.1.3 *French version*

Statistical analysis— The multivariate analysis on the French version also shows statistical significance [$Pillai's\ Trace = 1.351, F(105, 1169) = 2.661, p < 0.001$], while univariate analyses show statistic-

TABLE 15: NORMALIZED AVERAGED ACOUSTIC MEASURES OF THE STORYTELLING EXPRESSIVE CATEGORIES OF THE FRENCH VERSION. NEU: NEUTRAL CATEGORY OF STORYTELLING, P-C: POST-CHARACTER, DES: DESCRIPTIVE, SUS: SUSPENSE, N/P: NEGATIVE/PASSIVE, N/A: NEGATIVE/ACTIVE, P/P: POSITIVE/PASSIVE, P/A: POSITIVE/PASSIVE.

Parameter	Category							
	P-C	**DES**	**NEU**	**N/P**	**N/A**	**P/P**	**P/A**	**SUS**
Nsp	-0.45	0.52	-0.30	-0.12	-0.27	1.34	-0.43	0.29
AR	0.45	-0.73	0.15	0.48	0.09	-0.10	0.40	-0.63
F0mean	-0.16	0.69	0.06	-0.59	1.00	-0.36	0.73	-0.14
F0IQR	-0.40	0.45	-0.05	-0.41	0.93	-0.19	0.80	0.18
intmean	0.36	0.13	0.18	-0.46	1.12	-0.11	0.75	-0.68
Jitter	-0.59	-0.06	0.02	0.27	0.44	0.35	-0.08	0.17
Shimmer	-0.25	0.10	-0.31	0.26	-0.03	0.14	0.25	0.18
HNRmean	0.54	0.29	0.15	-0.53	0.46	-0.19	0.47	-0.58
pe1000	-0.23	0.00	-0.14	0.13	-0.17	0.08	-0.35	0.32
HammI	0.68	0.04	0.19	-0.46	0.36	-0.14	0.65	-0.65
SS	-0.47	-0.02	-0.20	0.33	-0.46	0.01	-0.65	0.62
NAQ	-0.07	0.71	-0.24	-0.26	0.69	-0.16	0.11	-0.18
PSP	0.16	0.16	0.07	-0.19	-0.02	-0.19	0.17	-0.11
MDQ	-0.88	0.45	-0.25	0.11	0.48	0.01	0.19	0.54
H1H2	-0.75	0.20	0.05	0.03	0.43	0.08	0.21	0.42

TABLE 16: MEANS (M) AND STANDARD DEVIATIONS (SD) OF THE CANONICALLY DERIVED SUPERVARIABLE FOR EACH CATEGORY AND LANGUAGE.

Language	Statistic	P-C	DES	NEU	N/P	N/A	P/P	P/A	SUS
Spanish	M	-1.64	1.51	0.76	-0.74	2.10	-0.49	1.91	-1.17
	SD	0.54	1.12	0.86	0.89	0.96	0.71	0.83	1.01
English	M	-2.08	1.24	0.18	-0.40	1.55	-0.39	1.38	0.37
	SD	1.01	0.95	1.12	0.82	0.99	1.01	0.61	1.13
German	M	-1.61	0.81	0.29	-0.84	1.51	-0.59	1.80	-0.06
	SD	0.79	0.82	0.78	0.84	0.74	0.71	0.86	1.27
French	M	-1.84	0.66	-0.36	0.18	0.62	0.67	-0.20	1.08
	SD	1.17	0.75	0.96	0.81	0.99	1.00	0.31	0.96

ally significant results on all parameters except in shimmer, PE1000, and PSP (see Table 11). Post-hoc tests results are reported on MDQ, H1H2, jitter, int_{mean}, F0_{mean}, F0_{IQR}, HNR_{mean}, HAMMI, SS, NSP, and AR, as they are the most relevant parameters. Normalized averaged values of all parameters for the French version can be observed in Table 15.

Similarly to the English narrator, all prosodic features show relevance for differentiating storytelling expressive categories in the French version of the story. However, fewer statistically significant differ-

TABLE 17: RESULTS OF THE POST-HOC TESTS ON THE CANONICALLY DERIVED SUPERVARIABLE BY LANGUAGE. THE MATRIX IS HALF EMPTY BECAUSE IT IS SYMMETRICAL ON THE DIAGONAL. †: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$.

Language	Category	P-C	DES	NEU	N/P	N/A	P/P	P/A	SUS
Spanish	P-C	-	†	†	†	†	**	†	0.525
	DES	-	-	*	†	0.304	†	0.905	†
	NEU	-	-	-	†	†	†	**	†
	N/P	-	-	-	-	†	0.987	†	0.618
	N/A	-	-	-	-	-	†	0.998	†
	P/P	-	-	-	-	-	-	†	0.342
	P/A	-	-	-	-	-	-	-	†
	SUS	-	-	-	-	-	-	-	-
English	P-C	-	†	†	†	†	†	†	†
	DES	-	-	**	†	0.957	†	0.999	*
	NEU	-	-	-	0.223	†	0.704	*	0.996
	N/P	-	-	-	-	†	1.000	†	*
	N/A	-	-	-	-	-	†	0.999	†
	P/P	-	-	-	-	-	-	†	0.344
	P/A	-	-	-	-	-	-	-	0.066
	SUS	-	-	-	-	-	-	-	-
German	P-C	-	†	†	**	†	0.157	†	†
	DES	-	-	0.376	†	0.068	†	*	*
	NEU	-	-	-	†	†	0.082	†	0.873
	N/P	-	-	-	-	†	0.990	†	*
	N/A	-	-	-	-	-	†	0.983	†
	P/P	-	-	-	-	-	-	†	0.707
	P/A	-	-	-	-	-	-	-	†
	SUS	-	-	-	-	-	-	-	-
French	P-C	-	†	†	†	†	†	*	†
	DES	-	-	†	0.634	1.000	1.000	0.729	0.731
	NEU	-	-	-	0.410	*	0.079	1.000	†
	N/P	-	-	-	-	0.849	0.854	0.807	0.807
	P/P	-	-	-	-	-	-	0.790	0.939
	P/A	-	-	-	-	-	-	-	0.200
	SUS	-	-	-	-	-	-	-	-

ences are observed. Regarding $F0_{mean}$, it shows the largest number of significant contrasts for negative/active or descriptive categories. Specifically, these categories significantly differ from all categories in terms of $F0_{mean}$ with the exception of neutral, positive/active and between each other. The lowest normalized averaged $F0_{mean}$ value of negative/passive utterances only significantly differs from descriptive and negative/active categories. $F0_{IQR}$ values are quite similar among categories according to the post-hoc tests. The post-character category shows the largest number of statistically significant contrasts in terms of $F0_{IQR}$, concretely, when compared to descriptive, negative/active and suspense categories. With regard to int_{mean}, the suspense category shows the lowest intensity value, differing at a statistically significant level from post-character, descriptive, neutral, negative/active, and positive/active categories. On the contrary, negative/active utterances show the highest value of int_{mean}, which is similar to post-character and positive/active categories. The most remarkable observation from NSP results is its

high value in positive/passive utterances. This value is significantly different from the ones obtained for each category with the exception of descriptive utterances. In the French version of the story, the descriptive category also show a relatively high NSP value (although the only significant contrasts are obtained when compared to neutral and post-character utterances) together with the lowest AR measure (significant contrasts against post-character, neutral and negative/passive utterances). Even though jitter, HNR_{mean}, HAMMI, and SS do not show any specific pattern per category due to the low number of statistically significant differences, MDQ proves specially useful for characterizing the post-character category (although the distribution is similar to that of neutral and positive categories). This lowest value can be associated with a tenser voice (Kane and Gobl, 2013). In this sense, the H1H2 follows a practically identical behaviour as MDQ. However, H1H2 also shows statistically significant differences with respect to the neutral category. According to the VOQ parameters, the breathiest phonation appears to be present in the suspense utterances, although few significant contrasts among categories are obtained.

Discriminant analysis— Three significant canonical discriminant functions can be also derived from the French version [Function 1: $Wilks'\ \Lambda = 0.183,\ \chi^2(105) = 289.872,\ p < 0.0001$; Function 2: $Wilks'\ \Lambda = 0.396,\ \chi^2(84) = 158.110,\ p < 0.0001$; Function 3: $Wilks'\ \Lambda = 0.566,\ \chi^2(65) = 97.089,\ p = 0.006$]. The first canonical function explains 52.1% of the variance and correlates positively with three VOQ paramters: MDQ ($r = 0.55$), H1H2 ($r = 0.44$), and jitter ($r = 0.304$). The second function accounts for 19.2% of the variance and correlates positively with int_{mean} ($r = 0.76$), F0_{mean} ($r = 0.71$), F0_{IQR} ($r = 0.51$), HNR_{mean} ($r = 0.47$), HAMMI ($r = 0.45$), and negatively with SS ($r = -0.44$). The third function explains 16.9% of the variance and correlates with speech tempo measures: AR ($r = 0.51$) and NSP ($r = -0.42$).

The French narrator shows the second best F_1^M (see Table 13), but also including a category with no correct classifications: the positive/active category. In this case, it is clear that this lack of correctly classified instances is due to the very low number of samples of positive/active utterances (see Table 10). Post-character utterances are again well classified and the negative/active category achieves a F1 value of 0.48. Consistently with the previous results, the best F_1^M is obtained using all parameters, improving in a substantial 33.6% yielded by the LDA classifier trained using only prosodic parameters. Finally, VOQ features by themselves are not enough to achieve a good F_1^M, but it is crucial when combined with prosodic parameters.

The univariate test performed on the canonically derived data (see Fig. 17 to observe the canonically derived supervariable) show a statistically significant result [$F(7, 28.16) = 30.446,\ p < 0.001,\ \eta^2 = 0.552$]. However, the subsequent post-hoc tests show fewer statistically significant results than the Spanish, English, and German versions of the story (see Table 17). Only the post-character utterances are quite well differentiated.

5.4.2 *Storytelling expressive categories across languages*

Once the acoustic characterization for each language has been conducted, the similarities across languages at both perceptual (Section 5.4.2.1) and acoustic (Section 5.4.2.2) levels are analysed in this Section.

5.4.2.1 *Perceptual-level similarities*

In this Section, an analysis regarding to what extent the different storytellers used the same (or different) storytelling expressive categories for each utterance is conducted. All categories are considered in the analysis, including the 'Other' category in order to avoid loosing information of similarity/dissimilarity between narrators. Note that the post-character and descriptive categories are included in this analysis (even though they have been annotated only from text), as they have shown very similar acoustic patterns across languages, specially post-character utterances. An overall visual representation of the similarity

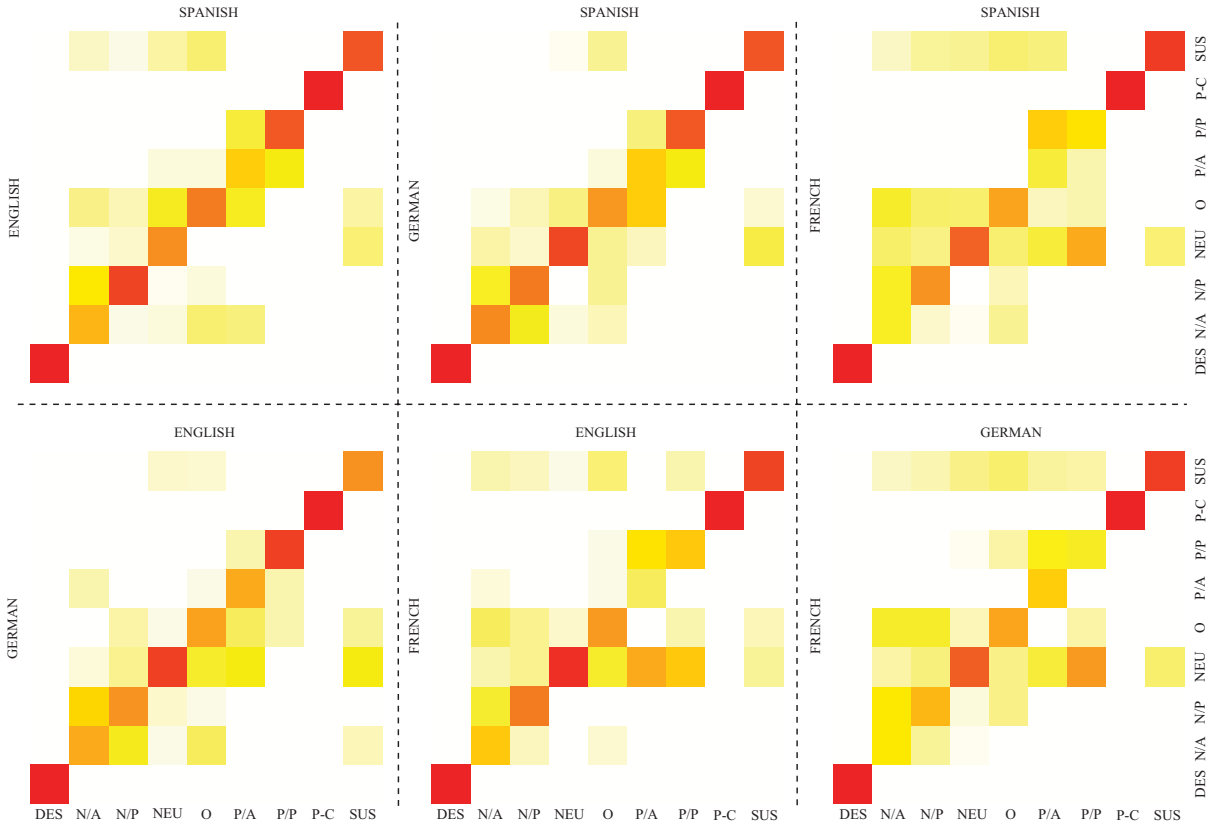

FIGURE 18: HEATMAPS COMPUTED FROM THE CONTINGENCY TABLES FOR EACH PAIR OF NARRATORS, SHOWING THE RELATIONSHIP REGARDING THEIR USE OF EXPRESSIVE CATEGORIES. A WARMER COLOUR REPRESENTS MORE COMMON INSTANCES. O: 'OTHER'.

between narrators regarding the use of storytelling expressive categories for each sentence can be observed in Fig. 18.

In order to measure the similarity between narrators in terms of the use of expressiveness, Cramér's V coefficients (ϕ_C) were computed as nominal variables (category labels) are used (Cramér, 1946):

$$\phi_C = \sqrt{\frac{\chi^2}{n(k-1)}} \tag{4}$$

where χ^2 is the Pearson's Chi-square, n is the total number of cases (263 in this case), and k is the number of categories (9 in this case). In Table 18, the Cramer's V coefficients for each pair of narrators are shown. Since all these magnitudes of association belong to the interval (0.60–0.80), they can be regarded as strong (Rea and Parker, 1992). Thus, there is a high similarity in the use of expressiveness by the four narrators, being the Spanish and German narrators the ones showing the greatest resemblance. Contrarily, the French narrator is the one who shows the most different use of storytelling expressive categories with respect to the rest of narrators.

Furthermore, measures of perceived similarity have already been obtained from the cross-language perceptual tests conducted in Sections 5.2.1 and 5.2.2 for neutral and suspenseful utterances, respectively. As observed in both Sections, users perceived most of the utterances comparisons as "roughly the same expressiveness" with a substantial agreement ($p_{\text{"roughly the same"}} = 0.93$ for neutral utterances, $p_{\text{"roughly the same"}} = 0.90$ for suspense utterances), evidencing that the storytellers shared similar expressiveness for the same utterances (those utterances under evaluation). Although it is a remarkable result, it should be corroborated with more subjective tests (e.g., with native users).

TABLE 18: RELATIONSHIP BETWEEN NARRATORS IN TERMS OF THE USE OF EXPRESSIVENESS FOR EACH UTTERANCE. THE MATRIX IS HALF EMPTY BECAUSE IT IS SYMMETRICAL ON THE DIAGONAL.

	Cramér's V coefficient			
	Spanish	**English**	**German**	**French**
Spanish	-	0.728	0.753	0.657
English	-	-	0.731	0.691
German	-	-	-	0.660
French	-	-	-	-

5.4.2.2 *Acoustic-level similarities*

With the objective of evaluating the degree of acoustic similarity across narrators, firstly, the parameters that manifested as relevant in most languages are discussed. Then, similarly to previous cross-language studies (Grawunder and Winter, 2010; Pell et al., 2009b; Liu and Pell, 2014), the most consistent acoustic tendencies in storytelling expressive categories across languages are evaluated.

The parameters that have manifested as relevant in all languages following the criteria detailed in Section 5.3.1 are $F0_{mean}$ and int_{mean} (see Table 19). Nonetheless, the French narrator used less $F0_{mean}$ variability among categories, as this parameter correlates with the second canonical function instead of the first one (see Section 5.4.1.3) and shows a larger Wilks' Lambda value than the rest of narrators (see Table 11). Similarly, the English narrator conveyed int_{mean} with less variability, as this parameter correlates with the third canonical function (see Section 5.4.1.1) and shows the largest int_{mean} Wilks' Lambda result among narrators in Table 11. The rest of considered prosodic parameters entail relevance (at different levels) in at least three languages. In relation to $F0_{IQR}$, no relevance is found in the German narrator (see Section 5.4.1.2), whereas the rest of narrators conveyed the storytelling expressive categories with considerable different $F0_{IQR}$. To conclude with prosodic parameters, the measures extracted from the Spanish version of the story related to speech tempo do not belong to the relevant set of parameters, i.e., the narrator used almost the same speech tempo among all the expressive categories. Nevertheless, it is to note the large NSP value in descriptive utterances. In the rest of the languages, tempo parameters typically show correlation with the third canonical function, which explains 10-17% of variance among categories, although AR correlates with the second canonical function in the French version (see Section 5.4.1.3). Finally, the considered narrators modified their VOQ across categories to different extents. For instance, the French narrator showed strong correlations with the first canonical function (accounting for 52.1% of the variance) exclusively in terms of MDQ, H1H2 and jitter. In contrast, the rest of narrators manifested at least two prosodic parameters correlating with the first function in all cases. Moreover, the English narrator only showed two relevant VOQ features (besides HAMMI being questionably relevant as highlighted in Section 5.4.1.1). In contrast, the French, Spanish and German narrators showed six, four, and six relevant VOQ parameters, respectively. That is, the English narrator mainly makes use of prosodic variations to convey the different expressive categories of the story. As a consequence, the LDA classification result of English has yielded to the worst performance among the bunch of results shown in Table 13.

In what concerns acoustic similarities within storytelling expressive categories across languages, an overall view can be obtained using the canonically derived supervariables. To that effect, they are depicted in Fig. 17 together with post-hoc comparisons of the distributions across languages. As it can be observed, there is a similar acoustic trend across categories. Specially, notice the post-character category, which is the only category that shows similar acoustic distributions across all languages. The rest

TABLE 19: RELEVANT PARAMETERS IN THE DISCRIMINATION AMONG STORYTELLING EXPRESSIVE CATEGORIES BY LANGUAGE ACCORDING TO THE DEFINED CRITERIA.

Parameter	Spanish	English	German	French
Nsp	-	X	X	X
AR	-	X	X	X
$F0_{mean}$	X	X	X	X
$F0_{IQR}$	X	X	-	X
int_{mean}	X	X	X	X
Jitter	X	-	-	X
Shimmer	-	-	-	-
HNR_{mean}	X	-	X	X
pe1000	-	-	X	-
HammI	-	X	X	X
SS	X	-	X	X
NAQ	X	-	X	-
PSP	-	-	X	-
MDQ	-	-	-	X
H1H2	-	X	-	X

of categories show some statistically significant differences between languages, mostly involving the French narrator. Post-character utterances, in general, were expressed with lower NSP because of their typical short duration and faster AR, probably due to the fact that the conveyed information tends to be expected a priori and it is very concrete. Nonetheless, the Spanish narrator used a slower AR in general when expressing these utterances. Concerning other prosodic parameters, post-character utterances tended to be transmitted with a muffler voice that implies lower f0 and intensity values. Nevertheless, the French narrator used a higher intensity and just a slight decrease in $F0_{mean}$. Probably, this may be caused by emotional traces of the previous character's intervention that the narrator could not (or did not wanted to) hold, confirmed through informal subjective tests. Due to the higher intensity used by the French narrator in this category, jitter and HNR_{mean} also show different patterns with respect to the other narrators (Brockmann et al., 2008). The last common tendency is the very low values of MDQ, which entail a tenser phonation in this category according to previous literature (Kane and Gobl, 2013), although spectral features do not corroborate this assumption entirely. Regarding the neutral category, in general, it is located somewhere in the middle among all categories within each language (see Fig. 17), with few subtle variations across languages. However, there is a significant difference between the Spanish and French narrators in terms of how they expressed these utterances. Concretely, the Spanish narrator shows significantly larger $F0_{mean}$, int_{mean}, HNR_{mean} and NAQ than his French counterpart. Descriptive utterances were generally expressed quite similar to neutral utterances but with higher NSP and slower AR. This might be a consequence of the relevant information transmitted to the audience in such utterances, which needs to be deeply internalized in order to picture the scenarios during the course of the story. In addition, the moderate increase of $F0_{mean}$ can be attributed to greater emphasis in stressed vowels of certain adjectives (e.g., *"huuuge"*, *"enooormous"*, etc.), as implicitly suggested by Theune et al. (2006). The greatest differences between narrators are observed in the Spanish narrator when compared to his French and German counterparts (see Fig. 17). Specifically, the French narrator used a significantly lower AR than the Spanish narrator, whereas the German narrator also shows signi-

