Speech Communication 88 (2017) 1–16

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

The role of prosody and voice quality in indirect storytelling speech: A cross-narrator perspective in four European languages

Raúl Montaño, Francesc Alías*

GTM - Grup de recerca en Tecnologies Mèdia, La Salle - Universitat Ramon Llull, Quatre Camins, 30, Barcelona 08022, Spain

ARTICLE INFO

Article history:
Received 16 October 2015
Revised 5 January 2017
Accepted 8 January 2017
Available online 11 January 2017

Keywords:
Storytelling
Speech analysis
Prosody
Voice quality
Cross-narrator
Expressive categories

ABSTRACT

During the last decades, the majority of works devoted on expressive speech acoustic analysis have focused on emotions, although there is a growing interest in other speaking styles such as storytelling. In this work, we analyze indirect storytelling speech extracted from audiobooks corpora. Specifically, we study to what extent the results obtained in a previous work centered on a Spanish narrator are generalizable to other narrators telling the same story in English, French, and German. We analyze the indirect speech of a story oriented to a young audience in terms of prosody and voice quality through statistical and discriminant analyses, after classifying the sentences of the story in several expressive categories: neutral, descriptive, post-character, suspense, negative/passive, negative/active, positive/passive, and positive/active. The results confirm the existence of the storytelling categories already observed in the tale's Spanish version across the considered narrators, besides establishing a set of acoustic parameters that are useful to discriminate them. Moreover, a strong relationship is observed in the selection of the expressive category per utterance across the narrators. The analyses also show that both prosody and voice quality contribute significantly to the discrimination among storytelling expressive categories, being conveyed with similar acoustic patterns across narrators in the considered four European languages.

1. Introduction

During the last decades, there has been a main focus on the analysis, synthesis and recognition of vocal expression of emotions in the field of speech technologies (see, Schröder, 2004, and references therein). However, more recently there has been a growing interest in diverse context-specific speaking styles oriented to different audiences, such as radio news, politics and conversations (Roekhaut et al., 2010), polite and informal styles (Grawunder and Winter, 2010), or catholic mass ceremonies and sport commentaries (Obin et al., 2011), among others. A particular speaking style rich in expressiveness is storytelling oriented to an audience. Due to its variety of expressiveness, audiobooks containing a story have been widely used during the last years to generate and evaluate expressive synthetic speech (Prahallad and Black, 2011; Székely et al., 2012; Charfuelan and Steiner, 2013; King and Karaiskos, 2013; Jauk et al., 2015), while other works have used different approaches for synthetic storytelling speech applications (Silva et al., 2001; 2004; Burkhardt, 2011). However, the thorough analysis of this particular speaking style with the objective of understanding its inherent acoustic characteristics has received less attention. Nonetheless, the specific prosodic characteristics of stories and tales have been studied in the literature indeed, although following quite different approaches. For instance, some authors have focused on specific expressive situations that may occur within storytelling speech like suspense situations (Theune et al., 2006), or have analyzed the structure of tales (Doukhan et al., 2011; Adell et al., 2005). On the contrary, emotional categories have been considered to annotate storytelling speech (Alm and Sproat, 2005; Sarkar et al., 2014), although this last approach has been questioned for the annotation of indirect storytelling speech recently by Montaño and Alías (2016). The aforementioned works analyzed storytelling speech corpora for a particular language. Up to our knowledge, there are no studies specifically focused on the acoustic analysis of storytelling speech considering different languages and narrators.

In this work, we conduct a cross-narrator study on indirect storytelling speech in terms of prosody and voice quality (henceforth, VoQ) considering the same story interpreted by four professional male storytellers in four European languages. Concretely, we explore two Germanic languages (British English and German) and two Romance languages (European Spanish and French), following the annotation methodology introduced by Montaño and Alías (2016). Our goal in this paper is to perform a preliminary cross-narrator analysis of indirect storytelling speech, studying if the ev-

* Corresponding author.
E-mail addresses: raulma@salleurl.edu (R. Montaño), falias@salleurl.edu (F. Alías).

http://dx.doi.org/10.1016/j.specom.2017.01.007
0167-6393/

idences obtained in our previous work for the Spanish version of the story can be further generalized to other storytellers and languages. After all the conducted analyses, we would like to have some evidences to answer the following questions: (i) do the previously defined storytelling expressive categories exist across the different versions of the story?; (ii) do narrators use the same expressive category for each utterance?; (iii) are the acoustic characteristics of the storytelling expressive categories comparable across narrators?; (iv) is VoQ as important as prosody to discriminate among storytelling expressive categories?.

This paper is structured as follows. Section 2 reviews previous works related to cross-language analyses of speaking styles, and introduces the considered acoustic parameters. Section 3 describes the storytelling expressive categories and the annotation methodology. Next, all the analyses and their results are detailed in Section 4. Firstly, the annotation process is explained step by step in Section 4.1, and then, the acoustic analyses results are reported in Section 4.2. Finally, after some discussion (Section 5), this paper ends with the conclusions and future work in Section 6.

2. Related work

Although, as far as we know, no cross-language studies have been conducted specifically focused on the acoustic analysis of storytelling speech, several works have dealt with emotions, other speaking styles, or conversational situations from a cross-language perspective. After describing these works in Section 2.1, Section 2.2 introduces the acoustic parameters selected from the literature to conduct the subsequent analyses.

2.1. Expressiveness analysis from a cross-language perspective

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. The results highlighted that the expression of the emotions under analysis (conveyed via meaningless utterances) showed global tendencies for the considered acoustic and perceptual attributes across those languages, regardless of its 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 already 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 of 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 yielded similar conclusions (Andreeva et al., 2015).

Some context-specific speaking styles such as infant-directed speech and polite/informal speech have also been analyzed from a cross-language perspective. Infant-directed speech was prosodically analyzed 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). 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 American counterparts. Furthermore, vocalic hesitations were prosodically analyzed 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. The 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 also showed different timbre qualities across languages. 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.

2.2. Acoustic parameters used in the analysis of storytelling speech

As one of the key goals of this work is to study to what extent the results obtained by Montaño and Alías (2016) in Spanish can be generalized to other narrators and languages, the same acoustic parameters are considered in the present work. Such parameters have also been used in previous works related to storytelling speech analysis (Adell et al., 2005; Theune et al., 2006; Doukhan et al., 2011; Alm and Sproat, 2005), and automatic classification of expressive audiobooks (Székely et al., 2011; Eyben et al., 2012; Chen and Gales, 2012; Jauk et al., 2015).