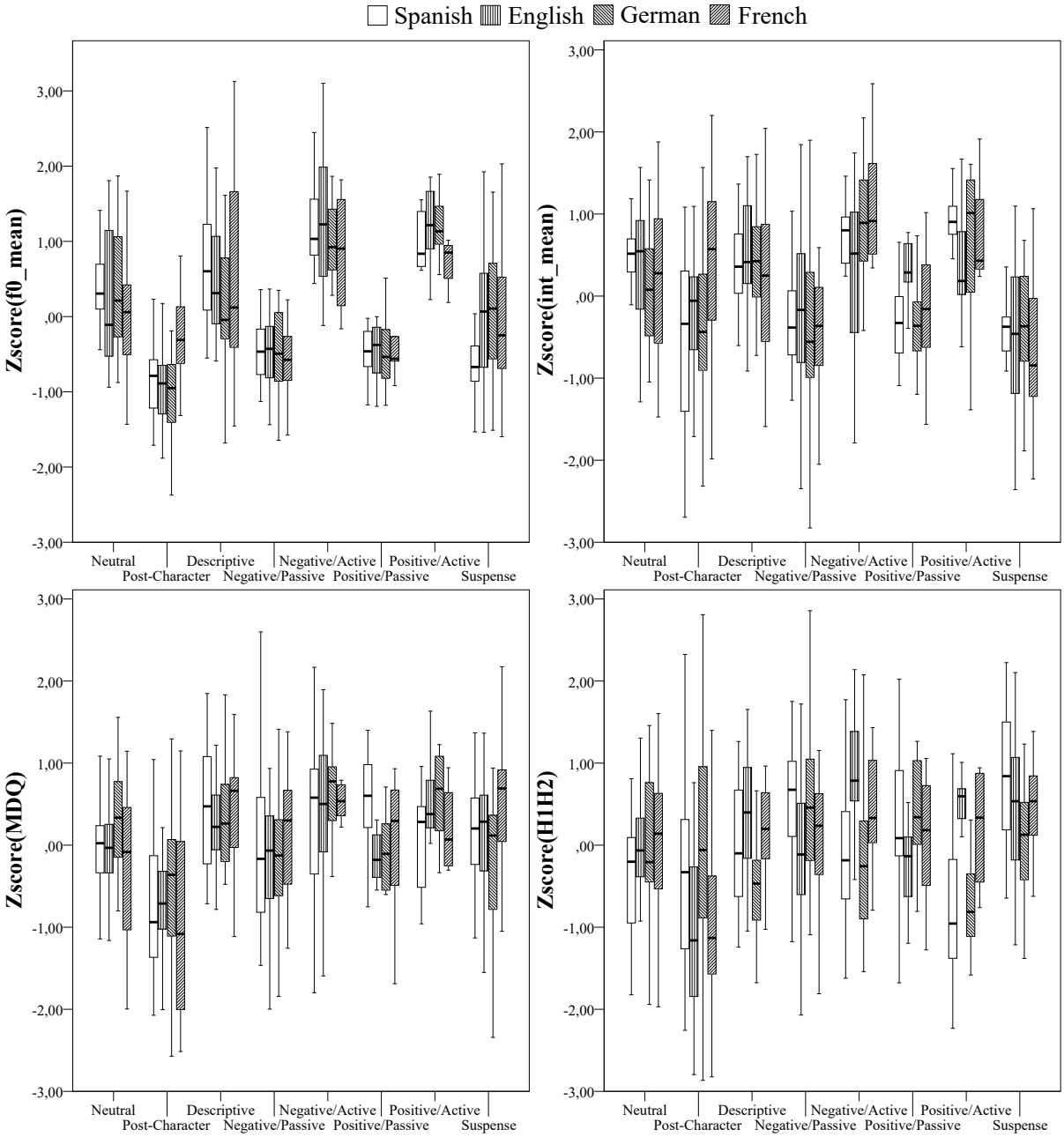

FIGURE 19: Z-SCORES DISTRIBUTIONS OF MEAN FUNDAMENTAL FREQUENCY, MEAN INTENSITY, MAXIMA DISPERSION QUOTIENT, AND H1H2 PARAMETERS BY LANGUAGE.

ficantly lower AR together with lower H1H2 and higher SS. With respect to affective categories, it can be observed from Fig. 17 that the French narrator used a specific expressive pattern across expressive categories when compared to the rest of narrators. This difference is specially striking in positive/active utterances, which made difficult the finding of these utterances in the French audiobook (see Table 10). In both active categories, regarding prosodic parameters, there is a global pattern of high $F0_{mean}$, $F0_{IQR}$ and int_{mean} and low values of such parameters in their passive counterparts, although the pattern is less consistent in passive categories. Concretely, the Spanish narrator expressed negative/passive with high $F0_{IQR}$ while the English narrator expressed positive/passive utterances with a moderately high int_{mean}. These general patterns are consistent with prior findings (cf. Schröder, 2004, and references therein), although the slight deviations may be attributed to the milder expressiveness of the indirect storytelling speech with respect to other previously analysed states that entail more extreme expressiveness. In rela-

tion to these previous studies, a flatter spectral slope has also been obtained in terms of the SS parameter except in the French narrator, which shows the opposite behaviour. Ultimately, suspenseful utterances show a different pattern in the Spanish version of the story compared to its counterparts. The most common prosodic patterns observed across languages are the slower AR and the lower int_{mean}. This decrease in int_{mean} is related to other authors' suggestion that a lower intensity may induce suspense (Doukhan et al., 2011). In contrast, the narrators used VoQ in quite diverse ways to express suspense, thus, being difficult to define which phonation was used in general, even though a breathier phonation could be the most appropriate option. To support this last claim, as most common tendencies, H1H2 is quite high across languages with the exception of the German narrator that shows a mid-range value, whereas SS results in low values across all narrators except in the French one, which shows high values. Probably, these common acoustic patterns are enough to awake a suspenseful feeling in the audience, as many utterances have been perceived as similar across languages in terms of generation of a suspense feeling in Section 5.2.2.

In summary, the prosodic features that, in general, showed more common patterns across languages are $F0_{mean}$ and int_{mean}, while in terms of VoQ features, MDQ and H1H2 also show relatively similar patterns (see Fig. 19).

5.5 DISCUSSION

In this Section, the results obtained along this part of the thesis are discussed besides recapitulating the objectives introduced in Section 1.2.

5.5.1 *Annotation of storytelling expressive categories*

The annotation of the English, German, and French versions of the story has followed the methodology used in part one with some adaptations, achieving a comparable amount of successfully classified utterances (around 85% in average). Although this manually-based approach has been useful to analyse several expressive categories within storytelling speech across languages, it is a tedious and time consuming task. Thus, the development of an automatic version would be very interesting for the annotation of sentences from a story with appropriate expressiveness. Note that the adopted valence/activation scheme for affective categories has already showed automation potential in Section 5.2.3. Although the text input has not been parametrized, the addition of linguistic features (e.g., part of speech, sentence length, information from affective dictionaries and context, etc.) could also be used to improve the automatic classification (Planet and Iriondo, 2013).

5.5.2 *Do the previously defined storytelling expressive categories exist?*

After all the conducted analyses, the cross-language results demonstrate the existence of the storytelling expressive categories introduced in part one of the thesis. Although presenting different distributions within the story, all the expressive categories already observed in the Spanish version of story have also been identified in the English, German, and French counterparts. Moreover, there are particular acoustic patterns within each category that are comparable across languages, specially, across the Spanish, English, and German versions of the story. Furthermore, it is to note the similar acoustic patterns observed in the post-character and descriptive categories, as they have only been identified from a text-based annotation perspective.

5.5.3 *Do narrators use the same expressiveness for each utterance?*

A strong relationship has been observed between narrators regarding the use of expressiveness for each utterance, according to the outcome of the annotation process beyond their personal styles. As it could be expected, narrators sometimes used different expressiveness to express the same sentence. However, the relationship is remarkable since, as far as we know, no expressive indications were given to the narrators, highlighting the fact that professional storytellers make use of similar expressiveness in spite of their personal styles.

The results of the cross-language perceptual tests have also shown a high degree of similarity in the use of expressiveness across narrators and languages. In general, the evaluators, although being neither native nor experts, perceived a similar expressiveness in most of the utterances under evaluation across languages. However, it is to note that the semantic content of the utterances under analysis (the Spanish version was included as reference) might have been of help (Borod et al., 2000).

5.5.4 *Are the acoustic characteristics of each storytelling expressive category comparable across languages?*

The level of $F0_{mean}$ differentiates storytelling expressive categories in a very similar way, although the French narrator used less variability of this parameter across them. This fact, highlights that the F0 level is a crucial parameter in storytellers, similarly to what has been observed after analysing attitudes and emotions (Mozziconacci, 2001; Pell et al., 2009b). Nonetheless, it has been proved that F0 interacts with other parameters, such as int_{mean}, MDQ, and H1H2. The MDQ parameter has arisen as the VoQ parameter showing the largest number of cross-language similarities in the differentiation of the storytelling expressive categories under analysis. Such result could somehow be related to the fact that MDQ has previously shown a significant improvement in the detection of the phonation types within running speech when compared to other glottal flow parameters such as the NAQ, the quasi-open quotient, and the difference between the two first harmonics of the narrowband voice source spectrum (Kane and Gobl, 2013). Regarding spectral parameters, in general, few common patterns across languages have been onserved, although the H1H2 parameter is the one showing the greatest resemblances.

The four storytellers made use of the $F0_{mean}$, the int_{mean}, the MDQ, and the H1H2 parameters in a relatively equal measure to differentiate the storytelling expressive categories under analysis. However, as expected a priori, several differences have also been observed in the sense of proportionality (direct or inverse) or degrees (e.g., much higher vs. higher), specially in the French narrator. Such differences can be attributed to personal styles within storytelling, similarly to what occurs when using actors to develop emotional corpora (Wallbott and Scherer, 1986). Different individuals may have had different previous experiences, and their capability or expertise to convey a story in an expressive way may not be the same. Nevertheless, a remarkable amount of similarities in the use of expressiveness among different storytellers has been observed beyond their personal preferences, specially, across the Spanish, English, and German versions of the story.

5.5.5 *Is Voice Quality as important as prosody to discriminate among storytelling expressive categories across languages?*

In part one of the thesis, three prosodic parameters ($F0_{mean}$, $F0_{IQR}$, and int_{mean}) and four VoQ parameters (SS, H1H2, jitter, and HNR_{mean}) proved as relevant parameters in the discrimination among storytelling categories of the Spanish version of the story. In the present part of the thesis, $F0_{mean}$ and int_{mean} have proved as relevant for the discrimination among those categories in all four versions of the story. Similarly, $F0_{IQR}$ has turned out relevant in the Spanish, English, and French versions of the story, whereas

the German narrator introduced less F0 range variability across categories in his interpretation. In the same way, the speech tempo measures are relevant in the discrimination among categories when they are conveyed by the English, German, and French narrators (an evidence not observed in the Spanish version). However, the AR and the NSP correlate with the significant canonical functions that explain a low portion of the variance among categories. Finally, regarding VOQ parameters, it is to note that H1H2, SS, and HNR_{mean} manifest as a relevant parameters in three out of the four versions of the story. Thus, only jitter has to be removed from the relevant VOQ features identified in the Spanish narrator to characterize expressive categories in storytelling speech.

Furthermore, the results from the LDA classifications have shown that both prosody and VOQ contribute in a relatively equal way to the discrimination among storytelling expressive categories for the considered languages. Although the English narrator relied more on prosodic variations to convey the story under analysis, this might be intended by the narrator (i.e., a personal style), thus, it is recommendable to take always into account VOQ when dealing with storytelling speech.

Finally, in part one it was concluded that the expressiveness contained in the Spanish version of the story could not be related to emotional speech, as it contained subtler speech nuances. This finding can be generalized to the rest of versions analysed in part two because of the following reasons. On the one hand, the conducted post-hoc analyses have shown few significant results in all versions of the story. In contrast, when analysing emotions, more significant differences can be observed. On the other hand, the macro-averaged LDA F1 scores obtained for each language (which range between 0.3–0.5) are more modest and distant than the classification performances obtained in many works focused on the analysis of emotions. For instance, the macro-averaged F1 score obtained for the emotional corpus analysed in part one was 0.93. An implication of this finding is that when considering the indirect speech from audiobooks to generate synthetic emotional speech, the synthetic output might need and extra post-processing that boosts the expressiveness to achieve an excellent resemblance to basic emotions.

5.5.6 *Language-specific characteristics vs. personal styles*

An informal evaluation of the acoustic data was conducted before the statistical and discriminant analyses, showing well-known language-specific characteristics. For example, articulation rate is language-dependent (Fenk-Oczlon and Fenk, 2010), and tends to be much faster in Spanish than in English (de Johnson et al., 1979), a characteristic observable in the corpora at hand. However, the computation of z-scored relative differences removes this language-specific characteristics. This can also be extrapolated to the rest of parameters. Hence, the use of different tempo profiles by the narrators can be attributed to their personal style, as speech tempo can be changed by speakers whenever they want (Trouvain, 2004).

Considering the results obtained from the analysis of language communities from Europe with similar cultural worldview (Scherer et al., 1988), it can be concluded that exploring relative differences between storytelling expressive categories within each language has been sufficient enough to cancel language-specific patterns out. Therefore, it seems plausible to consider that the observed differences in the way of expressing the different storytelling expressive categories by each narrator can be attributed to personal styles rather than language-specific factors. In this sense, it has been also argued that differences in the way of conveying expressiveness can be entirely accounted for by individual differences in personality (Matsumoto, 2006). Notice, for instance, that several contradictory acoustic patterns have been observed in the French version of the story when compared to the other versions. However, there are no evidences that can relate this fact to language-dependent acoustic characteristics, thus, it can be considered that they can be explained by the French narrator personal style.

Furthermore, few works have dealt with the cross-language analysis of VOQ features. For most of the parameters considered in this work, it is difficult to hypothesize about potential language-dependent VOQ profiles as, up to our knowledge, there are no works focused on studying such parameters across languages. In any case, if the narrators decided to modify their phonation from one storytelling express-

ive category to another (e.g., from a breathier to a tenser phonation), this information would be reflected in the z-scored relative differences under study and should be considered a personal decision.

5.6 CONCLUSIONS OF PART II

In this second part of the thesis, it has been studied to what extent the results of a previous part can be generalized to other languages. To that effect, the analysis has been extended by considering the corresponding English, German and French version of the same story under analysis. Moreover, the same annotation methodology (with some adaptations) has been applied with the objective of confirming the existence of several storytelling expressive categories (a total of eight categories) and the role that both prosodic and VoQ features play in indirect storytelling speech through several statistical and discriminant analyses.

The different narrators have shown a strong relationship regarding the use of expressiveness for each utterance of the story in terms of the Cramer's V coefficient ($\phi_C \in [0.6–0.8]$). The lowest values are present when comparing the different narrators with respect to the French narrator, as he used neutral and suspenseful speech in more passages of the story. Moreover, the high level of similarity indicated by the evaluators through the cross-language perceptual tests is another evidence. This is a remarkable result as they are non-experts and non-native.

In what concerns acoustic characterization of the different storytelling expressive categories, subtle variations have been observed across categories. However, significant common patterns have also been observed across languages. Regarding prosodic features, narrators expressed the different categories similarly in terms of F0 and intensity levels, whereas the most common VoQ patterns have been found in the MDQ and H1H2 parameters.

Furthermore, both prosody and VoQ have shown a relatively equal importance in the discrimination among storytelling expressive categories. The most relevant prosodic parameters in the discrimination have resulted in $F0_{mean}$, $F0_{IQR}$, and int_{mean}, whereas the spectral slope, H1H2, and HNR_{mean} have resulted in the most relevant VoQ parameters. Last but not least, it is worth highlighting that these results have been found beyond the observed personal styles of the four narrators.

Part III

ANALYSES ORIENTED TO SYNTHESIS

The purpose of this part of the thesis is to define first steps towards developing a storytelling Text-To-Speech synthesis system capable of driving a speaking avatar by means of modelling the specific prosodic patterns (pitch, intensity, and tempo) of this speaking style together with a study of the relationship between speech and gestures. The first Chapter of this part is related to the Text-To-Speech synthesis stage, and presents two experiments. Firstly, a very preliminary experiment is described, which consists of deriving a series of global fixed rules from the analysis of the Spanish narrator of Part I and applying them to neutral synthetic speech. Secondly, a refined synthesis framework that integrates a prosodic rule-based model derived from a small but representative set of utterances to generate storytelling speaking style from neutral speech is presented. The latter experiment is conducted on increasing suspense as a proof of concept in order to show the viability of the proposal in terms of naturalness and storytelling resemblance. Finally, the second Chapter is related to the Speech-To-Speaking Avatar synthesis stage. It includes a study of the relationship between speech and gestures that derives in synchrony and emphasis rules to drive a 3D avatar in a gesture synthesis system.

6 STORYTELLING SPEECH ANALYSIS ORIENTED TO SYNTHESIS

6.1 INTRODUCTION TO THE CHALLENGE

There is a growing interest in generating expressive synthetic speech containing particular speaking styles. However, building ad-hoc speech corpora for each and every specific expressive style becomes a very daunting task. This is of special relevance for storytelling speech, where many subtle speech nuances and characters impersonations may take place. The approach at which this thesis aims is to generate the expressive storytelling categories from neutral speech by means of rule-based acoustic models derived from the analysis of small but representative corpora. To that effect, a preliminary experiment applying averaged rules derived from the previous analyses to a neutral synthesis was conducted. Later on, a more refined analysis-oriented-to-synthesis methodology and an improved synthesis framework were employed to generate synthetic increasing suspense as a proof of concept.

6.2 A FIRST STEP TOWARDS DEVELOPING A STORYTELLING SPEECH SYNTHESIZER

A preliminary step was conducted to subjectively validate if resynthesizing neutral synthetic speech with averaged rules using the TTS synthesizer of LS-URL is sufficient to generate synthetic storytelling speech. In the following Sections, the synthesis framework and the perceptual evaluations are explained.

6.2.1 *The storytelling US-HNM synthesis framework*

A diagram showing the synthesis framework to generate storytelling speech is depicted in Fig. 20. The storytelling text is fed to the US TTS system of LS-URL in order to retrieve a neutral synthetic utterance. This utterance passes through a HNM analysis step that parametrizes the speech signal. Then, the averaged prosodic rules are applied to the HNM parameters as a constant for the whole utterance. The preliminary prosodic rules were derived by averaging the $F0_{mean}$, the int_{mean}, and the AR of each expressive category and then dividing each category prosodic value by the neutral category of storytelling. In the end, the final resynthesis with the desired storytelling expressive category is obtained.

The modifications and final signal resynthesis were done using a MATLAB implementation (cf., Calzada and Socoró, 2012) of the HNM technique (Laroche et al., 1993). The HNM is based on the fact that the speech signal is composed of a deterministic (or harmonics) and a stochastic (or noise) component. On the one hand, the lower band of the spectrum is modelled by means of a sum of harmonically related sinusoids, which characterizes the voiced part of speech with amplitudes, frequencies, and phases of the sinusoids. On the other hand, the unvoiced parts of speech are modelled by the stochastic component. Specifically, spectral and temporal fluctuations are represented by LPC coefficients and energy. This parametrization into two components allows for a flexible manipulation of the acoustic features.

In contrast to other implementations where the maximum voiced frequency is allowed to vary (Stylianou, 1996), the considered implementation fixes it at 5Khz based on (Erro, 2008). Only prosodic modifications are considered since the current TTS does not allow VoQ modifications.

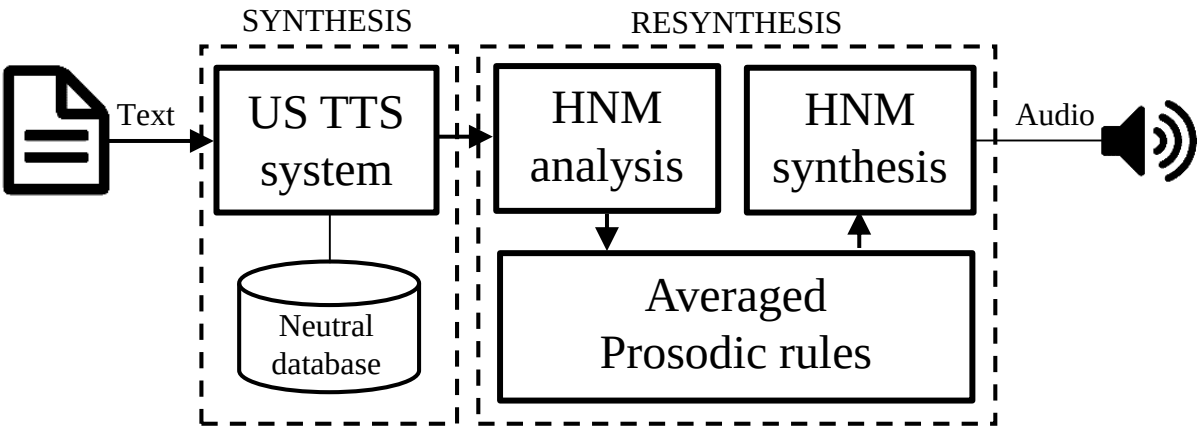

FIGURE 20: STORYTELLING UNIT SELECTION+HARMONIC PLUS NOISE MODEL TEXT-TO-SPEECH SYNTHESIS FRAMEWORK.

6.2.2 *Speech synthesis evaluation*

The averaged prosodic rules were applied to a randomly selected set of sentences from the corpus at hand. 52 sentences (4 sentences for each category) were resynthesized with the obtained prosodic rules (PR) and the same 52 sentences applying the original prosody (OP) of each utterance (the baseline for comparison).

The synthetic results were evaluated using the TRUE online platform (Planet et al., 2008). The subjective test was performed by 15 people, from which 9 are male and 6 female with a mean age of 34 (only 5 people are familiar with the field of speech technologies). The question proposed to those who took the test was: *"Which audio do you think it resembles most to the reference audio in terms of expressiveness?"*, where the reference audio was the original utterance from the audiobook. This perceptual test is designed considering a 5-level Comparative Mean Opinion Score (CMOS) scheme (OP much better, OP better, no difference, PR better, and PR much better), and it is composed of 52 comparisons of the same sentence resynthesized with PR and OP, which are compared including the original sentence of the audio book as a reference.

As a general result, it can be observed from Fig. 21 that the most extreme cases of the 5-level CMOS range are the least chosen options, showing that both transformations (PR and OP) are perceived similarly. However, post-character sentences tend to be preferred when the original prosody is applied. This can be due to the fact that sometimes the narrator maintains emotional traces from the previous character intervention whereas in other post-character sentences he barely is expressive. Suspense sentences have obtained very good results when synthesized with the PR.

6.2.3 *Discussion*

In general, the results add further evidence that there are expressive categories inside the storytelling speaking style that show specific prosodic cues and can be modelled for synthesis purposes. However, informal perceptual evaluations comparing the resynthesized speech (using the prosodic rules) to the original utterances of the professional storyteller still showed that there is a long way to accomplish a strong resemblance. In the following Section, a new synthesis framework and an analysis-oriented-to-synthesis methodology that aims to bridge this gap is presented.