The selected set of acoustic parameters describe prosodic, perturbation, spectral, and glottal flow information. Below, all of them are descibed:

- *Fundamental frequency*: In order to study the fundamental frequency (F0) of the analyzed speech, we consider $F0_{mean}$ and its range. Although the latter can be regarded as the difference between some top and bottom values from the F0 dynamics used by a speaker, the definition of these top and bottom values is rather controversial (Patterson, 2000). In this work, we use the inter-quartile range ($F0_{IQR}$), computed as the difference between the 0.95 and 0.05 quantiles following Doukhan et al. (2011), as we reckon that this measure filters out potential incorrect extreme values of the F0 distribution. Both F0 parameters are measured in Hz.
- *Intensity*: Loudness is related to the volume of speech, and it is often referred as intensity (int) when measuring the speech signal amplitude. We consider int_{mean} measured in dB.
- *Speech tempo*: Speech tempo (how fast or slow a person is speaking) is frequently reported in the form of speaking and/or articulation rates. The former includes pauses in the measure while the latter does not. The syllable is usually considered as the 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 rates is syllables per second (Trouvain and Möbius, 2014). In this work, we consider the articulation rate (AR) in syllables/s, but we

also extract pauses information in the form of number of silent pauses within the utterance (Nsp).

- *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 type of normalization. The jitter parameter used is jitter local, i.e., the average absolute difference between consecutive periods, divided by the average period (in %).
- *Shimmer*: Similarly to jitter, shimmer can be defined as cycle-to-cycle variations of the speech waveform amplitude. In this case, the short-term perturbations affect the signal's amplitude. The shimmer parameter used is shimmer local, i.e., the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude (in %).
- *Harmonic-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). We compute HNR (HNR_{mean}) with the method of Boersma (1993), i.e., based on the signal autocorrelation (in dB).
- *Relative amount of energy above 1000 Hz (pe1000)*: 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)*: It is computed as the energy band difference between [0, 500] Hz and [500, 4000] Hz in dB.
- *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 (see Eq. (1)). The parameter is robust against distortion and it is capable of differentiating among different phonation types and emotions (Alku et al., 2002; Airas and Alku, 2007).

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

- *Maxima Dispersion Quotient (MDQ)*: This parameter measures following Eq. (2) how impulse-like the glottal excitation is, via wavelet analysis of the linear prediction residual (Kane and Gobl, 2013). The dispersion of peaks across the different frequency bands are measured in relation to the glottal closure instant (GCI) 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)), and it has been claimed as more robust in changing F0 conditions (Alku et al., 1997).

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

Although the annotation methodology of this work has been conducted at the utterance level, the parameters are only extracted from vowels to ensure its reliability (Montaño and Alías, 2016). Then, they are averaged at the utterance level.

All acoustic measures are extracted using a Praat script specifically developed for this task (Boersma and Weenink, 2014), except for the glottal flow parameters (NAQ, PSP, and MDQ), which are extracted using COVAREP (version 1.3.1) 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 WebMAUS service are used (Bigi and Hirst, 2012; Kisler et al., 2012), respectively. All these automatic segmentations are manually revised in order to dispose of reliable data for subsequent analyses (nearly 64,000 phonemes have been revised and corrected if necessary). This way, the potential inconsistencies that could arise as a consequence of using different labelling tools per language are removed.

3. Annotation methodology for indirect storytelling speech across narrators

In our previous work (Montaño and Alías, 2016), we introduced an annotation methodology for storytelling speech corpora. Such annotation methodology is based on storytelling discourse modes, i.e., narrative, descriptive, and dialogue modes (Adam, 1992; Calsamiglia and Tusón, 1999), and narrative sub-modes, resulting in different categories which we denoted as storytelling expressive categories.

The narrative and descriptive modes belong to the indirect discourse, while the dialogue mode represents the direct discourse parts. In tales and stories, the narrative mode is generally the predominant mode. It is used to inform the listener/reader about the actions that are taking place in the story. Differently, the descriptive mode has the function of describing characters, environments, objects, etc. The dialogue mode manifests when the characters have a conversation and their turns explicitly appear in the story.

The annotation process of our previous work resulted in eight sentence-level categories (Montaño and Alías, 2016), which are described as follows:

- *Post-character (P-C)*: Sentence-level utterances that specify a character intervention, e.g., *"said Peter to his mother"*, which belong to the narrative mode. They can be identified from text as they usually start with a declarative verb in the third person or "speaking word" ("said", "answered", "murmured", "asked", etc.) and, sometimes, the character who intervened in the previous direct discourse is named.
- *Descriptive (DES)*: Sentence-level utterances that compose the aforementioned descriptive mode. They can be identified from text because of 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"*).
- *Suspense (SUS)*: Sentence-level utterances where it is evident that the storyteller wants to elicit uncertainty in the audience. They belong to the narrative mode and can be identified at a perceptual level.
- *Neutral (NEU)*: Sentence-level utterances where the storyteller does not use expressive speech. Even though neutral is a highly subjective and diffuse term, 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*