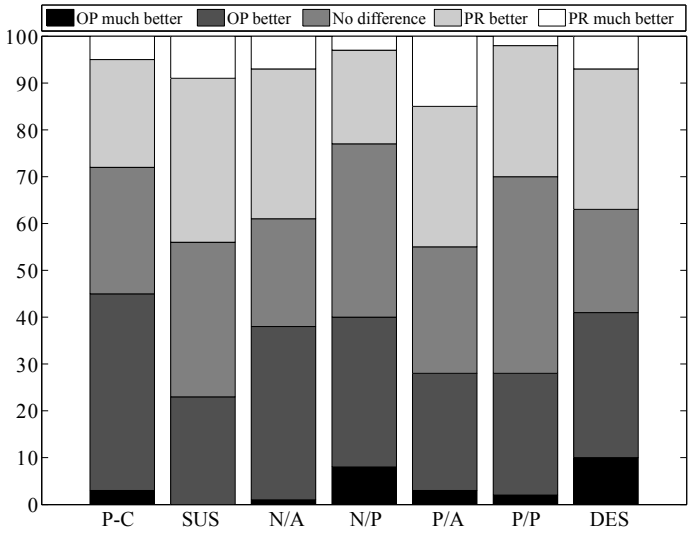

FIGURE 21: PERCENTAGES BARS OF THE RESULTS FROM THE INDIRECT DISCOURSE SYNTHESIS EVALUATION. P-C: POST-CHARACTER, SUS: SUSPENSE, N/P: NEGATIVE/PASSIVE, N/A: NEGATIVE/ACTIVE, P/P: POSITIVE/PASSIVE, P/A: POSITIVE/ACTIVE, DES: DESCRIPTIVE.

6.3 A STEP FURTHER: STORYTELLING RULE-BASED PROSODIC MODELS IN A US-AHM FRAMEWORK

In this Section, an analysis-oriented-to-synthesis methodology that derives in a rule-based prosodic model is applied to one storytelling expressive category as a proof of concept: the increasing suspense. This category observed and defined by Theune et al. (2006) as those situations where the dramatic event is expected in advance and the suspense is built up until a pause, which is followed by the revelation of the important information.

The authors defined a set of fixed prosodic rules for the increasing suspense based on the analysis of *one* sentence uttered by a professional actor. The acoustic characteristics observed in that utterance were a gradual increase in pitch and intensity, accompanied by a decrease in tempo. Then, a pause was present before the description of the actual dramatic event. Thus, this type of suspense was divided into to zones: before (zone 1) and after (zone 2) the pause. Next, the prosodic modifications were applied to a neutral synthetic utterance generated with the Fluency Dutch TTS system. In the first zone, a sinusoidal function applied to stressed syllables was proposed to model the gradual increase of pitch (from +25 to +60Hz), whereas a constant increase up to +10 DB (on the whole signal) and +150% (on stressed vowels) was considered for intensity and duration transformations, respectively. In the second zone, pitch and durations gradually decreased to their normal values, whereas for intensity an increase of +6 DB was applied to the first word with no further modifications afterwards.

In this work, the rule-based prosodic model of increasing suspense is derived by considering a reduced set of representative utterances (around 30 sec. of speech). Such model is included in a hybrid US plus adaptive Harmonic Model (AHM) (cf., Degottex and Stylianou, 2013) synthesis framework. The AHM technique is considered instead of HNM (see previous Section) as it has been proved to provide better synthesis quality than the HNM technique (Kafentzis et al., 2014; Hu et al., 2013).

6.3.1 *The storytelling US-aHM synthesis framework*

The US-AHM TTS synthesis system depicted in Fig. 22 builds on the idea of enabling US-TTS synthesis to manage different expressive styles within the same synthesis framework (Alías et al., 2008).

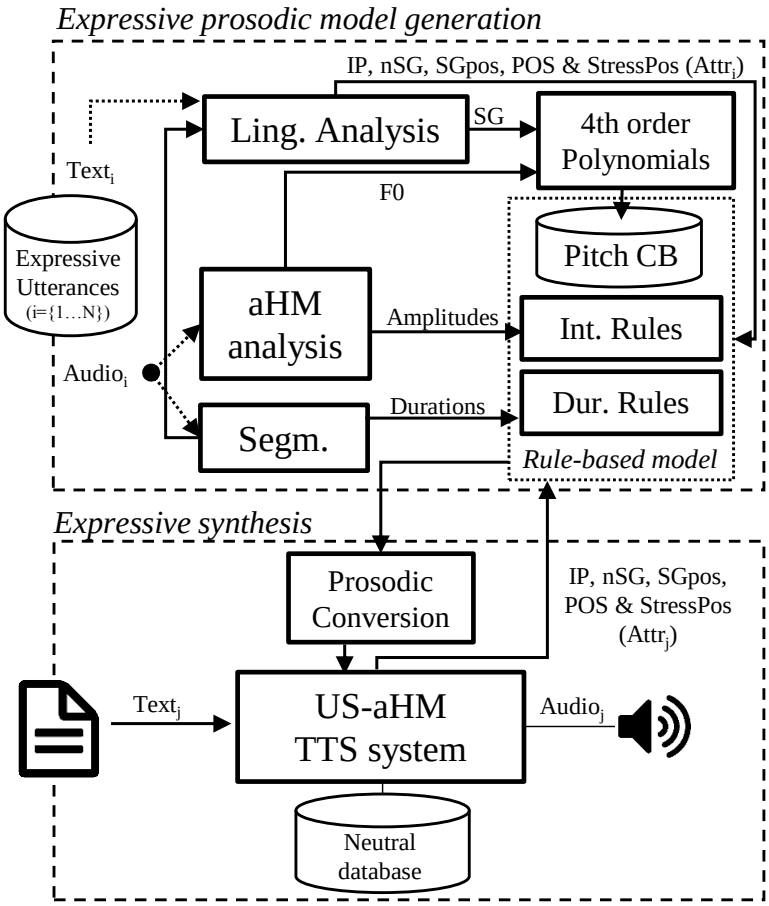

FIGURE 22: HYBRID UNIT SELECTION-ADAPTIVE HARMONIC MODEL TEXT-TO-SPEECH EXPRESSIVE SYNTHESIS FRAMEWORK BASED ON A RULE-BASED PROSODIC MODEL.

The process starts by building the rule-based prosodic model from utterances containing the desired expressive speaking style. During the synthesis stage, the TTS system converts any input text to the target expressive speaking style from a neutral Spanish female voice.

6.3.1.1 *Expressive prosodic model generation*

Firstly, it is worth remarking that the basic intonation unit considered in the introduced synthesis framework is the Stress Group (SG) (Erro et al., 2010; Iriondo et al., 2007). In this work, the SG is defined as a stressed syllable plus all succeeding unstressed syllables within the same compound sentence (Thorsen, 1978). This way, the definition favours having enough examples with relatively similar F0 contours when dealing with few utterances. As it can be observed in Fig. 22, each selected expressive utterance is linguistically analysed and segmented. As a result, the following SG-level attributes are extracted (see Figs. 22 and 23):

- **Intonational Phrase (IP)**: This attribute identifies to which IP within the utterance the SG belongs to.
- **nSGs**: Refers to the total number of SGs within each IP.
- **SGpos**: The SGpos indicates the position of the SG within each IP, differentiating PRE (unstressed SG of initial position), BEG (Beginning), MID (Middle), PEN (Penultimate), and END (Final SG).

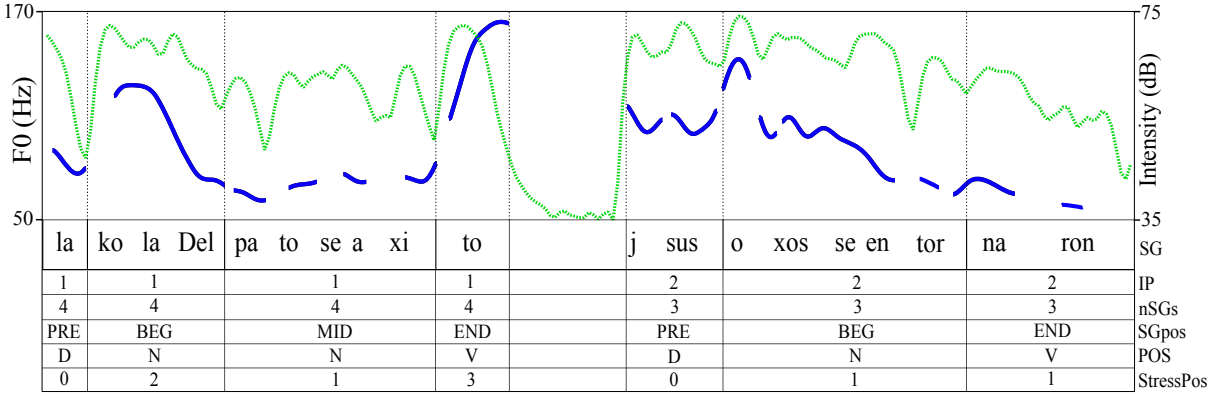

FIGURE 23: INCREASING SUSPENSE EXAMPLE: *La cola del pato se agitó, y sus ojos se entornaron* ("THE DUCK'S TAIL TWITCHED, AND ITS EYES NARROWED"). STRESSED SYLLABLES ARE IN BOLD. THE PHONETIC TRANSCRIPTION OF THE STRESS GROUP TIER IS IN SAMPA FOR SPANISH (WELLS, 1997). BLUE SOLID LINE: FUNDAMENTAL FREQUENCY. GREEN DOTTED LINE: INTENSITY.

- **Part Of Speech (POS)**: Freeling POS labels for Spanish are used (cf., Lloberes et al., 2010) to identify POS.
- **StressPos**: For the computation of this attribute, each SG is divided into 3 parts of equal duration. Then, each SG is labelled according to the position of the stress (middle of the stressed vowel), i.e., first (1), second (2), and third (3) SG part. Unstressed SGs before the first stressed syllable are represented with a value of 0.

The expressive utterances considered to derive the prosodic model are also analysed by means of the AHM technique implemented in the COVAREP (version 1.4.1) algorithms (Degottex et al., 2014) to extract the F0 and amplitude parameters. F0 curves are obtained for each SG by considering both the AHM F0 parameters and the SG segmentation. The SG-level attributes together with the 4th-order coefficients obtained from the polynomial fitting of SG F0 curves (Iriondo et al., 2007) are used to define each SG codeword (i.e., a vector containing attributes and polynomial coefficients). Finally, these codewords are stored in the F0 Codebook (CB). Regarding intensity and durations, a series of rules are also derived from a detailed analysis of the utterances.

6.3.1.2 *Expressive synthesis stage*

At run time, the input text to be synthesized is fed into the US-AHM TTS system. The TTS system extracts the aforementioned linguistic attributes and accesses the rule-based prosodic model to apply the corresponding expressive prosodic conversions (see Fig. 22). After retrieving the selected units from the neutral speech database, the corresponding AHM parameters are converted according to the target expressive style. Finally, the AHM-based synthesis generates the synthetic expressive speech. Pitch modification is performed by following the procedure described by Kafentzis et al. (2014), but with two slight different variations (Erro et al., 2007).

It is worth noting that the approach entails that the final curve may consist of a combination of F0 patterns from different utterances. Thus, a simple yet effective combination cost is defined to assess which combinations are more suitable to be concatenated. Concretely, if two consecutive SG F0 curves come from different utterances, the cost is increased by 1 and 0 otherwise. As a result, several combinations may contain the minimum cost. In these cases, following a similar approach to (Alías et al., 2005), the final combination is randomly chosen in order to increase synthesis variability, and a interpolation technique is also included to avoid discontinuities between F0 curves of consecutive SGs.

6.3.2 *Developing a rule-based prosodic model of increasing suspense*

In this Section, the increasing suspense speech material considered in the subsequent analyses is detailed, followed by the description of the actual analyses conducted to derive the rule-based model.

6.3.2.1 *Material*

The increasing suspense speech was obtained from an audiobook interpreted by a Spanish professional male storyteller. The storyteller interpreted a story that belongs to the fantasy and adventures genres (with children and pre-teenagers as its main target audience). The audiobook contains around 4 hours of storytelling speech. However, only eight utterances that fully fit the expressive profile of increasing suspense have been found. All the utterances were manually segmented to allow reliable subsequent analyses. Fig. 23 depicts an example of the complete labelling at the SG-level of an increasing suspense utterance.

6.3.2.2 *Analysis oriented to synthesis*

In this Section, the analysis of each prosodic parameter is described together with the resulting rules.

Fundamental Frequency—Similarly to Theune et al. (2006), a tendency consisting of a F0 increase along zone 1 and a gradual decrease in zone 2 has been observed in all the utterances. However, not all the utterances show a gradual F0 increase in all the stressed syllables of the first zone. For instance, in Fig. 23 it can be observed that the word *"pato"* ("duck") is not F0-accented in the stressed syllable. On the contrary, the F0 curve drops as if the storyteller wanted to emphasize even more the last SG "*agitó*" ("twitched"). This phenomenon also manifests in the rest of utterances without a gradual increase, being related to the POS of the SG. Other examples can also be an adjective complementing a verb, e.g., "***era*** *evidente*" ("it **was** clear"), or and adjective complementing a noun, e.g., "***hombre*** *alto*" ("tall **man**"). Another clear pattern observed in all the utterances is a substantial rise of F0 in the last SG of zone 1. This rise is preceded in all cases by a downfall except if the penultimate SG of zone 1 is a verb, e.g., *"**inundó** la habitación"* ("**flooded** the room"), where two F0 rises are present (reaching a higher point in the last SG). Finally, within zone 2 the only clear pattern observed is a F0 boost in the first SG whose POS corresponds to a verb, a noun, an adjective, or an adverb, accompanied with a gradual decrease until the end of the utterance. In short, the F0 rules derived from the study of the analysed utterances are the following for the first zone:

- The F0 curve of each SG of the input text is retrieved according to its position within the zone (note that the IP is equivalent to the zone in increasing suspense) and its stress position, in that order.
- In case of having more than one MID SG, the POS is also considered (before the stress position) in order to establish which SG should be F0-accented.
- A relevance score is assigned to each POS label according to the POS importance. From most to least relevant: verbs, nouns, adjectives/adverbs, and rest. If a SG complements another SG, its relevance score is degraded (except in verbs) to retrieve a not F0-accented curve.

For the second zone:

- The SG F0 curve is retrieved according to the number of SGs, the SG position, and the stress position, in that order.
  Finally, since two different speakers are considered, a scaling of F0 curves is mandatory.

Intensity—Similarly to what was observed in the analysis of F0, the gradual intensity increase reported by Theune et al. (2006) was not observed either within the analysed material. Therefore, we opted for modifying energy coherently with the F0 curve following (Sorin et al., 2015), which is based on the fundamental relationship between the instantaneous F0 and instantaneous energy of a speech signal. In

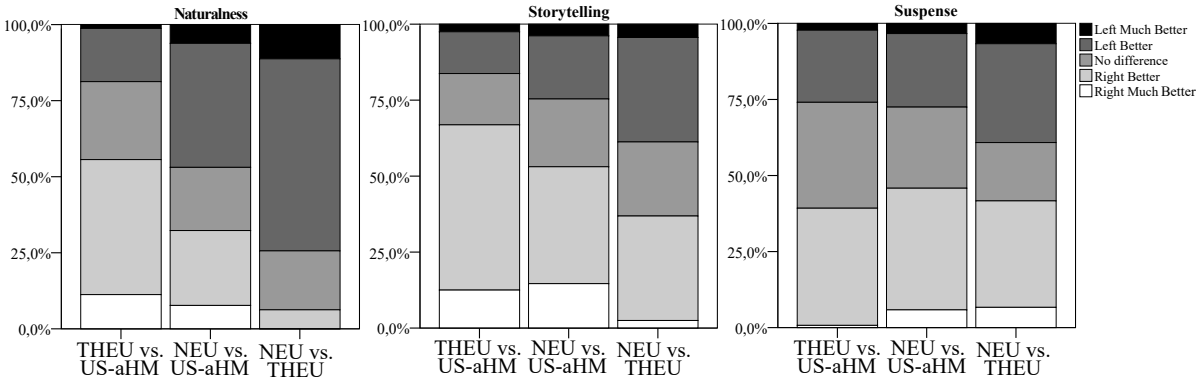

FIGURE 24: PERCENTAGE BARS REPRESENTING THE ANSWERS OF THE SUBJECTS FOR EACH EVALUATION. NEU: NEUTRAL; THEU: THEUNE ET AL. (2006)

order to validate this approach, a correlation analysis between F0 and intensity curves was performed in the set of increasing suspense utterances obtaining a value of $r = 0.654$ and a linear regression slope of 9.8 DB/octave. These values confirm the viability of the considered approach as they are very similar to the $r = 0.670$ and 9 DB/octave obtained in (Sorin et al., 2015).

Duration—Theune et al. (2006) observed a pause of 1.04 s between both zones in their utterance. However in the considered set of utterances, such pause duration is much lower (mean duration of 0.4 s$\pm$0.1 s). Furthermore, Theune et al. (2006) observed a progressive increase of stressed vowels durations in the first zone. This pattern was detected in one of the eight increasing suspense utterances. Nevertheless, as 7 out of the 8 sentences did not present that pattern, this rule of Theune et al. (2006) was not included in our rules. Despite further detailed analyses of rhythm patterns and changes of speech tempo between both zones, no clear patterns whatsoever were found. Therefore, in this work, the only duration rule included in the synthesis framework is to apply a value of 0.4 s to the pause between both zones.

6.3.3 *Perceptual evaluation*

The perceptual evaluation was conducted by means of a 5-point scale ($[-2, +2]$) CMOS on 5 synthetic utterances using the TRUE online platform (Planet et al., 2008). Such utterances, were generated from made-up sentences with a semantic content related to stories, e.g., *"Caperucita llamó a la puerta, pero nadie contestaba"* ("Little Red Cap knocked on the door, but no one answered"). In each comparison, two utterances synthesized through the AHM-US TTS framework were presented to the evaluator (randomly ordered in each comparison), using either the introduced rule-based prosodic model, the fixed rules of Theune et al. (2006), or the neutral synthetic speech as baseline (5 utterances x 3 methods = 15 comparisons).

All subjects were asked to relatively grade both speech fragments in terms of naturalness, storytelling resemblance, and expression of suspense. As no specific target was available, no reference audio was included to avoid biasing the CMOS towards the presented method if some of the of the prosodic patterns of the analysed utterances were included. It is worth noting that three control points were added to remove unreliable evaluators from subsequent analyses (18 comparisons plus a final survey in total). From the total of 32 subjects (mean age 34$\pm$10), 4 were discarded for the aforementioned reliability criterion. The results from the subjective test were analysed in terms of percentage scores (see Fig. 24) and differences in the CMOS Median (MDN) values. The latter, were analysed by means of a one-sample Wilcoxon signed-rank test with significance level $p < 0.05$.

Regarding naturalness, our approach significantly[12] outperforms Theune et al. (2006) (MDN = 1; 55% US-AHM better/much better) and it is perceived equal to the neutral synthetic counterpart (MDN = 0; 53% US-AHM no difference/better/much better). On the contrary, the method of Theune et al. (2006) obtains significantly lower results than the neutral synthetic speech (MDN = -1; 74% neutral better/much better). Moreover, storytelling quality results indicate that the proposed method outperforms both Theune et al. (2006) (MDN = 1; US-AHM 63% better/much better) and the neutral synthetic speech (MDN = 1; US-AHM 53% better/much better). Differently, Theune et al. (2006) is perceived similar to neutral in this evaluation (MDN = 0; neutral 63% no difference/better/much better). Finally, results regarding the expression of suspense show that all methods are perceived similarly, even though it is to note that the proposed method is perceived as slightly better with respect to Theune et al. (2006) (26% preferred Theune et al. (2006) and 40% preferred the US-AHM method) together with a significant preference in front of the neutral synthesis (MDN = 1; US-AHM 48% better/much better).

6.3.4 *Discussion*

The perceptual evaluation has shown that the introduced approach is specially successful in terms of naturalness and storytelling resemblance. Regarding naturalness, no significant degradation in speech quality has been observed with respect to the neutral synthetic counterpart, even when dealing with three consecutive prosodic modifications (F0, duration, and intensity). Concerning storytelling resemblance, even not presenting an audio reference, some evaluators remarked that they took into account the rich expressiveness used by storytellers. However, in relation to suspense expression, others commented that a warmer and more whispery voice could improve the suspenseful feeling, which can be related to the little suspense observed in the results. From these comments and the results, it is concluded that voice quality could be also necessary to fully resemble suspense. Another interesting observation is that some subjects suggested to increase the pause duration whereas others stated the opposite, which may indicate that more than one pause duration profile exist for this type of suspense. In this sense, a deeper analysis considering more data with several narrators/languages might clear up this and other aspects. Finally, note that the defined rules might only apply for Spanish, although comparable acoustic patterns among storytellers of similar linguistic communities have already been observed in the second part of the thesis.

6.4 CONCLUSIONS OF PART III-6

In this part of the thesis, a first attempt to generate synthetic storytelling speech has been described. Firstly, a series of global averaged prosodic rules derived from the analyses of Parts I & II for each storytelling expressive category have been implemented in a HNM synthesis phase. The synthesis evaluation has shown a first confirmation that there are expressive categories inside the storytelling speaking style that show specific prosodic cues, which can be used as rules for synthesis purposes. However, the obtained synthetic speech still showed that there were quite room for improvement to fully resemble a real storyteller. Being so, a hybrid speech synthesis framework based on US and AHM has been employed to generate storytelling speech using a rule-based prosodic model derived from the analysis of few but representative utterances of increasing suspense (less than 1 min of speech). The US-AHM approach has been evaluated on a subjective test comparing it to the fixed prosodic rules introduced by Theune et al. (2006), using the neutral synthetic speech as basline. The proposed method obtains good naturalness and storytelling resemblance, although it is similar to the other methods in terms of suspense arousal.

[12] When the words significant/outperform are used in the text we refer to $p < 0.05$ and if the words similar/equal are used we refer to $p > 0.05$.

7

ANALYSIS OF THE INTERACTION BETWEEN SPEECH AND GESTURES ORIENTED TO SYNTHESIZING ANIMATIONS

In this Chapter, a study of the relationship between gestures and expressive speech is performed with the objective of enhancing gesture animations of a virtual storyteller intended to be developed in future works within the GTM of LS-URL. This work builds on the idea of generating appropriate gestures to a speech input (see Fig. 25) with varying emphasis, so as to increase the credibility of a storyteller avatar.

7.1 AN INTRODUCTION TO GESTURES

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 they entail communicative goals. They are part of the body language

According to McNeill (1992), gestures can be classified in a four class taxonomy:

- **Iconic**: Represent concrete entities or actions.
- **Metaphoric**: Refer to abstract concepts.
- **Deictic**: Indicate locations in space
- **Beats**: Do not contain semantic content.

Beats usually involve simple movements such as up-and-down or back-and-forward motions, and are the most used gestures constituting about half of all gestures (McNeill, 1992). The present study is based on the generation of gestures based solely in prosody, ignoring the semantic content of the accompanying speech signal. The semiotic value of a beat is similar to emphasis (McNeill, 1992), and it is known that prosody correlates well with emphasis (Terken, 1991). This suggests a possible relationship between beats and prosody.

In order to study gestures in detail, Kendon (1980) proposed a segmentation scheme. He defined the terms “gesture unit”, Gesture Phrase (GP) and “gesture phase”. A gesture unit starts with a rest pose, contains one or several consecutive gestures, and ends with another rest pose. A GP is what it is

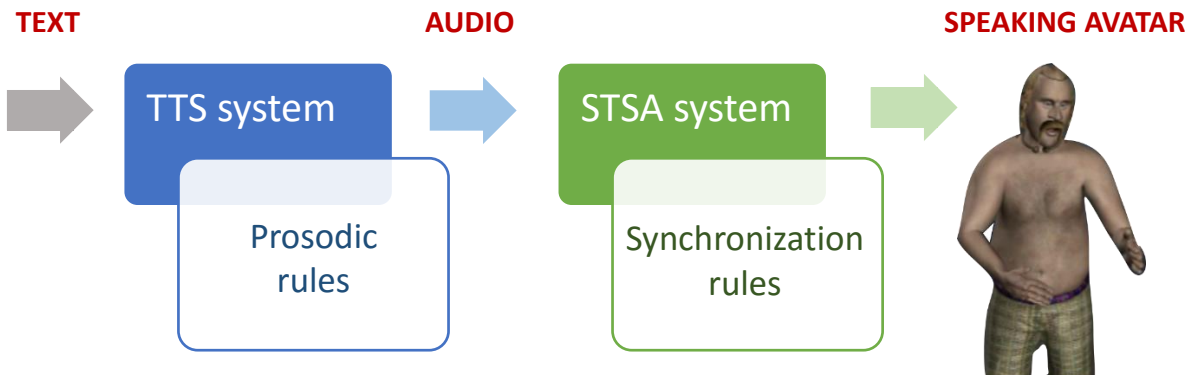

FIGURE 25: DIAGRAM REPRESENTING THE ENVISIONED STAGES OF THE PROCESS TO CREATE A STORYTELLER AVATAR.

generally called a gesture, and it is composed by several phases, some of them being mandatory and some optional:

- **Preparation** (Optional): 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.
- **Stroke** (Mandatory): It is the phase that contains the 'expression of the gesture', whatever it may be. This phase involves greater effort than any other phases. A GP must contain a stroke.
- **Retraction** (Optional): Body parts are moved to the rest position. This phase can not be present if the speaker concatenates a stroke phase with another stroke.

  Later works added other phases. For instance, Kita et al. (1998) identified another phase:

- **Pre-stroke or post-stroke holds** (Optional): These are temporary cessations of movement that occur immediately before or after a gesture stroke.

  Finally, Kipp (2004) suggested another one:

- **Recoil phase** (Optional): Occurs after a forceful stroke when the hand lashes back from the end position of the stroke.

7.2 GESTURES AND SPEECH MODELLING

In this section the analysis of gestures and speech carried out in order to determine synchrony and intensity rules is detailed. Section 7.2.1 describes the corpus used in the analysis. Next, Sections 7.2.2 and 7.2.3 explain the studies focused on defining intensity variables for gestures and speech. Finally, after a correlation analysis (Section 7.2.4), the synchrony and intensity rules are obtained.

7.2.1 *Audiovisual Corpus*

In order to perform the following experiments it is necessary to capture synchronized gesture and speech data. A corpus composed of motion capture data and video was recorded to obtain animations and their corresponding video recordings. The corpus consists of 6 clips that last slightly more than one minute each, in which an amateur actor with MOtion CAPture (MOCAP) recording experience was asked to perform an improvised monologue with a concrete speaking style and performing only beat gestures. The selected styles were aggressive and neutral, as they cover a considerable spectrum of the activation dimension (Russell, 1980). These 2 styles were recorded in order to obtain a rich gestures database, as some specific gestures are correlated to certain emotions (Kipp and Martin, 2009), and gestures differences between emotions can be explained by the dimension of activation (Wallbott, 1998).

For each style, three clips were recorded. The recording session took place at the Medialab, Llull., 2012. This laboratory has 24 Vicon MX3 cameras that allow full-body motion capture with a frame rate of 120 samples per second. The recording set-up is illustrated in Fig. 26.

The video channel was captured with a fixed-position video camera, whereas the speech signal with a clip-on microphone attached to the flap on the actor's suit (see Fig. 26). Both channels were saved in a single video file. Hereafter, the video is used for the following purposes: to extract the prosody parameters, temporally aligning motion data with audio data, and to analyse the shape and timing of gestures.

The data files obtained by both equipments start at slightly different times as they are manually activated. Later, these data channels are aligned in time so as to reproduce the original sequence on the original animation mode. Anchor points in both video and motion data were created thanks to a T-pose (see Fig. 26) in order to make this alignment possible.

FIGURE 26: SET-UP OF THE RECORDING SESSION. AN OPTICAL MOTION CAPTURE SYSTEM WITH 24 INFRA-RED CAMERAS, A CLIP-ON WIRELESS MICROPHONE, AND A VIDEO CAMERA (NOT IN THE IMAGE). THE ACTOR IS WEARING A BLACK MOTION CAPTURE SUIT WITH REFLECTIVE MARKERS.

7.2.2 *Gesture analysis*

Before defining the synchrony rules between gesture and speech, it is necessary to define the level of intensity or strength of the gestures. Having done this, it is possible to search for intensity correlations between speech and gestures. First of all, a video annotation phase to extract relevant information for the analysis was conducted. Then, a measure from the literature that could be suitable for the problem at hand was taken into account. After computing that measure with the extracted data, some tests were conducted using other methods in an effort to validate the results which were obtained. The validation process consisted of extracting parameters from the MOCAP data that were combined into one single measure indicating the strength level of the gesture using linear classification methods. In order to achieve the goal of the analysis, a perceptual test was previously conducted. The final measure was tested in a correlation analysis against its speech analogous.

7.2.2.1 *Video annotation*

The tool selected for the annotation of the videos was Anvil (Kipp, 2001), as it was developed for video analysis and it had already been used in gesture analysis works, e.g., by Kipp et al. (2007). Moreover, Anvil allows to load 3D MOCAP data, which was necessary for the annotation process. The annotation was divided into three tiers: left hand, right hand, and both hands. Furthermore, four annotation levels per hand were considered:

- **Phase**: In this annotation tier all gesture phases explained in Section 7.1 are annotated.
- **Stroke**: This tier contains the tags assigned in the perceptual test (see Section 7.2.2.2). There are two options: weak or strong stroke.
- **Apex**: Indicates the apex time of the stroke (the point of maximum extension) that will be compared with its speech analogous. Usually, this time coincides with the end of the stroke for discrete gestures[13], although a little recoil may occur. In continuous gestures, the apex time might not be at the end of the stroke but around the middle. As the apex time has to coincide with the point

[13] A discrete gesture can be defined as a motion that can not be decomposed into several units, e.g., a single stroke

of maximum extension of the stroke (Leonard and Cummins, 2011), the annotation differentiates between discrete and continuous gestures.

- **Velocity**: The velocity curve obtained from MOCAP data is loaded into this tier, so this process did not entail an annotation task. It is helpful in the annotation process, specially during the location of apex times. In discrete gestures, the point of maximum extension must coincide with a minimum of the curve. As mentioned before, continuous gestures may have the point of maximum extension around the middle. This is also reflected in the velocity curve as a minimum. For discrete gestures, we have also taken into account a technique that consists of analysing consecutive frames to spot the apex time by detecting a change in the blurriness of the hand (Shattuck-Hufnagel et al., 2007). Basically, the image of the hand sharpens in the point of maximum extension. However, using the velocity curve was prioritized as it is faster, easier and more reliable.

As the gesture synthesis system that will be used in the future considers one gesture as a combination of both hands, the "phase" annotation and the "apex" tier of the left and right hands are merged in the group of "both hands". If the annotations of each hand match, the corresponding annotation is introduced in the "both hands" tier. However, sometimes the annotations of the left and right hand tiers are not equal. When this occurs, a decision has to be made. In general, in cases of obtaining different phases for each hand, those of greater relevance (oriented to the gesture synthesis) have preference, i.e., in decreasing order of relevance: stroke - preparation - hold. In addition, there are two more situations regarding only the stroke phases where a decision is also needed:

1. The actor sometimes performs a strong stroke with one hand that tends to produce an almost involuntary stroke of the other hand that is significantly less important. In these cases, we take the most significant stroke.
2. Two strokes from different hands may come from the same gesture, but they start and/or end at slightly different times. The considered procedure was to select all the range covered by both strokes.

Following the adopted definition of apex (point of maximum extension), each stroke can only have one apex time.

The complete annotation process of a particular video fragment is depicted in Fig. 27.

7.2.2.2 *Classification of stroke strength*

The main goal of classifying strokes according to their strength is to evaluate if stronger accents are associated with stronger gestures, and vice versa. To that effect, the strength level of each stroke is represented as a numerical variable, which is denoted as Gesture Strength Indicator (GSI). This variable results from the computation of the FMDistance (Onuma et al., 2008). FMDistance is a distance function that approximates the kinetic energy of the rigid body (or set of bodies) attached to a joint, and it is used for classifying motion capture sequences. In this work, only the upper-body movement is taken into account. Thus, both forearms, arms and shoulders are selected. It is also worth noting that logarithmic computation of FMDistance reported by Onuma et al. (2008) is computed, following the authors recommendations.

In order to validate this variable, a perceptual test was conducted. Specifically, two experts on video annotation carried out the first stage of the evaluation by tagging a total of 792 stroke phases, which were annotated along all the videos as weak or strong strokes according to their perception. After this process was finished, a third expert annotator determined the final tag in case of disagreement.

In order to quantify how well the FMDistance discriminated between weak and strong strokes, we selected the Matthews Correlation Coefficient (MCC) (Matthews, 1975), as it measures the quality of

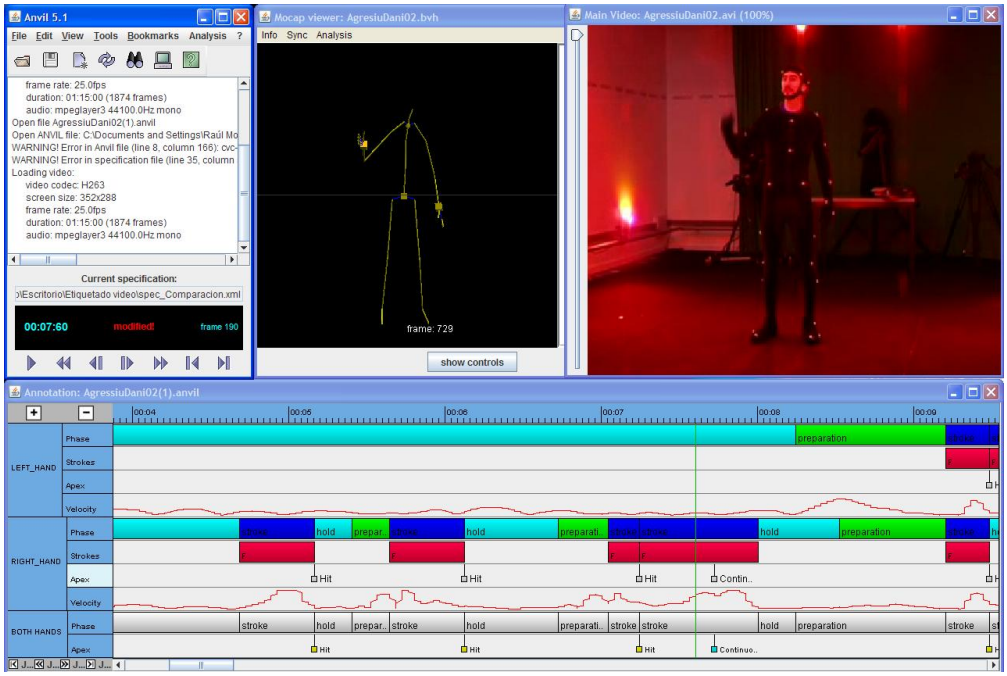

FIGURE 27: EXAMPLE OF THE ENTIRE ANNOTATION OF A PARTICULAR VIDEO FRAGMENT IN ANVIL. THE UPPER THREE WINDOWS ARE (FROM LEFT TO RIGHT): THE COMMAND WINDOW, THE MOTION CAPTURE VIEWER AND THE VIDEO. THE LATTER TWO ARE MANUALLY SYNCHRONIZED. THE BOTTOM WINDOW IS THE COMPLETE VIDEO ANNOTATION.

binary classifications and, unlike the F-measure (van Rijsbergen, 1974), it also takes into account "true negatives" of the resulting confusion matrix. MCC returns a value between -1 and +1, being -1 total disagreement between prediction and observation, +1 a perfect prediction, and 0 no better than random classification. The threshold value to distinguish between weak and strong strokes was experimentally determined for each method in order to maximize the MCC.

The MCC obtained for the FMDistance is 0.563. This coefficient reaches 0.5 when 75% of cases are correctly predicted, so we can say that this GSI surpasses that percentage.

As well as considering the FMDistance, some other available parameters from the MOCAP data were analysed in order to validate the FMDistance. The parameters were: trace (covered distance of the hand), duration, maximum velocity, accumulated kinetic energy, maximum acceleration, maximum deceleration, and velocity range. More than 100 classification configurations mixing these parameters were generated using logistic regression in Weka (Hall et al., 2009), to choose the best combination of parameters that could lead to the most appropriate measure of gesture strength. The F-Measure was observed so as to evaluate which parameters performed better. F-Measure measures the classification accuracy and it considers both the precision and the recall and, in this case, results as the harmonic mean of both. FMDistance, velocity range, and maximum velocity obtained the highest scores so, eventually, classification were conducted using these parameters. Different linear classification methods were used, as Fisher's Linear Discriminant Analysis and some methods implemented in Weka: Logistic Regression, SVM using Sequential Minimal Optimization (SMO) (Platt, 1999), and the stochastic variant of the Primal Estimated sub-GrAdient SOlver for SVM (SPegasos) (Shalev-Shwartz et al., 2007). In Weka, these methods return the coefficients used for the linear combination, so the implementation of the classification algorithms could be developed and, eventually, compute the MCC.

TABLE 20: MATTHEWS CORRELATION COEFFICIENT RESULTS OBTAINED BY CLASSIFIER AFTER INTRODUCING FMDISTANCE, VELOCITY RANGE, AND MAXIMUM VELOCITY IN EACH ONE OF THEM.

Classifier	MCC
Fisher discriminant	0.562
SPegasos	0.554
Logistic regression	0.551
SMO	0.525

Table 20 shows the results obtained by classifier after introducing FMDistance, velocity range, and maximum velocity in each one of them in order to classify according stoke strength. It can be observed that FMDistance is the best GSI candidate as no method surpasses its MCC value of 0.563. Although the differences are not significant, it is to note that the motivation of this experiment was only to validate FMDistance as a measure for quantifying gesture strength. Hence, FMDistance is used in the following correlation analysis.

7.2.3 *Intonation analysis*

A similar analysis to the one described in the previous Section was carried out to analyse the intonation of the speaker. To that effect, the prosodic parameters (pitch, intensity and duration) were extracted and combined, resulting in a measure for quantifying pitch accent strength. Again, this goal was accomplished after a perceptual evaluation phase.

7.2.3.1 *Speech corpus annotation*

Praat was used for the annotation of the Spanish speech corpus at hand. First, the speech corpus was hierarchically segmented into sentences, words, syllables, and phonemes. The EasyAlign tool was used and manual corrections were applied afterwards if necessary.

The intonation system used to describe pitch accents was the ToBI system. As this annotation may entail between 100 or 200 times the real duration of the corpus, some works in the literature have tried to automate (or semi-automate) the annotation process (Syrdal et al., 2001). However, a manual tagging was performed in order to obtain reliable results while being aware of some delicate situations (e.g. pitch - doubling / halving) which, if not well processed, could distort the analysis.

Finally, two more annotation tiers were manually added to Praat. One of them is an annotation of the syllable nucleus inside the pitch accent. The other consists of a point tier which indicates the time where the pitch achieves its maximum value inside the syllable, as usually the peak is taken as a point of reference (Leonard and Cummins, 2011; Loehr, 2004). This point is denoted as the Pitch Accent Peak Time (PAPT). This labelling point is used to analyse the correlation between gestures and speech (distance between apex time and PAPT). As an example, the complete annotation of a speech fragment is depicted in Fig. 28.

7.2.3.2 *Classification of pitch accent strength*

In order to measure prosodic prominence, two measures are considered (Silipo and Greenberg, 1999; 2010), which could be appropriate for reaching the present goal. Silipo and Greenberg (1999) proposed an Evidence Variable (EV) for marking prosodic stress computed with the multiplication of mean energy

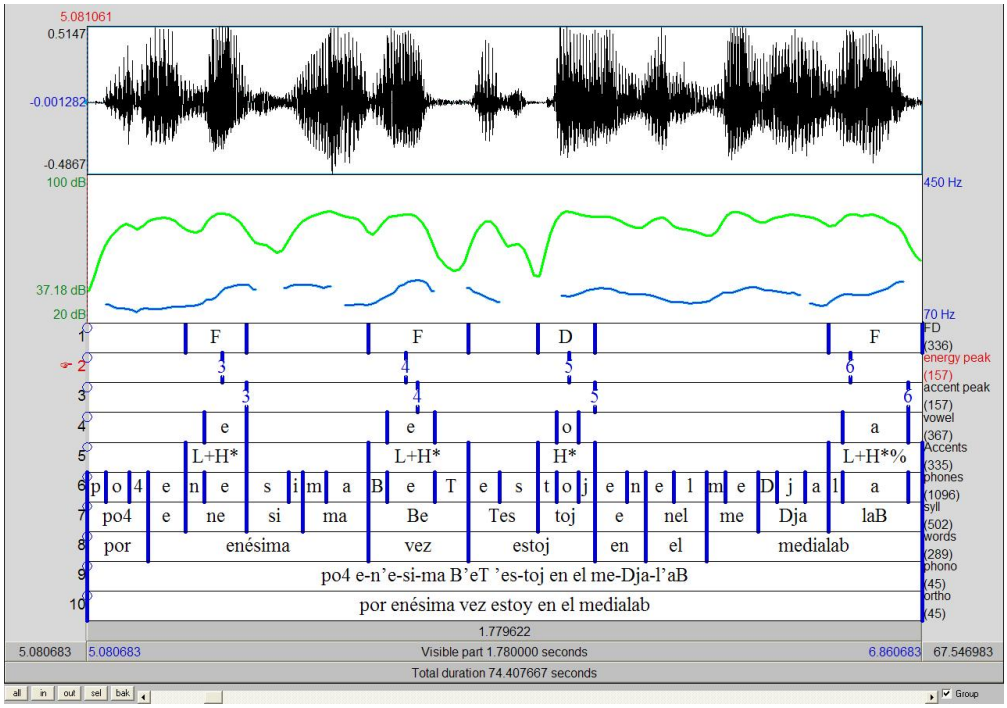

FIGURE 28: EXAMPLE OF THE ENTIRE ANNOTATION OF A PARTICULAR AUDIO FRAGMENT. THE FD ANNOTATION TIER CONTAINS THE TAGS FROM THE PERCEPTUAL TEST. SENTENCE TRANSLATION: "FOR THE UMPTEENTH TIME I AM AT THE MEDIALAB."

and duration of the syllable nucleus. More recent work has added to the EV a new parameter (called "m") which described the contribution of the pitch range (Silipo and Greenberg, 2010). The latter variable will be referred from now on as EV2. This parameter rewards situations were pitch variation inside the nucleus has a rising pattern. In our experience, situations with a decreasing pitch pattern were perceptually weak, so these cases (labelled with H+L* ToBI tags) have been penalized in the same way as Silipo and Greenberg (2010) did, which consists in forcing these cases to the inverse maximum pitch range value.

The major acoustic correlates of prosodic prominence reported in the literature are pitch movements, global energy and duration (Streefkerk et al., 1999), although some works obtained good results leaving out the pitch (Silipo and Greenberg, 1999), or adding spectral emphasis (Tamburini and Wagner, 2007). In this work, the spectral emphasis is omitted as it does not carry information about stress in Spanish (Ortega-Llebaria and Prieto, 2011).

The perceptual test to evaluate how well these variables can discriminate between weak and strong pitch accents was carried out by two experts on audio annotation separately. They were asked to annotate each pitch accent of the neutral and aggressive audiovisual corpora as weak or strong, according to their perception. Later, a third expert annotator made the final decision in the case of disagreement. The annotators based their classification on context, as it is important for the perception of prominence (Tamburini and Wagner, 2007).

Once the annotations were finished, the evaluation of the appropriateness of the evidence variables (EV and EV2) to discriminate between weak and strong pitch accents was conducted. It is worth pointing out that the objective of this work is not a classification task by itself, but this classifications will help to identify, if possible, what was perceived acoustically. MCC results for the variables can be observed in Table 21. The conclusion is that EV2 performs considerably better than the original EV.

TABLE 21: MATTHEWS CORRELATION COEFFICIENT RESULTS FOR ALL VARIABLES.

Variable	MCC
PSI	0.671
EV2	0.630
EV	0.575

TABLE 22: MATTHEWS CORRELATION COEFFICIENT RESULTS FOR ALL THE PITCH ACCENT CLASSIFICATIONS.

Classifier	MCC
SPegasos	0.662
Logistic regression	0.652
SMO	0.640
Fisher discriminant	0.636

In spite of this test, the numeric values that did not fit its correspondent strength tag were validated. It was noticed that in some situations when the speaker spoke loudly, the mean pitch was very high but the pitch contour was flat (pitch range close to zero and, therefore, EV2 close to zero). So, with the aim of improving the obtained results, the mean pitch was included in the pitch term (see Equation 7) in the EV2 equation (Silipo and Greenberg, 2010). These weights were assigned to the pitch parameters to maintain the variable between zero and one[14], and this new variable is denoted as Prosody Strength Indicator (PSI):

$$PSI = D \cdot E \cdot (0.5PR + 0.5MP) \tag{7}$$

where D is duration of the syllable nucleus, E is mean energy of the syllable nucleus, PR is pitch range of the syllable nucleus (with the applied penalty for H+L* pitch accents as explained before), and MP is mean pitch of the syllable nucleus. As it can be observed in Table 21, the MCC value of the PSI is higher than the one obtained by the EV2. It can be concluded that the use of mean pitch improves the discrimination between weak and strong pitch accents as perceived in the perceptual test.

Finally, the capability of the PSI to distinguish between weak and strong pitch accents was evaluated using the same linear classification methods as in Section 7.2.2.2. None of these methods outperformed the MCC achieved by the PSI (see Tables 21 and 22). Therefore, the PSI was selected for the subsequent analyses.