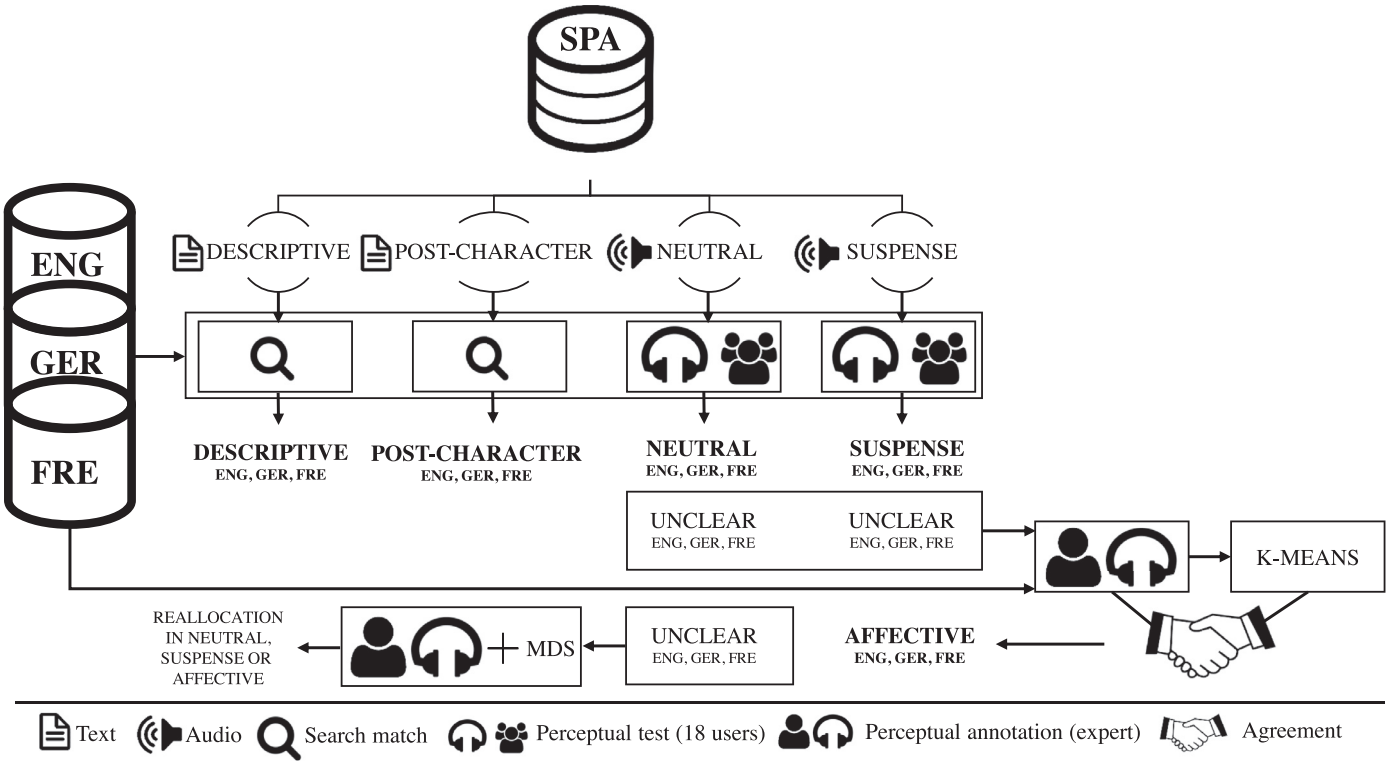

Fig. 1. Diagram showing the annotation process adapted from Montaño and Alías (2016) for the cross-language scenario, considering the Spanish (SPA), English (ENG), German (GER) and French (FRE) versions of the same story. MDS: Multi-Dimensional Scaling.

came into the living room") that can be identified from text. However, in order to avoid mislabellings, its neutral expressiveness has to be subsequently assessed at a perceptual level if we consider them as the baseline for relative comparisons with other expressive categories (Montaño and Alías, 2016; Braunschweiler and Buchholz, 2011).

- *Affective categories*: Since there are no standard set of labels to annotate the rest of sentence-level utterances of the narrative mode, we use a valence/activation representation to annotate them. While discerning valence is relatively feasible from text, posterior perceptual validations are needed to evaluate the expressiveness. This approach generates four categories: *Negative/passive (N/P)*, *Negative/active (N/A)*, *Positive/passive (P/P)* and *Positive/active (P/A)*.

In order to repeat the same manual annotation methodology of (Montaño and Alías, 2016), it would be necessary 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, but we could take advantage of annotating the same story. In order to address this limitation, the original annotation process has been adapted for the English, German, and French corpora by including several extra perceptual tests (see Fig. 1) taking the Spanish annotation 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) suggested that similarity among linguistic communities is not a relevant factor. In our case, the languages are from similar linguistic communities close to each other. Nonetheless, the results obtained from the process should be interpreted with caution since other authors have observed a cultural dependency, at least between some countries (Altrov et al., 2013).

Note that these cross-language tests are only conducted on the neutral and suspense categories. Applying the same approach to the annotation of affective categories was discarded after observing the inherent difficulty of asking to non-native 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 in text and speech analysis (advanced English learner and beginner French and German learner) together with a clustering stage for the annotation of the storytelling affective categories. As (differently to valence) 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.

4. Analysis of indirect storytelling speech across narrators

For the cross-narrator analysis of storytelling speech conducted in this work, we consider parallel corpora of audiobooks where the same story is interpreted by native professional male storytellers in four European languages: Spanish, English, French and German. 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. The story is from the recent past and belongs to the fantasy and adventures genres, with children and preteenagers as its main target audience. 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 in Spanish (Montaño and Alías, 2016), is borrowed for the English, French and German versions of the story. However, the annotation of the perception-dependent categories has been conducted for English, French and German languages as described in the following Sections.

4.1. Classification with the annotation methodology

4.1.1. Neutral category annotation

We performed three different tests confronting the neutral utterances of the Spanish narrator against the other three narrators (one test per language) using the online platform TRUE (Planet et al., 2008). 18 native Spanish speakers (14 males, 4 females; mean age: 33 ± 8.6) were recruited to take the tests. Although their mother tongue was Spanish, 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"*. We randomly selected 30 Spanish neutral utterances from the original corpus of 53 sentences (Montaño and Alías, 2016), 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, we computed the free-marginal Kappa to measure the inter-rater agreement (Randolph, 2005). 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; Fleiss, 1971), showing substantial agreement on choosing the option *"Roughly the same"*. 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 (1,2, and 2 for the English, French and German narrators, respectively) were left aside for their further reallocation (see Fig. 1).

It is worth noting that the averaged value across tests of proportion of category assignment (Fleiss, 1971), resulted in 0.93 for the *"Roughly the same"* category. Thus, this result together with the substantial values of κ_{free} are a first encounter of cross-narrator and cross-language similarities in storytelling speech since, according to the results of the test, the four narrators used a neutral expressiveness for most of the sentences evaluated perceptually. Nonetheless, this result should be corroborated through further studies with more subjects (e.g., native subjects), since they might have projected each expression into the Spanish expressive space.

4.1.2. Suspense category annotation

Considering the same approach followed to annotate the neutral utterances, the 21 suspense utterances identified in the Spanish version of the story 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 what has been already observed for 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 (5,3, and 5 for the English, French and German narrators, respectively) were collected for the subsequent process (see Fig. 1). We also want to note that, although different suspense levels may be present in storytelling (Cheong and Young, 2006; Theune et al., 2006), all the suspenseful utterances are analyzed together, leaving their possible subclassification left for further investigations.

Finally, we would like to remark 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 is the second indication of cross-language similarities in storytelling speech at a perceptual level, at least as perceived by the considered set of evaluators.

Table 1
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

4.1.3. 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. Then, the clustering stage is executed. It keeps the positive and negative labels of the expert annotator, and adds activation labels (passive and active). To that effect, we opted to use k-means (computed using the SPSS software, IBM Corp., 2013) because of three reasons (Jain et al., 1999): (1) its simplicity, (2) the number of clusters is known a priori ($k = 2$), 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. All the acoustic parameters described in Section 2.2 were considered in the k-means clustering step. It was run following a cross-validation approach because of its random initialization. As we only observed very few deviations from the first run, we retained that run to perform the agreement with the human annotator, and then filter out unclear instances.

The annotation outcome for each language is shown in Table 1. 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.

Table 2
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

4.1.4. Reallocation of unclear utterances

In contrast to the remaining 44 utterances at this point of the annotation of the Spanish version (Montaño and Alías, 2016), 87, 94, and 108 utterances remain to be classified within the English, German, and French versions of the story, respectively, as a logical consequence of not having native expert annotators at our disposal. For the reallocation of these unclear utterances the same methodology applied by Montaño and Alías (2016) is followed. Specifically, the expert annotator listened again to these utterances and classified them considering a Multi-Dimensional Scaling (MDS) acoustic representation of the utterances among the different storytelling expressive categories (see Fig. 1). After the reallocation process, some utterances were discarded for the subsequent acoustic analyses for similar reasons 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 showed in Table 2. As can be observed, the categories are quite balanced across languages with some exceptions. For instance, the annotation process has obtained few positive/active utterances from the French narrator, whereas more neutral and suspenseful utterances have been obtained compared to the rest. Moreover, 37 neutral utterances have been collected from the English narrator while from his Spanish, German, and French counterparts, the number of utterances have resulted in 53, 58, and 64, respectively. In order to obtain reliable results, we only consider the validated neutral utterances of the perceptual cross-language analyses to have a similar amount of samples in all languages (see Table 2). We also want to remark that the low number of samples obtained for the positive expressive categories may entail some difficulties in the subsequent statistical and discriminant analyses.

Finally, we want 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 (cf., Montaño and Alías, 2016), English, German, and French corpora, respectively, have been annotated within a storytelling expressive category.

4.2. Acoustic analysis of storytelling expressive categories

4.2.1. Acoustic characteristics of storytelling expressive categories per narrator

Acoustic results are normalized within each speaker using z-scores, and relative acoustic differences between storytelling categories are studied in order to reduce speaker-dependent profiles. Next, within each language, statistical and discriminant analyses are conducted (using the statistical software SPSS, IBM Corp., 2013) 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). Specifically, a multivariate analysis of variance (MANOVA) is performed considering all acoustic parameters as dependent variables, previously to a series of univariate analyses to evaluate differences among categories for each parameter. Statistical significance is assumed to be achieved at $p < 0.05$. 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 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 reader is referred to the Appendix A for a detailed analysis of these pairwise comparisons and the discriminant analysis that defines the relevant parameters. We also report Wilks' lambda 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).

Following Montaño and Alías (2016) and Monzo et al. (2007), a Linear Discriminant Analysis (LDA) is also conducted for each corpus to assess the role of both prosody and VoQ in the discrimination among the storytelling expressive categories. The LDA classification results are reported using the F1 measure, as it combines both precision and recall into a single metric giving a compact vision about the classification performance (Sebastiani, 2001). 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).

Spanish narrator –Firstly, 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 categories. The reader is referred to A.1 for the Tukey's HSD post-hoc tests conducted on all relevant parameters according to the chosen relevance criterion. These parameters resulted in $F0_{mean}$, int_{mean}, SS, H1H2, $F0_{IQR}$, jitter, and HNR_{mean}. The corresponding normalized averaged results can be observed in Table A.7.

The results of the LDA classification are shown in Table 3. When considering all the acoustic features, 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 2) 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 3). However, combining prosodic and VoQ features leads to the highest F_1^M when classifying the expressive categories (a value of 0.503), which corresponds to 31% and 38.6% of relative improvements when compared to considering only prosody and VoQ, respectively. These results highlight 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

Table 3
LDA 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

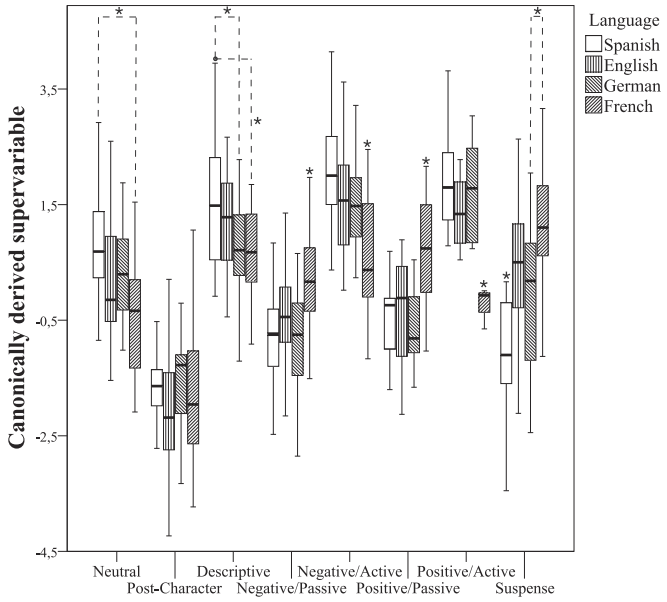

Fig. 2. Canonically derived supervariable for each narrator. Distributions with one asterisk on top are statistically different from the rest of distributions ($p < 0.05$), and those linked by a dashed line and an asterisk also differ significantly. No statistically significant difference otherwise.

Table 4
Wilks' lambda values of each parameter by language. Note that the lower the value, the more discriminative the parameter. The asterisk (*) indicates p< 0.05 in the univariate analysis, i.e., there is at least one significant difference between two of the categories under 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*

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.

The canonically derived supervariable (see Fig. 2) for a boxplot representation of this supervariable) using the raw discriminant function coefficients significantly differentiates among categories [$F(7, 52.39) = 68.823$, $p < 0.001$, $\eta^2 = 0.720$]. Note that the value of η^2 implies that 72% of the variance in the canonically derived supervariable is accounted for by the different categories. Overall, descriptive and both active categories show high values while the rest of categories show lower values, specially the post-character category.

English narrator— The MANOVA on the English version of the story reveals statistically significant results [$Pillai'sTrace = 1.405, F(105, 1365) = 3.265, p < 0.001$] but, differently from the Spanish narrator, the univariate analyses do not show statistically significant differences among categories on all parameters (see Table 4). The most relevant parameters in the English version of the story result in $F0_{mean}$, H1H2, $F0_{IQR}$, AR, Nsp, HammI and int_{mean}. Thus, more prosodic parameters than VoQ features. Normalized averaged values of all parameters for the English version can be observed in Table A.8.

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 3), 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 2). 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 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 differentiates among categories [$F(7, 42.53) = 44.321$, $p < 0.001$, $\eta^2 = 0.606$]. The results are similar to the Spanish narrator with the exception of the suspense category, which shows a higher value (see Fig. 2).

German narrator— The MANOVA on the German version shows a statistically significant result [$Pillai'sTrace = 1.429, F(105, 1253) = 3.060, p < 0.001$], with only shimmer as acoustic parameter with no statistically significant differences among categories in the subsequent univariate analyses (see