7.2.4 *Gestures and Speech Correlation*

Once the gestures and speech analyses described in the previous Sections was performed, the next step consisted of the specific analysis of gesture and speech correlations, both in terms of synchrony and intensity. The main motivation of this analysis is the definition of synchrony and intensity rules to be implemented in a gesture synthesizer.

[14]Note that the prosodic values are normalized by subtracting the minimum value of both corpora and dividing by the difference between the maximum and minimum

7.2.4.1 *Correlation analysis*

A total of 792 strokes were identified in the corpus (sum of right and left hand strokes). However, in the "both hands" group there were 484 strokes. As the number of pitch accents resulted in 882, the number of apex times is almost half of this. Therefore, the most appropriate pitch accent for each apex time had to be identified. Moreover, the GSI value for the cases where a stroke in "both hands" tier came from left and right hand strokes simultaneously had to be established. The highest GSI value was chosen in these cases as it is usually the parameter which carries more information about the speaker gesture intention.

In order to perform the correlation analysis it was necessary to undertake a new annotation phase. For this annotation stage, the "words" and "pitch accents" tiers from the Praat textgrids (see Fig. 28) were imported to Anvil (over the previous annotation). Then, two more tiers were created, one to indicate if the pitch accents had an associated stroke (YES/NO), and another one to indicate if the stroke in the "both hands" tier came from left hand (L), right hand (R) or both hands (B). As a result, each gesture could be linked with its corresponding pitch accent.

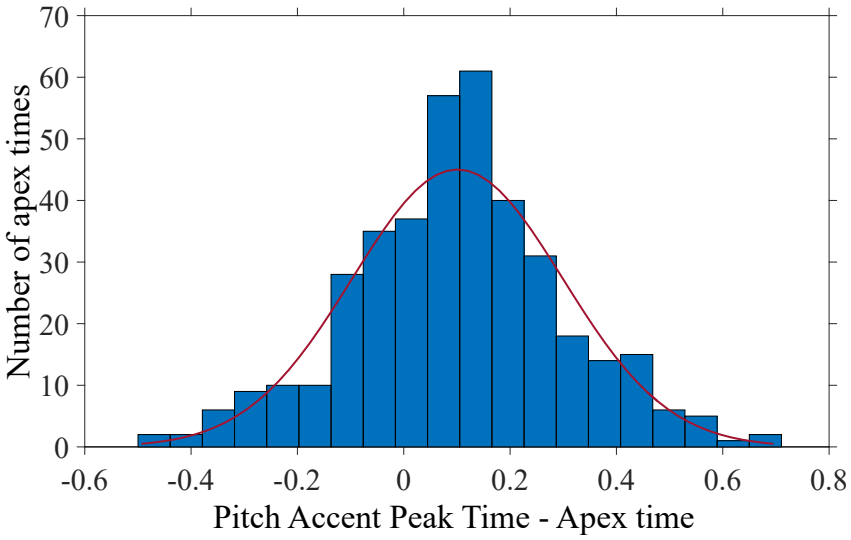

FIGURE 29: HISTOGRAM WITH THE DIFFERENCES IN SECONDS BETWEEN PITCH ACCENT PEAK AND APEX TIMES. POSITIVE VALUES INDICATE THAT THE APEX COMES BEFORE THE PITCH ACCENT PEAK. NEGATIVE VALUES INDICATE THE OPPOSITE. THE MEAN VALUE IS 0.1 SECONDS. THE STANDARD DEVIATION IS 0.2 SECONDS.

The first analysis considers differences between PAPTs and apex times. As it can be observed in Fig. 29, the distribution of differences is gaussian-like and it can be concluded that, in most cases (71.72%), the gesture precedes the pitch accent. This is in agreement with Kendon (1980) and McNeill (1992) observations, although the current approach differs as concrete speech signal and gesture anchor points are considered. Moreover, there are contradictory results in the related literature, as in (Leonard and Cummins, 2011), where the point of maximum extension of the gesture tended to occur after a pitch accent peak with a mean of (approximately) -0.1 seconds, whereas our mean is +0.1 seconds. However, Loehr (2004) observed a mean of 0 seconds and a standard deviation of 0.27 seconds.

For the definition of the synchrony rule values greater than the 75th percentile and smaller than the 25th percentile are ignored. Therefore, differences between PAPTs and apexes must fall between -0.03 and 0.22 seconds. There reason for this selection is two-fold. On the one hand, it is preferable discarding potential outliers. More extreme percentiles could have been chosen (e.g., 90th and 10th) to deal with this but, on the other hand, ensuring a well established pitch accent-gesture association in the future synthesis is primordial. As the mean difference between pitch accent peaks is 0.5 seconds, values greater

than 0.22 seconds could have brought the gesture too close to a neighbour pitch accent. This could lead to perceiving the gesture associated to a wrong pitch accent, which would result in a less natural animation.

The following analysis evaluates the relationship between pitch accent and stroke strengths. In order to compute the correlation between the GSI and the PSI, a scatter plot was generated (see Fig. 30) and the Pearson's correlation coefficient was computed. This coefficient measures the correlation of two variables (PSI and GSI, in this case), so the linear dependency can be observed. Results can be between -1 and 1. Being 1 a perfect positive correlation, -1 a perfect negative correlation and 0 means that there is no linear relationship between the variables. This coefficient resulted in a value of 0.525, which is not a high correlation but it can be interpreted as a tendency. A natural logarithm transformation was applied to the PSI variable so as to have similar values in both axes.

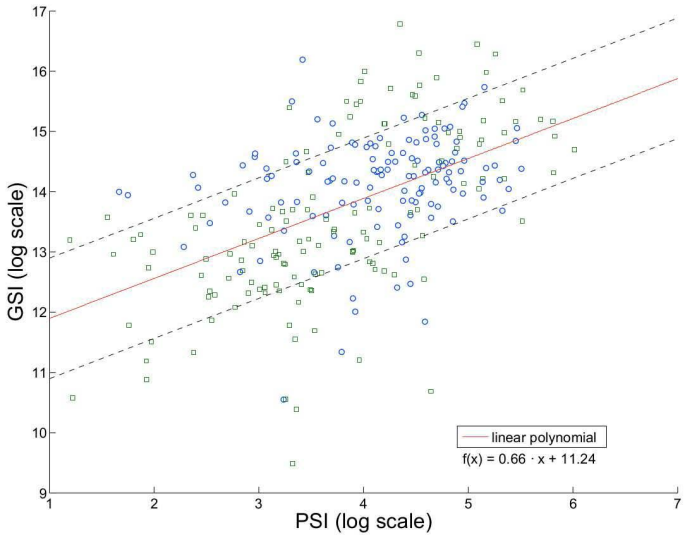

FIGURE 30: SCATTER PLOT OF THE PROSODY STRENGTH INDICATOR AND GESTURE STRENGTH INDICATOR VALUES THAT COME FROM AGGRESSIVE (CIRCLES) AND NEUTRAL (SQUARES) STYLES. THE LINEAR POLYNOMIAL STRAIGHT LINE REPRESENTS THE CORRELATION BETWEEN THE PSI AND THE GSI. THE DASHED LINES REPRESENT THE MARGIN, EXPERIMENTALLY DETERMINED, FOR THE INTENSITY RULE.

At first sight, some outliers are noticeable in Fig. 30. These outliers come from two different situations: the strength indicators do not reflect precisely what it is perceived or simply there is no relationship between the gesture and the prosodic characteristics of speech signal. The PSI and the GSI have been selected as they resulted in the best candidates to quantify speech and gesture strength in terms of the MCC, respectively. However, the PSI can be less precise in situations were the duration of the syllabic nucleus is stretched by a hesitation of the speaker (greater duration leads to a greater PSI), or when a short duration in a pitch accent perceived as strong also penalises the pitch range parameter (as it cannot increase significantly in a short duration range), for example. This proves that perceived and computation differences exist. In the same way, the GSI can also be misleading if a slow gesture has a long trace, for example, as it will most likely be perceived as a weak gesture but the kinetic energy may grow considerably. It can be concluded that there are situations where the prosodic strength and the gesture strength do not have a clean match. It is worth pointing out that in Fig. 30 the distinction between aggressive and neutral does not imply that all gestures from the aggressive style are "aggressive gestures" and the same happens for the neutral style. This distinction is only based on the data obtained from each style but, for example, some gestures in the neutral style may be considered as more "aggressive gestures" than some gestures in the aggressive style and vice versa. In spite of this, as it can be observed in Fig.

30, most of the aggressive data is spread along the top-right zone of the scatter plot while the opposite occurs for the neutral data.

7.2.4.2 *Synchrony and intensity rules*

In this section, the synchrony and intensity rules extracted from the data are detailed. These rules are obtained using all the corpus data and can be applied to both aggressive and neutral styles. These styles were selected in order to cover different levels of emphasis, but the intention is not to restrict the rules to these styles only. Other speaking styles could be suitable to work with these rules. However, this would require additional testing.

The synchrony rule is derived from the correlation analysis. A synchrony window of -0.03 to 0.22 seconds was obtained, being positive values an anticipation of the stroke apex relative to the pitch accent peak, whereas negative values indicate the opposite. This window defines the synchrony rule.

For the intensity rule, two interpretations based on the results are derived. If the linear polynomial line is taken as reference (see Fig. 30), it can be said that, given a PSI value, the GSI has to be near this straight line. However, according to the scatter plot, a margin of GSI-PSI correspondences is defined. Experimentally, this margin has been set to 1 above and below the linear polynomial line (see Fig. 30). The constants from these equations come from the analysis of the data at hand. The intensity rule is defined by the following equation:

$$GSI \in (0.6633PSI + 10.24, 0.6633PSI + 12.24) \tag{8}$$

These rules have been extracted after a meticulous analysis of the corpus at hand, and they will be used in the synthesis phase in order to recreate synthetic animations. These rules will be implemented in the cost function, which selects a GP according to the speech input. On one hand, the synchrony rule will define how the GPs and the pitch accents are synchronized and which are the desired GP durations. On the other hand, the intensity rule will drive which GP is more appropriate according to the strength indicators.

7.3 DISCUSSION

The considered speech corpora is in Spanish, so the analyses cannot be generalized to other languages at the moment. It is also worth noting that the derived rules are restricted to the corpus at hand, i.e., they are not general rules applicable to any voice, or at least this has not been tested in this work. The goal was not to search for a general rule but to look for correlations between gesture and speech and test if an intensity rule together with a synchrony rule improves the naturalness of synthetic animations.

Another interesting conclusion after finishing all the gestures and speech analysis is that it is easier to classify speech than gestures (according to their respective perceived strengths), as the MCC value of the speech classification is considerably higher. Finally, it also interesting that quantifying speech prominence is not an easy task as perceived and computed differences exist. This has also been stated for other languages like German (Tamburini and Wagner, 2007).

7.4 CONCLUSIONS OF PART III-7

In this Chapter, a study of the relationship between gestures and expressive speech in terms of temporal alignment and emphasis has been presented. For the emphasis analysis, strength indicators for gesture and speech have been introduced and validated as good candidates for representing perceived strength.

After a correlation analysis of both indicators, an intensity rule has been derived. Regarding the temporal alignment analysis, a synchrony rule has been obtained after a meticulous analysis of anchor points from gesture (apexes) and speech (pitch accent peaks).

Part IV

CONCLUSIONS OF THE THESIS AND FUTURE WORK

8

CONCLUSIONS AND FUTURE WORK

This Chapter contains the conclusions of the thesis, together with some discussion and future work. The structure of this Chapter is based on the three main objectives of the thesis. Section 8.1 is related to the first objective (defining and validating a methodology suitable for annotating indirect storytelling speech in terms of expressive categories), Section 8.2 relates to the second objective (evaluating the relevance of both prosody and VoQ to characterize the expressive categories of indirect storytelling speech across languages), and finally, Section 8.3 contains the conclusions linked to the third objective (defining a TTS synthesis framework capable of generating storytelling speech from rule-based acoustic models, together with the first steps towards developing a storytelling speaking avatar using speech-gestures synchrony rules to drive its animation).

8.1 CROSS-LANGUAGE ANALYSIS OF EXPRESSIVE CATEGORIES IN INDIRECT STORYTELLING SPEECH CORPORA: ANNOTATION METHODOLOGY

This thesis aimed from the very beginning to shed some light regarding the acoustic characteristics of a particular expressive speech: the storytelling speaking style. More specifically, the work has been focused on the analysis of the speech produced by a professional storyteller narrating fictional tales and/or stories to a young audience. We have focused on the indirect discourse parts within storytelling speech, since it usually contains subtler expressive nuances roughly studied by previous works. In contrast to direct speech storytelling (i.e., the storyteller impersonating a character), which may entail acted emotional speech, modelling the subtle expressive nuances of the indirect speech entails specific challenges when it comes to improve the naturalness of a TTS system designed to synthesize storytelling speaking style.

An annotation methodology based on storytelling discourse modes that blends text-dependent and perception-dependent categories has been proposed to deal with storytelling speech annotation at the sentence level. The lack of standardized analysis guidance to address the acoustic analysis of indirect storytelling speech made necessary to evaluate which annotation method should be considered for the aforementioned purpose. Our initial hypothesis was that the indirect speech should not be annotated with basic emotions labels, since storytellers tend to use such in some of the characters interventions, being different from the expressiveness employed when performing the indirect speech. In the indirect discourse parts, the hypothesis was that storytellers aimed at eliciting emotions in the audience rather than acting them, thus, subtler expressive nuances may appear in their narration. In any case, again, this were hypotheses to be tested in the conducted experiments.

The annotation process has resulted in several sub-modes denoted as expressive categories: neutral, descriptive, post-character, suspense, negative/passive, negative/active, positive/passive, and positive/active, which have been identified in the four versions of the story. This annotation methodology has resulted in 84.8%, 83.7%, 83.3%, and 87.8% successfully classified utterances from the Spanish, English, German, and French corpora, respectively. Although these percentages do not measure how good is the annotation methodology against other counterparts, they show that even facing a problem without standard guidance, the different inter-annotator agreements, etc., the retained corpora for the subsequent analyses have been considerable and comparable across languages. Note that those discarded utterances that have been classified as “Other” could also be tackled in the future to analyse, e.g., the paralinguistic elements that they contain such as, laughter, yawns, etc. Furthermore, the preliminary synthesis evaluation performed in Part III has shown a first confirmation that there are expressive categories inside the

storytelling speaking style that show specific prosodic cues, since the derived averaged prosodic rules have been perceived similarly to the original prosody.

To assess the difference between storytelling expressive categories and basic emotions, the former and the emotional categories of LS-URL LAICOM-UAB speech corpora have been represented together in a 3D common acoustic space via MDS using the same prosodic and VoQ features. The results have shown that emotions are placed in more extreme regions of such space, while the position of the indirect speech categories graphically supports the subtler expressiveness they contain. In line with this result, the macro-averaged F1 obtained for the classification of storytelling expressive categories is much lower ($F_1^M = 0.503$) than the one obtained from the analysis of the emotional corpus ($F_1^M = 0.926$). Low macro-averaged F1 values of 0.335, 0.356, and 0.364 have also been obtained for the English, German, and French versions of the story, respectively. These results evidence that relating the indirect speech of storytelling corpora to basic emotions may not be the best solution, since it seems that the latter show more exaggerated acoustic characteristics.

Overall, the results presented in this work have validated the proposed annotation methodology as one viable solution to tackle the analysis of storytelling speech corpora. Nevertheless, other annotation methods could also be explored in future works and a comparison between them could also be performed since, obviously, the introduced methodology has been defined and proved for a particular type of story.

Furthermore, it is worth to remark that some of the storytelling expressive categories observed in the parallel corpora at hand might not be present in all tales and stories, whereas other expressive situations may be yet to be modelled. For instance, some tales do not contain explicit dialogues between characters and, thus, post-character utterances could not be studied. Some of the phenomena (laughter, yawns, etc.) observed in utterances currently denoted as 'Other' could be specifically analysed in future works. To this aim, it might be necessary to collect more evidences to obtain reliable results. Concerning potential cross-genre generalization of the obtained results, it seems logical to think that the farther the genre and target audience are from the analysed ones (i.e., fantasy-adventures and young people, respectively) the more arguable it would be to export the obtained results to it.

Although the annotation of the indirect discourse has been manually addressed in order to avoid potential errors of a fully automatic method, a counterpart automatic approach could be developed in the future. It is worth to say that the valence/activation scheme for affective categories was selected bearing in mind this automation in future works. Concretely, a valence-based scheme has already been satisfactorily used to classify affective text (see, Trilla and Alías, 2013, and references therein), whereas the activation dimension can be quite well discriminated in terms of acoustic patterns (Schröder, 2004; Schuller, 2011; Nicolaou et al., 2011). For instance, the development of such automatic annotation methodology could be tackled using the following hierarchical approach. Firstly, the input story/tale would be initially analysed at the text level, being classified into storytelling discourse modes. Next, a second text analysis step, this time only focused on the narrative mode (as the descriptive mode is already composed of descriptive sentences), would be conducted to classify sentences according to valence sub-modes and post-character sentences, which has already been proved feasible (Trilla and Alías, 2013; Zhang et al., 2003). Finally, the analysis at the acoustic level would be asked to disambiguate the expressive categories according to a set of acoustic rules.

In the future, we plan to include such automation of the storytelling annotation process for the analysis of more speech corpora in a cross-genre scenario. In addition, we find particularly interesting to further investigate possible personal styles of storytellers. In this work, we have assumed (and observed to a considerably degree) that storytellers perform the storytelling expressive category in a unimodal way, which may not be the case for other corpora. Hence, a thorough study regarding this matter would be certainly clarifying. Moreover, a different storytelling annotation philosophy could be defined and compared to the current introduced methodology. For instance, such philosophy could consist in drifting away from the assignment of categorical values to the utterances. In such scenario, the annotators would

position the utterances in some continuous space, which could be spanned for example by the emotional corpus, i.e, taking the expressiveness of basic emotions as reference.

8.2 CROSS-LANGUAGE ANALYSIS OF EXPRESSIVE CATEGORIES IN INDIRECT STORYTELLING SPEECH CORPORA: THE RELEVANCE OF BOTH PROSODY AND VOICE QUALITY

Another research gap where we placed the focus is the study of VoQ features in storytelling speech. In order to bridge this gap, a larger set of VoQ parameters than the one of typical prosodic features has been considered in the analyses. They have been conducted on nearly 80 minutes of storytelling speech, which has been manually revised at the phoneme-level (64,000 phonemes approximately) in order to ensure the reliability of the results.

Both prosody and VoQ have shown a relatively equal importance in the discrimination among storytelling expressive categories of indirect storytelling speech across languages. The most relevant prosodic parameters in their discrimination have resulted in $F0_{mean}$, $F0_{IQR}$, and int_{mean}, whereas the SS, H1H2, and HNR_{mean} have resulted in the most relevant VoQ parameters. Moreover, these results have been found beyond the observed personal styles of the four narrators. Although the English narrator relied more on prosodic variations to convey the story under analysis, this might be intended by the narrator (i.e., a personal style), thus, it is recommendable to take into account VoQ when dealing with storytelling speech.

In what concerns acoustic characterization of the different storytelling expressive categories, significant common patterns have also been observed across languages. Regarding prosodic features, narrators expressed the different categories similarly in terms of F0 and intensity levels, whereas the most common VoQ patterns have been found in the MDQ and H1H2 parameters.

In the future, prosodic and VoQ transplantation experiments could also be conducted with the objective of exploring the role of both prosody and VoQ from a perceptual point of view. It is worth noting that for the aforementioned future automatic annotation approach, a more balanced representation of prosodic and VoQ parameters should probably be considered in order to avoid potential classification biases.

8.3 ANALYSES ORIENTED TO A TEXT-TO-SPEECH-TO-SPEAKING-AVATAR SYNTHESIS FRAMEWORK

The long-term vision of this thesis was focused on enhancing two key aspects of audiovisual storytellers in HCI applications: the quality of their expressive speech (TTS synthesis) and the naturalness of the non-verbal language of the 3D speaking avatar (STSA synthesis).

Regarding the TTS synthesis experiments to generate storytelling speech, the preliminary US-HNM approach was useful to confirm the existence of expressive categories within the storytelling speaking style through perceptual tests. However, the low resemblance with a real storyteller showed that there was still room for improvement. Being so, a hybrid speech synthesis framework based on US and AHM has been developed to generate storytelling speech using a rule-based prosodic model derived from the analysis of few but representative utterances of increasing suspense as a proof of concept. The US-AHM approach has been evaluated on a subjective test comparing it to the fixed prosodic rules introduced by Theune et al. (2006), using the neutral synthetic speech as baseline. The proposed method obtains good naturalness and storytelling resemblance, although it is similar to the other methods in terms of suspense arousal.

In the future, we will try to increase the suspenseful feeling of the increasing suspense category by including VoQ in the rule-based acoustic model and generalize the model considering more speech

data from several narrators and languages. Furthermore, we plan to include the rest of the identified storytelling expressive categories in the synthesis framework after the creation of their respective rule-based acoustic models. Note that the synthesis framework allows the inclusion of such models from very small and representative speech corpora, which offers a great scalability when compared to other synthesis approaches.

In what concerns analyses oriented to the STSA synthesis stage, a study of the relationship between gestures and expressive speech in terms of temporal alignment and emphasis has been conducted. An expressive corpus of both gestures and speech (recorded concurrently) has been analysed to extract synchrony and intensity (or emphasis) rules. The synchrony rule has been able to define temporal correspondences between speech and gesture, and the intensity rule relates speech and gesture strength levels. These rules have been derived from two introduced variables, which contain the prominence (or strength) of each pitch accent and gesture: the PSI and the GSI, respectively. Such variables have been computed using prosodic parameters of the speech signal and kinematic parameters of gestures, and have been validated as good candidates for representing the perceived strength. After a correlation analysis of both indicators, an intensity rule has been derived. Regarding the temporal alignment analysis, a synchrony rule has been obtained after a meticulous analysis of anchor points from gesture (apexes) and speech (pitch accent peaks).

The next step to develop a TTS-STSA synthesis system will be to include these rules in a STSA synthesis stage that drives the animation of a 3D speaking storyteller. Obviously, the appropriateness of the resulting gesture synthesis should be validated by means of perceptual tests. It is also worth to note that further analyses to derive specific rules are still necessary to improve the naturalness of the synthetic non-verbal output. Currently, only beat gestures are taken into consideration but the analysis of the relationship of speech with other types of gestures may also benefit the naturalness of the synthetic output. In order to perform these analyses, the audiovisual recording of a professional storyteller narrating a story in a MOCAP environment would be ideal.