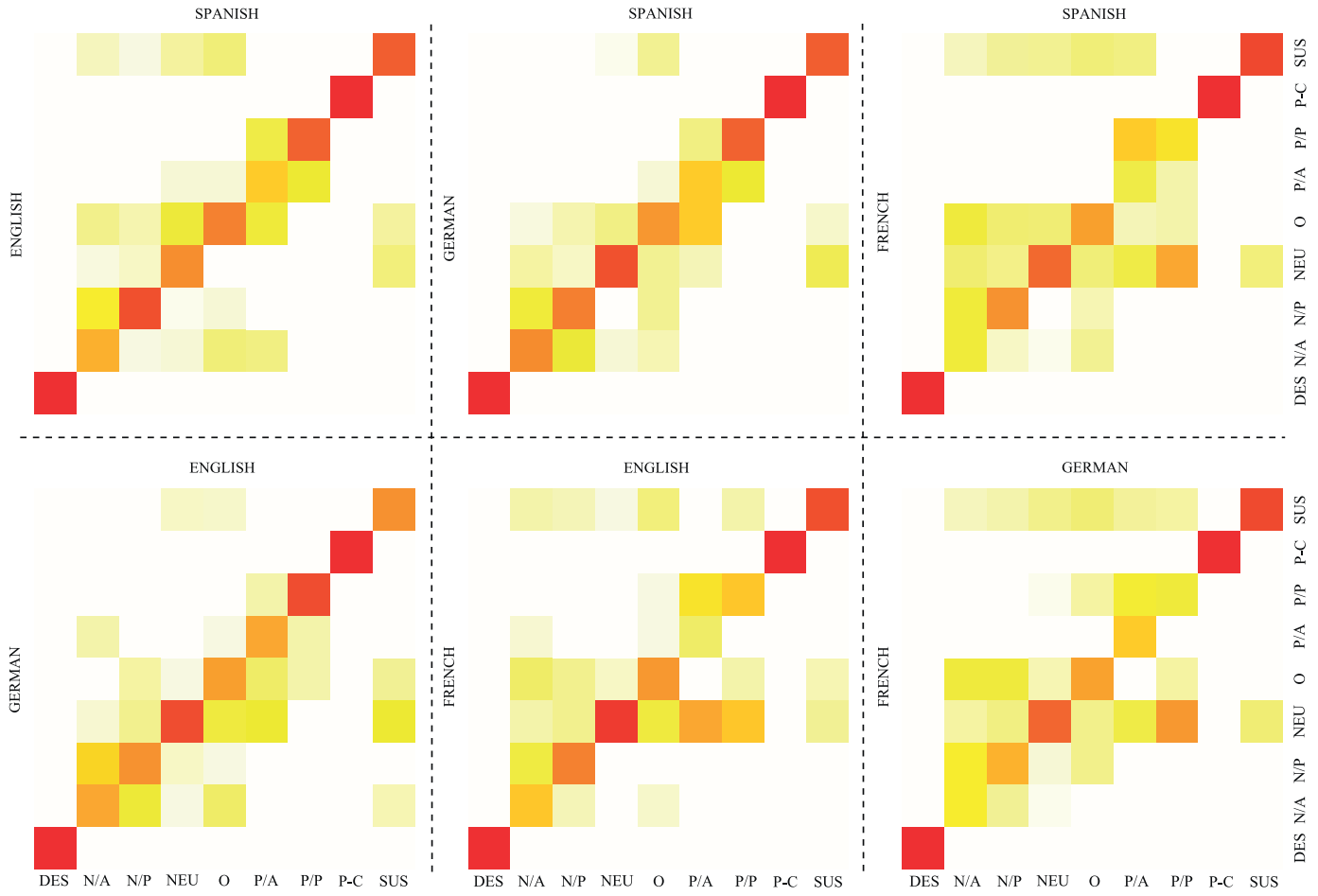

Fig. 3. Heatmaps computed from the contingency tables for each pair of narrators, showing the relationship regarding their use of expressive categories. A warmer colour represents more common instances. Post-character and descriptive categories are included as a visual cue of maximum correlation. O: 'Other'.

Table 4). Relevant parameters consist of $F0_{mean}$, PSP, int_{mean}, NAQ, SS, pe1000, HammI, HNR_{mean}, Nsp and AR. The normalized results can be observed in Table A.9.

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 3). 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. 2. Similarly to the other versions, in general, active and descriptive categories entail higher values than the rest of expressive categories.

French narrator— The multivariate analysis on the French version also shows statistical significance [$Pillai'sTrace = 1.351, F(105, 1169) = 2.661, p < 0.001$], while univariate analyses show statistically significant results on all parameters except in shimmer, pe1000, and PSP (see Table 4). 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. The normalized results can be observed in Table A.10.

The French narrator shows the second best F_1^M (see Table 3), 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 2). 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. 2 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, we informally observed fewer statistically significant results than the Spanish, English, and German versions of the story by means of univariate tests. Only the post-character category is quite well differentiated.

4.2.2. Similarities among narrators

Perceptual-level similarities—In this analysis, we have included the 'Other' category in order to avoid loosing information of similarity/dissimilarity between narrators. An overall visual representation of the similarity between narrators regarding the use of storytelling expressive categories for each sentence can be observed in Fig. 3. Note that the post-character and descriptive categories are included in Fig. 3 just as a visual cue of maximum correlation, but they are not included in the following analysis since they are assigned based only on text.

Table 5
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.

	Cramer's V coefficient			
	Spanish	English	German	French
Spanish	–	0.680	0.710	0.598
English	–	–	0.684	0.641
German	–	–	–	0.602
French	–	–	–	–

Table 6
Relevant parameters in the discrimination among storytelling expressive categories by language according to the defined criteria, i.e., 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.

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

In order to measure the similarity between narrators in terms of the use of expressiveness, Cramer's V coefficients (ϕ_C) were computed as nominal variables (category labels) are used (Cramér, 1946). The resulting coefficients for each pair of narrators can be observed in Table 5. Since all these magnitudes of association belong to the interval (0.6–0.8), 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 according to the output of the annotation process (as already observed in Table 2).

Acoustic-level similarities—The parameters that have manifested as relevant in all languages following the previously defined criteria are $F0_{mean}$ and int_{mean} (see Table 6). Nonetheless, the results for the French narrator shows less $F0_{mean}$ variability among categories, as this parameter correlates with the second canonical function instead of the first one, and shows a larger Wilks' Lambda value than the rest of narrators (see Table 4). Similarly, the English narrator conveyed int_{mean} with less variability, as this parameter correlates with the third canonical function and shows the largest int_{mean} Wilks' Lambda result among narrators in Table 4.

In relation to $F0_{IQR}$, no relevance is found in the German narrator, whereas the rest of narrators show 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.

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. 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 3.

Concerning 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. 2 together with post-hoc comparisons of the distributions across languages. As 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 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. The last common tendency is the very low values of MDQ, which entail a tenser phonation in this category, 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. 2), with few subtle variations across narrators (see Tables A.7–A.10). 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.). The greatest differences between narrators are observed in the Spanish narrator when compared to his French and German counterparts (see Fig. 2). Specifically, the French narrator used a significantly lower AR than the Spanish narrator, whereas the German narrator also shows significantly lower AR together with lower H1H2 and higher SS. With respect to affective categories, it can be observed from Fig. 2 that the French narrator used a specific expressive pattern across expressive categories when compared to the rest of nar-

rators. This difference is specially striking in positive/active utterances, which made difficult the finding of these utterances in the French audiobook (see Table 2). 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 analyzed states that entail more extreme expressiveness. In relation to these previous studies, in our work we have also obtained a flatter spectral slope in terms of the SS parameter except in the French narrator, which shows the opposite behavior. 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} (see Tables A.7–A.10). 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 4.1.2.

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. 4).

5. Discussion

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 our previous work with some adaptations, achieving a comparable amount of successfully classified utterances (around 85% in average). Although this manually-based approach has been useful to analyze 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 4.1.3. Although we have not parametrized the text input, 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).

5.2. Do the previously defined storytelling expressive categories exist?

After all the conducted analyses, the cross-narrator results demonstrate the existence of the storytelling expressive categories introduced in our previous work (Montaño and Alías, 2016). Although presenting different distributions within the story, all the expressive categories already observed in the Spanish version of story have been also 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.

Nevertheless, we also want 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 modeled. On the one hand, for instance, some tales do not contain explicit dialogues between characters and, thus, post-character utterances could not be studied. On the other hand, some of the phenomena (laughter, yawns, etc.) observed in utterances currently denoted as 'Other' could be specifically analyzed in future works. To this aim, it might be necessary to collect more evidences to obtain reliable results. Furthermore, a previous study has already proposed 'increasing' vs. 'sudden' suspense situations (Theune et al., 2006), although they have not been observed in our corpora. Finally, 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 analyzed ones (i.e., fantasy-adventures and young people, respectively) the more arguable it would be to export the obtained results to it.

5.3. Do narrators use the same expressiveness for each utterance?

We have obtained a strong relationship (Cramer's V coefficients between 0.6 and 0.8) among narrators regarding the use of expressiveness for each utterance according to the outcome of the annotation process beyond their personal styles. As 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 (substantial values of κ_{free}) 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 worth noting 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). Therefore, further studies considering native evaluators are needed to corroborate this result, as the current evaluators might have also projected each expression into the Spanish expressive space.

5.4. Are the acoustic characteristics of each storytelling expressive category comparable across narrators?

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 analyzing attitudes and emotions (Mozziconacci, 2001; Pell et al., 2009b). Nonetheless, we have 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-narrator 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 com-

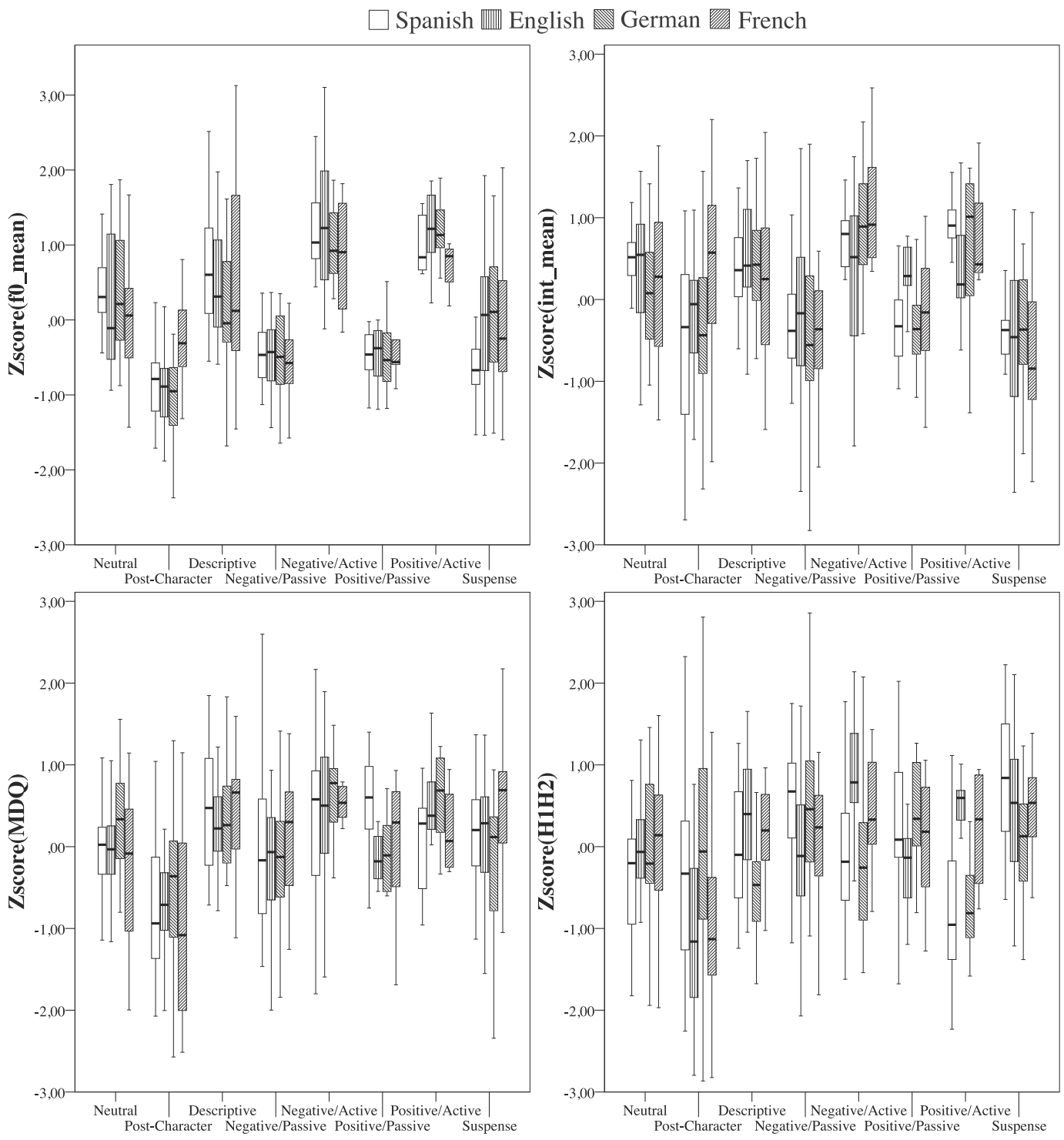

Fig. 4. Z-scores distributions of mean fundamental frequency (F0), mean intensity, MDQ, and H1H2 parameters by language.

pared 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, we have observed few common patterns across languages, 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 could be 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. We attribute such differences 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, we have observed a remarkable amount of similarities in the use of expressiveness among differ-

ent storytellers beyond their personal preferences, specially, across the Spanish, English, and German versions of the story.

5.5. Is VoQ as important as prosody to discriminate among storytelling expressive categories?

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 (see Table 3). Although the English narrator relied more on prosodic variations to convey the story under analysis, we consider that this is intended by the narrator (i.e., a personal style), thus, it is recommendable to take always into account VoQ when dealing with storytelling speech.

6. Conclusions and future work

In this paper, we have studied to what extent the results of a previous proof-of-concept work proving the existence of storytelling expressive categories in Spanish can be generalized to other narrators and/or languages. To that effect, we have extended that research by considering the corresponding English, German and French version of the same story under analysis. Moreover, we have applied the same annotation methodology (with some adaptations due to the lack of having native evaluators) with the objective of confirming the existence of several storytelling expressive categories (a total of eight categories), and studying the role that both prosodic and VoQ features play in indirect storytelling speech through several statistical and discriminant analyses.

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, we want to highlight that these results have been found beyond the observed personal styles of the four narrators.

Concerning acoustic characterization of the different storytelling expressive categories, subtle variations have been observed across categories. However, significant common patterns have also been observed across narrators. 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. Note that cross-language conclusions should be further corroborated since we only used one speaker per language.

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 (substantial values of the free marginal Kappa) is another evidence.

In the future, we plan to analyze more speech corpora, in order to perform cross-genre analyses following a similar methodology but including native evaluators and exploring automatic processes. Other acoustic parameters and consonant information could also be explored to expand the analysis. Moreover, 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. Finally, we plan to derive an acoustic model for the identified storytelling expressive categories to be later included within a Text-To-Speech synthesis system.