To conclude, in this thesis several steps have been performed that will turn the *CuentaCuentos* 2.0 application represented in Fig.1 into an improved 3.0 version. Although the enhancement of the avatar personalization has been put aside for the moment, the outcomes of the thesis can be included in the system to improve the naturalness and appropriateness of the storytelling expressive speech and its associated gestures. In a nutshell, the main contributions of the thesis are the following:

- An annotation methodology to analyse storytelling speech that strips down the discourse modes of storytelling into sentence-level expressive categories.
- A thorough analysis of both prosody and VoQ in storytelling speech across languages that reveals a relatively equal importance of both set of features.
- An analysis-oriented-to-synthesis methodology that can derive acoustic rule-based models from a small but representative set of utterances.
- Synchrony and intensity (or emphasis) rules that describe the interaction between expressive speech and gestures, which can be used within a STSA synthesis system.

Part V

APPENDIXES

A

SPEECH ANALYSIS OF THE MAIN CHARACTER OF THE STORY

Although the main focus of the thesis was the analysis and synthesis of the indirect discourse parts of storytelling speech, a preliminary analysis of the direct discourse parts has also been addressed (see Fig. 12). In particular, the main character of the Spanish version of the story has been analysed. The Spanish narrator interprets this character without changing his voice too much, but it is noticeable that he tries to imitate the voice of a pre-teenager according to informal perceptual tests.

A.1 EMOTIONAL ANNOTATION OF THE MAIN CHARACTER

If a narrator interprets the characters, he/she typically modifies his/her voice into a more exaggerated register of expressions, where (acted) full-blown emotions may manifest (Buurman, 2007). As a consequence, for the classification of the dialogue mode a basic emotions annotation scheme is used.

From the speech corpus at hand (the same of Parts I & II), two experts on speech technologies were asked to classify the main character's utterances into six basic emotions (hot anger, cold anger, joy, sadness, surprise, and fear) besides a neutral category. The final number of utterances identified in the speech corpus can be observed in Table 23.

A.2 ANALYSIS OF DIRECT VS. INDIRECT NEUTRAL SPEECH

The first experiment compares the neutral indirect speech of the narrator with his neutral direct speech when interpreting the main character. For this analysis, the intensity Standard deviation (STD) (int_{Std}) was also extracted to check if, in the corpus at hand, the neutral indirect speech of the narrator is associated to greater intensity variation (Theune et al., 2006).

It can be observed from Table 24 that the narrator's $F0_{mean}$ is slightly higher when performing the the neutral speech of the main character, which is a logical result as the narrator is interpreting a more infantile voice (it is well-known that women and children have a higher mean pitch). In spite of this, the difference in $F0_{mean}$ is small, maybe because of the aforementioned remark that the narrator does not change his tone significantly. The $F0_{Std}$ and and int_{Std} are higher in the indirect speech, which is consistent with the fact that narrators use much more variation of F0 and intensity than other speaking styles such as, the newsreader style (associated to neutral speech according to Theune et al., 2006). The

TABLE 23: TOTAL AMOUNT OF EMOTIONAL UTTERANCES IDENTIFIED IN THE SPEECH CORPUS.

Category	**# Utterances**
Neutral	14
Hot anger	18
Cold anger	15
Joy	14
Sadness	15
Surprise	12
Fear	18

TABLE 24: COMPARISON OF THE AVERAGED VALUES BETWEEN THE NEUTRAL INDIRECT SPEECH OF THE NARRATOR AND HIS NEUTRAL DIRECT SPEECH WHEN INTERPRETING THE MAIN CHARACTER.

Speaking style	$F0_{mean}$ [Hz]	$F0_{Std}$ [Hz]	int_{mean} [dB]	int_{Std} [dB]	AR [syll/sec]
Narrator's neutral indirect speech	104	31.0	71	16	7.7
Narrator's neutral direct speech	108	25.8	70	11	7.2

TABLE 25: AVERAGED RESULTS OF THE CHARACTER EMOTIONS ANALYSIS.

Emotion	$F0_{mean}$ [Hz]	$F0_{Std}$ [Hz]	AR [syll/sec]	int_{mean} [dB]
Neutral	108.0	25.8	7.2	70.0
Hot anger	+82.8%	+112.3%	-20.6%	+9.1%
Cold anger	+42.4%	+69.0%	-16.7%	+4.0%
Joy	+28.9%	+67.6%	-11.2%	+7.2%
Sadness	-11.5%	-28.5%	-21.6%	-3.3%
Surprise	+45.2%	+92.7%	-15.9%	+0.7%
Fear	+29.1%	+27.2%	-2.2%	+5.3%

int_{mean} is also higher in the indirect speech, which can also be associated to its greater expressiveness. In this case, the AR of the indirect speech is faster than the AR of the main character neutral voice. This may be due to the fact that what the main character is saying is more relevant for the development of the story than the neutral indirect passages, which tend to be merely informative, so the narrator slows down his AR a bit.

A.3 STORY CHARACTER'S EMOTIONS VS. EMOTIONAL PROFILES IN THE LITERATURE

In this Section, the conclusions obtained from the previous prosodic analyses are compared them with classic emotional acoustic profiles reported in the literature.

Regarding emotional prosodic results (see Table 25), it is remarkable that all the emotions have a slower AR than the character's neutral voice. In general, anger, joy, surprise, and fear tend to have a faster AR in the literature (Iriondo et al., 2004a; Burkhardt and Sendlmeier, 2000; Iriondo et al., 2000; Kienast et al., 1999). However, the difference between joy and happiness is not so clear. For example, Burkhardt and Sendlmeier (2000) clearly separated both and proposed a decrease in AR for happiness (similarly to Kienast et al., 1999) and an increase for joy. Other studies such as the one conducted by Banse and Scherer (1996), used the term elation to refer to a more intense form of happiness, although in that study both emotions obtained a higher AR. Results of AR from the emotions expressed by the narrator while interpreting the main character are the ones which have a larger divergence when compared to the general literature focused on emotion analysis. As a preliminary conclusion, it seems that storytellers speak slower even in the character emotions (besides the indirect discourse). This can be due to the fact that they need to draw the audience attention and allow them to be able to follow all the delivered information. Thus, this parameter may yield the main difference with respect to more natural or spontaneous emotions.

Hot anger shows the most exaggerated values of $F0_{mean}$, $F0_{Std}$ and int_{mean} of all the emotional catalogue. The raise of this prosodic parameters is quite coherent with previous studies focused on basic emotions (Iriondo et al., 2004a; Burkhardt and Sendlmeier, 2000; Iriondo et al., 2000). Cold anger has

TABLE 26: RELATIONSHIP BETWEEN THE PROSODIC PATTERNS OF EMOTIONS OF THE MAIN CHARACTER VOICE AND THE ONES REPORTED IN THE LITERATURE FOR BASIC EMOTIONS IN DIFFERENT LANGUAGES. "**" INDICATES THAT THERE IS A CLEAR RELATIONSHIP WHEREAS "X" STATES THE OPPOSITE. "*" INDICATES SOME KIND OF RELATIONSHIP..

Emotion	Parameter	Language				
		Spanish (Iriondo et al., 2000)	Catalan (Iriondo et al., 2004b)	English (Murray and Arnott, 1995)	German (Burkhardt and Sendlmeier, 2000)	Portuguese (Nunes et al., 2010)
Anger	$F0_{mean}$		**	**	**	**
	$F0_{IQR}$	**	**	**		
	$F0_{Std}$					**
	int_{mean}	**	**	**		
	AR	**	X	X	X	
Joy	$F0_{mean}$	**	**		**	**
	$F0_{IQR}$	**	X	**	**	
	$F0_{Std}$					**
	int_{mean}	X	**	**		
	AR	X	**	*	*	
Sadness	$F0_{mean}$	**	**	**	**	**
	$F0_{IQR}$	**	**	**	**	
	$F0_{Std}$					**
	int_{mean}	**	**	**		
	AR	**	**	**	**	
Surprise	$F0_{mean}$	**				
	$F0_{IQR}$	**				
	$F0_{Std}$					
	int_{mean}	**				
	AR	X				
Fear	$F0_{mean}$	**	**	**	**	**
	$F0_{IQR}$	X	X	**	**	
	$F0_{Std}$					**
	int_{mean}	**	X			
	AR	X	**	X	X	

the same changes as hot anger but not so wide. Joy shows the highest int_{mean} right after hot anger, and its F0-related values are quite high in general. Sadness is the emotion which has more relationship with the acoustic profiles reported in the literature, as it entails a decrease in all the prosodic parameters (Iriondo et al., 2004a; Burkhardt and Sendlmeier, 2000; Iriondo et al., 2000; Banse and Scherer, 1996). Surprise, which is usually related to an increase of the prosodic parameters with respect to a neutral reference, has also a relationship with other studies (except for AR as well) (Iriondo et al., 2000). Fear has a relative coherency with the literature. In general, F0, intensity and AR also increase in fear when compared to a neutral reference (Burkhardt and Sendlmeier, 2000; Iriondo et al., 2000). From Table 25, it can be observed that $F0_{mean}$, $F0_{Std}$, and int_{mean} increase. Finally, the AR obtained for fear is the highest of all the emotions, almost the same as the one for the character's neutral voice.

Table 26 summarizes the degree of relationship between the emotions of the main character and the profiles of basic emotions reported in the literature (clear relationship=78%, no relationship=19%, and some kind of relationship=3%). Cold anger has not been taken into account since usually hot anger is the reported emotion.

A.4 CONCLUSIONS

The analysis of the main character emotions has shown that, in contrast to the indirect storytelling speech, acted basic emotions are present in the characters of a tale and they share many prosodic similarities with

the patterns reported in the general literature related to emotion analysis (a global agreement of 78% was obtained), being the AR the most conflictive parameter. This may be due to the fact that storytellers speak slower in order to catch the attention of the audience (children, in general) and give them some extra time to comprehend the plot. This is an important conclusion of this work in order to differentiate spontaneous emotions from emotions in storytelling, but further analyses are needed to add more evidences.

B

ANALYSIS OF LS-URL LAICOM-UAB SPEECH CORPORA

In this Appendix, the same acoustic analysis framework applied to the storytelling corpora is used to analyse the emotional LS-URL LAICOM-UAB speech corpora. The Chapter is only informative, providing more details than those presented in the analysis of Part I of the thesis, in case the reader finds them helpful.

B.1 ACOUSTIC ANALYSIS

The statistical analysis on the emotional corpora can be summarized as follows. The MANOVA revealed statistically significant results $[Pillai's\ Trace = 2.420,\ F(60, 24472) = 624.869,\ p < 0.001]$. Subsequent univariate analyses showed statistical significance on all parameters, meaning that for each parameter there is one or more statistically significant differences among emotional categories (see Table 28). Following the same criteria as in Section X, we considered all parameters with the exception of glottal flow parameters and NSP as the most relevant features. The normalized results for each emotion can be observed in Table 27. All parameters showed statistically significant contrasts among all categories with the exception of NSP ($p_{neu-sen} = 0.999$), int_{mean} ($p_{neu-sad} = 0.800$), PE1000 ($p_{neu-sen} = 0.094$), SS ($p_{sad-sen} = 0.516$), PSP ($p_{hap-sad} = 1.000$), MDQ ($p_{neu-hap} = 0.908$), and H1H2 ($p_{hap-sad} = 1.000$). Even though there are three categories less in the emotional analysis than in the storytelling analysis (which could be one cause for better differentiation among categories), the substantial larger number of statistically significant differences could be attributed to the fact that storytelling entails subtler speech nuances than those present in emotions. In the following analyses, more evidences to support this last claim are investigated.

Regarding F0 results after the acoustic analysis of emotions, happiness was expressed with the highest F0_{mean} followed by aggressive, neutral, sadness and sensual in that order. However, the lowest F0_{IQR} was obtained in the neutral category. In relation to int_{mean}, it showed the same pattern as F0_{mean} but exchanging the position of sadness and sensual categories. The articulation rates results showed neutral as the category with fastest articulation rate surpassing aggressive, happiness, sensual and sadness in that order. Concerning perturbation measures, it is interesting to observe that the sensual category show high values of jitter and shimmer together with low values of HNR_{mean}, probably because of the soft phonation used. Sadness, however, is specially characterized by high values of jitter and HNR_{mean}. Spectral measures such as SS, PE1000 and HAMMI, which are slightly correlated, reflect a flatter spectral tilt, i.e. a tenser voice, in aggressive and happiness categories and the opposite in sadness and sensual categories. However, H1H2 (H2 was always lower than H1) does not follow the exact same pattern, e.g., higher H1H2 in the aggressive category than the one obtained in the sadness category. Moreover, the glottal parameters do not show clear patterns that could support this assumption.

B.2 DISCRIMINANT ANALYSIS

After the discriminant analysis in the emotional corpus, four canonical discriminant functions were obtained, all with significant results $[Function\ 1:\ Wilks'\ \Lambda = 0.011,\ \chi^2(60) = 27432.222,\ p < 0.0001;\ Function\ 2:\ Wilks'\ \Lambda = 0.105,\ \chi^2(42) = 13799.477,\ p < 0.0001;\ Function\ 3:\ Wilks'\ \Lambda = 0.326,\ \chi^2(26) = 6866.260,\ p < 0.0001;\ Function\ 4:\ Wilks'\ \Lambda = 0.642,\ \chi^2(12) = 2711.586,\ p < 0.0001]$. In Fig. 31, the combined groups plot with the first two discriminant functions is depicted. The first canonical function explained 69.5% of the variance and correlated positively with

TABLE 27: NORMALIZED ACOUSTIC MEASURES OF THE EMOTIONAL CORPUS.

Category	Prosodic parameter				
	Nsp	**AR**	**$F0_{mean}$**	**$F0_{IQR}$**	**int_{mean}**
NEU	-0.22	0.75	-0.47	-0.55	-0.35
AGR	0.26	0.23	1.23	0.47	0.98
HAP	-0.09	-0.39	1.41	1.47	1.14
SAD	0.54	-1.01	-0.54	-0.14	-1.06
SEN	-0.23	-0.74	-1.11	-0.47	-0.70
	VoQ parameter				
	Jitter	**Shimmer**	**HNR_{mean}**	**pe1000**	**HammI**
NEU	-0.31	-0.67	0.32	-0.24	0.06
AGR	-0.69	0.49	-0.23	1.06	-1.07
HAP	-0.42	0.35	-0.45	0.49	-0.93
SAD	0.54	0.15	0.57	-0.86	1.36
SEN	1.55	0.72	-0.80	-0.16	0.55
	VoQ parameter				
	SS	**NAQ**	**PSP**	**MDQ**	**H1H2**
NEU	-0.12	0.39	-0.31	-0.20	0.64
AGR	1.35	0.25	0.53	1.00	-0.19
HAP	0.84	-0.73	0.00	-0.23	-0.83
SAD	-1.01	0.01	-0.01	0.00	-0.83
SEN	-1.05	-0.61	0.22	-0.42	0.33

$F0_{mean}$ ($r = 0.83$) and SS ($r = 0.56$), and negatively with HAMMI ($r = -0.44$). The second function accounted for 17.7% of the variance and correlated positively with AR ($r = 0.65$) and H1H2 ($r = 0.42$), and negatively with jitter ($r = -0.43$) and shimmer ($r = -0.40$). The third function explained 8.2% of the variance and correlated positively with int_{mean} ($r = 0.64$), HNR_{mean} ($r = 0.37$), and negatively with PE1000 ($r = -0.35$) and PSP ($r = -0.16$). Finally, the fourth function represented the 4.7% of the variance and correlated with MDQ ($r = 0.52$), NAQ ($r = 0.43$), $F0_{IQR}$ ($r = -0.40$), and NSP ($r = 0.30$).

The overall hit-ratio of the LDA classification resulted in 93.4% and results per category are quite similar: $NEU = 95.9\%$, $AGG = 93.9\%$, $HAP = 92.6\%$, $SAD = 89.8\%$, $SEN = 91.1\%$. The the macro-averaged F1 scores (F_1^M) and results by category can be observed in Table 29). Interestingly, the results obtained when considering VoQ parameters exclusively outperform those obtained from prosodic features. Nonetheless, considering both set of features leads to the highest classification performance (improvement of 14.34% and 6.07% when adding VoQ parameters to the prosodic ones and vice versa, respectively).

TABLE 28: WILKS' LAMBDA VALUES FOR EACH PARAMETER OBTAINED FROM THE ANALYSIS OF EMOTIONAL SPEECH, ORDERED FROM LOWEST (BEST DISCRIMINATION CAPABILITY) TO HIGHEST (WORST DISCRIMINATION CAPABILITY). ALL SHOWED $p<0.05$ IN THE UNIVARIATE TESTS.

Parameter	**$F0_{mean}$**	**SS**	**HammI**	**$F0_{IQR}$**	**Jitter**	**AR**	**int_{mean}**	**H1H2**
Wilks' Lambda	0.147	0.262	0.336	0.478	0.491	0.517	0.566	0.607
Parameter	**pe1000**	**Shimmer**	**HNR_{mean}**	**MDQ**	**NAQ**	**PSP**	**Nsp**	
Wilks' Lambda	0.630	0.694	0.781	0.782	0.800	0.908	0.915	

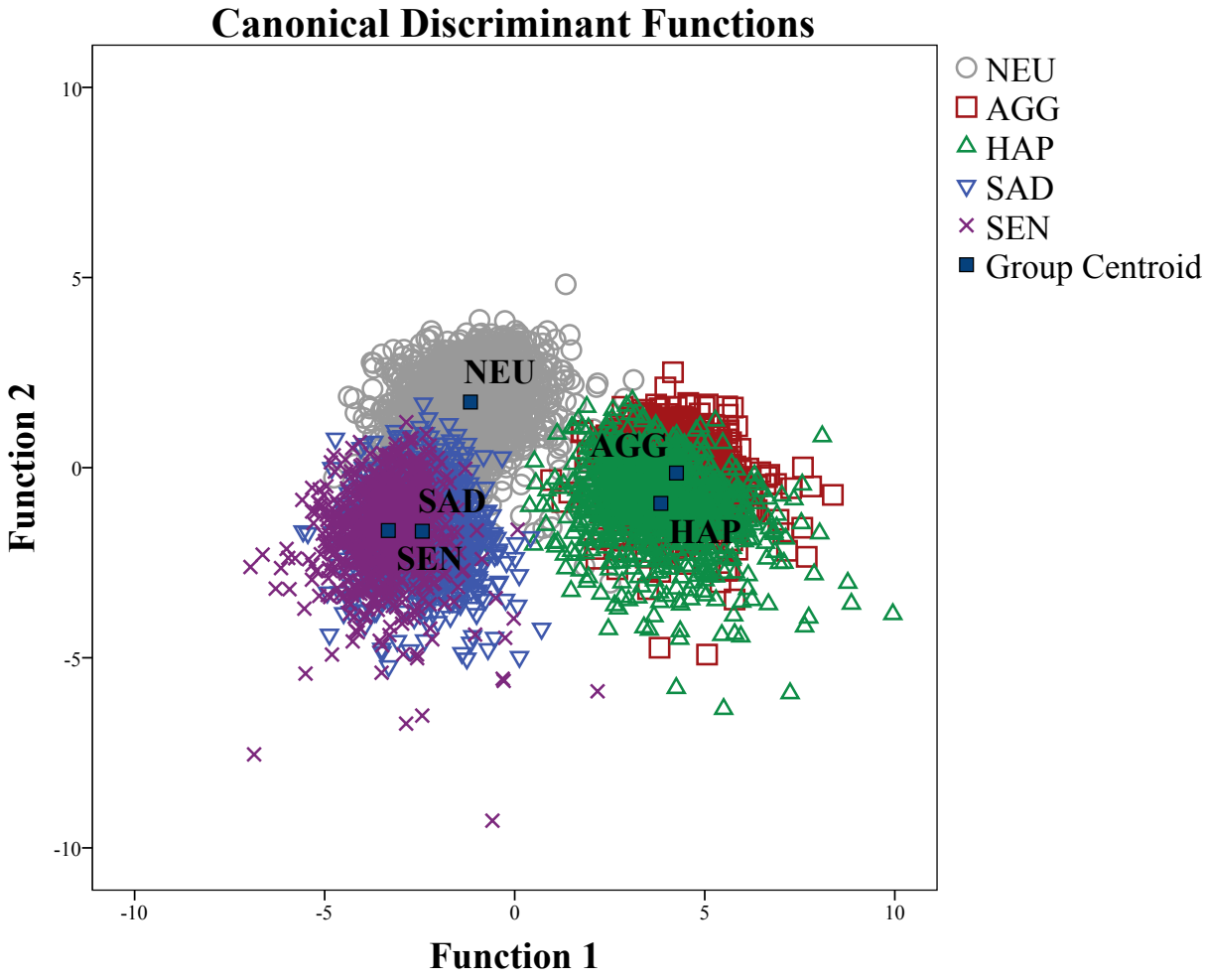

FIGURE 31: LINEAR DISCRIMINANT ANALYSIS COMBINED-GROUPS PLOT OF THE EMOTIONAL CORPUS.

TABLE 29: LDA F1S SCORES PER EMOTION CATEGORY. P: PROSODY.

Features	**NEU**	**HAP**	**SAD**	**SEN**	**AGG**	F_1^M
P	0.892	0.793	0.703	0.824	0.834	0.810
VoQ	0.933	0.815	0.880	0.834	0.898	0.873
P+VoQ	0.926	0.921	0.898	0.902	0.942	0.926

BIBLIOGRAPHY

D. Abercrombie. *Elements of general phonetics*. Edinburgh University Press, 1967.

J.-M. Adam. *Les textes: types et prototypes: récit, description, argumentation, explication et dialogue*. Paris, Nathan, 1992.

J. Adell, A. Bonafonte, and D. Escudero. Analysis of prosodic features towards modelling of emotional and pragmatic attributes of speech. *Proces. Leng. Nat.*, 35:277–283, 2005.

M. Airas and P. Alku. Comparison of multiple voice source parameters in different phonation types. In *Proc. Interspeech*, pages 1410–1413, Antwerp, Belgium, 2007.

F. Alías, I. Iriondo, L. Formiga, X. Gonzalvo, C. Monzo, and X. Sevillano. High quality Spanish restricted-domain TTS oriented to a weather forecast application. In *Proc. Interspeech*, pages 2573–2576, Lisbon, Portugal, 2005.

F. Alías, X. Sevillano, J. Socoró, and X. Gonzalvo. Towards High-Quality Next-Generation Text-to-Speech Synthesis: A Multidomain Approach by Automatic Domain Classification. *IEEE Trans. Audio, Speech & Lang. Process.*, 16(7):1340–1354, 2008.

F. Alías, L. Formiga, and X. Llorá. Efficient and reliable perceptual weight tuning for unit-selection text-to-speech synthesis based on active interactive genetic algorithms: A proof-of-concept. *Speech Commun.*, 53(5):786–800, 2011.