Acknowledgments

Raúl Montaño thanks the support of the European Social Fund (ESF) and the Catalan Government (SUR/DEC) for the pre-doctoral FI grant No. 2015FI_B2 00110. We also thank the SUR/DEC for the grant (2014-SGR-0590), the annotators and the people that took the perceptual test for their help, and Dr. Oriol Guasch and Marc Freixes for their support when needed.

Appendix A. Post-hoc and discriminant analyses per narrator

In this appendix, the pairwise comparisons between storytelling expressive categories are detailed per language. Moreover, we report the discriminant analysis that defines the relevant parameters.

A1. Spanish narrator

Concerning $F0_{mean}$, post-character, negative/passive, positive/passive and suspense utterances, they were expressed with significantly lower $F0_{mean}$ than the rest of categories, but with no statistically significant differences 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 show similar intensity levels 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.

Concerning 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 can be sensed at a perceptual level. The H1H2 parameter does not show 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, only four statistically significant differences are obtained from 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.

The conducted discriminant analysis shows 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

Table A1
Normalized averaged acoustic measures of the storytelling expressive categories of the Spanish 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	−0.67	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

Table A2
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
$F0_{mean}$	−0.93	0.46	0.17	−0.47	1.27	−0.46	1.21	0.06
$F0_{IQR}$	−0.35	0.38	−0.07	−0.42	0.51	−0.41	1.07	0.18
int_{mean}	−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
HNR_{mean}	−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

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 4.

A2. English narrator

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 (Montaño and Alías, 2016). 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. Concerning speech tempo parameters, descriptive utterances were expressed with slow AR and large Nsp. On the contrary, post-character utterances show the opposite behavior. 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 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

Table A3
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
$F0_{mean}$	−0.96	0.15	0.33	−0.44	1.06	−0.40	1.20	0.01
$F0_{IQR}$	−0.38	−0.10	−0.16	0.04	0.81	−0.36	0.17	−0.04
int_{mean}	−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
HNR_{mean}	−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

Table A4
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
$F0_{mean}$	−0.16	0.69	0.06	−0.59	1.00	−0.36	0.73	−0.14
$F0_{IQR}$	−0.40	0.45	−0.05	−0.41	0.93	−0.19	0.80	0.18
int_{mean}	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
HNR_{mean}	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

not enough statistically significant results in order to derive clear conclusions.

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} ($r = 0.37$).

A3. German narrator

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.

Concerning 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.

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

A4. French narrator

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 differences 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 behavior 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.

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 parameters: 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 ($r = -0.42$).

References

Adam, J.-M., 1992. Les Textes: Types et Prototypes: Récit, Description, Argumentation, Explication et Dialogue. Paris, Nathan.

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

Airas, M., Alku, P., 2007. Comparison of multiple voice source parameters in different phonation types. In: Proceedings of the Interspeech. Antwerp, Belgium, pp. 1410–1413.

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

Alku, P., Strik, H., Vilkman, E., 1997. Parabolic spectral parameter - a new method for quantification of the glottal flow. Speech Commun. 22 (1), 67–79.

Alm, C.O., Roth, D., Sproat, R., 2005. Emotions from text: machine learning for text-based emotion prediction. In: Proceedings of the HLT/EMNLP, pp. 579–586.

Alm, C.O., Sproat, R., 2005. Perceptions of emotions in expressive storytelling. In: Proceedings of the Interspeech. Lisbon, Portugal, pp. 533–536.

Altrov, R., et al., 2013. Aspects of cultural communication in recognizing emotions. Trames (2) 159–174.

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

Andreeva, B., Möbius, B., Demenko, G., Zimmerer, F., Jügler, J., 2015. Linguistic measures of pitch range in slavic and germanic languages. In: Proceedings of the Interspeech. Dresden, Germany, pp. 968–972.

Bigi, B., Hirst, D., 2012. SPeech phonetization alignment and syllabification (SPPAS): a tool for the automatic analysis of speech prosody. In: Proceedings of the Speech Prosody. Shanghai, China.

Boersma, P., 1993. 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.

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

Borod, J.C., Pick, L.H., Hall, S., Sliwinski, M., Madigan, N., Obler, L.K., Welkowitz, J., Canino, E., Erhan, H.M., Goral, M., Morrison, C., Tabert, M., 2000. Relationships among facial, prosodic, and lexical channels of emotional perceptual processing. Cogn. & Emot. 14 (2), 193–211.

Braunschweiler, N., Buchholz, S., 2011. Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality. In: Proceedings of the Interspeech. Florence, Italy, pp. 1821–1824.

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

Burkhardt, F., 2011. An affective spoken storyteller. In: Proceedings of the Interspeech. Florence, Italy, pp. 3305–3306.

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

Charfuelan, M., Steiner, I., 2013. Expressive speech synthesis in MARY TTS using audiobook data and EmotionML. In: Proceedings of the Interspeech, pp. 1564–1568. Lyon, France.

Chen, L., Gales, M., 2012. Exploring rich expressive information from audiobook data using cluster adaptive training. In: Proceedings of the Interspeech, pp. 959–962.

Cheong, Y.-G., Young, R.M., 2006. A computational model of narrative generation for suspense. In: AAAI Computational Aesthetics Workshop, pp. 1906–1907. Boston, MA, USA.

Cramér, H., 1946. Mathematical Methods of Statistics. Princeton, NJ: Princeton University Press.

Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S., 2014. COVAREP - A collaborative voice analysis repository for speech technologies. In: Proceedings of the IEEE ICASSP. Florence, Italy, pp. 960–964.

Doukhan, D., Rilliard, A., Rosset, S., Adda-Decker, M., d'Alessandro, C., 2011. Prosodic analysis of a corpus of tales. In: Proceedings of the Interspeech. Florence, Italy, pp. 3129–3132.

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

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

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

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

Goldman, J.P., 2011. EasyAlign: An automatic phonetic alignment tool under Praat. In: Proceedings of the Interspeech. Florence, Italy, pp. 3233–3236.

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

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

IBM Corp., 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.
Jackson, M., Ladefoged, P., Huffman, M., Antoñanzas-Barroso, N., 1985. Measures of spectral tilt. J. Acoust. Soc. Am. 77 (S1) S86-S86.
Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Comput. Surv. 31 (3), 264–323.
Jauk, I., Bonafonte, A., Lopez-otero, P., Docio-fernandez, L., 2015. Creating expressive synthetic voices by unsupervised clustering of audiobooks. In: Proceedings of the Interspeech. Dresden, Germany, pp. 3380–3384.
Kane, J., Gobl, C., 2013. Wavelet maxima dispersion for breathy to tense voice discrimination. IEEE Trans. Audio, Speech & Lang. Process. 21, 1170–1179.
King, S., Karaiskos, V., 2013. The blizzard challenge 2013. In: Proceedings of the Blizzard Challenge Workshop. Barcelona, Catalonia.
Kisler, T., Schiel, F., Sloetjes, H., 2012. Signal processing via web services: the use case WebMAUS. In: Proceedings of the Digital Humanities. Hamburg, Germany, pp. 30–34.
Klecka, W., 1980. Discriminant Analysis, 19. SAGE Publications.
Ladd, D.R., Campbell, N., 1991. Theories of prosodic structure: evidence from syllable duration. In: Proceedings of the 12th International Congress of Phonetic Sciences, 2, pp. 290–293.
Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biom. 33 (1), 159–174.
Liu, P., Pell, M.D., 2014. Processing emotional prosody in Mandarin Chinese: A cross-language comparison. In: Proceedings of the Speech Prosody. Dublin, Ireland, pp. 95–99.
Montaño, R., Alías, F., 2016. The role of prosody and voice quality in indirect storytelling speech: annotation methodology and expressive categories. Speech Commun. 85, 8–18.
Monzo, C., Alías, F., Iriondo, I., Gonzalvo, X., Planet, S., 2007. Discriminating expressive speech styles by voice quality parameterization. In: Proceedings of the 16th International Congress on Phonetic Science. Saarbrücken, Germany, pp. 2081–2084.
Mozziconacci, S.J.L., 2001. Modeling emotion and attitude in speech by means of perceptually based parameter values. User Model. User-Adapt. Interact. 11 (4), 297–326.
Nicolaou, M.A., Gunes, H., Pantic, M., 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Trans. Affect. Comput. 2 (2), 92–105.
Obin, N., Lanchantin, P., Lacheret, A., Rodet, X., 2011. Discrete/Continuous modelling of speaking style in HMM-based speech synthesis: Design and evaluation. In: Proceedings of the Interspeech. Florence, Italy, pp. 2785–2788.
Patterson, D.J., 2000. Linguistic Approach to Pitch Range Modelling. Edinburgh University, Scotland, United Kingdom Ph.D. thesis.
Pell, M.D., Monetta, L., Paulmann, S., Kotz, S.A., 2009a. Recognizing emotions in a foreign language. J. Nonverbal Behav. 33 (2), 107–120.
Pell, M.D., Paulmann, S., Dara, C., Alasseri, A., Kotz, S.A., 2009b. Factors in the recognition of vocally expressed emotions: a comparison of four languages. J. Phonetics 37 (4), 417–435.
Pépiot, E., 2014. 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 305–309.
Planet, S., Iriondo, I., 2013. Children'S emotion recognition from spontaneous speech using a reduced set of acoustic and linguistic features. Cognit. Comput. 5 (4), 526–532.
Planet, S., Iriondo, I., Martínez, E., Montero, J.A., 2008. TRUE: an online testing platform for multimedia evaluation. In: Proceedings of the 6th Conference on Language Resources and Evaluation, p. 61.
Prahallad, K., Black, A.W., 2011. Segmentation of monologues in audio books for building synthetic voices. IEEE Trans. Audio Speech Lang. Process. 19 (5), 1444–1449.
Randolph, J.J., 2005. Free-marginal multirater Kappa: An alternative to Fleiss' fixed–marginal multirater Kappa. Learning & Instruction Symposium. Joensuu, Finland.
Rea, L.M., Parker, R.A., 1992. Designing and Conducting Survey Research. San Francisco: Jossey-Boss.
Roekhaut, S., Goldman, J., Simon, A., 2010. A model for varying speaking style in TTS systems. In: Proceedings of the Speech Prosody. Chicago, IL, USA, pp. 11–14.
Sarkar, P., Haque, A., Dutta, A.K., Reddy, G.M., Harikrishna, M.D., Dhara, P., Verma, R., Narendra, P.N., Sunil, B.K.S., Yadav, J., Rao, K.S., 2014. Designing prosody rule-set for converting neutral TTS speech to storytelling style speech for indian languages: Bengali, Hindi and Telugu. In: Proceedings of the 7th International Conference on Contemporary Computing (IC3). Noida, India, pp. 473–477.
Scherer, K.R., 1989. Vocal Correlates of Emotional Arousal and Affective Disturbance. In: Wagner, H., Manstead, A. (Eds.), Handbook of Social Psychophysiology: Emotion and Social Behavior. Wiley & Sons, Oxford, UK.
Scherer, K.R., Banse, R., Wallbott, H.G., 2001. Emotion inferences from vocal expression correlate across languages and cultures. J. Cross-Cult. Psychol. 32 (1), 76–92.
Schröder, M., 2004. Speech and Emotion Research: An overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis. Saarland Univ, Germany Ph.D. thesis.
Schuller, B., 2011. Recognizing affect from linguistic information in 3D continuous space. IEEE Trans. Affect. Comput. 2 (4), 192–205.
Sebastiani, F., 2001. Machine learning in automated text categorization. ACM Comput. Surv. 34 (1), 1–47.
Silva, A., Raimundo, G., Paiva, A., Melo, C., 2004. To tell or not to tell... Building an interactive virtual storyteller. In: Proceedings of the AISB, pp. 53–58.
Silva, A., Vala, M., Paiva, A., 2001. The storyteller: Building a synthetic character that tells stories. In: Proceedings of the Workshop Multimodal Communication and Context in Embodied Agents, pp. 53–58.
Székely, E., Cabral, J.P., Cahill, P., Carson-Berndsen, J., 2011. Clustering expressive speech styles in audiobooks using glottal source parameters. In: Proceedings of the Interspeech. Florence, Italy, pp. 2409–2412.
Székely, E., Csapo, T.G., Toth, B., Mihajlik, P., Carson-Berndsen, J., 2012. Synthesizing expressive speech from amateur audiobook recordings. In: IEEE Spoken Language Technology Workshop (SLT), pp. 297–302.
Theune, M., Meijs, K., Heylen, D., Ordelman, R., 2006. Generating expressive speech for storytelling applications. IEEE Trans. Audio, Speech Lang. Process. 14 (4), 1137–1144.
Thompson, W.F., Balkwill, L.L., 2006. Decoding speech prosody in five languages. Semiotica 158 (Brown 2000), 407–424.
Trouvain, J., Möbius, B., 2014. Sources of variation of articulation rate in native and non-native speech: comparisons of french and german. Proc. Speech Prosody 275–279.
Van Bezooijen, R., Otto, S.a., Heenan, T.a., 1983. Recognition of vocal expressions of emotion: a three-Nation study to identify universal characteristics. J. Cross-Cultural Psychol. 14 (4), 387–406.
Vasilescu, I., Adda-Decker, M., 2007. A cross-language study of acoustic and prosodic characteristics of vocalic hesitation. In: Fundamentals of Verbal and Nonverbal Communication and the Biometric Issue. IOS Press, pp. 140–148.
Wallbott, H.G., Scherer, K.R., 1986. Cues and channels in emotion recognition. J. Pers. Soc. Psychol. 51 (4), 690–699.
Yaeger-Dror, M., 2002. Register and prosodic variation, a cross language comparison. J. Pragmat. 34 (10–11), 1495–1536.