P. Alku, T. Bäckström, and E. Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *J. Acoust. Soc. Am.*, 112(2):701–710, 2002.

C. O. Alm. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proc. ACL/HLT*, HLT '11, pages 107–112, Stroudsburg, PA, USA, 2011.

C. O. Alm and R. Sproat. Perceptions of emotions in expressive storytelling. In *Proc. Interspeech*, pages 533–536, Lisbon, Portugal, 2005a.

C. O. Alm and R. Sproat. Emotional sequencing and development in fairy tales. In *Affective Computing and Intelligent Interaction*, pages 668–674. Springer, 2005b.

C. O. Alm, D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proc. HLT/EMNLP*, pages 579–586, 2005.

T. Alofs, M. Theune, and I. Swartjes. A tabletop interactive storytelling system: Designing for social interaction. *Int. J. Arts & Technol.*, 8(3):188–211, 2015.

B. Andreeva, G. Demenko, B. Möbius, F. Zimmerer, J. Jügler, and M. Oleskowicz-Popiel. Differences of pitch profiles in Germanic and Slavic languages. In *Proc. Interspeech*, pages 1307–1311, Singapore, 2014.

B. Andreeva, B. Möbius, G. Demenko, F. Zimmerer, and J. Jügler. Linguistic Measures of Pitch Range in Slavic and Germanic Languages. In *Proc. Interspeech*, pages 968–972, Dresden, Germany, 2015.

M. Antonijoan. Gesture-prosody correlations analysis. Master's thesis, La Salle - Ramon Llull University, 2012.

J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez. Automatic Detection of Pathological Voices Using Complexity Measures, Noise Parameters, and Mel-Cepstral Coefficients. *IEEE Trans. on Biomed. Eng.*, 58(2):370–379, 2011.

M. G. Bal. *Narratologie: Essais sur la signification narrative dans quatre romans modernes*. Paris: Editions Klincksieck, 1977.

R. Banse and K. R. Scherer. Acoustic profiles in vocal emotion expression. *J. Pers. and Soc. Psychol.*, 70(3):614–636, 1996.

R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guarasa. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. *Speech Commun.*, 52(5):394–404, 2010.

R. Barthes. Introduction to the structural analysis of narratives. *Image Music Text. Essays Selected and Translated by Stephen Heath*, pages 79–124, 1977.

S. Baumann, M. Grice, and R. Benzmüller. GToBI - A phonological system for the transcription of German intonation. In *Prosody 2000: Speech Recognition and Synthesis*, pages 21–28, Krakẃ , Poland, 2000.

M. E. Beckman and G. Ayers. Guidelines for ToBI labelling. In *Technical report*, pages 1–43, Linguistics Department, Ohio State University, 1997.

M. E. Beckman, M. Díaz-Campos, J. T. McGory, and T. A. Morgan. Intonation across Spanish, in the Tones and Break Indices framework. *Probus*, 14(1):9–36, 2002.

B. Bigi and D. Hirst. SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody. In *Proc. Speech Prosody*, Shanghai, China, 2012.

A. W. Black. Unit selection and emotional speech. In *Proc. Interspeech*, Geneva, Switzerland, 2003.

P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proc. Inst. Phonetic Sci.*, 17(1193):97–110, 1993.

P. Boersma and D. Weenink. Praat: doing phonetics by computer [Computer program]. (v.5.4.02). retrieved 26 November 2014 from http://www.praat.org/, 2014.

D. L. M. Bolinger. *Intonation and Its Parts: Melody in Spoken English*. Stanford University Press, Stanford, CA, 1986.

J. C. Borod, L. H. Pick, S. Hall, M. Sliwinski, N. Madigan, L. K. Obler, J. Welkowitz, E. Canino, H. M. Erhan, M. Goral, C. Morrison, and M. Tabert. Relationships among Facial, Prosodic, and Lexical Channels of Emotional Perceptual Processing. *Cogn. & Emot.*, 14(2):193–211, 2000.

A. Botinis, B. Granström, and B. Möbius. Developments and paradigms in intonation research. *Speech Commun.*, 33(4):263–296, 2001.

B. Bozkurt and T. Dutoit. Mixed-Phase Speech Modeling and Formant Estimation, Using Differential Phase Spectrums. In *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, pages 21–24, Geneva, Switzerland, 2003.

N. Braunschweiler and S. Buchholz. Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality. In *Proc. Interspeech*, pages 1821–1824, Florence, Italy, 2011.

C. Bremond. La logique des possibles narratifs. *Communications*, 8(1):60–76, 1966.

R. L. Brennan and D. J. Prediger. Coefficient kappa: Some uses, misuses, and alternatives. *Educ. Psychol. Meas.*, 41(3):687–699, 1981.

W. F. Brewer and E. H. Lichtenstein. Stories are to entertain: A structural-affect theory of stories. *J. Pragmat.*, 6(5-6):473–486, 1982.

M. Brockmann, C. Storck, P. N. Carding, and M. J. Drinnan. Voice loudness and gender effects on jitter and shimmer in healthy adults. *J. Speech, Lang., Hear. Res.*, 51(5):1152–1160, 2008.

F. Burkhardt. Emofilt : the simulation of emotional speech by prosody-transformation. *Proc. of Interspeech*, pages 509–512, 2005.

F. Burkhardt. An affective spoken storyteller. In *Proc. Interspeech*, pages 3305–3306, Florence, Italy, 2011.

F. Burkhardt and W. Sendlmeier. Verification of acoustical correlates of emotional speech using formant-synthesis. In *Proc. of the ISCA Workshop on Speech and Emotion*, pages 151–156, 2000.

H. Buurman. Virtual storytelling: Emotions for the narrator. Master's thesis, Univ. Twente, The Netherlands, 2007.

J. Cabral, L. Oliveira, G. Raimundo, and A. Paiva. What voice do we expect from a synthetic character? In *Proc. SPECOM*, pages 536–539, 2006.

H. Calsamiglia and A. Tusón. Los modos de organización del discurso (Chapter 10). In *Las Cosas del decir: manual de análisis del discurso*, pages 269–323. Ariel, 1999.

À. Calzada and J. C. Socoró. Vocal effort modification through harmonics plus noise model representation. In *Advances in Nonlinear Speech Processing*, pages 96–103. Springer, 2011.

À. Calzada and J. C. Socoró. Voice quality modification using a Harmonics Plus Noise Model. *Cognitive Computation*, pages 1–10, 2012.

X. Cao, S. E. Lindley, J. Helmes, and A. Sellen. Telling the whole story: anticipation, inspiration and reputation in a field deployment of TellTable. In *Proc. ACM Conf. Computer Suport. Cooperative Work*, pages 251–260, 2010.

W. Chafe. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press, 1994.

P. Charaudeau. *Grammaire du sens et de l'expression*. Hachette, 1992.

M. Charfuelan and I. Steiner. Expressive speech synthesis in MARY TTS using audiobook data and EmotionML. In *Proc. Interspeech*, pages 1564–1568, Lyon, France, 2013.

H. S. Cheang and M. D. Pell. The sound of sarcasm. *Speech Commun.*, 50(5):366–381, 2008.

Y.-G. Cheong and R. M. Young. A computational model of narrative generation for suspense. In *AAAI Comput. Aesthet. Workshop*, pages 1906–1907, Boston, MA, USA, 2006.

D. G. Childers and C. Lee. Vocal quality factors: Analysis, synthesis, and perception. *J. Acous. Soc. Am.*, 90(5):2394–2410, 1991.

S. Choi, J. Lee, A. Sprecher, and J. Jiang. The effect of segment selection on acoustic analysis. *J. Voice*, 26(1):1–7, 2012.

S. Cohan and L. M. Shires. *Telling Stories: A Theoretical Analysis of Narrative Fiction*. New York: Routledge, 1988.

R. Cornelius. *The Science of Emotion: Research and Tradition in the Psychology of Emotions*. Prentice Hall, Upper Saddle River, NJ, 1996.

H. Cramér. *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press, 1946.

A. Cruttenden. *Intonation*. Cambridge Textbooks in Linguistics. Cambridge University Press, 1986.

T. H. de Johnson, D. C. O'Connell, and E. J. Sabin. Temporal analysis of English and Spanish narratives. *Bull. Psychon. Soci.*, 13(6):347–350, 1979.

G. Degottex and Y. Stylianou. Analysis and Synthesis of Speech Using an Adaptive Full-Band Harmonic Model. *IEEE Trans. Audio, Speech & Lang. Process.*, 21(10):2085–2095, 2013.

G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer. COVAREP - A collaborative voice analysis repository for speech technologies. In *Proc. IEEE ICASSP*, pages 960–964, Florence, Italy, 2014.

P. H. Dejonckere, M. Remacle, E. Fresnel-Elbaz, V. Woisard, L. Crevier-Buchman, and B. Millet. Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. *Revue de laryngologie-otologie-rhinologie*, 117(3):219–224, 1995.

E. Delais-Roussarie, B. Post, M. Avanzi, C. Buthke, A. Di Cristo, I. Feldhausen, S. Jun, P. Martin, T. Meisenburg, A. Rialland, et al. Intonational Phonology of French: Developing a ToBI system for French. In *Intonation in Romance*, pages 63–100. OUP Oxford, 2015.

E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. Emotional speech: Towards a new generation of databases. *Speech Commun.*, 40(1-2):33–60, 2003.

D. Doukhan, A. Rilliard, S. Rosset, M. Adda-Decker, and C. d'Alessandro. Prosodic analysis of a corpus of tales. In *Proc. Interspeech*, pages 3129–3132, Florence, Italy, 2011.

B. Doval, C. d'Alessandro, and N. Henrich. The voice source as a causal/anticausal linear filter. In *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, pages 15–20, Geneva, Switzerland, 2003.

C. Drioli, G. Tisato, P. Cosi, and F. Tesser. Emotions and voice quality: Experiments with sinusoidal modeling. In *ISCA Tutor. and Res. Workshop on Voice Qual.: Funct., Anal. and Synth.*, pages 127–132, Geneva, Switzerland, 2003.

T. Drugman, B. Bozkurt, and T. Dutoit. A comparative study of glottal source estimation techniques. *Computer Speech & Language*, 26(1):20–34, 2012.

C. K. Enders. Performing multivariate group comparisons following a statistically significant MANOVA. *Meas. Eval. Couns. Dev.*, 36:40–56, 2003.

D. Erro. *Intra-lingual and cross-lingual voice conversion using Harmonic plus Stochastic Models*. PhD thesis, Technical University of Catalonia, 2008.

D. Erro, A. Moreno, and A. Bonafonte. Flexible harmonic/stochastic speech synthesis. In *ISCA Workshop on Speech Synth.*, pages 194–199, 2007.

D. Erro, E. Navas, I. Hernáez, and I. Saratxaga. Emotion conversion based on prosodic unit selection. *IEEE Trans. Audio, Speech, & Lang. Process.*, 18(5):974–983, 2010.

D. Escudero and V. Cardenoso. Corpus based extraction of quantitative prosodic parameters of stress groups in Spanish. *ICASSP*, 1:481–484, 2002.

L. Eskenazi, D. G. Childers, and D. M. Hicks. Acoustic correlates of vocal quality. *Journal of Speech, Language, and Hearing Research*, 33(2):298–306, 1990.

F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M.J.F. Gales, and K. Knill. Unsupervised clustering of emotion and voice styles for expressive TTS. In *Proc. IEEE ICASSP*, pages 4009–4012, 2012.

F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Computer Society*, 2015.

G. Fenk-Oczlon and A. Fenk. Measuring basic tempo across languages and some implications for speech rhythm. In *Proc. Interspeech*, pages 1537–1540, Makuhari, Japan, 2010.

A. Fernald, T. Taeschner, J. Dunn, M. Papousek, B. de Boysson-Bardies, and I. Fukui. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *J. Child Lang.*, 16(3):477–501, 1989.

A. Fernández-Baena, R. Montaño, M. Antonijoan, A. Roversi, D. Miralles, and F. Alías. Gesture synthesis adapted to speech emphasis. *Speech Communication*, 57:331–350, 2014.

J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychol. Bull.*, 76(5):378–382, 1971.

L. Formiga, A. Trilla, F. Alıas, I. Iriondo, and J. C. Socoró. Adaptation of the URL-TTS system to the 2010 Albayzin Evaluation Campaign. In *Proc. FALA*, pages 363–370, 2010.

V. Francisco, R. Hervás, F. Peinado, and P. Gervás. EmoTales: creating a corpus of folk tales with emotional annotations. *Lang. Resour. and Eval.*, 46(3):341–381, 2011.

J. M. Garrido. *Modelling Spanish Intonation for Text-to-Speech Applications*. PhD thesis, Universitat Autònoma de Barcelona, Barcelo,a Spain, 1996.

R. Gelin, C. d'Alessandro, O. Deroo, Q. A. Le, D. Doukhan, J.-C. Martin, C. Pelachaud, A. Rilliard, and S. Rosset. Towards a storytelling humanoid robot. In *AAAI Fall Symposium Series on Dialog with Robots*, pages 137–138, 2010.

G. V. Glass, P. D. Peckham, and J. R. Sanders. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42(3):237–288, 1972.

C. Gobl and A. Ní Chasaide. The role of voice quality in communicating emotion, mood and attitude. *Speech Commun.*, 40(1-2):189–212, 2003.

J. P. Goldman. EasyAlign: An automatic phonetic alignment tool under Praat. In *Proc. Interspeech*, pages 3233–3236, Florence, Italy, 2011.

X. Gonzalvo, I. Iriondo, J. C. Socoró, F. Alías, and C. Monzo. Mixing HMM-based Spanish speech synthesis with a CBR for prosody estimation. In *Advances in Nonlinear Speech Processing*, pages 78–85. Springer, 2007.

X. Gonzalvo, P. Taylor, C. Monzo, I. Iriondo, and J. C. Socoró. High quality emotional HMM-based synthesis in Spanish. *Lecture Notes in Computer Science*, 5933 LNAI:26–34, 2010. doi: 10.1007/978-3-642-11509-7_4.

D. Govind and S. Prasanna. Expressive speech synthesis: A review. *Int. J. Speech Technol.*, 16(2): 237–260, 2013.

S. Grawunder and B. Winter. Acoustic correlates of politeness: prosodic and voice quality measures in polite and informal speech of Korean and German speakers. In *Proc. Speech Prosody*, Chicago, IL, USA, 2010.

E. Greene, T. Mishra, P. Haffner, and A. Conkie. Predicting Character-Appropriate Voices for a TTS-based Storyteller System. *Proc. Interspeech*, pages 2210–2213, 2012.

A. J. Greimas. *Sémantique structurale*. Paris: Larousse, 1966.

I. Grichkovtsova, M. Morel, and A. Lacheret. The role of voice quality and prosodic contour in affective speech perception. *Speech Commun.*, 54(3):414–429, 2012.

O. Grynszpan, J. Martin, and J. Nadel. Designing educational software dedicated to people with autism. In A. Pruski and H. Knops, editors, *Assistive Technology: From Virtuality to Reality*, pages 456–460. IOS Press, 2005.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11, 2009.

B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin. Perceptual and acoustic correlates of abnormal voice qualities. *Acta Otolaryngol.*, 90(1-6):441–451, 1980.

D. Hirst. A Praat plugin for MOMEL and INTSINT with improved algorithms for modelling and coding of intonation. In *Proc. 16th Int. Congr. Phonetic Sci.*, pages 1233–1236, 2007.

D. Hirst and A. Di Cristo. *Intonation Systems: A Survey of Twenty Languages*. Cambridge University Press, 1998.

D. Hirst and R. Espesser. Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix*, 15:75–85, 1993.

H. Hollien and J. F. Michel. Vocal fry as a phonational register. *Journal of Speech, Language, and Hearing Research*, 11(3):600–604, 1968.

Q. Hu, K. Richmond, J. Yamagishi, and J. Latorre. An experimental comparison of multiple vocoder types. In *8th ISCA Workshop on Speech Synth.*, pages 135–140, Barcelona, Spain, 2013.

C. J. Huberty and M. D. Petoskey. Multivariate analysis of variance and covariance. *Handbook of applied multivariate statistics and mathematical modeling*, pages 183–208, 2000.

IBM Corp. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp., 2013.

I. Iriondo. *Producción de un corpus oral y modelado prosódico para la síntesis del habla expresiva*. PhD thesis, La Salle – Universitat Ramón Llull, Barcelona, Spain, 2008.

I. Iriondo, R. Guaus, A. Rodríguez, P. Lázaro, N. Montoya, J. M. Blanco, D. Bernadas, J. M. Oliver, D. Tena, and L. Longhi. Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. In *Proc. of the ISCA Workshop on Speech and Emotion*, pages 161–166, 2000.

I. Iriondo, F. Alías, J. Melenchón, and M. A. Llorca. Modeling and synthesizing emotional speech for Catalan Text-to-Speech synthesis. In *Tutorial and Research Workshop on Affective Dialog Systems*, pages 197–208, 2004a.

I. Iriondo, F. Alías, J. Melenchón, and M. Á. Llorca. Modeling and synthesizing emotional speech for Catalan text-to-speech synthesis. *Affective Dialogue Systems*, pages 197–208, 2004b.

I. Iriondo, J. C. Socoró, and F. Alías. Prosody modelling of spanish for expressive speech synthesis. In *IEEE Int. Conf. Acoust., Speech & Signal Process. (ICASSP)*, volume 4, pages 821–824, Honolulu, HI, 2007.

I. Iriondo, S. Planet, J. C. Socoró, E. Martínez, F. Alías, and C. Monzo. Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification. *Speech Commun.*, 51(9):744–758, 2009.

M. Jackson, P. Ladefoged, M. Huffman, and N. Antoñanzas-Barroso. Measures of spectral tilt. *J. Acoust. Soc. Am.*, 77(S1):S86–S86, 1985.

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

I. Jauk, A. Bonafonte, P. Lopez-otero, and L. Docio-fernandez. Creating Expressive Synthetic Voices by Unsupervised Clustering of Audiobooks. In *Proc. Interspeech*, pages 3380–3384, Dresden, Germany, 2015.

O. Jokisch, H. Kruschke, and R. Hoffmann. Prosodic reading style simulation for text-to-speech synthesis. In *Affective Computing and Intelligent Interaction*, pages 426–432, 2005.

S. T. Jovičić, M. Rajković, M. Đorđević, and Z. Kašić. Perceptual and statistical analysis of emotional speech in man-computer communication. In *SPECOM 2006*, St. Petersburg, Russia, 2006.

G. P. Kafentzis, G. Degottex, O. Rosec, and Y. Stylianou. Pitch Modifications of speech based on an Adaptive Harmonic Model. In *IEEE Int. Conf. Acoust., Speech & Signal Process. (ICASSP)*, Florence, Italy, 2014.

J. Kane and C. Gobl. Identifying regions of non-modal phonation using features of the wavelet transform. In *Proc. Interspeech*, pages 177–180, Florence, Italy, 2011.

J. Kane and C. Gobl. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Trans. Audio, Speech & Lang. Process.*, 21:1170–1179, 2013.

H. Kasuya, S. Ogawa, and Y. Kikuchi. An adaptive comb filtering method as applied to acoustic analyses of pathological voice. In *IEEE Int. Conf. Acoust., Speech & Signal Process. (ICASSP)*, volume 11, pages 669–672, 1986.

A. Kendon. Gesture and speech: two aspects of the process utterances. *Nonverbal Communication and Language*, pages 207–227, 1980.

M. Kienast, A. Paeschke, and W. F. Sendlmeier. Articulatory reduction in emotional speech. In *EUROSPEECH*, pages 117–120, 1999.

M. Kipp. Anvil: A Generic Annotation Tool for Multimodal Dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, pages 1367–1370, Aalborg, 2001.

M. Kipp. *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. PhD thesis, Boca Raton, Florida, 2004.

M. Kipp and J. C. Martin. Gesture and emotion: Can basic gestural form features discriminate emotions? *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII-09), IEEE Press.*, 2009.

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

T. Kisler, F. Schiel, and H. Sloetjes. Signal processing via web services: the use case WebMAUS. In *Proc. Digit. Humanit.*, pages 30–34, Hamburg, Germany, 2012.

S. Kita, I. van Gijn, and H. van der Hulst. Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth and M. Frhlich, editors, *Gesture and Sign Language in Human-Computer Interaction*, volume 1371 of *Lecture Notes in Computer Science*, pages 23–35. Springer Berlin / Heidelberg, 1998.

W. Klecka. *Discriminant Analysis*. 19. SAGE Publications, 1980.

L. F. Kready. *A study of fairy tales*. Houghton Mifflin Co. The Riverside Press, 1916.

J. Kreiman and B. R. Gerratt. Validity of rating scale measures of voice quality. *J. Acoust. Soc. Am.*, 104 (3):1598–1608, 1998.

J. Kreiman and B. R. Gerratt. Perception of aperiodicity in pathological voice. *The Journal of the Acoustical Society of America*, 117(4):2201–2211, 2005.

J. Kreiman, D. Vanlancker-Sidtis, and B. R. Gerratt. Defining and measuring voice quality. In *In Proceedings of From Sound To Sense: 50+ Years of Discoveries in Speech Communication*, pages 115–120, 2004.

J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

J. B. Kruskal and M. Wish. *Multidimensional scaling*, volume 11. Sage, 1978.

W. Labov. The transformation of experience in narrative syntax. *Language in the Inner City: Studies in the Black English Vernacular*, pages 354–96, 1972.

W. Labov and J. Waletzky. Narrative analysis. *Essays on the Verbal and Visual Arts*, pages 12–44, 1997.

D. R. Ladd and N. Campbell. Theories of prosodic structure: evidence from syllable duration. In *Proceedings of the 12th International Congress of Phonetic Sciences*, volume 2, pages 290–293, 1991.

J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biom.*, 33 (1):159–174, 1977.

J. Laroche, Y. Stylianou, and E. Moulines. HNS: Speech modification based on a harmonic+noise model. In *IEEE Int. Conf. Acoust., Speech & Signal Process. (ICASSP)*, volume 2, pages 550–553, 1993.

J. Latorre, V. Wan, M. J. Gales, L. Chen, K. Chin, K. Knill, M. Akamine, et al. Speech factorization for HMM-TTS based on cluster adaptive training. In *Proc. Interspeech*, pages 971–974, Portland, OR, USA, 2012.

J. Laver. *The phonetic description of voice quality*. Cambridge University Press, 1980.

J. Laver. *The gift of speech*. Edinburgh University Press, 1991.

J. Laver. Three semiotic layers of spoken communication. *Journal of Phonetics*, 31(3-4):413–415, 2003.

I. Leite, M. McCoy, M. Lohani, D. Ullman, N. Salomons, C. K. Stokes, S. Rivers, and B. Scassellati. Emotional Storytelling in the Classroom: Individual versus Group Interaction between Children and Robots. In *HRI*, pages 75–82, 2015.

T. Leonard and F. Cummins. The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, 26(10):1457–1471, 2011.

S. Levine, C. Theobalt, and V. Koltun. Real-time prosody-driven synthesis of body language. *ACM Trans. Graph.*, 28(5):172:1–172:10, Dec. 2009. ISSN 0730-0301.

S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun. Gesture controllers. *ACM Trans. Graph.*, 29(4): 124:1–124:11, July 2010.

P. Lieberman and S. Blumstein. *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge Studies in Speech Science and Communication. Cambridge University Press, 1988.

P. Liu and M. D. Pell. Processing emotional prosody in Mandarin Chinese: A cross-language comparison. In *Proc. Speech Prosody*, pages 95–99, Dublin, Ireland, 2014.

L. M. Lix, J. C. Keselman, and H. J. Keselman. Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance "F" test. *Review of Educational Research*, 66(4):579–619, 1996.

M. Lloberes, I. Castellón, and L. Padró. Spanish freeling dependency grammar. In *Proc. 7th Lang. Resour. and Eval. Conf.*, La Valletta, Malta, 2010.

L. S. U. R. Llull. MediaLab. Motion Capture, RV + RA, Animation, Videogames and CAD. http://www.salleurl.edu/medialab, May 2012.

D. Loehr. *Gesture and Intonation*. PhD thesis, Georgetown University, 2004.

E. Lopez-Gonzalo, J. M. Rodriguez-Garcia, L. Hernandez-Gomez, and J. M. Villar. Automatic prosodic modeling for speaker and task adaptation in text-to-speech. In *ICASSP*, pages 927–930, 1997.

P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo. iVectors for Continuous Emotion Recognition. In *Proc. Iberspeech*, pages 31–40, 2014.

J. Lorenzo-Trueba, R. Barra-Chicote, R. San-Segundo, J. Ferreiros, J. Yamagishi, and J. M. Montero. Emotion transplantation through adaptation in HMM-based speech synthesis. *Comput. Speech & Lang.*, 34(1):292–307, 2015.

J. Louw and E. Barnard. Automatic intonation modeling with INTSINT. In *Proceedings of the Pattern Recognition Association of South Africa*, pages 107–111, 2004.

N. Mamede and P. Chaleira. Character identification in children stories. In J. L. Vicedo, P. Martínez-Barco, R. Muñoz, and M. Saiz Noeda, editors, *Adv. in Nat. Lang. Process.*, volume 3230 of *Lect. Notes in Comput. Sci.*, pages 82–90. Springer Berlin Heidelberg, 2004.

E. Marchetti. Micro culture: Interactive storytelling and learning in the museum. In *IEEE 4th International Conference on Digital Game and Intelligent Toy Enhanced Learning (DIGITEL)*, pages 84–88, Takamatsu, 2012.

D. Matsumoto. Are Cultural Differences in Emotion Regulation Mediated by Personality Traits? *J. Cross-Cult. Psychol.*, 37(4):421–437, 2006.

B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405(2):442–451, 1975.

D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, 1992.

M. Merabti, A. E. Rhalibi, Y. Shen, J. Daniel, A. Melendez, and M. Price. Interactive storytelling: Approaches and techniques to achieve dynamic stories. *Trans. on Edutainment*, 1:118–134, 2008.

D. Michaelis, T. Gramss, and H. W. Strube. Glottal to Noise Excitation ratio - A new measure for describing pathological voices. *Acta Acustica united with Acustica*, 83(4):700–706, 1997.

H.-J. Min, S.-C. Kim, J. Kim, J.-W. Chung, and J. C. Park. Speaker-TTS voice mapping towards natural and characteristic robot storytelling. In *IEEE RO-MAN*, pages 793–800, 2013.

H. Mixdorff. Speech technology, ToBI, and making sense of prosody. *Speech Prosody*, pages 31–37, 2002.

R. Montaño and F. Alías. The role of prosody and voice quality in text-dependent categories of storytelling across languages. In *Proc. Interspeech*, pages 1186–1190, Dresden, Germany, 2015.

R. Montaño and F. Alías. The Role of Prosody and Voice Quality in Indirect Storytelling Speech: Analysis Methodology and Expressive Categories. *Speech Commun.*, 2016a. Second Round of Revisions.

R. Montaño and F. Alías. The Role of Prosody and Voice Quality in Indirect Storytelling Speech: A Cross-language Perspective. *Speech Commun.*, 2016b. First Round of Revisions.

R. Montaño, F. Alías, and J. Ferrer. Prosodic analysis of storytelling discourse modes and narrative situations oriented to Text-to-Speech synthesis. In *8th ISCA Workshop on Speech Synthesis*, pages 171–176, Barcelona, Spain, 2013.

R. Montaño, M. Freixes, F. Alías, and J. C. Socoró. Generating Storytelling Speech from a hybrid US-aHM Neutral TTS synthesis framework using a rule-based prosodic model. In *Proc. Interspeech*, San Francisco, USA, 2016. Submitted.

N. Montoya. El papel de la voz en la publicidad audiovisual dirigida a los niños. *Zer: Revista de Estudios de Comun.*, (4):161–177, 1998.

C. Monzo, F. Alías, I. Iriondo, X. Gonzalvo, and S. Planet. Discriminating expressive speech styles by voice quality parameterization. In *Proc. 16th Int. Congr. Phonetic Sci.*, pages 2081–2084, Saarbrücken, Germany, 2007.

C. Monzo, I. Iriondo, and J. C. Socoró. Voice quality modelling for expressive speech synthesis. *The Scientific World Journal*, 2014, 2014.

S. J. L. Mozziconacci. Modeling emotion and attitude in speech by means of perceptually based parameter values. *User Model. User-Adapt. Interact.*, 11(4):297–326, 2001.

I. R. Murray and J. L. Arnott. Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Commun.*, 16(4):369–390, 1995.

B. Mutlu, J. K. Hodgins, and J. Forlizzi. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *Proc. of the 6th IEEE-RAS International Conference on Humanoid Robots*, pages 518–523, Genova, 2006.

R. W. Neufeld and R. C. Gardner. Data aggregation in evaluating psychological constructs: Multivariate and logical deductive considerations. *J. Math. Psychol.*, 34(3):276–296, 1990.

M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. Affect. Comput.*, 2(2):92–105, 2011.

S. Nobe. *Representational Gestures, Cognitive Rhythms, and Acoustic Aspects of Speech: A Network/-threshold Model of Gesture Production*. University of Chicago, Department of Psychology, 1996.

N. Norrick. *Conversational Narrative: Storytelling in Everyday Talk*. J. Benjamins, 2000.

A. Nunes, R. L. Coimbra, and A. Teixeira. Voice quality of european portuguese emotional speech. In *Computational Processing of the Portuguese Language*, pages 142–151, 2010.

N. Obin, J. Beliao, C. Veaux, and A. Lacheret. SLAM: Automatic Stylization and Labelling of Speech Melody. In *Proc. Speech Prosody*, pages 1–5, Dublin, Ireland, 2014.

M. Oliveira. *Prosodic features in spontaneous narratives*. PhD thesis, Simon Fraser University, Burnaby, Canada, 2000.

K. Onuma, C. Faloutsos, and J. K. Hodgins. FMDistance: A fast and effective distance function for motion capture data. In *Short Papers Proceedings of EUROGRAPHICS*, 2008.

M. Ortega-Llebaria and P. Prieto. Acoustic correlates of stress in central catalan and castilian spanish. *Lang Speech*, 54(1):73–97, 2011. ISSN 0023-8309.

S. Patel, K. R. Scherer, J. Sundberg, and E. Björkner. Acoustic markers of emotions based on voice physiology. In *Proc. Speech Prosody*, Chicago, IL, USA, 2010.

D. J. Patterson. *Linguistic approach to pitch range modelling*. PhD thesis, Edinburgh University, Scotland, United Kingdom, 2000.

M. I. Pegoraro Krook. Speaking Fundamental Frequency Characteristics of Normal Swedish Subjects Obtained by Glottal Frequency Analysis. *Folia Phoniatr.*, 40(2):82–90, 1988.

M. D. Pell, L. Monetta, S. Paulmann, and S. A. Kotz. Recognizing emotions in a foreign language. *J. Nonverbal Behav.*, 33(2):107–120, 2009a.

M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz. Factors in the recognition of vocally expressed emotions: A comparison of four languages. *J. Phonetics*, 37(4):417–435, 2009b.

E. Pépiot. Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers. *Proc. Speech Prosody*, pages 305–309, 2014.

J. B. Pierrehumbert. *The phonology and phonetics of English intonation*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 1980.

K. C. S. Pillai. Some New Test Criteria in Multivariate Analysis. *The Annals of Mathematical Statistics*, 26(1):117–121, 1955.

S. Planet and I. Iriondo. Children's emotion recognition from spontaneous speech using a reduced set of acoustic and linguistic features. *Cognit. Comput.*, 5(4):526–532, 2013.

S. Planet, I. Iriondo, E. Martínez, and J. A. Montero. TRUE: an online testing platform for multimedia evaluation. In *Workshop Corpora for Res. Emot. & Affect*, page 61, 2008.

J. C. Platt. *Fast training of support vector machines using sequential minimal optimization*, pages 185–208. MIT Press, Cambridge, MA, USA, 1999.

K. Prahallad and A. W. Black. Segmentation of monologues in audio books for building synthetic voices. *IEEE Trans. Audio, Speech & Lang. Process.*, 19(5):1444–1449, July 2011.

P. Prieto. The intonational phonology of Catalan. In S. A. Sun, editor, *Prosodic typology 2. The phonology of intonation and phrasing*, pages 43–80. Oxford: Oxford University Press, 2014.

V. A. Propp. *Morphology of the Folktale*. University of Texas Press (1968), 2nd edition, 1928.

J. J. Randolph. Free-marginal multirater Kappa: An alternative to Fleiss' fixed-marginal multirater Kappa. In *Learn. & Instruct. Symp.*, Joensuu, Finland, 2005.

L. M. Rea and R. A. Parker. *Designing and conducting survey research*. San Francisco: Jossey-Boss, 1992.

M. A. Redford. A comparative analysis of pausing in child and adult storytelling. *Applied psycholinguistics*, 34(3):569–589, 2013.

M. Renwick, Y. Yasinnik, and S. Shattuck-Hufnagel. The timing of speech-accompanying gestures with respect to prosody. *From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, pages 97–102, 2004.

S. Rimmon-Kenan. *Narrative fiction: Contemporary poetics*. London: Methuen, 1983.

P. Roach. Some languages are spoken more quickly than others. *Language Myths.*, L. Bauer and P. Trudgill (eds.). London: Penguin:150–159, 1998.

S. Roekhaut, J. Goldman, and A. Simon. A model for varying speaking style in TTS systems. In *Proc. Speech Prosody*, pages 11–14, Chicago, IL, USA, 2010.

J. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.

P. Sarkar and K. Sreenivasa Rao. Data-driven pause prediction for speech synthesis in storytelling style speech. In *21st Nat. Conf. Commun. (NCC)*, pages 1–5, 2015.

P. Sarkar, A. Haque, A. K. Dutta, G. M. Reddy, M. D. Harikrishna, P. Dhara, R. Verma, P. N. Narendra, B. K. S. Sunil, J. Yadav, and K. S. Rao. Designing prosody rule-set for converting neutral TTS speech to storytelling style speech for indian languages: Bengali, hindi and telugu. In *7th Int. Conf. Contemp. Comput. (IC3)*, pages 473–477, Noida, India, 2014.

K. R. Scherer. Vocal correlates of emotional arousal and affective disturbance. In H. Wagner and A. Manstead, editors, *Handb. Psychophysiol.: Emot. and Soc. Behav.* Wiley & Sons, Oxford, UK, 1989.

K. R. Scherer, H. G. Wallbott, D. Matsumoto, and T. Kudoh. Emotional experience in cultural context: A comparison between Europe, Japan, and the US. *Facets Emot.*, pages 5–30, 1988.

K. R. Scherer, R. Banse, and H. G. Wallbott. Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. *J. Cross-Cult. Psychol.*, 32(1):76–92, 2001.

M. Schröder. *Speech and Emotion Research: An overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*. PhD thesis, Saarland Univ., Germany, 2004.

B. Schuller. Recognizing affect from linguistic information in 3D continuous space. *IEEE Trans. Affect. Comput.*, 2(4):192–205, 2011.

F. Sebastiani. Machine Learning in Automated Text Categorization. *ACM Comput. Surv.*, 34(1):1–47, 2001.

T. Seo, T. Kanda, and Y. Fujikoshi. The effects of nonnormality of tests for dimensionality in canonical correlation and manova models. *Journal of Multivariate Analysis*, 52(2):325–337, 1995.

S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *24th International Conference on MachineLearning*, pages 807–814, 2007.

K. Shama, A. Krishna, and N. U. Cholayya. Study of Harmonics-to-Noise Ratio and Critical-Band Energy Spectrum of Speech as Acoustic Indicators of Laryngeal and Voice Pathology. *EURASIP Journal on Advances in Signal Processing*, 2007:1–9, 2007.

S. Shattuck-Hufnagel, Y. Yasinnik, N. Veilleux, and M. Renwick. A method for studying the time alignment of gestures and prosody in american english: 'hits' and pitch accents in academic-lecture-style speech. In Anna Esposito, Maja Bratanić, Eric Keller and Maria Marinaro, editor, *Fundamentals of Verbal and Nonverbal Communication and the Biometric Issue*, volume 18 of *NATO Publishing Sub-Series E: Human and Societal Dynamics*. Washington, DC, 2007.

R. Silipo and S. Greenberg. Automatic transcription of prosodic stress for spontaneous English discourse. In J. J. Olds, Y. Hasegawa, M. Ohala, and A. C. Bailey, editors, *Proceedings of the XIVth International Congress of Phonetic Sciences (ICPhS99)*, pages 2351–2354. The Regents of the University of California, 1999.

R. Silipo and S. Greenberg. Pitch behavior detection for automatic prominence recognition. *Speech Prosody*, 2010.

A. Silva, M. Vala, and A. Paiva. The storyteller: Building a synthetic character that tells stories. In *Proc. Workshop Multimodal Commun. and Context in Embodied Agents*, pages 53–58, 2001.

A. Silva, G. Raimundo, A. Paiva, and C. Melo. To tell or not to tell... Building an interactive virtual storyteller. In *Proc. AISB*, pages 53–58, 2004.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: A Standard for Labeling English Prosody. In *2nd International Conference on Spoken Language Processing (ICSLP 92)*, pages 867–870, Banff, Alberta, Canada, 1992.

A. Sorin, S. Shechtman, and V. Pollet. Coherent modification of pitch and energy for expressive prosody implantation. In *IEEE Int. Conf. Acoust., Speech & Signal Process. (ICASSP)*, pages 4914–4918, 2015.

B. M. Streefkerk, L. C. W. Pols, and L. ten Bosch. Acoustical features as predictors for prominence in read aloud dutch sentences used in ann's. In *EUROSPEECH*. ISCA, 1999.

Y. Stylianou. *Harmonic plus Noise Models for Speech, combined with Statisticals Methods, for Speech and Speaker modification*. PhD thesis, École Nationale Supérieure des Télécommunications, 1996.

A. Suchato, T. Pongkittiphan, S. Suntornwanitkit, N. Suesattabongkot, and P. Punyabukkana. Digital storytelling book generator with customizable synthetic voice styles. In *Proc. of the 4th International Convention on Rehabilitation Engineering & Assistive Technology*, pages 1–4, Shanghai, China, 2010.

J. Sundberg, S. Patel, E. Bjorkner, and K. R. Scherer. Interdependencies among Voice Source Parameters in Emotional Speech. *IEEE Trans. Affect. Comput.*, 2(3):162–174, 2011.

A. K. Syrdal, J. Hirschberg, J. McGory, and M. Beckman. Automatic ToBI prediction and alignment to speed manual labeling of prosody. *Speech Commun.*, 33(1-2):135–151, 2001.

E. Székely, J. P. Cabral, P. Cahill, and J. Carson-Berndsen. Clustering expressive speech styles in audiobooks using glottal source parameters. In *Proc. Interspeech*, pages 2409–2412, Florence, Italy, 2011.

B. G. Tabachnick and L. S. Fidell. *Using multivariate statistics*. New York: Harper & Row, 1983.

F. Tamburini and P. Wagner. On automatic prominence detection for german. In *INTERSPEECH*, pages 1809–1812. ISCA, 2007.

A. Taylor. The biographical pattern in traditional narrative. *J. Folklore Inst.*, pages 114–129, 1964.

J. Terken. Fundamental frequency and perceived prominence of accented syllables. *The Journal of the Acoustical Society of America*, 89(4):1768–1776, 1991.

M. Theune, K. Meijs, D. Heylen, and R. Ordelman. Generating expressive speech for storytelling applications. *IEEE Trans. Audio, Speech & Lang. Process.*, 14(4):1137–1144, 2006.

W. F. Thompson and L. L. Balkwill. Decoding speech prosody in five languages. *Semiotica*, 158(Brown 2000):407–424, 2006.

N. Thorsen. An Acoustical Investigation of Danish Intonation. *Journal of Phonetics*, 6(3):151–175, 1978.

T. Todorov. Les catégories du récit littéraire. In *Communications*, volume 8 (1), pages 125–151. Seuil, 1966.

M. Toolan. *Narrative: A critical linguistic introduction*. London: Routledge, 1988.

H. Traunmüller. Evidence for demodulation in speech perception. *Proceedings of the 6th ICSLP*, pages 790–793, 2000.

A. Trilla and F. Alías. Sentence-based sentiment analysis for expressive Text-To-Speech. *IEEE Trans. Audio, Speech, & Lang. Process.*, 21(2):223–233, 2013.

A. Trilla, F. Alías, and I. Lozano. Text classification of domain-styled text and sentiment-styled text for expressive speech synthesis. *Proc. VI Jornadas en Tecnología del Habla (FALA2010)*, pages 75–78, 2010.

J. Trouvain. *Tempo Variation in Speech Production. Implications for Speech Synthesis*. PhD thesis, Saarland Univ., Germany, 2004.

K. P. Truong and D. A. van Leeuwen. Visualizing acoustic similarities between emotions in speech: An acoustic map of emotions. In *Proc. Interspeech*, pages 2265–2668, Antwerp, Belgium, 2007.

P. Tsiakoulis, S. Karabetsos, A. Chalamandaris, and S. Raptis. *An Overview of the ILSP Unit Selection Text-to-Speech Synthesis System*, chapter Artificial Intelligence: Methods and Applications, pages 370–383. Lecture Notes in Computer Science. Springer International Publishing, 2014.

L. Valbonesi. *Multimodal Signal Analysis of Prosody and Hand Motion: Temporal Correlation in Speech and Gestures*. University of Illinois at Chicago, 2002.

R. Van Bezooijen, S. a. Otto, and T. a. Heenan. Recognition of Vocal Expressions of Emotion: A Three-Nation Study to Identify Universal Characteristics. *J. Cross-Cultural Psychology*, 14(4):387–406, 1983.

C. van Rijsbergen. Foundation of evaluation. *Journal of Documentation*, 30(4):365–373, 1974.

J. van Santen, L. Black, G. Cohen, A. Kain, E. Klabbers, T. Mishra, J. de Villiers, and X. Niu. Applications of computer generated expressive speech for communication disorders. In *Proc. of Eurospeech*, pages 1657–1660, Geneva, Switzerland, 2003.

I. Vasilescu and M. Adda-Decker. A cross-language study of acoustic and prosodic characteristics of vocalic hesitation. *Fundam. Verbal & Non-verbal Commun. & Biom. Issue*, IOS Press, Esposito, A., Bratanic, M., Keller, E., and Marinaro, M. (eds.):140–148, 2007.

H. G. Wallbott. Bodily expression of emotion. *Eur. J. Soc. Psychol.*, 28(6):879–896, 1998.

H. G. Wallbott and K. R. Scherer. Cues and channels in emotion recognition. *J. Pers. Soc. Psychol.*, 51 (4):690–699, 1986.

T. Wang, H. Ding, J. Kuang, and Q. Ma. Mapping emotions into acoustic space: the role of voice quality. In *15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014.

J. Wells. SAMPA computer readable phonetic alphabet. In D. Gibbon, R. Moore, and R. Winski, editors, *Handbook of Standards and Resources for Spoken Language Systems*, pages Part IV, section B. Berlin and New York: Mouton de Gruyter, 1997.

J.-F. Weng, H.-l. Kuo, and S.-S. Tseng. Interactive storytelling for elementary school nature science education. In *Proc. of the 11th ICALT*, pages 336–338, Athens, GA, 2011.

C. W. Wightman. ToBI or not ToBI? In *Proc. Speech Prosody*, pages 25–29, Aix-en-Provence, France, 2002.

V. Wolfe and D. Martin. Acoustic correlates of dysphonia: type and severity. *Journal of Communication Disorders*, 30(5):403–416, 1997.

M. Yaeger-Dror. Register and prosodic variation, a cross language comparison. *J. Pragmat.*, 34(10-11): 1495–1536, 2002.

J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi. Acoustic Modelling of Speaking Styles and Emotional Expressions in HMM-based Speech Synthesis. *IEICE Trans. Inf. & Syst.*, E88-D(3):502–509, 2005.

J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano. Model adaptation approach to speech synthesis with diverse voices and styles. In *IEEE Int. Conf. Acoust., Speech & Signal Process. (ICASSP)*, volume 4, pages 1233–1236, 2007.

E. Yumoto, W. J. Gould, and T. Baer. Harmonics-to-noise ratio as an index of the degree of hoarseness, 1982.

H. Zen, K. Tokuda, and A. W. Black. Statistical parametric speech synthesis. *Speech Commun.*, 51(11): 1039–1064, 2009.

J. K. Zhang, A. W. Black, and R. Sproat. Identifying Speaker in Children's Stories for Speech Synthesis. In *Proc. Eurospeech*, pages 2041–2044, Geneva, Switzerland, 2003.

E. Zovato, A. Pacchiotti, S. Quazza, and S. Sandri. Towards Emotional Speech Synthesis: A Rule Based Approach. In *5th ISCA Workshop on Speech Synth.*, pages 219–220, Pittsburgh, PA, USA, 2004.

Esta Tesis Doctoral ha sido defendida el día ____ d_____ de 201__

En el Centro_____

de la Universidad Ramon Llull, ante el Tribunal formado por los Doctores y Doctoras abajo firmantes, habiendo obtenido la calificación:

Presidente/a

_____

Vocal

_____

Vocal *

_____

Vocal *

_____

Secretario/a

_____

Doctorando/a

(*): Sólo en el caso de tener un tribunal de 5 miembros